KIPS Tr. Software and Data Eng.
Vol.4, No.8 pp.301~308 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2015.4.8.301

TRED : Twitter based Realtime Event-location Detector

Junyeob Yim† · Byung-Yeon Hwang††

ABSTRACT

SNS is a web-based online platform service supporting the formation of relations between users. SNS users have usually used a desktop or laptop for this purpose so far. However, the number of SNS users is greatly increasing and their access to the web is improving with the spread of smart phones. They share their daily lives with other users through SNSs. We can detect events if we analyze the contents that are left by SNS users, where the individual acts as a sensor. Such analyses have already been attempted by many researchers. In particular, Twitter is used in related spheres in various ways, because it has structural characteristics suitable for detecting events. However, there is a limitation concerning the detection of events and their locations. Thus, we developed a system that can detect the location immediately based on the district mentioned in Twitter. We tested whether the system can function in real time and evaluated its ability to detect events that occurred in reality. We also tried to improve its detection efficiency by removing noise.

Keywords : Social Network Analysis, Twitter, Natural Language Processing, Realtime Event Detection

트위터 기반의 실시간 이벤트 지역 탐지 시스템

임 준 엽† · 황 병 연††

요 약

SNS는 사용자들의 관계 형성을 도와주는 웹 기반의 온라인 플랫폼 서비스이다. 기존의 SNS 사용자들은 이를 이용하기 위해 주로 데스크톱이나 노트북을 이용하였다. 그러나 최근 스마트폰의 보급으로 인해 웹 접근성이 확대되면서 SNS 사용자가 크게 증가하였다. SNS를 이용하는 사용자들은 주로 자신의 일상이나 경험한 일들을 다른 사용자들과 공유한다. 이때 사용자 개인을 하나의 센서로 가정하고 그들이 남긴 콘텐츠를 분석할 수 있다면, 이를 이용해 현실에서 발생한 이벤트를 탐지할 수 있다. 이러한 시도는 이미 많은 연구에서 진행되고 있다. 특히 트위터의 경우 이벤트 탐지에 적합한 구조적 특징들로 인해 관련 분야에서 다양하게 활용되고 있다. 그러나 대부분이 제한적인 이벤트 탐지 및 이벤트가 발생한 지역을 탐지하는 것에 있어 명확한 한계점을 지니고 있다. 이에 본 논문에서는 트위터에서 언급 빈도가 급증한 지역들을 기반으로 이벤트가 발생한 지역을 실시간으로 탐지하는 TRED(Twitter based Realtime Event-location Detector) 시스템을 제안하였다. 이후 성능평가를 통해 제안하는 시스템이 실시간으로 동작할 수 있는지를 확인하였고, 실제 발생한 이벤트를 탐지함으로써 효율성을 입증하였다. 또한 노이즈 제거를 통한 탐지율 향상을 언급하였으며 향후 보다 높은 성능의 시스템에 대한 가능성을 보였다.

키워드 : 소셜 네트워크 분석, 트위터, 자연어 처리, 실시간 이벤트 탐지

1. 서 론

소셜 네트워크 서비스(Social Network Service; SNS)는 온라인 상의 사용자들이 서로 정보를 공유하고 소통할 수 있도록 도와주는 온라인 플랫폼 서비스이다. 이를 이용하는 사용자들은 자신들이 경험하거나 새로 얻게 된 정보들을 다른 사용자들과 공유함으로써 사회적인 관계망을 형성해나간다[1]. 오늘날 대부분의 SNS들은 웹(Web) 상에서 구현되어 제공된다. 이와 더불어 최근 스마트폰의 도입으로 인한 웹 접근성의 확대로 인해 이용자가 크게 급증하였다. 다양한 통계자료를 제공하고 있는 [2]의 자료에 의하면, 2014년 1월 1일 기준 약 645백만 명에 이르는 사용자가 SNS 중의 하나인 트위터(Twitter)를 이용한다고 조사되었다. 이들이 생산하는 하루 평균 트윗(Tweet) 수는 5,800만 건이며, 이는 매초 약 9,000건에 달하는 양이다. 이러한 추세를 반영하듯 최근 들어 다양한 SNS들이 생겨나고 있으며, 이들을 응용하기 위한 학제적·산업적인 연구들이 다방면으로 진행되고 있다.

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업(No. 2011-0009407)의 연구비와 2015년 가톨릭대학교 교비연구비의 지원으로 수행되었음.
† 준 회 원 : 가톨릭대학교 컴퓨터공학과 석사
†† 종신회원 : 가톨릭대학교 컴퓨터정보공학부 교수
Manuscript Received : April 14, 2015
First Revision : May 26, 2015
Accepted : June 2, 2015
* Corresponding Author : Byung-Yeon Hwang(byhwang@catholic.ac.kr)

그중 트위터는 다른 SNS와는 구별되는 여러 가지 특징들로 인해 여러 연구들이 진행되고 있다. 트위터가 가진 가장 주요한 특징은 개방적인 네트워크 구조이다. 대부분의 SNS에서는 사용자들 간의 관계를 표현하기 위해 다양한 방법을 사용한다. 주된 방법으로는 페이스북이나 인스타그램 등에서와 같은 친구 관계가 있다. 예를 들어, 사용자 A와 사용자 B가 동일한 SNS에서 서로 소통하기를 원한다면 사용자 A와 B는 친구 관계를 맺어야 한다. 이를 위해 친구 관계를 원하는 사용자 A는 사용자 B에게 친구 요청을 하고, 사용자 B는 사용자 A와 친구 관계를 맺기 위해 사용자 A가 보낸 친구 신청을 수락한다. 이와 같은 과정을 거치면 비로소 사용자 A와 B는 친구 관계가 되어 서로가 SNS 상에서 작성한 게시글이나 각종 콘텐츠를 공유할 수 있다. 그러나 트위터는 앞선 방식처럼 상호합의하에 진행되는 것과는 달리 다른 한쪽의 일방적인 요청만으로도 친구 관계 성립이 가능하다. 이를 트위터에서는 팔로어(Follower)-팔로잉(Following) 관계로 표현한다. 팔로어는 자신을 친구로 신청한 사용자를 의미하고, 팔로잉은 친구 신청을 요청한 사용자를 의미한다. 이러한 트위터의 특징은 보다 자유로운 사용자 간의 관계 형성을 제공하며, 이로 인해 보다 넓은 범위의 정보 확산이 이루어진다.

