KIPS Tr. Software and Data Eng.
Vol.5, No.7 pp.345~350 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.7.345

Keyword Filtering about Disaster and the Method of Detecting Area in Detecting Real-Time Event Using Twitter

Hyunsoo Ha† · Byung-Yeon Hwang††

ABSTRACT

This research suggests the keyword filtering about disaster and the method of detecting area in real-time event detecting system by analyzing contents of twitter. The diffusion of smart-mobile has lead to a fast spread of SNS and nowadays, various researches based on studying SNS are being processed. Among SNS, the twitter has a characteristic of fast diffusion since it is written in 140 words of short paragraph. Therefore, the tweets that are written by twitter users are able to perform a role of sensor. By using these features the research has been constructed which detects the events that have been occurred. However, people became reluctant to open their information of location because it is reported that private information leakage are increasing. Also, problems associated with accuracy are occurred in process of analyzing the tweet contents that do not follow the spelling rule. Therefore, additional designing keyword filtering and the method of area detection on detecting real-time event process were required in order to develop the accuracy. This research suggests the method of keyword filtering about disaster and two methods of detecting area. One is the method of removing area noise which removes the noise that occurred in the local name words. And the other one is the method of determinating the area which confirms local name words by using landmarks. By applying the method of keyword filtering about disaster and two methods of detecting area, the accuracy has improved. It has improved 49% to 78% by using the method of removing area noise and the other accuracy has improved 49% to 89% by using the method of determinating the area.

Keywords : Twitter, Real-Time Event Detect, Detecting Area, Keyword Filtering

트위터를 활용한 실시간 이벤트 탐지에서의 재난 키워드 필터링과 지명 검출 기법

하 현 수† · 황 병 연††

요 약

본 논문에서는 트위터를 활용하여 이벤트를 실시간으로 탐지하는 시스템에서의 재난 키워드 필터링과 지명 검출 기법을 제안한다. 스마트폰의 보급이 SNS의 빠른 확산을 이끌었고, 최근 SNS를 활용하여 다양한 연구들이 진행되고 있다. SNS 중에서 트위터는 140자의 단문으로 작성되어 빠르게 확산되는 특성을 가지고 있다. 따라서 트윗 사용자들이 작성하는 트윗은 하나의 센서 역할을 수행할 수 있다. 이러한 특성들을 이용하여 발생한 이벤트를 탐지하는 연구가 진행되었다. 그러나 최근 개인 정보 유출 사례가 증가해 자신의 위치 정보를 공개하기 꺼려함에 따라 재난이 발생한 지역을 파악하는데 어려움이 있다. 또한 맞춤법을 따르지 않은 게시글의 내용을 분석하는 과정에서 정확성과 관련된 문제가 발생한다. 따라서 이벤트 발생 탐지 과정에 재난 관련 키워드 필터링과 지명 검출 기법이 추가적으로 적용되어야 한다. 본 논문에서는 재난 관련 키워드 필터링의 적용과 두 가지 지명 검출 기법을 제안한다. 지명을 검출하는 두 가지 기법은 지명 단어에서 발생되는 노이즈를 제거하는 지명 노이즈 제거 기법과 랜드 마크를 이용하여 지명 단어를 확정하는 지명 확정 기법이다. 재난 관련 키워드와 두 지명 검출 기법을 적용한 결과 기존 시스템의 정확도 49%에서 지명노이즈 제거기법은 78%, 지명확정기법은 89%로 향상되었다.

키워드 : 트위터, 실시간 이벤트 탐지, 지명 검출, 키워드 필터링

1. 서 론

최근 들어 SNS의 영향력은 점차 확대되고 있다. 다양한 SNS들 중에서 트위터는 최대 140자의 단문 텍스트 업로드를 제공한다. 따라서 게시글을 단기간에 작성하여 정보의 빠른 확산이 가능하다. 또한 팔로잉-팔로워 구조로 이루어져 있어서 개방적인 네트워크를 형성하고 있다. 팔로잉-팔로워 구조는 트위터 이용자가 일방적으로 트위터 상대방에게 요청을 보내면 상대방의 콘텐츠를 공유하게 되는 구조이다. 따라서 트위터 이용자들이 자신이 경험한 일들을 트윗으로 작성하면 다른 이용자들과 작성된 트윗을 공유하게 된다.

