Journal of Korea Spatial Information Society Vol.23, No.5 : 1-7, October 2015
http://dx.doi.org/10.12672/ksis.2015.23.5.001
ISSN 2287-9242(Print)
ISSN 2287-9250(Online)

트윗 텍스트의 유사 키워드 추출을 통한 이벤트 지역 탐지 기법

A Method for Detecting Event-Location based on Similar Keyword Extraction in Tweet Text

임준엽* · 하현수 · 황병연*****

Junyeob Yim · Hyunsoo Ha · Byung-Yeon Hwang

요 약 트위터는 다른 SNS와 대비되는 정보의 빠른 전파력과 확산성을 갖고 있다. 따라서 트위터를 이용하여 현실에서 발생한 이벤트를 탐지하는 여러 연구가 진행되고 있다. 트위터 사용자 개개인을 하나의 센서로 가정하고 그들이 작성한 트윗 텍스트를 분석하여 이벤트 탐지에 이용하는 것이다. 이와 관련된 연구들은 이미 많은 성과를 보이며 진행되어 왔으나 여러 가지 문제점들로 인해 새로운 한계에 직면했다. 특히 선행 연구의 대다수가 이벤트의 발생 위치를 추적하기 위해 GPS좌표를 이용한다. 그러나 이는 최근 트위터 사용자들이 위치정보 공개에 회의적인 점을 감안하면 명확한 한계점으로 제시될 수 있다. 이에 본 논문에서는 트위터에서 제공하는 위치정보를 이용하지 않고 트윗 텍스트에서 위치정보를 추적하는 방법을 제시하였다. 트윗 텍스트에서 키워드를 추출하여 키워드간의 관계를 고려해 연관단어를 군집화 하였다. 본 논문에서 제안한 알고리즘을 적용한 실험을 통해 이벤트가 발생한 지역과 실제로 발생한 이벤트의 탐지 여부를 확인하였다. 또한 본 논문에서 제안한 기법이 기존 매체들보다 빠른 탐지를 보임으로써 제안된 기법의 우수성을 입증하였다.

키워드 : 소셜 네트워크 분석, 트위터, 지역 이벤트 탐지, 유사 키워드

Abstract Twitter has the fast propagation and diffusion of information compare to other SNS. Therefore, many researches about detecting real-time event using twitter are progressing. Twitter real-time event detecting system assumes every twitter user as a sensor and analyzes their written tweet in order to detect the event. Researches that are related to this twitter have already obtained good results but confronted the limits because of some problems. Especially, many existing researches are using the method that can trace an event location by using GPS coordinate. However, it can be suggested a definite limitation through the present user's skeptical responses about making personal location information public. Therefore, this paper suggests the method that traces the location information in tweet contents text without using the provided location information from twitter. Associated words were grouped by using the keyword that extracted in tweet contents text. The place that the events have occurred and whether the events have surely occurred are detected by this experiment using this algorithm. Furthermore, this experiment demonstrated the necessity of the suggested methods by showing faster detection compare to the other existing media.

Keywords : Social Network Analysis, Twitter, Location Event Detection, Similar Keyword

1. 서 론

최근 스마트폰의 보급으로 인한 웹 접근성의 확대로 인해 소셜 네트워크 서비스(Social Network Service; SNS)를 이용하는 사용자들이 급증하고 있다. SNS란 인터넷 공간에서 사용자들 사이의 커뮤니케이션 공간을 생성할 수 있도록 도와주는 서비스이다. 그 공간 안에서는 이미 다수의 사용자들 간의 빠르고 광범위한 정보의 전달이 이루어지고 있다. 또한, 모든 정보가 기록된다는 주요한 특징으로 인해 과거에 발생한 대규모 데이터를 분석하여 패턴을 찾은 후 미래를 예측할 수 있다는 가능성이 제시되고 있다. 이에 따라 SNS를 분석하는 연구는 많은 연구자들로부터 이목을 끌고 있다.

† This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0009407).