이와 더불어 트위터 사용자들은 140자 제한의 단문 텍스트 서비스인 트윗(Tweet)을 작성함으로써 다른 사용자들과 자신의 의견 및 정보 등을 공유한다. 트윗은 다른 SNS에서 사용자 개인이 남길 수 있는 게시물과 같은 역할을 한다. 그러나 대부분의 경우에는 장문의 텍스트 위주로 작성되는 것이 아닌 짧은 단문의 형태로 작성이 된다. 따라서 사용자로 하여금 비교적 가벼운 내용의 글을 간편하게 작성하도록 유도한다. 또한 트위터를 이용하는 사용자가 특정 정보에 대해 다른 사용자와의 공유를 원할 때 다른 SNS에 비해 보다 빠른 정보 확산이 이루어진다. 이러한 트위터의 특징인 빠르고 넓은 범위의 정보 확산은 트위터를 이용한 많은 연구들의 근간이 된다.

한편 트위터에서 주로 작성되는 트윗들은 사용자들의 일상이나 새로운 경험, 정보 등에 관한 내용이다. 트윗의 내용적 분류로는 [3]에서 언급한 바와 같이 새로운 정보나 뉴스, 개인의 의견이나 감정, 기업의 광고나 홍보, 캠페인 등이 주를 이룬다. 그중 새로운 정보나 개인이 경험한 새로운 사건에 대한 내용들은 이벤트 탐지의 관점에서 볼 때 이벤트 탐지를 위한 하나의 도구로써 활용될 수 있다. 이와 관련된 대다수의 연구들은 주로 특정 사건에 대해 평소보다 트윗 발생량이 많을 때를 이벤트로 탐지해낸다[4]. 예를 들어, 이벤트로 발생될 수 있는 사건들 중 감기의 경우 사회적인 신호로서 이를 탐지할 수 있다[5]. 이를 위해 트위터 내에서 감기와 관련된 트윗들을 수집하여 트윗들이 발생된 위치를 추적한다면, 현재 감기가 성행하는 지리적인 위치를 찾을 수 있다. 만약 해당 지역을 조기에 탐지하여 그에 맞는 대응책을 취한다면 감기 예방 효과를 최대화할 수 있다.

현실에서 발생하는 이벤트들은 대부분이 지리적인 위치를 가진다[6]. 이에 관한 예로는 각종 재난상황이나 지역 축제, 선거 등이 있다. 이때 특정 이벤트가 발생할 경우 이벤트를 경험한 사람들은 자신이 겪은 일을 주변 사람들에게 알리는 경향이 있다. 또한 이벤트 규모에 따라 차이가 있기는 하나 이벤트 경험자 중 트위터 사용자들은 자신이 경험한 일들을 트위터 내에서 다른 사용자들에게 전파시킨다. 따라서 그들이 작성한 콘텐츠를 분석할 수 있다면 각각의 사용자 개인을 이벤트 탐지를 위한 하나의 센서로 활용하는 것이 가능하다. 이에 본 논문에서는 트위터를 이용해 현실에서 발생한 각종 이벤트를 탐지하기 위한 TRED(Twitter based Realtime Event-location Detector) 시스템을 제안한다. 이벤트의 탐지는 이벤트가 발생한 지명을 기반으로 하여 각종 이벤트들의 종류와 상관없이 탐지할 수 있도록 설계하였다. 또한 시스템의 구현이 비교적 용이한 국내를 기준으로 이벤트 탐지를 시도하였다.

논문의 구성은 다음과 같다. 2절에서 본 연구와 관련된 선행 연구들을 살펴본다. 이후 3절에서 제안하는 시스템의 구조와 각 모듈들을 소개하고, 4절에서의 실험을 통해 제안하는 시스템의 성능과 실제 탐지된 이벤트들을 살펴본다. 마지막 5절에서 결론과 향후 연구계획에 대해 기술한다.

2. 관련 연구

[7]은 트위터를 이용하여 이벤트를 탐지하기 위한 Toretter 시스템을 제안하였다. Toretter 시스템은 일본에서 발생한 지진이나 태풍의 지리적인 위치를 실시간으로 탐지해낸다. 논문에서 밝힌 바로는 각종 재난상황에 놓인 트위터 사용자들이 자신이 겪고 있는 재난의 종류와 규모 등을 트윗을 통해 작성하였다. 따라서 미리 입력한 지진이나 태풍에 관한 키워드를 이용해 특정 재난과 관련된 트윗들을 필터링하는 방법을 택하였다. 이러한 과정을 거친 후, 트윗 작성 시 입력되는 트윗의 위치좌표(Geocode)를 이용하여 재난이 발생한 지역을 감지한다. 최종적으로 재난이 발생된 지역이 감지되면, 해당 지역에 위치한 사용자들에게 전자메일을 보낸다. 실험 결과 96%의 지진 감지율을 보였으며 일본의 기상청보다 빠르게 재난지역에 속보를 전파하였다. 이와 유사한 이벤트 탐지 시스템으로 [8]에서 제안한 TEDAS가 있다. TEDAS는 미국에서 발생한 재난이나 범죄에 관한 이벤트를 탐지하기 위해 실시간으로 발생하는 트윗들을 이용하였다. Toretter와 마찬가지로 미리 지정한 키워드와 트윗의 위치좌표를 활용하였으며 최종적으로 이벤트의 종류와 이벤트가 발생한 지역을 탐지하였다.

