Journal of Korea Multimedia Society Vol. 15, No. 7, July 2012(pp. 910-919)

트위터 사용자가 제공한 위치정보의 신뢰성 분석

이범석[†], 김석중[††], 황병연[†††]

요 약

트위터와 페이스북 같은 소셜 네트워크 서비스가 급격히 성장하면서, 소셜 네트워크 분석에 관련된 연구들도 많은 관심을 받고 있다. 특히 최근에는 트위터 상에 사용자가 관찰한 방대한 양의 정보가 실시간으로 생산된다는 점에 착안하여, 트위터 데이터 분석을 통한 이벤트 감지를 시도하는 연구가 진행되어왔다. 이를 통해 지진 발생을 감지하여 알려주는 시스템이나 지역 축제를 탐지하는 시스템의 개발 등 다양한 연구가 있었다. 그러나 이러한 시스템은 이벤트 발생위치를 탐지할 때 사용자가 제공한 위치정보나 트윗 작성위치를 사용하면서도 그 정확성에 대한 분석은 수행하지 않았다. 본 논문에서는 이벤트 감지 시스템 개발의 사전연구로써, 사용자가 입력한 프로필의 위치정보와 트윗에 포함된 GPS 좌표 사이의 관계와 신뢰성을 분석한다. 이 실험을 위해 52 만개 이상의 국내 사용자 계정과 280 만개 이상의 해외 사용자 계정을 분석하였고, 그 결과 국내 사용자의 경우 49.73%가, 해외 사용자의 경우 90.64%가 프로필 위치에서 주로 트윗을 작성한 것으로 나타났다. 이러한 분석 결과를 통해 사용자 위치정보의 신뢰성 수준을 알 수 있었으며, 이 결과는 추후 트위터의 위치정보를 활용하는 응용을 개발할 때 참고할 수 있을 것으로 기대한다.

Analyzing the Credibility of the Location Information Provided by Twitter Users

Bumsuk Lee[†], Seokjung Kim[††], Byung-Yeon Hwang[†††]

ABSTRACT

We have observed huge success in social network services like Facebook and Twitter, and many researchers have done their analysis on these services. As massive data observed by users is produced on Twitter, many researchers have been conducting research to detect an event on Twitter. Some of them developed a system to detect the earthquakes or to find the local festivals. However, they did not consider the credibility of location information on Twitter although their systems were using the location information. In this paper, we analyze the credibility of the profile location and the correlation between the spatial attributes on Twitter as the preliminary research of the event detection system on Twitter. We analyzed 0.5 million Twitter users in Korea and 2.8 million users around the world. 49.73% of the users in Korea and 90.64% of the users in the world posted tweets in their profile locations. This paper will be helpful to understand the credibility of the spatial attributes on Twitter when the researchers develop an application using them.

Key words: Social Network Analysis(소셜 네트워크 분석), Twitter(트위터), Location Information(위치정보)

※ 교신저자(Corresponding Author) : 황병연, 주소 : 경기도 부천시 원미구 지봉로 43 다솔관 417호(420-743), 전화 : 02)2164-4363, FAX : 02)2164-4777, E-mail : byhwang@catholic.ac.kr
접수일 : 2012년 1월 17일, 수정일 : 2012년 4월 4일
완료일 : 2012년 5월 9일
[†] 정회원, 가톨릭대학교 연구원
(E-mail : bumsuk@catholic.ac.kr)
[††] 준회원, 가톨릭대학교 컴퓨터공학과 석사과정
(E-mail : typicalkorean@catholic.ac.kr)
[†††] 종신회원, 가톨릭대학교 컴퓨터정보공학부 교수
※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업(No. 2011-0009407)의 연구비와 2011년 가톨릭대학교 교비연구비의 지원으로 수행되었음.

1. 서 론

최근 몇 년 동안 트위터와 페이스북 서비스가 큰 성공을 거두면서 소셜 네트워크 서비스에 대한 사회과학 분야 또는 정보검색 분야 연구자들의 관심이 높아지고 있다. 특히 트위터는 단문 텍스트를 기반으로 하는 마이크로블로그(micro-blog) 서비스로써 소형 포터블 스마트 기기에 아주 적합한 구조를 가지고 있다. 트위터는 2011년 3월 일본지진과 같은 현실 세계의 사건에 민감하게 반응할 뿐만 아니라[1], 튀니지의 재스민 혁명을 포함한 중동지역의 민주화 시위를 실시간으로 전 세계에 알리고 여러 국가들의 지지를 이끌어내는데 기여했던 것처럼 현실 세계에 미치는 영향 또한 커지고 있다[2].

트위터는 구조적으로 사용자들 사이의 관계가 명확하고, 사용자가 작성한 트윗에는 개인의 일상, 사회 이슈, 사건 및 사고와 같은 내용과 함께 GPS 좌표 정보도 선택적으로 포함할 수 있기 때문에 이를 이용하여 다양한 위치 기반 응용 시스템을 개발할 수 있다. Takeshi Sakaki 등의 논문[3]과 Meenakshi Nagarajan 등의 논문[4]에서는 트위터 데이터를 분석하여 미리 지정한 이벤트나 지역별 이슈를 감지하는 시스템을 개발하였다. 특히 Takeshi Sakaki 등은 일본에서 발생한 지진을 대상 이벤트(target event)로 설정하고 연구를 진행하였는데, 주목할 것은 일본 내에 스마트기기를 이용하여 트위터에 접속하는 사용자가 많기 때문에 가능한 연구였다는 점이다. 이러한 연구를 진행하기 위해서는 트위터 사용자가 직접 경험하거나 사용자 주변에서 일어나는 사건에 대해 실시간으로 트윗(tweet)을 작성할 다수의 사용자가 존재한다는 조건이 선행되기 때문이다.