이러한 트위터의 빠른 정보 전파력을 이용하여 트위터의

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업(No. 2011-0009407)의 연구비와 2016년 가톨릭대학교 교비연구비의 지원으로 수행되었음.
† 준 회 원 : 가톨릭대학교 컴퓨터공학과 학부생
†† 종신회원 : 가톨릭대학교 컴퓨터정보공학부 교수
Manuscript Received : January 26, 2016
First Revision : May 9, 2016
Accepted : May 10, 2016
* Corresponding Author : Byung-Yeon Hwang(byhwang@catholic.ac.kr)

각 이용자들을 하나의 센서로 판단하고, 이용자가 작성하는 트윗 내용을 분석하여 이벤트를 탐지하는 TRED(Twitter Based Realtime Event-Location Detector) 시스템[1]을 구축하였다. 그러나 이벤트가 발생된 지역을 검출하는 과정에서 낮은 정확도를 보이는 문제점이 발견되었다. 즉, 실제 이벤트가 발생하지 않은 지역을 검출하는 사례가 발생하였다. 첫 번째 원인은 이벤트의 정확한 범위 지정에 관련된 문제점이었다. 이벤트의 정확한 범위 지정은 특정 이벤트 관련한 키워드를 지정하여 해결하였다. 본 논문에서는 특정 이벤트를 재해, 사건, 사고를 포함한 재난으로 정하였다.

두 번째 원인은 트위터를 비롯한 SNS 이용자들이 게시글을 작성할 때 맞춤법과 띄어쓰기를 제대로 지키지 않는 문제점이다. 이는 트윗의 내용을 분석하는 과정에서 실제 지명이 아닌 단어들을 지명으로 판단하여 검출하는 결과를 초래했다. 결국 이벤트 탐지에서도 오류가 발생했다. 또한 최근 개인 정보 유출 사건이 자주 발생함에 따라 SNS 이용자들이 개인 정보에 민감해졌다. 그에 따라 SNS 이용자들이 게시글에 자신의 위치 정보를 태그하는 경우가 현저히 감소했다. 결과적으로 트윗 내용을 분석하는 방법의 차선책으로 연구한 게시글에 태그된 위치 정보를 이용하여 이벤트 발생 지명을 파악하는 방법의 결과가 기대치에 미치지 못했다.

따라서 본 논문에서는 이벤트를 탐지하는 TRED 시스템의 키워드 추출 과정과 지명 검출 과정에 추가적인 알고리즘을 설계하여 제안한다. 우선 재난 이벤트 관련 키워드 필터링 과정을 위해 트위터에서 실제로 사용되어지는 재난 관련 단어를 수집하고 분류과정을 거친 뒤 데이터베이스에 저장하였다. 다음으로 정확한 지명 검출을 위해 지명 단어의 노이즈 유형에 따라 지명 노이즈 제거 기법과 지명 확정 기법, 두 가지 기법을 고안했다. 두 가지 지명 검출 기법을 기존의 시스템에 적용하여 탐지율과 정확도를 기준으로 두 기법의 결과를 비교해보았다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해서 소개하고, 3장에서는 전체적인 시스템의 흐름과 키워드 필터링 그리고 지명 검출 기법에 대해서 살펴본다. 이후 4장에서 실험 데이터를 통해 키워드 필터링과 지명 검출 기법을 적용한 실험결과를 보이고 5장에서 결론과 향후 계획을 설명한다.

2. 관련 연구

[1]에서 제안하는 TRED 시스템은 트위터 사용자를 이벤트 탐지의 센서로 이용한 실시간 이벤트 탐지 시스템이며 [2]와 비슷하게 실시간으로 트윗을 수집하여 정제하는 과정을 거친다. TRED 시스템은 [2]와 다르게 TF(Term Frequency), VT(Variety of Tweets), DA(Document Average) 수식을 이용하여 평소보다 자주 언급되는 지명에서 이벤트가 발생했다고 판단하는 알고리즘을 적용시켰다. 그러나 지명에 관한 노이즈를 제거하는 과정이 없어 이벤트 탐지 정확도가 낮다. 지명 단어와 형태는 같으나 의미가 다른 동형이의어와 지명을 포함하고 있는 단어에 의해 노이즈가 발생한다. 이에 따라 정확하게 이벤트를 탐지하기 위해서 본 논문에서 제안하는 추가적인 기법들이 필요하다고 판단된다.

[2]에서는 범죄, 사고, 재해관련 이벤트 발생 위치를 탐지하고 이벤트 발생 시간을 표시하는 시스템을 제안하였다. 트윗을 실시간으로 수집하여 이벤트의 발생 위치와 시간 패턴을 분석하는 시스템 과정을 거친다. 또한 GPS 데이터를 이용하여 이벤트 발생 지역을 파악할 수 있다. 그리고 트윗의 반환시간을 계산하여 이벤트 발생 시각을 알 수 있다. 이벤트가 발생한 위치와 시각을 판단하는 근거에 대한 정보를 얻었다.