* Junyeob Yim, Master, Dept. of Computer Science and Engineering, The Catholic University of Korea. junyeob1205@catholic.ac.kr
** Hyunsoo Ha, Bachelor's Student, Dept. of Computer Science and Engineering, The Catholic University of Korea. hss0924@catholic.ac.kr
*** Byung-Yeon Hwang, Professor, Dept. of Computer Science and Engineering, The Catholic University of Korea. byhwang@catholic.ac.kr (Corresponding Author)

따라서 많은 SNS를 대상으로 연구가 진행되고 있는데, 국내에서 서비스가 이루어지고 있는 SNS로는 여러 종류가 있다. 그 중, 트위터는 다른 SNS에 비해 여러 가지 특징을 가지고 있으며, 개발자들에게 다양한 API (Application Program Interface)를 제공하기 때문에 연구적 활용가치가 매우 높다. 우선, 트위터의 경우 트윗(Tweet)이라는 140자의 단문텍스트 서비스를 제공한다. 이로 하여금 트위터 사용자는 비교적 가벼운 내용의 트윗을 보다 간편하게 작성하여 다른 사용자들에게 전파한다. 이는 정보의 생산과 전달을 매우 빠르게 하는 결정적 요인으로 작용한다. 따라서 기존 대다수의 SNS와 같이 장문의 텍스트를 요구하거나 사진이나 동영상 등을 활용하는 게시물을 작성하는 방식에 비해 보다 나은 신속성을 제공한다. 이와 더불어 트위터는 개방적인 네트워크 구조를 지니고 있다. 대다수의 SNS는 사용자간의 정보 전달을 위해 친구 관계가 되어야 한다. 트위터에서는 팔로워(Follower)-팔로윙(Following)이라는 개념이 이용되는데, 이는 한쪽 사용자의 단 방향적인 요청만으로 관계가 성립된다. 따라서 보다 유동적이고 개방적인 네트워크 구조가 형성되며, 넓은 범위의 정보 확산을 유도해 낸다.

앞서 언급한 특징들로 인해 트위터를 이용한 연구는 매우 다양하다[1]. 그 중 이벤트를 탐지하고 전파하기 위한 시도들이 존재한다[2,3,4]. 이와 같은 대다수의 이벤트 탐지 방법들은 사용자가 제공한 위치정보를 이용한다[5]. 이에 따라 GPS 좌표의 신뢰성 파악에 대한 연구도 진행되었다[6]. GPS 좌표를 이용하여 이벤트가 발생한 위치를 추적하는 방법은 다양하다. 그러나 결국 특정 키워드를 미리 입력하여 해당 키워드가 급증하는 지역을 탐지하는 것이다. 이러한 방법의 경우 두 가지 문제점이 있다. 첫째, 미리 입력하지 않은 키워드에 관한 이벤트는 탐지할 수 없다는 것이다. 누락된 키워드는 지속적인 업데이트과정이 필요하며, 새로운 형태의 이벤트에 대한 재정의가 필요해진다. 둘째, 최근 들어 SNS 이용자들이 개인정보에 관심이 늘어나면서 위치정보를 공개하는 사용자 수가 급격히 줄고 있다. 즉, 이벤트의 발생위치를 추적하기 위한 GPS 좌표의 이용가치가 매우 떨어진 것이다. 이와 같은 문제를 해결하기 위해 본 논문에서는 사용자들이 작성하는 트윗 내용을 이용하여 이벤트가 발생한 위치를 추적하는 기법을 제안한다. 트윗 내용을 직접 분석하여 명사들을 추출하였고, 추출된 명사간의 연관관계를 고려하여 각각의 군집을 형성한다. 연관 단어가 중복되어 추출되면 그에 따른 가중치를 적용하였고, 지역마다 임의의 임계점을 두어 가중치의 합이 넘었을 경우 이벤트가 발생한 지역이라고 판단하는 기법이다. 이와 더불어 제안된 기법의 실효성을 검증하기 위해 실제 발생한 이벤트에 대한 실험을 진행하였다.

이 논문의 구성은 다음과 같다. 2장의 관련 연구에서는 제안하는 기법과 관련된 연구를 살펴보고, 3장에서 제안된 기법의 구조와 실험 방법을 소개한다. 이후 4장에서는 제안된 기법의 성능을 평가하고 5장에서 결론과 향후 연구과제에 대해 기술한다.