이러한 방식의 이벤트 탐지는 트위터 사용자 개개인을 하나의 센서로써 활용하여 보다 빠른 이벤트의 탐지가 가능하다. 또한 실시간적인 탐지가 가능하기 때문에 특정 재난에 대한 전파가 효과적으로 이루어진다면 재난에 대한 피해를 최소화할 수 있다. 그러나 사전에 미리 지정된 키워드를 이용했다는 점에서 탐지 가능한 이벤트가 키워드 내용에 국한된다는 한계점이 있다. 이와 더불어 앞서 소개된 연구에서는 트윗의 발생위치를 추적하기 위해 트윗의 위치좌표를 이용하였다. 하지만 트위터는 사용자가 원하지 않는다면 트

윗의 작성위치에 대한 정보를 저장하지 않는 것이 가능하다. 최근 트위터 사용자들이 트윗 작성위치를 공개하는 것에 대해 회의적이라는 점을 고려하면 이러한 트윗 발생위치 추정방식은 명확한 한계가 있다. 또한 트윗의 발생위치를 찾는다 하더라도 [7]에서 언급된 대도시 방향으로의 태풍 이동 경로 왜곡과 같은 문제점들은 개선하기 힘든 한계점으로 볼 수 있다. 한편 [8]에서는 트윗을 남긴 사용자의 프로필 위치를 이용하여 위치좌표가 포함되지 않은 트윗들의 발생위치 추정을 시도하였다. 그러나 정확한 프로필의 위치를 가진 사용자는 전체 사용자 중 12%에 불과하였기 때문에 이를 예측하기 위한 단계가 필요하다는 것을 언급하였다. 이와 같은 방식은 [9]에서도 언급하였으며 결론적으로 프로필 위치를 트윗의 작성위치로 확정하기에는 무리가 있다고 하였다. 따라서 폭넓은 범위의 이벤트 탐지와 이벤트의 발생위치를 탐지하기 위해서는 앞서 언급한 한계점들을 해결하기 위한 연구가 필요하다.

3. 실시간 이벤트 지역 탐지 시스템

본 연구에서는 트위터를 이용해 현실에서 이벤트가 발생한 지역을 실시간으로 탐지하는 시스템을 소개한다. 제안하는 시스템의 전체적인 구조는 Fig. 1과 같다.

3절에서는 Fig. 1의 각각에 해당하는 3개의 모듈을 각 절로 나누어 설명한다. 시스템의 구현과 성능평가는 국내를 기준으로 수행하였다. 만일 설계된 시스템을 다른 국가에서 구현할 경우 3.1절과 3.2절에 관한 부분을 해당 국가에 맞게 변형한다. 이와 관련된 자세한 설명은 해당된 각 절에서 하도록 한다.

3.1 트윗 데이터 수집

전체 시스템의 기본 입력 데이터인 트윗을 수집하기 위해 트위터에서 제공하는 API를 이용한다. 트위터에서는 인증된 사용자들에게 다양한 API들을 제공한다. 그중 스트리밍(Streaming) API[10]를 이용하면 트위터 사용자들이 작성한 트윗을 실시간으로 수집하는 것이 가능하다. 스트리밍 API는 전 세계에서 발생된 트윗들을 대상으로 수집이 이루어진다. 따라서 특정 국가에서 발생한 이벤트를 탐지하기 위해서는 트윗이 발생된 국가에 대한 정보가 필요하다. 트윗이 발생된 국가를 판단하는 방법은 여러 가지가 있다. 본 논문에서는 위치좌표를 사용하지 않고 트윗 내용만을 이용해 트윗이 발생된 국가를 판별한다. 이를 위해 트윗 문장 내에 한국어가 포함된 트윗을 국내에서 발생된 트윗으로 판단하였다. 만일 국내가 아닌 다른 국가에서 설계된 시스템을 구현할 경우 해당 국가에 맞는 언어가 포함된 트윗을 선별한다. 단, 영어와 같이 언어만으로는 트윗이 발생된 국가를 판단하기 어려운 경우가 있는데, 이는 추가적인 국가 판별에 대한 판단기준이 필요하다.

한편 스트리밍 API는 트위터 사(社)와의 별도의 파이어호스(Firehose) 계약이 없을 경우 전체 트윗 발생량의 무작위 1%만을 제공한다. 그러나 제안하는 시스템은 트윗의 수집량과는 관계없이 수집된 트윗들의 각 속성별 비율을 이용한다. 그러므로 트윗을 지속적으로 수집하고 저장할 수 있다면 무료로 제공되는 Streaming API를 이용하더라도 이벤트 탐지가 가능하다.

3.2 트윗 정제

특정 국가에 대한 트윗 수집이 완료되면 해당 국가의 이벤트 지역을 탐지하기 위해 지역과 관련된 트윗들을 추출한다. 이를 위해서는 우선적으로 자연어 처리가 필요하다. 따라서 본 논문에서는 형태소 분석을 위한 루씬 한국어 형태소 분석기[11]를 이용하였다. 형태소 분석을 수행하게 되면 트윗 내의 문장들이 문장성분 태그가 부착된 형태소 단위로 나누어진다. 그중 명사들을 추출하여 이벤트 탐지를 위한 후보 키워드로써 활용한다. 하나의 트윗에는 여러 개의 명사가 있을 수 있으며, 추출된 명사들은 하나의 집합으로써 트윗을 구성한다.