[3]은 위치와 시간을 특정 주제에 제한시켜 정보를 수집하는 LTT를 사용하여 콘텐츠로부터 시간, 위치 그리고 사회적인 메시지를 얻는 기법을 제안한다. LTT를 이용하여 얻어낸 사회적 메시지는 사회에서 발생하는 이벤트들의 전반적인 흐름을 나타내고 있다. 이를 통해 얻은 정보는 어떠한 이벤트가 어디서 누구에 의해 발생하였는지 파악할 수 있게 한다. 그럼으로써 적합한 큰 영향을 미칠 수 있는 결정을 내리는데 도움을 줄 수 있다. 이 연구를 통해 이벤트 범위 지정에 대한 도움을 받고, 키워드 필터링의 필요성을 인식하였다. 그러나 [2]와 [3]은 한국어로 진행된 연구가 아니기에 한국어가 적용된 연구가 필요하다고 판단하였다.

[4]와 [5]에서는 사전의 뜻풀이 말에서 추출한 통계적 의미정보에 기반한 동형이의어 중의성 해결 시스템을 제안한다. 우선 동형이의어를 포함하고 있는 사전의 뜻풀이 말에서 정확한 의미정보를 추출하기 위해서 사전 뜻풀이말의 유형을 분류하였다. 또한 용언과 체언이 같이 사용되는 경우들을 파악하여 통계적 방법을 이용해 동형이의어의 정확한 의미를 분류한다. 마지막으로 지명 단어와 동형이의어 관계에 있는 노이즈 단어를 분류하는 방안을 제시한다. 그러나 트위터와 같은 SNS의 게시글은 맞춤법을 제대로 따르지 않거나 인터넷 용어를 사용하는 경향이 강하다. 이러한 SNS 게시글의 특성 때문에 기존의 기법으로는 사전적 의미를 분석하기에 어려움이 있다. 따라서 본 논문에서는 SNS의 특성에 맞게 개선된 알고리즘을 지명 검출 기법에 적용하였다.

[6]과 [7]에서는 트위터 내용에서 지명 검출 정확도를 개선하는 방법을 제안한다. 그러나 [6]과 [7]의 기법은 지명 노이즈 필터링 처리를 할 수 있는 데이터양의 부족이라는 한계를 드러냈다. 또한 노이즈는 다양한 측면에서 발생하는데 노이즈 필터링 처리를 위한 규칙을 일반화하여 적용하였다는 한계점이 있다. 마지막으로 이벤트로 판단하는 정확한 범위 지정에 대한 내용이 부족하다.

3. 키워드 필터링 적용과 지명 노이즈 제거 기법

3.1 전체 시스템 흐름도

본 논문에서 제안하는 기법을 적용할 이벤트 탐지 시스템(TRED System)은 Fig. 1과 같이 구축되었다. TRED 시스템은 트윗 수집, 트윗 분석, 이벤트 탐지의 세 단계로 구성된다. 우선 트윗 수집 단계에서는 트위터에서 무료로 제공되는 API를 이용해 트윗을 수집한다[8]. ‘Firehose’라는 계약

을 통해 실시간으로 작성되어지는 모든 트윗을 수집할 수 있다. 그러나 TRED 시스템의 이벤트 탐지 알고리즘은 트윗의 양보다는 트윗의 비율이 중요하게 작용하므로 무료로 제공되는 API를 통해 일부 트윗만을 수집한다. 이후 크롤러(crawler)를 이용하여 한국어가 포함된 트윗을 선별하고 국내에서 발생한 트윗으로 판단하였다.

트윗을 분석하기 위해 수집된 트윗을 루씬 형태소 분석기를 통해 어절 단위로 나눈다[9]. 띄어쓰기를 기준으로 나누어진 어절 단위의 트윗 내용에서 키워드를 추출하고, 이벤트가 발생한 지명을 검출한다. 키워드와 지역이 포함된 트윗은 데이터베이스에 큐 형태로 저장하여 이벤트 탐지 단계에서 활용한다. 수집된 트윗에서 키워드를 추출하는 이유는 TRED 시스템이 이벤트가 발생되었다고 탐지했을 때, 트윗 내용을 전부 읽지 않고, 키워드를 통해 어떤 이벤트가 발생되었는지 알기 위함이다. 마찬가지로 지명을 검출하는 이유도 어디에서 이벤트가 발생하였는지 파악하기 위해서다. 지명 지정의 기준은 통계청에서 제공한 시·군·구 범위의 168개의 행정구역명을 참고하여 결정하였다[10].