2. 관련연구

Li et al.[7]는 문서 내 핵심 키워드를 추출하는 다양한 키워드 추출 모델을 소개하였다. 이러한 키워드 추출 모델은 전통적인 형태의 문서를 기준으로 개발된 것이다. 따라서 논문에서는 최근 인터넷에서 작성되는 블로그 형태의 문서에 그대로 적용하기에는 무리가 있다는 점을 언급하였다. 따라서 가장 적합한 모델을 찾기 위해 페이스북 사용자들이 작성한 게시글을 대상으로 실험하였다. 실험 결과 Friedman[8]이 소개한 GBM (Gradient Boosting Machine)이 가장 높은 성능을 보였다. 그러나 이는 기계학습을 위한 트레이닝 데이터(Training Data)가 필요하며, 영어로 된 문서를 기준으로 실험이 수행되었다. 따라서 한국어로 된 SNS 문서에 그대로 적용하기에는 무리가 있다.

Ku and Min[9]은 한국어로 된 문서 내의 핵심 키워드를 추출하기 위해 비사전기반의 키워드 추출 기법을 이용했다. 이를 위해 우선 문장의 형태소 분석을 통한 명사 추출을 수행하였다. 명사 추출의 경우 선제거와 정제, 후제거 과정을 통해 문장 내에서 명사를 획득한다. 선제거 과정에서는 명사가 아닌 품사들을 제거하는 과정이며 정제 과정은 조사와 어미를 제거하는 과정이다. 마지막으로 후제거는 무의미한 단어일 경우를 제거하는 과정이다. 이와 같은 명사 획득 과정을 통해 핵심 키워드일 확률이 높은 명사들을 추출하였다. 명사 추출 이후 우선순위에 따라 핵심 키워드를 선정하였으며, 실험을 통해 이를 입증하였다. 이러한 연구는 자연어 처리가 필요한 대다수의 연구에서 활용가치가 높으며, 관련 연구들의 실험을 위한 선행 연구로 볼 수 있다.

한편, Sakaki et al.[2]는 트위터를 이용하여 현실에서 발생한 이벤트를 탐지하는데 성공하였다. Sakaki et al.[2]가 제안한 Toretter 시스템은 지진이나 태풍과 같은 재난 상황이 발생할 경우 트위터 사용자가 남긴 트윗 메시지를 통해 이벤트가 발생한 지역을 탐지하는 시스템이다. 여기서는 재난과 관련된 키워드를 미

리 입력하였고, 해당 키워드를 포함한 트윗이 급증할 경우 이벤트로 판별한다. 이후 해당 이벤트의 위치를 탐색하기 위해 급증한 트윗들의 위치좌표(Geocode)를 이용하였다. 그러나 서론에서 잠시 언급했다시피, 최근 위치 정보 공개를 꺼리는 사용자들이 많아지면서 위치 좌표에 의존한 위치 정보 수집은 명확한 한계를 보인다. 이를 개선하기 위해 Yim et al.[10]은 트윗에 포함된 위치좌표를 이용하지 않고 트윗 텍스트 자체를 분석하여 위치 정보를 판단하였다.

3. 이벤트 지역 탐지 기법

3.1 전체 시스템 구조

본 논문에서는 이벤트가 발생된 지역을 탐지하기 위해 Figure 1과 같은 형태의 시스템을 구축하였다. 시스템의 전체적인 구성은 Yim et al.[10]이 소개한 내용을 참고하였다. 트위터를 사용하여 이벤트 발생을 실시간으로 탐지하는 시스템을 구현한 실험과정과 결과를 담고 있다. 그러나 Yim et al.[10]이 제안한 시스템 모듈에서는 많은 노이즈로 인해 탐지된 이벤트의 정확도가 낮았으며 이에 대해 개선이 시급함을 언급하였다. 따라서 이벤트 탐지의 정확도를 향상시키기 위해 연관 분포를 분석하는 알고리즘을 설계하였다. 알고리즘은 Yim et al.[10]이 제안한 전체적인 시스템을 기반으로 하였다. 트윗 내용을 수집하고 한글로 작성된 트윗만을 걸러내 말뭉치로 분리하는 트윗 수집 과정까지는 같다. 그러나 트윗 분석 과정에서 차이를 보인다. 기존 시스템에서는 지역 검출 빈도를 단순히 누적하여 이벤트를 탐지하였다. 반면 본 논문에서 제안한 클러스터링 기법을 적용한 시스템에서는 유사 키워드 추출을 통해 지명과 이벤트의 연관성을 수치로 나타내어 이벤트를 탐지한다. 즉 정확도가 향상된 이벤트 탐지가 가능하다.