형태소 분석이 완료되면 각각의 트윗이 어느 지역에 관한 트윗인지를 판별한다. 이를 위해 트윗의 키워드 집합 중 지명이 포함된 트윗들을 필터링한다. 지명에 대한 판단기준은

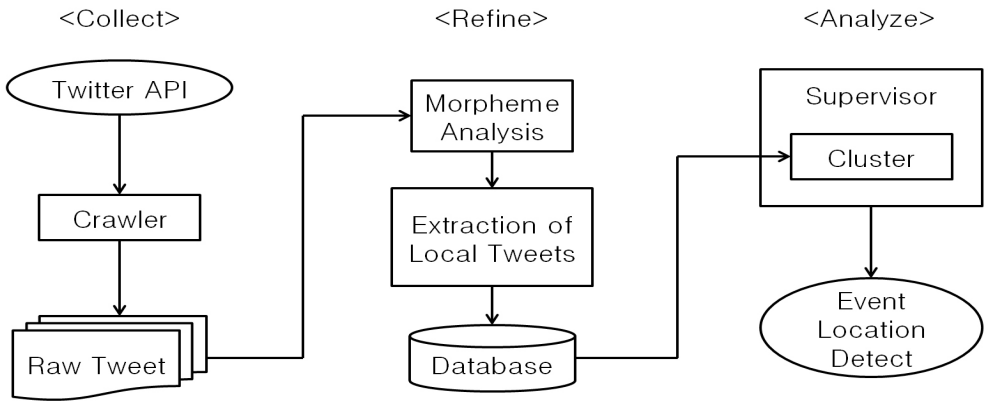


Fig. 1. Structure of System Detecting Events in Real-Time

[12]에서 정리한 행정구역 분류표를 이용하였으며, 각 트윗의 키워드들 중 행정구역명이 포함된 트윗을 지역과 관련된 트윗으로 판단한다.

행정구역 분류는 전체적인 트리 형태로 구성된다. 하나의 트윗 내에 여러 개의 행정구역명이 포함될 경우 상위에 위치한 행정구역의 트윗으로 보았다. 또한 트리 내에서 친척 노드 관계와 같이 연관 없는 두 개 이상의 지역이 함께 트윗에 포함된 경우 순서상 먼저 언급된 지역과 관련된 트윗으로 보았다.

행정구역명이 포함되지 않은 트윗의 경우 시스템에서는 활용하지 못하게 되어 전체적으로는 이벤트의 탐지율을 떨어트리는 요인이 된다. 따라서 이를 개선하고자 랜드마크의 개념으로 지역별 지하철역명을 추가로 이용하였다. 위의 과정을 Fig. 2에 정리하였다.

이 절에서도 3.1절과 마찬가지로 설계된 시스템을 다른 국가에서 구현할 경우 해당 국가 언어에 맞는 형태소 분석기와 행정구역 분류를 이용한다. 단, 작성된 트윗의 언어만으로 국가를 판단하기 어려운 경우가 있다. 이에 관한 판단 기준은 추가적인 연구가 필요하다.

3.3 트윗 분석

최종적으로 필터링 작업이 완료되면 지역별로 트윗을 나누는 클러스터링 작업을 수행한다. 제안하는 시스템은 실시간 시스템이므로 초기 클러스터 형성 후 지속적인 클러스터 스캔을 통해 이벤트를 탐지한다. 클러스터 형성 중에 발생되는 트윗들은 연결된 리스트 형태로 저장하여 시스템이 실행된 후 순서대로 분석한다. 스캔 시 매번 새로 발생된 트윗들과 오래된 트윗들에 대해 클러스터가 갱신되며, 각각의 클러스터인 지역에 대해 Table 1에 제시된 수치 값들을 계산한다.

Table 1에 제시된 수치 값들은 트윗을 실시간으로 분석하기 위해 과거의 트윗 발생량과 비교하기 위한 값이다. 여기서 말하는 각 구간은 40분이며 현재를 기준으로 구간이 나뉜다. TF는 현재로부터 40분 전까지 발생한 트윗들 중 하나의 지역과 관련된 트윗들의 개수를 의미한다. 또한 VT는 동일구간 동안 발생된 트윗들 중 하나의 지역에 대해 중복된 리트윗(Retweet)을 제거한 트윗들의 수를 의미한다. 마지막으로 DA는 현재로부터 2일 전까지 발생한 트윗들 중 하나의 지역과 관련된 트윗들의 평균 개수를 의미한다.

Table 1. Numeral Values Used for Detecting Twitter Events

수치명	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

예를 들어, Fig. 3과 같이 A라는 지역이 있고, A지역을 언급한 트윗이 발생된 시간별로 나누어져 있다. 이때 TF는 현재로부터 40분 전까지 A지역을 언급한 트윗 개수이므로 4이다. 또한 VT는 TF에서 중복을 제거한 것이므로 3이다. 마지막으로 DA는 2일 전까지 A지역을 언급한 트윗 개수 12를 구간 개수인 72로 나눈 값이므로 약 0.166이다.