마지막으로 이벤트 탐지 단계는 분석 단계에서 저장되었던 트윗을 이용한 이벤트 탐지 및 전파 과정을 의미한다. TRED 시스템의 이벤트 탐지 알고리즘을 거쳐 이벤트가 발생되었다고 확정한다. 발생한 이벤트의 시각을 기준으로 10개의 지역의 우선순위를 정하며 순위는 실시간으로 갱신된다. 또한 탐지된 이벤트 내용과 지역을 시스템 이용자들에게 전파한다.

본 논문에서 제안하는 키워드 필터링 기법은 키워드 추출 과정에 추가되었다. 그리고 지명 검출 기법은 트윗 분석의 지명 탐지 과정에 추가 적용되었다.

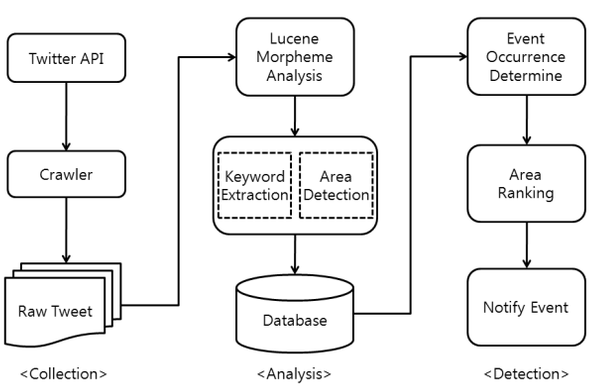

Fig. 1. System Flow Chart

3.2 재난 관련 키워드 필터링 적용

재난 사고 관련 키워드 셋은 트위터 사용자들이 실제 재난 발생 시 현장에서 이용하는 단어를 수집하여 구성하였다. 재난 사고 관련 키워드는 소방방재청이 지정한 재난 카테고리 중에서 자주 발생되는 20종류를 기준으로 하였다. 수집된 재난 관련 키워드의 예시로 '지진', '태풍', '홍수', '메르스', '침몰', '추돌', '붕괴', '산불', '폭발', '운행 정지', '원전', '집회', '시위', '지반 침하' 등을 들 수 있다. 재난 관련 키워드 수집 후 재난의 종류를 기준으로 자연재해, 전염병, 해양사고, 대형교통사고, 건물사고, 화재, 철도 및 전철사고, 원전사고, 범죄, 싱크홀 총 10가지로 세분화시켜 데이터베이스에 저장하였다. 키워드의 세분화는 재난 발생 탐지를 전파할 때 보다 정확한 내용을 전달하기 위함이다.

그러나 재난 관련 키워드에서 노이즈가 발생할 수 있다. 예를 들면 '지진'과 관련하여 '동공지진', '지진희'와 같이 키워드를 포함한 단어가 있다. 각 키워드 별로 생길 수 있는 노이즈 단어를 선별하여 키워드 노이즈 데이터베이스에 저장하였다. 수집된 트윗은 키워드 데이터베이스와 키워드 노이즈 데이터베이스를 거치며 1차적으로 정제된다. 재난 관련 키워드 필터링의 흐름도는 Fig. 2와 같다.

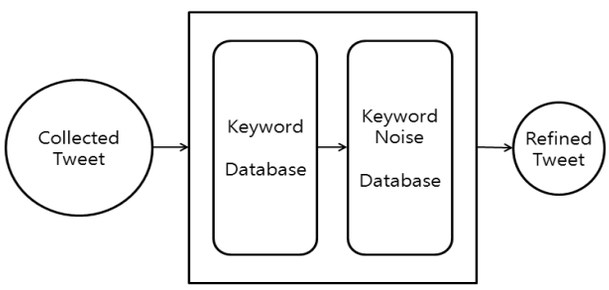

Fig. 2. Keyword Filtering Process

3.3 지명 노이즈 제거 기법

지명 검출 과정에서 발생하는 노이즈는 두 가지로 분류된다. 첫 번째 지명 노이즈는 지명 바로 뒤에 조사가 붙어 동형이의어로 사용되는 경우이다. 동형이의어는 형태는 같으나 의미가 다른 관계에 있는 단어를 의미한다. 예를 들어 '강화'라는 단어가 '강화한', '강화되어'와 같이 조사가 함께 사용되면 '강화하다'라는 의미로 단어가 사용된다. 두 번째 지명 노이즈는 지명을 포함하고 있는 단어로 인해 발생된다. '여주인공'이라는 단어를 지명 '여주'라고 판단하거나, '너구리'라는 단어를 지명 '구리'로 판단해 검출되는 것을 예로 들 수 있다.