동일한 이유에서 한국은 관련 연구를 진행하기 좋은 조건을 가지고 있다. 2010년 여름에 발표된 comScore의 자료[5]를 보면 일본 전체 인구 약 1억 2천만 명 중 약 1천 6백만 명이 트위터를 사용하는데, 가장 최근(2011년 9월 기준)의 조사결과[6]에 따르면, 한국은 전체 인구 약 5천만 명 중 한국인 사용자로 추정할 수 있는 트위터 계정의 수가 약 392만 개에 이른다. 또한, 국제전기통신연합(ITU: International Telecommunication Union)의 2011년 연간 보고서[7]에 따르면 2010년 한국의 가구당 인터넷 보급률이 96.8%로 세계 1위, 인구 100명당 무선 인터넷 접속자 수가 91명으로 세계 1위를 각각 차지하고 있고, 2011년 방송통신위원회 국정감사 자료[8]에 따르면 2011년 7월 국내 스마트폰 가입자가 1,626만 명으로 전체 인구의 33.6%가 스마트폰을 사용하고 있다. 다시 말하면, 트위터 사용자 수가 많고, 무선 인터넷 인프라가 잘 갖추어져 있으며, 스마트폰 보급률이 높은 한국은 트위터 분석을 통한 실시간 이벤트 감지 시스템의 개발에 관련된 연구를 하기에 아주 좋은 환경을 가지고 있다.

Takeshi Sakaki 등의 연구와 Meenakshi Nagarajan등의 연구는 서로 접근 방법이 조금 다르지만, 모두 각 트위터 사용자를 개별적인 센서로 가정하고 사용자가 작성한 트윗과 트윗에 포함된 위치정보를 분석하여 어떤 일이 발생하는지를 추출해냈다는 공통점이 있다. 이처럼 위치정보를 활용한 이벤트 감지 시스템에서는 센서로 가정된 사용자의 위치가 얼마나 잘 정의되느냐에 따라 성능의 차이가 발생할 수 있는데, Takeshi 등의 연구에서는 트윗에 포함된 GPS 좌표와 프로필에 정의된 주소를 이용해 센서의 위치를 지정한 반면, Meenakshi 등의 연구에서는 트위터 사용자를 국가별로 분류를 하였다. 국가별로 위치를 분류할 경우 국민투표나 대형사건과 같이 누구나 알고 있고, 누구나 관심을 가지는 국가적 관심사에 대해서는 추출이 가능하지만, 지역행사나 교통정체, 화재와 같이 특정 지역 사람들이 관심을 가지고 유용하게 사용할 수 있는 정보는 오히려 이벤트 추출 결과에서 제외되는 등의 문제점이 있었다. 또한 프로필에 정의된 주소는 사용자가 마음대로 작성할 수 있기 때문에, 이러한 정보를 이용하여 이벤트 발생지점을 예측할 경우 정확성이 낮아질 수 있다[9,10].

본 논문에서는 트위터 사용자의 프로필 위치 정보와 트윗이 작성된 위치정보 사이의 관계 분석을 통하여, 사용자가 제공한 프로필 위치가 어느 정도 신뢰성을 가지는지 확인하였다. 기존의 이벤트 감지에 대한 연구들은 단순히 사용자의 위치정보를 이용할 뿐 그에 대한 신뢰도가 어느 정도 수준인지 확인하지 않았기 때문에, 본 연구의 결과는 트위터에서 위치정보를 사용하는 응용을 개발할 때 연구의 기반을 제시해 줄 수 있을 것으로 기대한다.

이 실험을 위해 트위터에서 제공하는 Search API를 이용하여 수집한 52 만개 이상의 국내 사용자 계정과 Streaming API를 이용하여 수집한 280만개 이

상의 해외 사용자 계정을 분석하였다. 그 결과 국내 사용자의 경우 49.73%가, 해외 사용자의 경우 90.64%가 프로필 위치에서 주로 트윗을 작성한 것으로 나타났다. 본 논문의 의의는 다음과 같다.

1) 이 논문은 트위터에서 사용자가 제공한 위치정보의 신뢰성을 분석하였다.
2) 국내 트위터 사용자와 해외 사용자의 위치정보에 대한 신뢰성을 비교하였다.
3) 일반적인 거리 기반 클러스터링 기법이 아닌 행정구역 구분을 고려한 클러스터링 방법을 고안하여 실험을 수행하였다.
4) 본 논문의 결과는 트위터의 위치정보를 이용하는 응용을 개발할 때 기반이 되는 연구로써 활용될 수 있다.