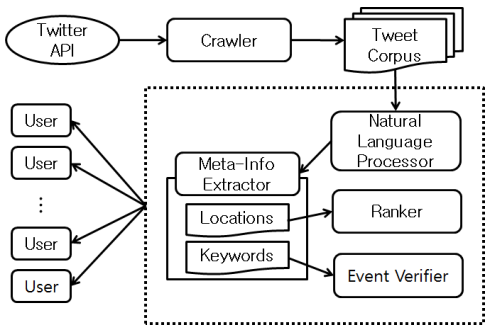

Figure 1. Twitter Event Area Detecting System

3.2 트윗 데이터 수집 및 분석

트위터는 다양한 API를 제공한다. Search API, Streaming API 등이 제공되는데 그 중에서 Streaming API[11]를 이용하여 실시간으로 발생되는 트윗 데이터를 수집한다. Streaming API는 전세계의 트윗을 모두 받아오기 때문에 크롤러(Crawler)를 통해 한국어로 작성된 트윗만을 수집한 후, 데이터베이스에 저장한다. Streaming API는 트위터 사와 'Firehose'라는 별도의 계약을 해야 많은 양의 트윗을 받을 수 있다. 그러나 본 논문에서 제안하는 시스템은 양이 중요하기보다 트윗의 비율이 중요하다. 따라서 무료로 제공되는 Streaming API를 사용해도 결과에 큰 영향을 끼치지 못한다고 판단된다. 트윗 데이터베이스 저장 이후 루씬(Lucene) 형태소 분석기[12]를 이용하여 자연어 처리를 수행한다. 이를 통해 트윗 텍스트에서 다수의 명사를 추출하였고, 추출된 명사를 지명과 그 외의 단어들로 구분하였다. 지명의 판단 기준은 통계청에서 2010년에 발표한 대한민국의 행정구역명 자료를 토대로 수립하였다[13].

이러한 과정을 거치면 다수의 트윗 코퍼스(tweet corpus)에서 각각의 지명들이 언급된 횟수를 알 수 있다. 코퍼스는 말뭉치를 의미하며 본 논문에서는 띄어쓰기를 기준으로 말뭉치를 지정하여 실험했다. 이후 각 지역별로 언급 빈도가 급증한 지역을 순위로 매긴 뒤 정렬 작업을 수행한다. 언급 빈도가 급증한 지역은 최근 트위터 사용자들에게 이슈가 된 지역이며, 이벤트가 발생했을 확률이 높다. 따라서 언급 빈도가 급증한 지역을 이벤트 지역으로 판단한다. 이때 시스템에서 이벤트 지역으로 판단한 지역을 검증하기 위해 트윗 문장에서 유사한 키워드를 추출하여 이벤트와 연관된 단어를 군집화 한다.

최종적으로 이벤트가 발생된 지역이 탐지되면 해당 지역과 연관 키워드를 사용자들에게 전달한다. 결국 제안하는 시스템의 사용자들은 이벤트가 발생된 지역과 이벤트 키워드를 수신하여 해당 이벤트에 따른 올바른 대처를 할 수 있을 것이다.

3.3 클러스터링을 이용한 유사 키워드 추출

이 절에서는 트윗 텍스트에서 유사한 키워드를 추출하여 이벤트를 검증하는 기법을 소개한다. 유사한 키워드를 추출하기 위해 트윗 내의 연관어들을 군집화 하고 각 연관어들의 개념적 거리를 측정하였다. 개념적 거리는 트윗에서 검출되는 연관어들의 관련성을 수치로 나타내는 것을 의미한다.