이러한 지명 검출 과정에서 생기는 노이즈를 제거해 나가는 기법이 '지명 노이즈 제거기법'이다. 지명 노이즈 제거 기법의 흐름도는 Fig. 3과 같다. 우선 수집된 트윗은 지명이 저장되어 있는 데이터베이스를 거치면서 1차적으로 걸러진다. 그 다음은 여러 조사들이 저장된 데이터베이스를 거친다. 조사 데이터베이스에는 '-을, -를, -이, -가, -한, -된' 등의 조사들이 저장되어 있다. 조사들이 지명 단어 뒤에 붙어있을 경우는 대개 동형이의어 관계에 있는 단어일 확률이 높다. 그러나 '-에서, -에'와 같은 조사는 지명 단어 뒤에 붙어있을 경우 실제 지명을 가리키는 단어일 확률이 매우 높아 예외로 지정했다. 마지막으로 지명이 포함되어 있는 단어들이 저장되어있는 데이터베이스를 거친다. 지명을 포함하고 있는 단어의 모든 사례들은 직접 트윗을 모니터링하고 분석하면서 저장한 것이다.

시간이 다소 오래 걸리며 지속적인 모니터링을 통해 사례들을 데이터베이스에 추가시켜야 하는 단점이 있다. 그러나

지명 노이즈를 정밀하게 제거할 수 있다고 판단된다. 정확한 지명 검출은 이벤트의 탐지율 향상에 영향을 미치기 때문에 노이즈 제거 기법을 적용하였다.

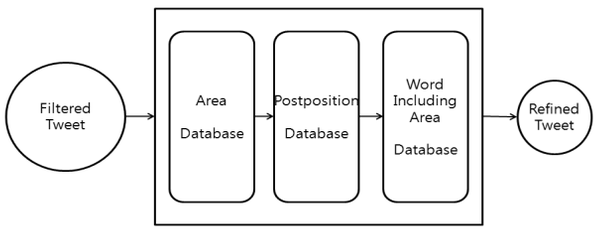

Fig. 3. Method of Removing Area Noise

3.4 지명 확정 기법

지명 노이즈 제거 기법은 실제 지명이 아닌 단어들을 걸러내는 기법이다. 반면에 지명 확정 기법은 랜드 마크 데이터베이스를 통해 실제 지명인 단어들만 검출하여 지명으로 확정한다. 랜드 마크란 지역의 이미지를 대표하는 특이성 있는 시설이나 건물을 의미한다. 이벤트가 발생한 위치를 지명보다 더 자세하게 검출하려는 목표를 두고 고안한 기법이다. 지명 확정 기법의 흐름도는 Fig. 4와 같다.

지명 노이즈 제거 기법과 마찬가지로 수집된 트윗에서 지명 데이터베이스를 거쳐 지명 단어가 포함된 트윗을 남긴다. 그 다음 랜드 마크 데이터베이스를 거친다. 랜드 마크 데이터베이스에는 각 지역의 학교, 건물, 다리, 공항, 지하철역, 항구, 공원, 산, 유적지가 저장되어 있다. 예시로 서울 '송파구'의 '제2 롯데월드', 인천의 '영종대교', 진도의 '팽목항'을 들 수 있다. 각 지역에 있는 모든 학교, 다리, 공항, 지하철역, 항구, 공원, 산, 유적지는 랜드 마크로 저장한다. 그러나 건물의 랜드 마크 기준을 객관적으로 제시하는데 어려움이 있다. 따라서 건물에 관한 랜드 마크는 트윗 상에서 특정 건물이 자주 거론되면 랜드 마크 데이터베이스에 추가 저장하였다. 또한 '주택', '아파트', '상가'와 같이 건물에 관한 추상적인 어휘를 랜드 마크로 이용하여 이벤트를 탐지하고 있다.

지명 확정 기법은 랜드 마크들을 데이터베이스에 저장시키기 위한 시간이 다소 소요된다. 그러나 지명 노이즈 제거 기법보다는 적은 시간이 소요되며, 이벤트가 발생하였을 시에 정확한 위치를 검출할 수 있다. 장기적인 관점에서 랜드 마크 데이터베이스가 다양하고 정확하게 구축된다면, 지명 노이즈 제거 기법보다 탐지율 측면의 성능이 뛰어날 것으로 예측한다.

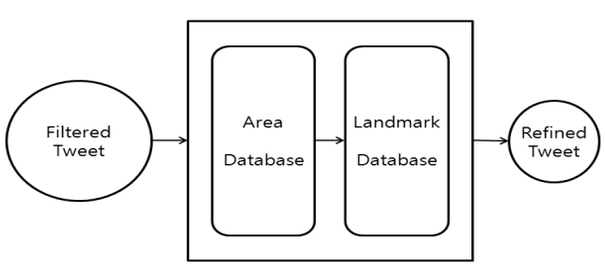

Fig. 4. Method of Determinating Area

4. 실험 결과

4.1 실험 환경

실험에 앞서 시스템이 실시간으로 동작함을 확인하였고 실험에 사용된 PC의 성능은 Table 1과 같다.