논문의 구성은 다음과 같다. 2장에서는 Takeshi 등의 연구와 Meenakshi의 연구를 관련연구로 소개하고 취약점을 지적한다. 3장에서는 트위터 위치정보의 신뢰성 분석을 위해 사용된 데이터의 수집과 정제과정에 대해 소개하고, 4장에서는 분석 방법 및 실험 결과를 제시한다. 마지막으로 5장에서는 이 연구 결과의 의미를 정리하고, 향후 연구계획을 소개한다.

2. 관련연구

2010년에 발표된 Takeshi 등의 연구는 트위터의 가장 중요한 속성 중 하나인 현재성을 활용하여 트위터 데이터가 실시간으로 동작하는 소셜 센서 데이터로써 다루어질 수 있음을 보여주었다. 그들은 지진 감지 시스템인 Toretter를 개발하였는데, 'earthquake'와 'shake'를 검색할 단어로 미리 지정해두고 해당 키워드가 갑자기 빈번해지는 지역을 실시간으로 감지하였다. 이 시스템을 통해 두 가지 중요한 실험을 했는데, 우선 각 단어의 검색 결과에 대해 의미론적 분석을 시도했다. 이 의미론적 분석은 트윗에 언급된 'earthquake'와 'shake'가 지진 때문인지, 아니면 다른 문맥에서 사용된 단어인지를 가려냈다. 이들의 실험결과는 전체적으로 약 80% 이상의 재현율(recall), 60% 이상의 정확도(precision), 70% 이상의 F-Value값을 가졌다. 두 번째 실험으로 해당 트윗들이 작성된 위치정보를 기준으로 Kalman Filter와 Particle Filter를 적용하여 지진이 발생한 진원을 예측하였다. 실험에서는 Particle Filter를 이용한 것이 진원을 예측하는데 좋은 성능을 보여주었다. 그러나 지진보다 넓은 범위를 포함하는 태풍 멜로르(Melor)의 진행경로를 예측한 결과는 그림 1과 같이 인구밀도가 높고 트위터 사용자가 많은 도심지에 가까운 형태로 이동경로가 예측되는 모습을 보였다. 이러한 특징은 위치정보를 사용하는 이벤트 감지 시스템은 넓은 지역을 포괄하는 이벤트보다는 좀 더 작은 지역을 기반으로 하는 이벤트 감지에 더 적합하다는 점을 보여준다.

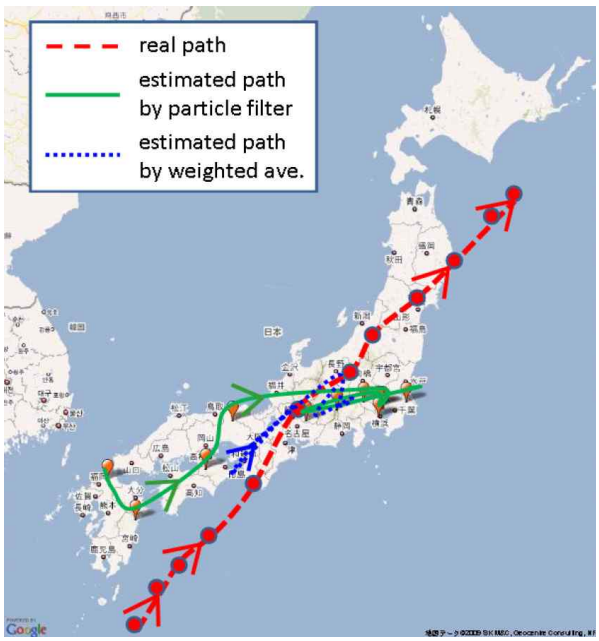

그림 1. Toretter 시스템이 예측한 태풍의 진행경로

Meenakshi 등의 연구는 날짜별, 국가별로 분할한 데이터 (spatio-temporal set)를 대상으로 TF-IDF (term frequency-inverse document frequency) 알고리즘을 적용하여, 시간, 공간, 주제의 세 요소로 구성된 이벤트를 추출하는 Twitris 시스템을 개발하였다. 그림 2는 지역, 시간, 이벤트 주제로 구성된 Twitris의 화면을 보여주는데, 상단 그림은 2009년 10월에 Swine Flu에 대해 언급한 트윗을 추출한 것으로 북미지역에서 관심을 가지고 있음을 보여준다. 하단의 그림은 2011년 3월에 Japan Tsunami에 대해 작성된 트윗을 보여주는데, 전 세계에서 고른 관심을 가지고 있음을 보여준다.