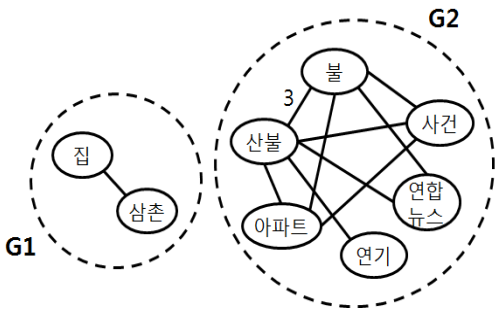

Figure 2. Associated Word Group

Table 1. Keyword Example

Tweet No.	Extracted Noun
Tweet 1	포항 / 산불 / 불 / 연합뉴스
Tweet 2	포항 / 산불 / 불 / 사건 / 아파트
Tweet 3	포항 / 산불 / 불
Tweet 4	포항 / 산불 / 연기
Tweet 5	포항 / 삼촌 / 집

연관어를 군집화하기 위해서는 같은 트윗에 함께 포함된 키워드를 연결하는 방식으로 수행한다. 이는 동시 다발적으로 발생된 어휘들이 연관성이 높다고 판단되기 때문이다. 따라서 Figure 2와 같이 각각의 키워드를 그래프 형태로 표현한다. 예를 들어 Table 1과 같이 '포항'이라는 지명을 포함한 5개의 트윗에서 각각의 명사들이 추출되었다고 가정하자. 여기서 지명을 제외한 각각의 추출된 키워드들은 하나의 노드로 표현하고, 같은 트윗에서 발생된 키워드일 경우 간선으로 연결한다. 트윗 1에서 추출된 명사가 '포항', '산불', '불', '연합뉴스'라고 할 때, '포항'을 제외한 명사 '산불'과 '불', '연합뉴스' 사이를 간선으로 연결한다.

이와 더불어 각 연관어들의 개념적 거리를 측정하는 작업이 필요하다. 따라서 순차적으로 트윗이 발생될 때 마다 간선을 그려가며 이미 간선으로 연결된 노드일 경우 가중치를 두어 값을 1씩 증가시킨다. 이후 Figure 2에서와 같이 G1, G2 등으로 간선으로 연결된 노드를 군집화 하고 생성된 군집 중 가장 높은 가중치를 가지는 그룹을 최종 이벤트 키워드 집합으로 판단한다. 예를 들어 트윗 3과 트윗 4 에서 추출된 명사 중에 '불'과 '산불'이 포함되어 있다면, 위에서 언급한 트윗 1의 추출명사까지 합쳐 총 3번이 언급된 것이다. 이러한 상황일 때 '불'과 '산불'의 간선의 가중치 값은 3이 된다. G1과 G2에서 간선들의 가중치의 합을 비교하여 높은 가중치의 합을 가지는 그룹인 G2를 선택할 수 있다. G2는 서로 연관된 단어들로만 연결되어 있기 때문에 이벤트 내용을 연상할 수 있다. 마찬가지로 노드와 간선으로 연결된 다수의 그룹이 생성될 때 각각의 그룹에서 가중치의 합을 비교하여 가장 높은 값을 가지는 그룹을 선택하여 이벤트 지역을 탐지할 수 있다.

최종 이벤트 키워드 집합이 정해지면 각 지역별로 임계값을 정하여 가중치의 합이 임계값을 넘는지를 살펴본다. 지역별 임계값은 실험을 통해 학습된 값이며, 임계값의 최댓값과 최솟값을 비교하여 시스템이 주기적으로 갱신한다. 이러한 이벤트 판별 방식은 단순히 발생빈도 등을 측정하는 것이 아니라 발생된 형태 및 연관 분포를 살펴보게 되므로 Yim et al.[10]의 한계점이었던 노이즈가 자동으로 제거되는 효과가 있다. 또한 이벤트가 발생된 지역뿐 아니라 이벤트의 핵심 키워드를 알 수 있다는 장점이 있다.