Table 1. Experimental Environment

CPU	INTEL(R) Core(TM) 2 Quad
HDD	298 GB
RAM	6.00 GB
OS	Windows 7

지명 검출 기법의 이벤트 탐지여부를 평가할 데이터는 2014년 12월 이후부터 2015년 7월까지 트위터를 통하여 수집하였다. 또한 재난 관련 키워드 필터링을 거쳐 정제된 트윗에 두 기법을 따로 적용하여 결과를 도출하였다.

4.2 지명 노이즈 제거 기법 결과

지명 노이즈 제거기법의 탐지할 대상 이벤트는 실제 발생된 다수의 이벤트 중 네이버 뉴스 속보에서 보도된 지역 관련 이벤트로 정하였다. 자세한 내용은 Table 2와 같다. 속보 2, 3, 4, 5, 7, 8, 10은 지명 노이즈 제거 기법에서 탐지되었다. 속보 2의 '안산', 속보 4의 '사당', 속보 5의 '세종' 세 지역은 지명 단어 노이즈가 많은 지명에 속했다. 그러나 지명 노이즈 제거기법을 이용하여 기존 TRED 시스템에서 보다 정확한 탐지를 가능하게 했다.

Table 2. Detecting Assessment Method of Removing Area Noise

속보	속보 내용	탐지 여부
1	잠실 제2 롯데월드 아쿠아리움 누수	X
2	안산 인질극	O
3	영종대교 다중 추돌사고	O
4	사당종합체육관 공사장 천장 붕괴	O
5	세종시 편의점 괴한 '총기 발사'	O
6	용인 도로공사 현장 붕괴	X
7	신촌 현대백화점 지반 침하	O
8	광화문 세월호 집회	O
9	구로구 아파트 도로 싱크홀	X
10	서울 내곡동 예비군 총기 난사	O

반면에 속보 1, 6, 9는 탐지되지 않았다. 속보 1은 '잠실'이라는 지명이 평소에도 자주 사용되었기 때문에 탐지할 수 없었다. 이벤트 탐지 시스템이 평소보다 지명 언급이 잦을 때를 기준으로 삼고 이벤트 발생으로 탐지하는 알고리즘 구조상 탐지 불가능하였다. 속보 6은 지명 단어 '용인'이 시스템에서 동형이의어로 판단되어 탐지되지 못했다. 트윗 내용을 분석한 결과 띄어쓰기가 제대로 이루어지지 않은 트윗이었다. '용인도로공사 현장 붕괴' 형태로 트윗이 작성되었고

조사 '-도'가 저장되어 있는 조사 데이터베이스를 거치면서 삭제되었다. 속보 9는 속보 1과 마찬가지로 '구로'라는 지명 단어가 평소에도 자주 검출되었기 때문에 이벤트 탐지가 불가했다. 그러나 '잠실'은 실제 지명 단어로 자주 사용되었고, '구로'는 지명을 포함한 단어의 사용이 잦았다. '잠실'은 시스템 알고리즘 구조상의 이유로 탐지에 실패했지만, '구로'는 노이즈 데이터베이스 추가로 탐지 가능하게 할 수 있다.

4.3 지명 확정 기법 결과

지명 확정 기법의 이벤트 탐지 대상은 지명 노이즈 제거 기법의 대상과 동일하다. 자세한 내용은 Table 3과 같다. 지명 확정 기법을 이용해 속보 1, 2, 3, 7, 8, 9는 탐지되었다. 속보 1은 '제2 롯데월드', 속보 3은 '영종대교', 속보 7은 '현대백화점', 속보 8은 '광화문'이 각각 지역의 랜드 마크로 저장되어있었기 때문에 탐지 가능하였다. 속보 2와 속보 9는 건물에 관한 추상적인 어휘인 '주택'과 '아파트'를 사용하여 탐지할 수 있었다. 작성된 트윗을 분석한 결과 속보 2는 '안산 주택가에서 인질극 발생'이라는 트윗이 있었다. 속보 9는 '구로 아파트 인근 도로 앞에서 싱크홀 발생'이라는 트윗이 작성되었다.