이전에도 블로그와 같은 소셜 미디어에서 이벤트를 감지해내려는 시도가 있었지만, 단문서비스인 트위터의 데이터를 이용한 시도는 이 연구가 처음이었

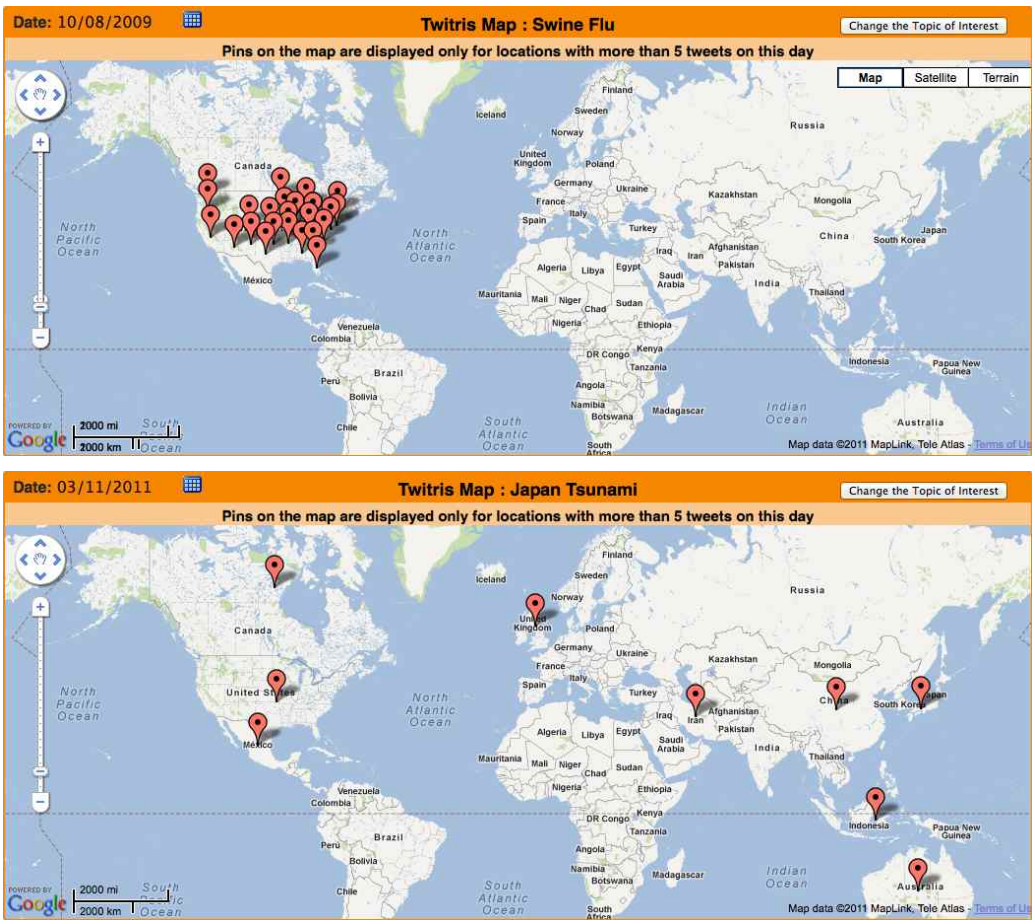

그림 2. 시간, 장소, 키워드를 보여주는 Twitris의 UI

다. Meenakshi 등의 연구는 트위터의 사용자들이 실제 발생한 사건들을 관심있게 다루고 있음을 실제로 확인했다는 점에 의미가 있다. 트위터의 데이터는 블로그 데이터와는 다르게 140자의 단문으로 이루어져 있기 때문에 내용 전체의 맥락을 고려하여 이벤트를 추출해내는 것이 쉽지 않다는 문제점이 있다. 대신 사용자의 프로필에 기본 위치를 설정할 수 있고, 스마트폰을 이용해서 트윗을 작성하게 되면 GPS 좌표를 위치정보로 포함할 수 있다는 점을 고려할 수 있다.

이 밖에도 Lee 등의 연구[11]는 트위터 기반의 이벤트 감지 시스템을 개발하면서 트윗 작성위치가 포함된 트윗만을 고려하였다. 하지만 실제 데이터를 분석해보면 위치정보가 포함된 트윗의 비율이 매우 낮기 때문에 이 데이터만으로는 만족스러운 결과를 얻는 것이 쉽지 않다. 따라서 다수의 트위터 기반의 이벤트 감지 시스템에 대한 연구에서는 충분한 데이터 확보를 위해 사용자 프로필의 위치를 사용하는 것이 필요하다. 프로필 위치정보를 사용한 대표적인 연구는 Takeshi의 연구와 Meenakshi의 연구에서 해당 위치정보의 신뢰성에 대한 분석은 수행되지 않았다. 하지만 이러한 점을 고려하지 않을 경우 이벤트 감지 시스템의 성능에 대한 객관적인 판단이 불가능하기 때문에, 연구 결과의 신뢰성을 확보하기 위해서는 트위터 사용자의 위치정보에 대한 신뢰성 분석이 선행되어야만 한다. 본 연구에서는 이러한 목적을 위해 실제 수집한 방대한 양의 데이터 분석을 통해, 트위터 위치정보의 신뢰성을 조사하였다.

3. 트위터의 위치정보와 데이터 정제

3.1 데이터의 수집

트위터에서 사용자의 위치정보를 수집하기 위해서는 먼저 사용자 계정을 수집해야 한다. 2011년 새롭게 변경된 트위터 API는 트위터 데이터에 대한 애플리케이션의 시간당 접근 횟수를 제한하기 때문에 이 실험에서는 28대의 클라이언트 서버를 이용하여

계정과 트윗을 수집하였다. 데이터의 수집을 위해 그림 3과 같은 단순한 형태의 트위터 사용자 수집기(crawler)를 구현하였는데, 시드(seed)로 입력된 사용자의 팔로워(follower) 정보를 XML 형태로 가져온 다음 데이터베이스에 저장하였다.