4. 실험 및 결과

4.1 실험 환경 분석

본 논문에서 제안한 시스템이 평균적으로 어느 정도의 처리량을 가지며 실시간으로 작동할 수 있는지 확인하기 위해 성능을 평가한다. 처음 시스템을 실행시키는 시간과 데이터베이스에 저장된 모든 지역을 1회 스캔하는 시간을 측정하였다. 실험 결과를 분석하기 전에 전체적인 시스템의 결과를 분석하기 위한 PC의 성능은 아래의 Table 2와 같다.

Figure 3은 본 논문에서 제안한 시스템이 실시간으로 작동하기 위해 걸린 시간을 측정한 결과를 나타낸다. Figure 4는 트윗 내용 안에 있는 지명을 검출하기 위해 데이터베이스에 저장되어있는 모든 지명들을 트윗 한 문장마다 스캔하는 시간을 의미한다. 10회 실험 측정 결과 시스템이 실시간으로 작동하기 위해 소요된 시간은 Figure 3과 같았으며, 평균은 21분 6초였다. 기존에 수집된 트윗을 분석하는데 시간이 소요되었기 때문에 이와 같은 결과가 나왔다. 그러나 시스템이 작동

Table 2. Experiment Environment

CPU	INTEL Quad Xeon 3.2GHz
RAM	4.00GB
HDD	1TB
OS	Window 8
Compiler	Java 1.7

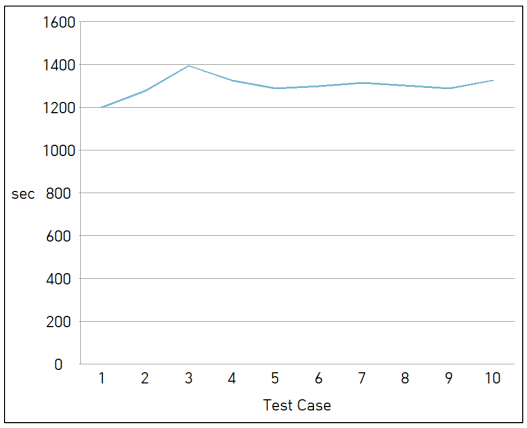

Figure 3. Time for Operating System in Real-time

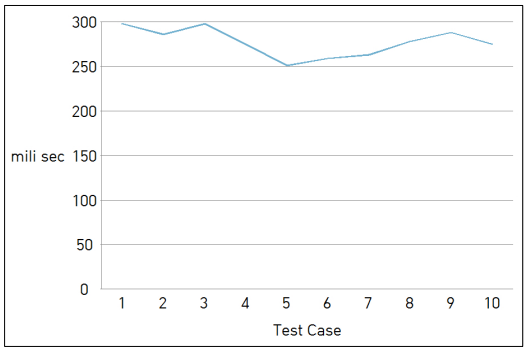

Figure 4. Time for Scanning all Area after Operating System

된 후 모든 지역을 스캔하는데 소요된 시간은 Figure 4와 같이 평균 0.2771초의 결과 나왔다. 이를 통해 본 논문에서 제안하는 시스템의 실시간 실행 가능 여부를 확인하였다.

4.2 실험 결과 분석

제안한 시스템에서 탐지한 이벤트와 실제 발생한 이벤트를 비교하여 시스템의 이벤트 탐지 가능 여부를 확인하였다. 시스템의 실시간 실행 가능함은 4.1절에서 확인되었다. 즉 이벤트 탐지 가능 여부를 알아내면 본 논문에서 제안한 실시간 이벤트 탐지가 가능함을 보일 수 있다. 실험을 위해 2013년 3월부터 13개월간 수집한 트윗 데이터를 시스템에 입력하여 이벤트 탐지 여부를 확인하였다. 지역별로 트윗을 수집하여 유사 키워드를 통해 이벤트 지역을 탐지하였고 실제 발생한 이벤트와 비교하여 실제 이벤트를 탐지할 수 있음을 확인하였다. 탐지할 이벤트 선정의 기준은 실