Table 3. Detecting Assessment Method Of Determinating Area

속보	속보 내용	탐지 여부
1	잠실 제2 롯데월드 아쿠아리움 누수	O
2	안산 인질극	O
3	영종대교 다중 추돌사고	O
4	사당종합체육관 공사장 천장 붕괴	X
5	세종시 편의점 괴한 '총기 발사'	X
6	용인 도로공사 현장 붕괴	X
7	신촌 현대백화점 지반 침하	O
8	광화문 세월호 집회	O
9	구로구 아파트 도로 싱크홀	O
10	서울 내곡동 예비군 총기 난사	X

그러나 속보 4, 5, 6, 10은 지명 확정 기법을 통해 탐지되지 않았다. 속보 4는 '사당종합체육관'이 기존 랜드 마크 데이터베이스에 저장되어 있지 않았기 때문이다. 속보 5, 6, 10도 마찬가지로 랜드 마크 데이터베이스에 저장되어 있지 않았기 때문에 탐지되지 못했다. 속보 4와 속보 6은 랜드 마크 데이터베이스에 랜드 마크를 추가한다면 탐지가능하다. 속보 5는 건물에 관한 추상적인 어휘 '편의점'을 추가하면 해결할 수 있다고 판단된다. 속보 10은 '군대'와 관련하여 대중에게 공개되어 있는 랜드 마크를 데이터베이스에 저장할 계획이다.

4.4 지명 검출 기법 성능 평가

본 논문에서 제안하는 기법의 성능을 평가할 기준으로 탐지율과 정확도를 제안한다. 탐지율은 실제로 발생한 이벤트의 수와 시스템에서 탐지한 이벤트 수의 비율을 의미한다. 즉 탐지율이 높을수록 시스템에서 탐지할 수 있는 이벤트의 범위가 넓어지는 것이다. 탐지율의 식은 Equation (1)과 같다.

$$\text{탐지율} = \frac{\text{시스템이 제대로 탐지한 이벤트 수}}{\text{실제로 발생한 이벤트 수}} \times 100 \quad (1)$$

정확도는 시스템에서 탐지한 이벤트 수와 탐지된 이벤트 중에서 실제로 발생한 이벤트 수의 비율을 나타낸다. 즉, 정확도는 실제로 발생한 이벤트를 얼마나 탐지했는가에 대한 척도이다. 정확도의 식은 (2)와 같다.

$$\text{정확도} = \frac{\text{실제로 발생한 이벤트수}}{\text{시스템에서 탐지한 이벤트 수}} \times 100 \quad (2)$$

본 논문에서 탐지율 성능을 평가할 기준 데이터는 2014년 12월 이후의 네이버 뉴스 속보로부터 선정한 100개의 이벤트이다[11]. 지명 노이즈 제거 기법은 실제 발생된 속보 이벤트 100개 중에 84개를 탐지하였다. 지명 확정 기법은 80개를 탐지하였다. 기존 시스템에서 98개를 탐지한 점을 감안하면 탐지율은 두 기법을 적용하였을 때 오히려 감소하였다. 그러나 지명 제거 기법은 노이즈 제거 데이터베이스를 구체화 시켜서 실제 지명 단어를 노이즈로 판단하는 상황을 줄일 수 있다. 지명 확정 기법은 랜드 마크 데이터베이스를 확장하여 이벤트 발생 지역 탐지 범위를 더 높일 수 있다고 판단된다. 또한 키워드 필터링 과정에서 키워드를 보완하여 해결할 수 있다고 생각된다.

정확도 성능을 평가할 기준은 지명 노이즈 제거 기법이 이벤트로 탐지한 100개와 지명 확정 기법이 이벤트로 탐지한 100개의 각각 다른 데이터이다. 데이터가 다른 이유는 키워드를 재난 관련 단어로 지정하고 필터링과정을 거치더라도 지명 검출 기법에 따라 탐지되는 이벤트가 다르기 때문이다. 지명 노이즈 제거 기법은 탐지한 100개의 이벤트 중에 78개가 실제로 발생된 이벤트였다. 지명 확정 기법은 탐지한 100개의 이벤트 중에서 89개가 실제로 발생된 이벤트였다. 각 지명 검출 기법에 따른 탐지율과 정확도, 그리고 기존 시스템에서의 탐지율과 정확도는 Fig. 5를 통해 비교할 수 있다. Fig. 5에서 'RAN'는 지명 노이즈 제거 기법, 'DAN'은 지명 확정 기법을 의미하고 'Original'은 기존 시스템을 의미한다.

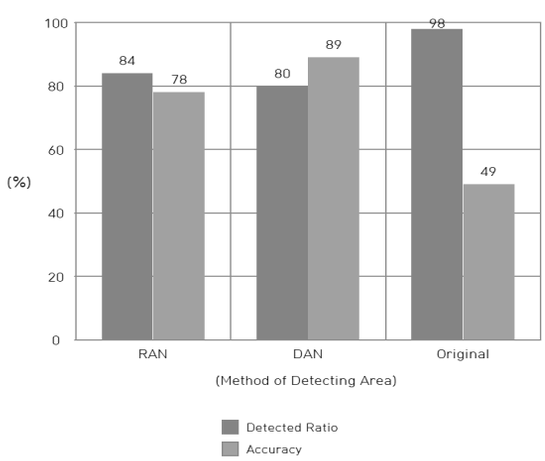

Fig. 5. Detected Ratio and Accuracy of each Method

5. 결론 및 향후 연구

본 논문에서는 트위터를 이용한 이벤트 탐지 시스템의 정확성을 높이기 위해 재난 관련 키워드 필터링과 지명 검출 기법을 제안하였다.