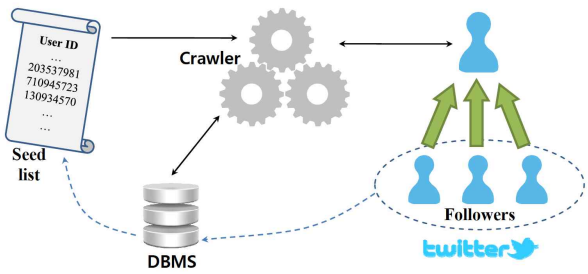

그림 3. 트위터 사용자 수집기

3.2 트위터의 위치정보

이 실험은 트위터에 존재하는 위치정보를 이용한다. 트위터 사용자 정보를 XML 형태로 받아오면 <user> 요소의 하위요소로써 <location> 요소가 있는데, 이 항목의 내용은 30자 이내의 글자 수 내에서 자유롭게 입력할 수 있다. 따라서 트위터를 사용하는 애플리케이션에 따라 GPS 좌표가 포함되어 있기도 하고, 사용자가 직접 입력한 위치정보를 포함하기도 한다. 또한 트위터의 컨텐츠에 해당하는 트윗에는 각 트윗을 작성한 위치를 선택적으로 추가할 수 있는데, <geo>태그에 좌표 형태로 포함되거나 <place>태그에 지명 형태로 포함될 수 있다. 국내 트위터 사용자만을 추려내기 위해 데이터 수집단계에서부터 <location>태그의 GPS 좌표가 대한민국 좌표범위 내에 있거나 한글 유니코드가 하나 이상 들어있는 계정을 별도로 수집하였다.

그림 4는 GPS 좌표를 기준으로 수집된 계정 샘플을 보여준다. <location>태그에 좌표가 입력된 경우는 사용자의 위치를 비교적 정확하게 알 수 있어서 이 실험에서뿐만 아니라 위치정보를 이용한 응용 시스템을 개발할 때 유용하게 활용할 수 있다.

```
USERID | Latitude | Longitude
#14385 | 36.6050974 | 127.5070896
#74425 | 35.2426033 | 128.9010575
#23117 | 37.5570288 | 126.9003614
#68032 | 35.1948194 | 128.0748978
#80298 | 37.565705 | 126.980624
#15451 | 35.127596 | 128.970786
#74346 | 34.9406968 | 127.6958882
#42731 | 33.501169 | 126.530751
```

그림 4. geoCode를 기준으로 수집된 계정 샘플

그림 5는 한글 유니코드가 포함된 계정의 샘플이다. 수집한 사용자의 수는 GPS 좌표를 기준으로 수집한 사용자의 수와 거의 비슷하였지만, 정확한 위치를 입력한 사용자의 비율이 매우 낮아 실험에 사용할 수 있는 데이터의 수가 매우 적었다. 그림 5의 예에서 볼 수 있듯이 1개 이상의 위치를 입력한 사용자들이나 "지구별"과 같이 위치정보로써 사용될 수 없는 무의미한 정보들이 다수를 차지하였다. 실제로 수집한 데이터에서 <location>태그의 내용에 한글 유니코드를 포함한 사용자들 중 약 3%정도에서만 시/도 및 구/군 단위까지 포함한 비교적 정확한 정보를 추출할 수 있었다. 이러한 문제점은 사용자 위치정보를 이용하는 응용시스템을 개발할 때 해결해야 할 어려움 중 하나이다.

```
USERID | LOCATION
#93944 | 대청역 4번출구
#33087 | 서울마포,노원,묵동(목동말구),전라남도빛고을광주
#06833 | 밤엔 잠실3동, 낮엔 정자동, 주말엔 압구정과 이태원
#66311 | 지구별_earth
#68213 | Korea suwon / 수원시 영통구 망포동
#13441 | 아루이 은하, 태양계의 제4성인 지구
#29630 | 계산,성남,분당,부천,남창,부산,대전,대구,광주 당구장
#41720 | Taegu(대구),South Korea
#33776 | Seoul,Korea_반짝반짝 우리지구별:)
#95744 | 연구공원단지 in SNU
#10644 | 용산구 이촌동, Wisconsin
#81578 | ソウル狎?亭/압구정
```

그림 5. 프로필의 사용자 정의 위치정보 샘플

마지막으로 트위터에서 사용할 수 있는 위치정보는 각 트윗에 포함된 위치정보가 있다. 이 정보는 트윗이 작성된 위치를 나타내는데, 주로 스마트폰과 같은 모바일 기기에서 작성할 때 선택적으로 포함이 되지만, 컴퓨터에서 트윗을 작성할 때에도 사용자의 필요에 따라 포함시킬 수 있다. 사용자 프로필의 <location> 정보는 주거지나 회사 위치를 넣는 경우가 많겠지만, 트윗의 위치는 사용자가 트윗을 작성할 때 실제로 머물던 위치를 의미하기 때문에 좀 더 정확한 사용자 위치라고 볼 수 있다. 특히 그림 6의 예에서 트윗 내용에서 지칭한 위치정보와 그 트윗에 포함된 GPS 좌표가 거의 대부분 일치함을 알 수 있다.