Table 3. Breaking News Contents and Detection

News	News Contents	Is detected
1	전국 산불 - 울산, 포항, 공주, 예산 등	O
2	중부지방 폭우	X
3	KTX 대구 부근 탈선 후 추돌	O
4	태풍 다나스 한국 접근 및 동해상 이동	X
5	삼성동 아이파크 아파트 헬기	X
6	중부지방 많은 눈	X
7	경주 마우나 리조트 붕괴	O
8	백령도 북 사격 훈련 및 대응 사격	△
9	진도 해상 여객선 침몰	O

제 발생된 다수의 이벤트 중 KBS 뉴스 속보에서 보도된 지역 관련 이벤트로 정하였으며, 자세한 내용은 Table 3과 같다. 본 논문에서 제안한 시스템은 웹을 통해 구현하였고, 정제된 트윗에서 추출한 지명과 이벤트 내용은 데이터베이스에 저장하였다. 저장된 지명과 이벤트 내용이 Table 3의 속보 내용과 일치하는지 확인하여 이벤트 탐지 기준으로 삼았다.

속보 1, 3, 7, 9는 속보보다 빠르게 이벤트 발생을 탐지하였다. 탐지된 이벤트 중에는 빠른 이벤트 발생 지역 탐지와 그에 따른 초동대처로 인명피해를 최소화 할 수 있는 성격의 이벤트들이 있다. 특히 재난 및 재해 관련 이벤트들이 그러한 성격을 가진다. 따라서 이벤트의 발생지역을 빠르게 탐지하는 것 뿐 아니라 탐지된 지역에 대한 정보를 빠르게 전파하는 것도 중요하다.

한편, 속보 2, 4, 5, 6, 8은 제안된 시스템에서 탐지하지 못했다. 속보 2, 6의 경우 특정 지명을 가지지 않는 전국적인 이벤트였기 때문에 이벤트가 발생된 세부적인 지명을 탐지결과로 반환한다는 특징은 오히려 탐지율을 떨어트리는 요인으로 작용했다. 이를 개선하기 위해 '중부지방', '수도권' 등 광역적인 지역 명칭에 대한 예외 처리가 필요할 것이다. 이와 더불어 속보 4, 8도 탐지하지 못했는데, 이는 관련 데이터가 없었기 때문인 것으로 확인되었다. 속보 4의 동해상과 속보 8의 백령도 인근은 인구밀도가 매우 적은 지역이다. 따라서 시스템에서 하나의 센서로 이용되는 트위터 사용자가 없었기 때문에 이러한 지역들은 시스템에서 탐지하기 어렵다. 한편 속보 8의 경우 속보가 발생된 이후에는 시스템에서 탐지가 되었는데, 이는 해당 속보를 접한 사용자들이 트윗을 작성하였고 이것이 시스템에 반영된 것으로 보인다. 마지막으로 속보 5 또

Table 4. The Time of First Detecting and Breaking News

News	Time of First Detection	Time of News Flash
1	2013.03.09. 18:09	2013.03.10. 02:50
2	X	2013.07.13. 09:30
3	2013.08.31. 08:51	2013.08.31. 10:30
4	X	2013.10.08. 13:58
5	X	2013.11.16. 09:15
6	X	2014.01.20. 05:00
7	2014.02.17. 21:24	2014.02.17. 22:50
8	△	2014.03.31. 13:35
9	2014.04.16. 08:50	2014.04.16. 10:04

한 시스템에서는 탐지하지 못하였는데, 이는 기존에 '삼성'이라는 단어가 자주 언급되어 발생 빈도의 변화가 미비했기 때문이다. 이와 같은 동음이의어에 따른 문제점들을 해결하기 위해서는 보다 정확한 형태소 분석과 문맥 내에서의 의미 판별을 위한 추가적인 작업이 필요할 것으로 생각된다.