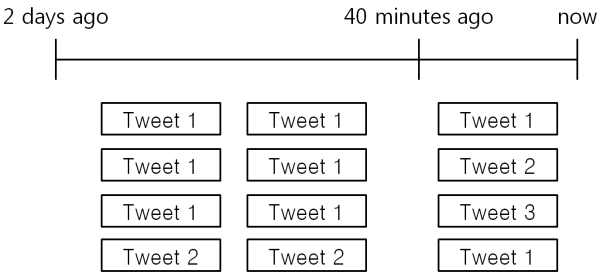


Fig. 3. Example of Tweets Mentioning Place A

각 지역별로 이벤트 탐지를 위한 각각의 수치 값들이 계산되면 시스템의 슈퍼바이저(Supervisor)가 모든 클러스터를 스캔하여 이벤트 후보지역들을 선별한 후 최종 이벤트로 결정할 지역들을 결정한다. 우선 이벤트 후보지역을 선별하기

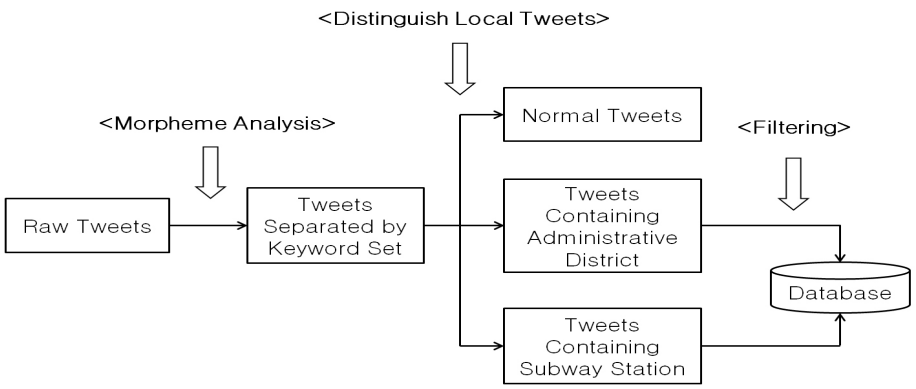


Fig. 2. The Process of Distinguishing Local Tweets

위해 각 지역별로 계산된 TF값과 DA값을 이용한다. 현실에서 이벤트가 발생할 경우 이벤트가 발생된 지역을 언급하는 트윗이 평소보다 증가하는 추세를 보인다[13]. 따라서 트윗 발생량이 평소보다 많은 지역을 이벤트 후보지역으로 선정한다. 이때 트윗 발생량이 비교적 불규칙적인 지역들이 있을 수 있다. 이러한 지역들은 이벤트의 발생 여부와 관계없이 이벤트 후보지역에 포함될 수 있다. 그러한 지역들을 줄이기 위해 트윗 증가량이 최소 p보다 높은 지역들을 이벤트 후보지역으로 선정한다. 본 연구에서는 p를 10으로 정하고 실험을 하였으나, 설계된 시스템에서 어느 정도 규모의 이벤트를 측정하느냐에 따라 10이라는 값은 변동될 수 있다. 단, 지정된 값을 너무 높게 설정할 경우 이벤트의 탐지가 늦어지는 역효과가 있을 수 있으므로 적절한 값을 조율하는 것이 필요하다. Equation (1)은 n개의 지역에 대해 각 지역들의 트윗 증가량을 계산하는 과정이다. 시스템에서는 계산 결과 Candidate(k)에 해당하는 지역들을 이벤트 후보지역으로 선정하였다.

$$Candidate(k) = \{k:\ TF_k - DA_k > p\} \quad (1)$$

마지막으로 선별된 후보지역들 중 최종 이벤트 지역으로 결정할 지역들을 선별한다. 이벤트 후보지역들 중 실제 이벤트가 발생되지 않은 지역의 경우 순간적으로 급증한 중복된 트윗으로 인해 이벤트 후보지역에 포함된 경우가 많다. 따라서 그러한 지역들을 판별하기 위해 현재 구간에서 발생된 트윗들이 얼마나 다양하게 발생되었는지를 고려한다. 이를 위한 판단기준으로 VT값과 DA값을 비교하여 VT가 DA보다 큰 지역들을 추출하여 최종 이벤트 지역으로 결정한다. Equation (2)는 n개의 이벤트 후보지역들의 트윗의 다양성을 계산하는 과정이다. 계산 결과 Event Decision(k)에 해당하는 지역들을 최종 이벤트 지역으로 결정한다.

$$Event\ Decision(k) = \{k:\ VT_k > DA_k\} \quad (2)$$

4. 실험 및 결과

실험을 위해 2013년 3월부터 2014년 4월까지의 이벤트들을 이용하였다. 이벤트는 동일 기간 KBS[14]에서 속보 및 특보로 다룬 내용들 중 특정 지역에서 일어난 사건들을 기준으로 하였다. 이 절에서는 시스템의 실시간 성능을 평가하고, 각각의 이벤트가 발생한 시간 순서로 다루도록 한다.

4.1 성능 평가

실험에 들어가기에 앞서 설계된 시스템이 평균적으로 어느 정도의 처리량을 가지며 그것이 실시간으로 작동할 수 있는지를 살펴본다. 이를 평가하기 위해 초기 시스템을 실행하는 데 걸리는 시간과 슈퍼바이저가 모든 지역을 1회 스캔하는 데 걸리는 시간을 측정하였다. 실험에 사용된 PC의 성능은 Table 2와 같다.

Table 2. Experimental Environment

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

실험 당시 국내에서 수집할 수 있었던 트윗은 평균적으로 매 시간 약 60,000개였다. Fig. 4와 같이 총 10회의 실험 측정 결과, 시스템이 실시간으로 작동하기 위해 소요된 시간은 평균 22분 7초였다. 시스템이 처음 작동될 때에는 2일간의 트윗을 분석해야 하기 때문에 비교적 많은 시간이 소요되었다. 그러나 시스템이 작동된 후 모든 지역을 스캔하는 데 소요된 시간은 Fig. 5와 같이 평균 0.277초에 불과했다. 따라서 분석을 위한 초기화 작업이 수행되고 나면 이후 실시간적인 분석이 충분히 가능함을 알 수 있다.