3.3 데이터 정제 및 정렬

트위터 사용자가 제공한 위치정보의 신뢰성을 분석하기 위해 프로필의 위치정보와 트윗 작성위치를

TWEETID	USERID	TWEET	LATITUDE	LONGITUDE
#798000	M1897	신림사거리 GS문고 앞에서 나쁜투표거부 캠페	37.482643	126.929772
#988000	M4805	밥 다먹고 찍기 (@ 무스쿠스 잠실점) http://t.co/	37.514197	127.100836
#339000	M8766	그림자...#instar @ 물치해수욕장 http://t.co/d	38.154125	128.60784
#908000	M8057	황태찜먹으러 ^^ (@ 황태희관) http://t.co/dFZh	37.671778	128.709368
#247000	M2755	지오캣에 발도장 쿡! 고양이짱많아 ㅋㅋㅋ @im	37.55431387	126.92337331
#024000	M6483	왕산 해수욕장 도착! (@ 왕산해수욕장 w/ 2 oth	37.455903	126.369485
#871000	M6072	I'm at 송도 센트럴공원 http://t.co/azbPckB	37.39244224	126.63869619
#663000	M7644	물 맑고 경치좋고 사람적당히있으며 맛있는 음	38.327714	128.52817
#001000	M2859	석미용실침산점에 발도장 쿡! 머리 펌 하러 ㅋ =	35.88946233	128.59033006
#672000	M8215	I'm at STARBUCKS COFFEE (중구 은행동 48-	36.3290121	127.42765725
#477000	M1024	@jetop2 어디신데요?!ㅋ 비가많이와요?!ㅌ	37.443865	127.01363269
#209000	M0471	안맞는다...ㅠㅠ (@ 효창골프연습장) http://t.co/	37.542609	126.962371
#972000	M7900	@KimQri 김규리님 정말 잘했어요 진정한 우승	37.5911729	127.0385667
#336000	M3235	I'm at Cold Stone Creamery (성동구 행당동16	37.5614011	127.0384038
#757000	M0133	여유로운 주말. 작지만 괜찮은 건프라샵 (@ 건	37.56075101	126.92590714
#730000	M3765	추카 웨딩댄스 페스티벌. with L (@ PJ호텔) htt	37.5646	126.995418

그림 6. 사용자 트윗과 트윗 작성 위치정보

비교하였다. 이 실험에는 두 개의 데이터셋이 사용되었는데, 첫 번째는 Lee 등의 논문[9, 11]에서 사용된 국내 사용자 계정 52만개에 대해 Search API를 이용하여 수집한 해당 사용자의 전체 트윗 1,113만개이고(한국인 계정 데이터셋), 두 번째는 @ladygaga 계정의 팔로워로 구성된 해외 사용자 계정 280만개에서 Streaming API를 이용하여 수집한 트윗 87만개이다(레이디가가 데이터셋). 레이디가가 데이터셋은 2011년 12월 26일부터 2012년 1월 25일까지 한 달 동안 수집하였다. 이 두 개의 데이터셋에서 실험에 사용할 수 있도록 잘 정의된 프로필 위치정보를 가진 사용자만 선별하고, 다시 트윗 작성위치가 GPS 좌표로 포함된 트윗 작성자를 걸러냈다. 이렇게 선택된 사용자의 프로필 위치와 트윗 작성위치에 대해 야후 API[12]를 이용하여 국내 사용자의 경우 서울시 및 6개 광역시는 구 단위로, 그 외 지역과 해외 사용자는 시 단위(또는 county)로 정리하였다. 이러한 데이터 정제과정을 통해 트위터에서 수집한 모든 위치정보를 구 단위 또는 시 단위의 행정구역명으로 변경하였다.

4. 실험 방법 및 결과

4.1 프로필 위치와 트윗 위치 매칭

우리는 이 실험에서 트위터 사용자 프로필에 입력된 위치와 트윗이 실제로 작성된 위치의 관계를 조사하기 위해 각 사용자별로 동일한 위치에서 작성된 트윗들끼리 그룹핑 한 다음 프로필의 위치가 어느 그룹과 매칭되는지 알아보았다. 우선 표 1과 같이 수

표 1. 사용자별 위치정보 문자열의 예

사용자ID#프로필위치#트윗작성위치
10001#서울시_양천구#서울시_서대문구
10001#서울시_양천구#서울시_양천구
10001#서울시_양천구#서울시_양천구
10001#서울시_양천구#서울시_양천구
10001#서울시_양천구#서울시_양천구
10001#서울시_양천구#서울시_양천구
10001#서울시_양천구#서울시_양천구
10001#서울시_양천구#서울시_영등포구
10001#서울시_양천구#서울시_영등포구
10001#서울시_양천구#서울시_영등포구
10001#서울시_양천구#서울시_중구
10001#서울시_양천구#서울시_중구
10002#경기도_의왕시#경기도_의왕시
10002#경기도_의왕시#경기도_의왕시
10002#경기도_의왕시#경기도_수원시
10002#경기도_의왕시#경기도_수원시
10002#경기도_의왕시#경기도_수원시
10002#경기도_의왕시#경기도_성남시

집한 데이터를 이용하여 “사용자ID#프로필위치#트윗작성위치”의 형태로 문자열을 생성하였다.