Table 4는 본 논문에서 소개된 시스템에서의 이벤트 최초 탐지 시각과 KBS 속보가 처음으로 보도된 시각을 나타내고 있다. 시스템에서는 탐지가 된 이벤트들의 최초 탐지 시각이 KBS 속보보다 평균적으로 1시간가량 빨랐다. 특히 속보 1의 경우 8시간 41분 빠른 탐지를 보였는데, 이는 뉴스 보도의 특성상 대규모 사건으로 확대되기 전까지는 보도되기 어렵다는 특성 때문이다. 또한 특정 사건이 발생하였을 때 실제 발생한 사건인지 사실 여부를 검증해야하기 때문에 보도를 준비하기 위한 시간이 소요되었을 것이다.

5. 결론 및 향후 연구

본 논문에서는 트위터를 이용하여 이벤트가 발생된 지역을 탐지하는 기법에 대해 소개하였다. 기존의 선행연구에서 진행되었던 위치 정보를 이용한 지역 탐지와는 다르게 사용자가 위치정보를 제공하지 않아도 지역 탐지에 있어서 문제를 발생시키지 않는다. 제안된 시스템에서는 트위터 데이터 내용 안에서 키워드를 추출하고 연관단어를 군집화 하여 이벤트가 발생한 지역을 탐지하였다. 연관단어의 중복이 있을 경우에 대해서는 가중치라는 예외적인 처리도 적용하였다. 이러한 이벤트가 발생된 지역 탐지에 대한 결과는 실험을 통해 탐지여부를 검증하였다. 실험 결과 탐지율 자체는 비교적 낮았으나, 이벤트를 탐지한 다른 매체들보다 매우 빠르다는 장점이 있었다. 한편, 시스템 전반에 걸쳐 여러 한계점들이 발견되었는데 주된 문제점은 지명과 동음이의어 관계에 있는 단어에 의한 것이었다.

따라서 향후 과제로 단문 텍스트 내의 동음이의어를 판별하기 위한 방안이 연구되어야 할 것이다. 또한, 본문에서 언급한 광역적인 지역 명칭에 대한 명확한 정의가 필요할 것이다. 이외에도 최초 탐지 이후 사실 여부를 검증하여 이벤트 발생 관련 지역의 사람들에게 전파하는 방법도 연구해야 할 것이다.

References

[1] Park, S. Y; Ha, Y. H; Kim, Y. H. 2010, Recent Studies on Twitter in the Field of Information Retrieval, KIISE Fall Conference, 25-29.

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

[3] Li, R; Lei, K. H; Khadiwala, R; Chang, K. 2012, TEDAS: a Twitter Based Event Detection and Analysis System, IEEE 28th International Conference on Data Engineering, 1273-1276.

[4] Lee, J; Baek, S; Lee, S; Bae, H. 2012, The Method for Real-time Complex Event Detection of Unstructured Big data, Journal of Korea Spatial Information Society, 20(5):99-109.

[5] Kim, M; Park S. 2014, Construction and Application of POI Database with Spatial Relations Using SNS, Journal of Korea Spatial Information Society, 22(4):21-38.

[6] Lee, B; Kim, S; Hwang, B. Y. 2012, Analyzing the Credibility of the Location Information Provided by Twitter Users, Journal of Korea Multimedia Society, 15(7):910-919.

[7] Li, Z; Zhou, D; Juan, Y. F; Han, J. 2010, Keyword Extraction for Social Snippets, The 19th International Conference on World Wide Web, 1143-1144.

[8] J. H. Friedman. 2001, Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics, 29(5):1189-1232.

[9] Ku, M; Min, D. 2009, Study on Keyword Extraction Method using Recursive Extracted Word Division, KIISE Fall Conference, 329-334.

[10] Yim, J; Yoon, J; Lee, B; Hwang, B. Y. 2014, Designing of Event Decision Module using Twitter, KIPS Spring Conference, 680-683.
[11] Twitter, 2012, The Streaming APIs Twitter Developers, Accessed Sep 24. https://dev.twitter.com/docs/streaming-apis.
[12] Lee, S. 2008, Lucean Korean Morph Analyzer, Accessed Oct 18. http://cafe.naver.com/korlucene.
[13] Republic of Korea National Statistical Office, 2010, Population and Housing Census 2010, http://www.kostat.go.kr.

Received : 2015.01.25
Revised : 2015.09.30
Accepted : 2015.10.13