제안된 키워드 필터링과 두 지명 검출 기법을 적용하였을 때 지명 노이즈 제거 기법의 정확도는 기존 시스템 보다 29% 향상되었다. 특히 지명 확정 기법의 정확도는 기존 시스템에 비해 40% 대폭 향상되었다. 지명 확정 기법의 정확도가 100%가 아닌 이유는 랜드 마크로 저장된 일부 단어에서도 노이즈가 발생했기 때문이다.

그러나 두 기법을 적용시켰을 때 기존 시스템보다 탐지율은 하락되었다. 탐지율을 확장시키기 위해서 지명 노이즈 제거 기법은 실제 지명 단어를 노이즈로 판단해 지우는 사례를 줄여야한다. 그리고 지명 확정 기법은 랜드 마크 데이터베이스를 추가적으로 확장해나가면서 탐지율을 확장시킬 수 있다. 랜드 마크는 시시각각 변하는 것이 아니고, 다소 정적으로 변하는 요소임을 감안할 때 탐지율 향상을 기대할 수 있다.

지명 노이즈 제거기법의 탐지율과 지명 확정 기법의 정확도가 결합 가능하다면 최적의 지명 검출 기법이 될 것이라 예상하고 시도해보았다. 그러나 지명 확정 기법의 알고리즘이 지명 노이즈 제거기법의 알고리즘을 무의미하게 만들었다. 지명 노이즈를 제거하더라도 랜드 마크가 트윗 내용에 언급되지 않으면 탐지가 불가능하기 때문이다. 따라서 지명 노이즈 제거 기법에 포함되어 있는 조사 데이터베이스를 부분적으로 지명 확정 기법에 적용할 계획이다.

향후 연구로는 이벤트 탐지 뒤 전파할 방법과 지명 확정 기법에서의 랜드 마크 추가 방법을 찾는 것이다. 전파할 방법은 웹과 스마트폰 애플리케이션을 사용하여 알림기능을 적용시킬 계획이다. 지명 확정 기법의 랜드 마크 데이터베이스의 추가 작업을 Google 지도 API에 등록되어 있는 위치를 이용하여 저장할 수 있는지에 관한 연구를 진행할 것이다. 또한 키워드 필터링부분에서 새로운 가치를 창출할 수 있는 키워드 관련 주제를 찾는 연구도 병행할 계획이다.

References

[1] J. Yim and B. Hwang, “Twitter Based Realtime Event-Location Detector,” *KIPS Transactions on Software and Data Engineering*, Vol.4, No.8, pp.301-308, 2015.

[2] 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.

[3] X. Zhou and L. Chen, “Event Detection over Twitter Social Media Streams,” *The VLDB Journal*, Vol.23, No.3, pp.381-400, 2014.

[4] J. Shin and C. Ock, “A Stage Transition Model for Korean Part-of-Speech and Homograph Tagging,” *Journal of KIISE : Software and Applications*, Vol.39, No.11, pp.889-901, 2012.

[5] J. Hur and C. Ock, “A Homonym Disambiguation System based on Semantic Information Extracted from Dictionary Definitions,” *Journal of KIISE : Software and Applications*, Vol.28, No.9, pp.688-698, 2001.

[6] J. Yim, H. Ha, and B. Hwang, The Method for Removing Noises from Event Detection using Twitter,” *Proc. of KSII Fall Conference*, pp.105-106, 2014.

[7] S. Woo and B. Hwang, “Geographical Name Denoising by Machine Learning of Event Detection Based on Twitter,” *KIPS Transactions on Software and Data Engineering*, Vol. 4, No.10, pp.447-454, 2015.

[8] Twitter Streaming API [Internet], http://dev.twitter.com/docs/streaming-apis.

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

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

[11] Naver Breaking News [internet], http://news.naver.com/main/list.nhn?mode=LSD&mid=sec&sid1=001.

하 현 수

e-mail : hss0924@catholic.ac.kr

2013년～현 재 가톨릭대학교 컴퓨터공학과 학부생

관심분야 : 소셜네트워크분석, 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 : 소셜네트워크분석, XML 데이터베이스, 정보검색, 데이터마이닝, 지리정보시스템