그 다음 이 문자열 목록을 동일한 것들끼리 병합하고 병합된 문자열의 개수에 따라 표 2와 같이 정렬하였다. 이처럼 정렬된 목록을 기반으로 각 사용자별 프로필의 위치가 정렬된 트윗 작성위치의 몇 번째(k) 그룹에 포함되는지를 확인하여 Top-k 그룹으로 분류하였다. 예를 들어 표 2의 예에서 사용자 10001은 프로필의 위치가 “서울시_양천구”인데, 양천구에서 가장 많은 트윗을 작성했기 때문에 이 사용자는 Top-1 그룹에 포함되었다. 사용자 10002의 경우에는 프로필 위치가 “경기도_의왕시”였는데, 트윗 작성위치 목록에서 의왕시가 두 번째에 랭크되었으므로 Top-2 그룹으로 분류하였다.

표 2. 병합 후 정렬한 문자열 목록

사용자ID#프로필위치#트윗작성위치
10001#서울시_양천구#서울시_양천구 (6) ← k=1
10001#서울시_양천구#서울시_영등포구 (3)
10001#서울시_양천구#서울시_중구 (2)
10001#서울시_양천구#서울시_서대문구 (1)
10002#경기도_의왕시#경기도_수원시 (3)
10002#경기도_의왕시#경기도_의왕시 (2) ← k=2
10002#경기도_의왕시#경기도_성남시 (1)

위의 과정을 통해 잘 정의된 프로필 위치정보와 트윗 위치정보를 모두 가진 계정을 선별하고, 프로필의 위치와 정리된 트윗 위치를 비교하여 서로 일치하는 것이 Top-k중 어떤 그룹에 속하는지를 확인하였다. 그림 7의 결과를 보면 국내 사용자 중 Top-1과 Top-2 그룹에 속한 사용자의 비율이 약 63%(각각 49.73%와 13.30%)이고, 해외 사용자는 약 96%(각각 90.64%와 5.85%)였다. 하지만 국내 사용자의 경우 Top-k의 어느 그룹에도 속하지 않는 사용자가 무려 29.61%에 달했다. 이러한 결과는 한국이 전 세계 평균의 10배, 전체 23위에 해당할 만큼 인구밀도가 높아 실제 거주지역과 활동지역이 다른 유동인구가 많기 때문으로 판단된다[13]. 많은 사람들이 프로필 위치에 회사나 학교의 위치보다는 주거지의 위치를 입력하는 경우가 많다는 점을 고려하면, 국내 사용자의 약 절반 정도는 주거지에서 가장 많은 트윗을 작성하지만, 30%정도의 사용자는 프로필에 입력한 위치에 전혀 인접하지 않은 다른 위치에서 주로 트윗을 작성한다는 사실을 알 수 있었다. 레이디가가 데이터셋의 경우 모든 데이터가 Top-4 이내에 속하였고, None 그룹에 포함된 사용자가 전혀 없었다. 이러한 비교를 통해 해외 사용자가 국내 사용자보다는 조금 더 명확한 프로필 위치정보를 제공한다는 사실을 알 수 있다.

다음으로는 단 하나의 트윗도 프로필 위치와 일치하지 않는 None 그룹을 제외한 나머지 그룹에 대해 몇 개씩의 지역을 가지는지 확인해보았고, 그 결과는 그림 8과 같다. 국내 사용자의 경우 Top-1 그룹은 평균 3.6개의 트윗 작성위치를 가지는데 반해, 해외 사용자의 경우 Top-1 그룹은 평균 1.4개의 트윗 작성위치를 가지는 것으로 나타났다. 특히 국내 사용자의 경우 해외 사용자보다 항상 두 배 이상 많은 트윗 작성위치를 가지고 있었으며, 국내 사용자 Top-6 이상의 그룹을 모두 포함한 Top-others의 경우 무려 12개 이상의 서로 다른 위치를 가지는 것을 알 수 있었다. 결과적으로 트윗 작성위치의 수는 k가 증가함에 따라 점점 늘어나고, 이러한 사실을 통해 더 다양한 지역에서 트윗을 작성하는 사용자일수록 프로필 위치정보의 신뢰성이 낮아진다는 점을 유추할 수 있었다.

4.2 각 그룹의 신뢰도 분석

앞의 실험에서는 동일한 위치에서 작성된 트윗들끼리 그룹핑한 다음 프로필의 위치가 어느 그룹과 매칭되는지 알아보았다. 하지만 이는 전체 계정 중 프로필의 위치에서 가장 많은 트윗을 남긴 Top-1 계정의 비율을 통해 프로필 위치와 트윗 위치의 정확