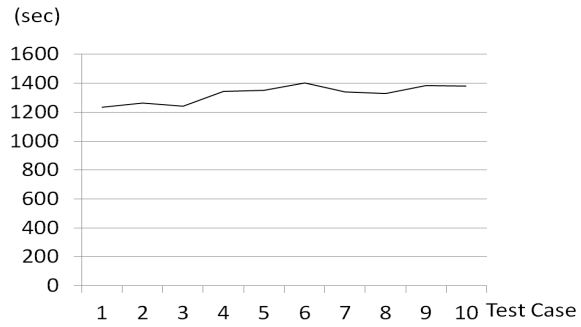


Fig. 4. The Time Required for the Starting Up of the System

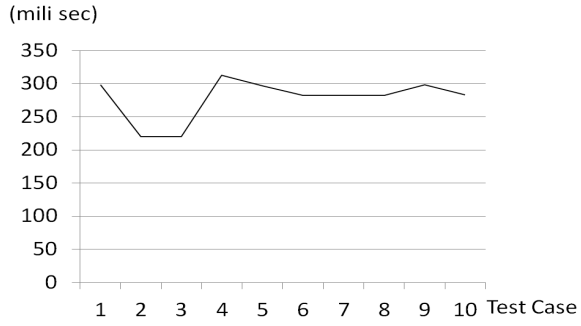


Fig. 5. The Time Required to Scan All Clusters Once

4.2 이벤트 분석

이 절에서는 제안하는 시스템이 실제 이벤트를 탐지할 수 있는가에 관해 실험을 하였다. [14]에서 발표한 특정 지역과 관련된 사건은 총 5개가 있었으며, 그중 4개를 탐지하였다. Fig. 6-9는 탐지된 4개의 타깃 이벤트에 대해 이벤트가 발생한 시간을 기점으로 4시간 동안의 각 지역별 수치 값에 대한 그래프이다.

Fig. 6은 2013년 3월 9일 15시 50분경 포항에서 산불 사

고가 발생했을 당시 포항지역 TF, DA, VT값들의 변화 추이 그래프이다. 시스템에서 최초로 탐지한 시각은 사건 발생 후 26분 후이었으며 탐지된 이벤트들 중 가장 빨랐다. 이는 포항에 관한 기존 트윗 발생량이 비교적 적었기 때문에, 이벤트 발생으로 증가하게 된 트윗 양이 즉각적으로 시스템에 반영된 것으로 판단된다. 또한 탐지된 다른 타깃 이벤트들과는 다르게 이벤트의 성격상 이벤트를 실제로 관찰한 사용자들이 상대적으로 많았다. 따라서 많은 사용자들이 트윗을 직접 작성할 수 있었고, 초기에 많은 트윗 양을 확보하여 다른 타깃 이벤트들보다 빠른 탐지가 가능했다.

Fig. 7은 대구역에서 발생한 KTX 열차 탈선 사고 시 대구역에 관한 트윗의 변화 추이이다. Fig. 7의 경우 다른 실험 결과와는 다르게 비교적 완만한 증가를 보였다. 이는 동일한 사건에 대한 키워드의 발생이 대구역뿐만이 아닌 대구, 동대구와 같이 유사한 키워드들로 분산되어 발생한 것으로 판단된다.

Fig. 8은 경주에서 발생한 마우나 리조트 붕괴사고 당시 트윗의 변화 추이이다. 이는 다른 이벤트들과는 비교적 다른 전파구조를 보였다. 우선 이벤트 발생 직후 이벤트 경험자들로부터 1차적인 트윗이 발생했다. 이는 VT값의 수치를 보면 알 수 있다. 이후 트윗 발생량이 줄어들다가 다시 한 번 크게 치솟았다. 이때의 VT값은 반대로 줄어드는 것을 볼 수 있는데 이는 기존에 작성된 트윗들이 리트윗 되기 시작하는 시점으로 해석된다. 또한 뒤늦게 작성된 기사 관련 트윗들도 리트윗 되어 다른 타깃 이벤트들보다 가장 높은 트윗이 발생되었다.

Fig. 9는 진도 주변에서 해상 여객선이 침몰된 사고이다. 이는 타깃 이벤트들 중 가장 늦게 탐지되었다. 다른 지역의 경우 이벤트를 경험하거나 관찰한 사용자가 직접 올린 트윗 메시지를 통해 전달되는 것이 일반적이다. 그러나 해상 여객선의 경우 여객선 내의 원활하지 않은 인터넷 사용과 주변 관찰자가 없었다는 점이 크게 작용했다. 따라서 최초로 발생한 진도 해상 여객선 침몰 관련 트윗 분석 결과, 인터넷 기사로 보도된 자료를 읽은 트위터 사용자들이 서로에게 전파시킴으로써 탐지가 된 것이다. 이러한 이벤트 탐지상의 특징은 본 연구의 한계점으로 제시될 수 있다. 다시 말해 하나의 센서 역할을 하는 트위터 사용자가 적은 지역은 빠른 이벤트 탐지가 어렵다.

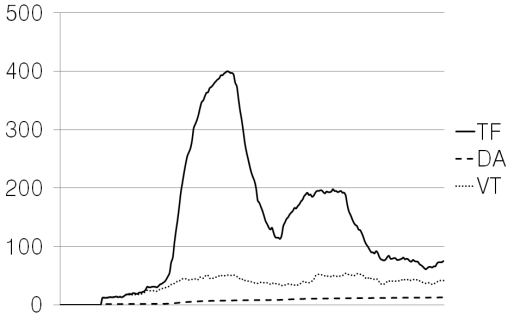


Fig. 6. Pohang at 15:50 on March 9, 2013

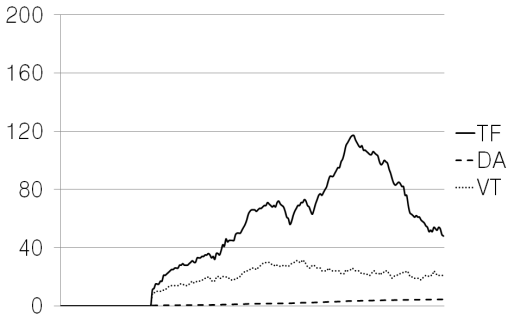


Fig. 7. Daegu Station at 7:15 on August 31, 2013

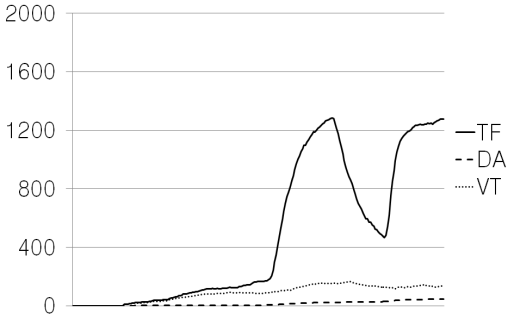


Fig. 8. Kyungju at 21:16 on February 17, 2014

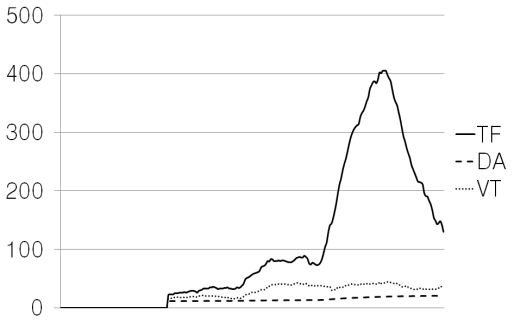


Fig. 9. Jindo at 8:55 on April 16, 2014

Table 3. Comparison of Times of Detection and News Report of Events

	포항 (산불)	대구역 (KTX 추돌 사고)	경주 (마우나 리조트 붕괴)	진도 (해상 여객선 침몰)
발생 시각	2013/03/09 15:50	2013/08/31 07:15	2014/02/17 21:16	2014/04/16 08:55
최초 탐지 시각	2013/03/09 16:16	2013/08/31 08:12	2014/02/17 21:49	2014/04/16 10:02
최초 보도 시각	2013/03/10 02:51	2013/08/31 10:26	2014/02/17 22:41	2014/04/16 10:04

Table 3은 이벤트가 발생한 시각과 제안된 시스템에서 이벤트를 탐지한 시각, 최초로 뉴스가 보도된 시각을 정리한

표이다. 탐지된 모든 이벤트들에 대해 제안된 시스템의 탐지 시각이 뉴스의 최초 보도 시각보다 빨랐음을 확인할 수 있다. 평균적으로는 이벤트가 발생한 후 약 45분 후에 이벤트를 탐지하였다. 이는 최초 이벤트가 발생한 후 이벤트를 경험 및 관찰한 트위터 사용자들이 트윗을 작성하고, 해당 트윗이 다른 사용자들에 의해 확산되기 위해 필요한 시간으로 해석된다.

뉴스 보도의 경우 특정 사건이 발생하게 되면 해당 이벤트에 관한 속보에 대해 실제 발생 여부 검증과 뉴스 보도를 위한 준비 시간이 필요하다. 따라서 제안하는 시스템과 같이 자동적인 이벤트 탐지가 상용화된다면 탐지 후 즉각적인 문자 메시지, 전자메일 등을 통해 현재보다 빠른 이벤트 전파가 가능해진다. 또한 특정 재난이나 사고에 대한 빠른 속보 전파와 그에 맞는 대처로 2차적인 피해를 최소화하는 데 기여할 수 있을 것으로 기대된다.

탐지된 4개의 이벤트들을 탐지한 시점에 시스템에서는 Table 4와 같이 다른 지역에서의 이벤트도 탐지하였다. Tabel 5를 보면 탐지한 지역의 개수가 서로 다른 것을 확인할 수 있다. 이는 [15]에서 언급한 것과 같이 시간대별로 트윗 발생량이 다르기 때문이었다. 따라서 트윗 발생량이 많은 시간대에는 이벤트가 발생하지 않았음에도 이벤트로 탐지된 지역들이 있었으며, 이러한 노이즈는 트윗 발생량이 많을수록 증가하는 추세를 보였다. 반면에 트윗 발생량이 적은 시간대에는 노이즈도 그만큼 적었던 것을 확인할 수 있었다.

Table 4. Places Detected as an Event with the Target Events

탐지 시간대	시스템에서 탐지한 이벤트 지역
2013/03/09 16:16	울산, 달성, 수원, 동대문, 동해, 포항
2013/08/31 08:12	대구역
2014/02/17 21:49	달성, 신기, 수정, 명동, 경주
2014/04/16 10:02	진도, 안산

한편 2013년 3월 9일 16시 16분에는 포항을 제외하고 울산, 달성, 수원, 동대문, 동해 지역이 탐지되었다. 여기서 울산과 동해는 포항과 지리적으로 인접한 지역으로서 산불에 관한 트윗들로 인해 탐지가 되었다. 동대문의 경우 트위터 사용자 중 한 명이 동대문에서 지갑을 잃어버리는 사소한 사건으로 트위터 내에서 화제가 되어 이벤트로 탐지되었다. 또한 수원은 축구경기에 관한 트윗들이 주를 이루었다.

반면에 달성이라는 지명의 경우 “달성하다”와 동음이의어 관계로 노이즈로 판별되었다. 이러한 노이즈는 2014년 2월 17일 21시 49분에서도 볼 수 있다. 여기서도 마찬가지로 “신기하다”, “수정하다”와 같은 동음이의어로 인한 노이즈가 관찰되었다. 이러한 노이즈는 시스템 성능에도 영향을 미칠 수 있다. 앞서 언급한 5개의 이벤트 중 탐지하지 못한 1개의 이벤트가 삼성동에서 헬기가 추락한 사고였다. 이 경우도 같은 종류의 노이즈로서 기존에 이미 “삼성”이라는 단어가 많이 발생되어 헬기 추락으로 인한 트윗의 증가가 미비한 것으로 판단된다. 따라서 이러한 동음이의어적 특징을 노이즈로 간주하고, 이를 제거할 수 있는 방안이 필요하다.

마지막으로 본 논문의 실험에서는 데이터의 특성상 재난과 관련된 이벤트가 주를 이루었다. 그러나 실험에서 다루지 않은 축제나 각종 새로운 지역 관련 정보들도 트위터 상에서 화제가 된다면 해당 지역을 탐지할 수 있다.

5. 결론 및 향후 연구 계획

본 논문에서는 트위터를 이용하여 이벤트가 발생한 지역을 실시간으로 탐지하는 시스템을 설계 및 구현하였다. 성능 평가 결과, 제안하는 시스템이 실시간으로 작동될 수 있음을 보였으며, 탐지된 모든 이벤트들이 뉴스의 보도보다 이른 시점에 탐지되었다. 이는 제안하는 시스템이 실제 상황에서 새로운 이벤트 탐지도구로써 활용될 수 있음을 의미한다. 또한 특정 시간에 다수의 이벤트를 실시간으로 탐지하기 때문에 각종 재난이나 사고에 대해 보다 빠른 초동대처의 가능성을 제시하였다.

향후 연구 과제로는 시스템의 성능 향상을 위해 가장 시급한 동음이의어 관계에 관한 노이즈 제거 연구가 있다. 또한 본 논문에서 언급하지 않았던 이벤트 발생 지역의 이벤트 내용을 알아내기 위한 방안과 탐지된 이벤트를 어떻게 전파할 것인가에 대한 방안을 연구할 계획이다. 한편 시스템의 한계점으로 언급되었던 내용 중 트위터 사용자가 없는 지역의 경우 조기에 탐지하기 어렵다는 단점이 있었다. 이와 더불어 미리 입력되지 않은 행정구역상의 지역은 이벤트가 발생하더라도 탐지할 수 없었다. 이는 보다 많은 위치 관련 데이터 구축을 통해 해결해나갈 계획이다.

References

[1] L. Barbosa and J. Feng, “Robust Sentiment Detection on Twitter from Biased and Noisy Data,” *Proc. of the 23rd International Conference on Computational Linguistics*, pp.36-44, 2010.

[2] Statistic Brain, “Twitter Statistics,” [Internet] http://www.statisticbrain. com/, 2014.

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

[4] B. Lee and B. Hwang, “A Study of the Correlation between the Spatial Attributes on Twitter,” *Proc. of the IEEE 28th International Conference on Data Engineering Workshop on Spatio-Temporal Data Integration and Retrieval*, pp.337-340, 2012.

[5] B. Lee, J. Yoon, S. Kim, and B. Hwang, “Detecting Social Signals of Flu Symptoms,” *Proc. of the 8th IEEE International Conference on Collaborative Computing:*

Networking, Applications and Worksharing, pp.544-545, 2012.

[6] J. Lee, S. Bengio, S. Kim, G. Lebanon, and Y. Singer, "Local Collaborative Ranking," *Proc. of the 23th International Conference on World Wide Web*, pp.85-95, 2014.

[7] T. Sakaki, M. Okzaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," *Proc. of the 19th International Conference on World Wide Web*, pp.851-860, 2010.

[8] R. Li, K. H. Lei, R. Khadiwala, and K. Chang, "TEDAS: a Twitter Based Event Detection and Analysis System," *Proc. of the IEEE 28th International Conference on Data Engineering*, pp.1273-1276. 2012.

[9] B. Lee, S. Kim, and B. 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.

[10] Twitter, "The Streaming APIs | Twitter Developers," [Internet] https://dev.twitter.com/docs/streaming-apis, 2014.

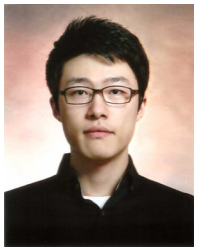
[11] S. Lee, "Lucean Korean Morph Analyzer," [Internet] http://cafe.naver.com/korlucene, 2014.

[12] Republic of Korea National Statistical Office, "Population and Housing Census 2010," [Internet] http://www.kostat.go.kr, 2010.

[13] P. Oh, J. Yim, J. Yoon, and B. Hwang, "Unspecified Event Detection System based on Contextual Location Name on Twitter," *KIPS Transactions on Software and Data Engineering*, Vol.3, No.9, pp.341-348, 2014.

[14] KBS, "24h Newsflash," [Internet] http://www.kbs.co.kr/.

[15] B. Lee, "A Temporal Analysis of Posting Behavior in Social Media Streams," *Proc. of the 6th International AAAI Conference on Weblogs and Social Media Workshop on Social Media Visualization*, pp.18-21, 2012.



임 준 엽

e-mail : junyeob1205@naver.com
2013년 가톨릭대학교 컴퓨터공학과(학사)
2015년 가톨릭대학교 컴퓨터공학과 석사
관심분야 : 소셜네트워크 분석, 데이터베이스, 데이터마이닝, 정보검색

황 병 연

e-mail : byhwang@catholic.ac.kr
1986년 서울대학교 컴퓨터공학과(학사)
1989년 KAIST 전산학과(석사)
1994년 KAIST 전산학과(박사)
1994년~현 재 가톨릭대학교 컴퓨터정보공학부 교수
1999년~2000년 (美) 미네소타대학교 방문교수
2007년~2008년 (美) 캘리포니아주립대학교 방문교수
관심분야 : 소셜네트워크 분석, XML 데이터베이스, 정보검색, 데이터마이닝, 지리정보시스템