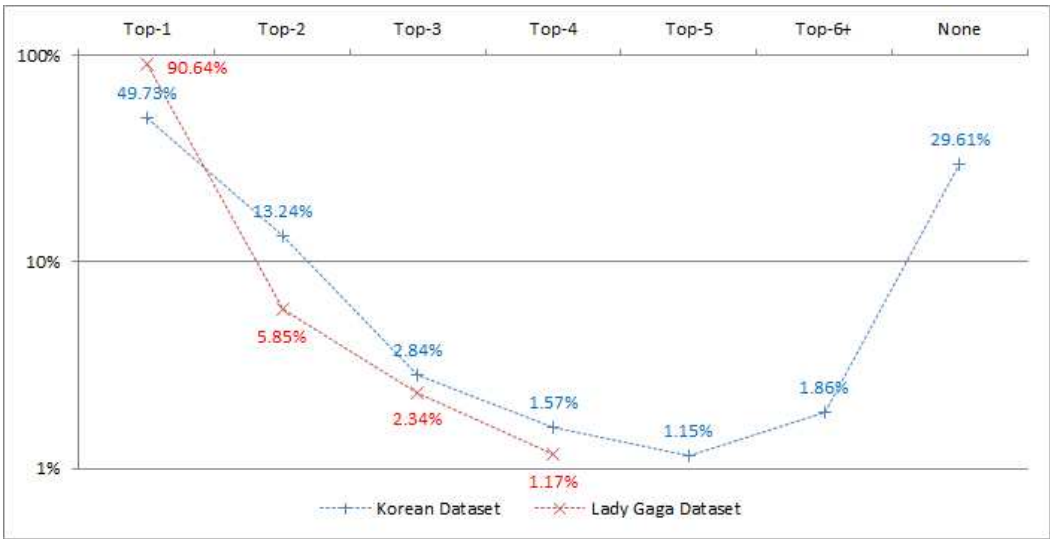

그림 7. Top-k 그룹별 사용자 수

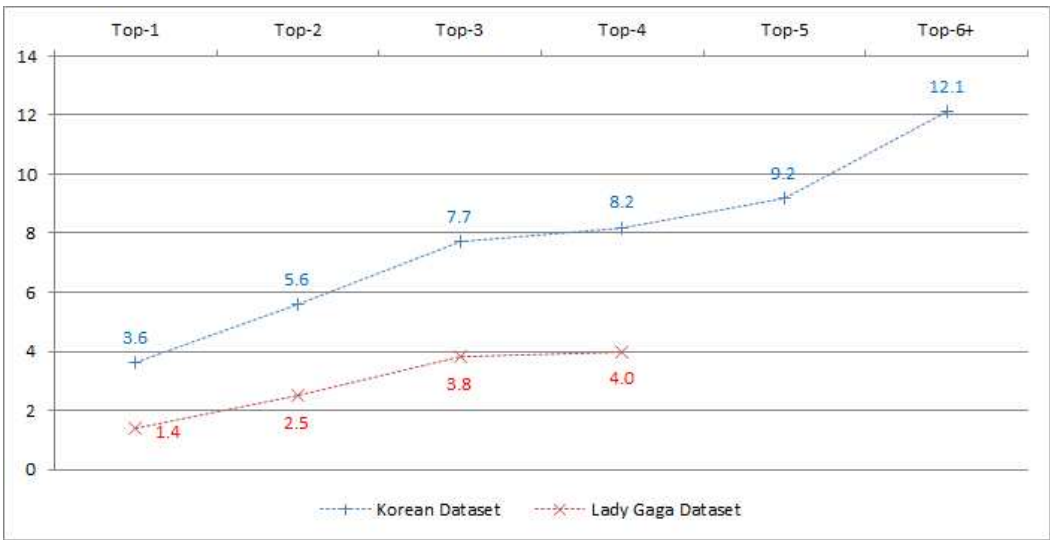

그림 8. Top-k 그룹별 사용자 위치의 평균 개수

도만을 확인한 것으로 Top-1 이외의 나머지 그룹은 고려하지 않았다. 그러나 표 3과 같은 예에서 사용자 51267은 프로필 위치와 Top-1이 일치하지만, 31개의 트윗을 남긴 Top-1 이외의 그룹을 고려할 경우 29개의 트윗을 남긴 "경기도_성남시"에서 가장 많은 트윗을 남겼다고 말할 수 없다. 따라서 우리는 각 Top-k의 신뢰도를 분석하기 위해 각 계정의 트윗이 그룹핑 된 후 각 Top-k 마다 몇 개의 트윗이 분류되었는지 해당 계정의 전체 트윗에 대한 비율을 계산해 보았다.

그림 9의 결과를 보면 국내 사용자의 경우 Top-1 그룹에 속한 사용자는 프로필의 위치에서 남긴 트윗

표 3. 사용자 51267의 정렬된 위치정보 문자열

사용자ID#프로필위치#트윗작성위치
51267#경기도_성남시#경기도_성남시 (29) ← k=1 51267#경기도_성남시#서울시_송파구 (14) 51267#경기도_성남시#서울시_마포구 (8) 51267#경기도_성남시#충청북도_청주시 (5) 51267#경기도_성남시#서울시_강남구 (4)

이 자신이 남긴 전체 트윗의 78.8%에 달했다. 이는 사용자의 프로필 위치가 L1이고, 가장 빈번하게 트윗을 남긴 위치가 L2라고 할 때, 'L1에서 트윗을 남기는 사용자'라는 가정의 신뢰도가 78.8%임을 말한다. 이로써 우리는 사용자로부터 L2의 정보를 가져올 수

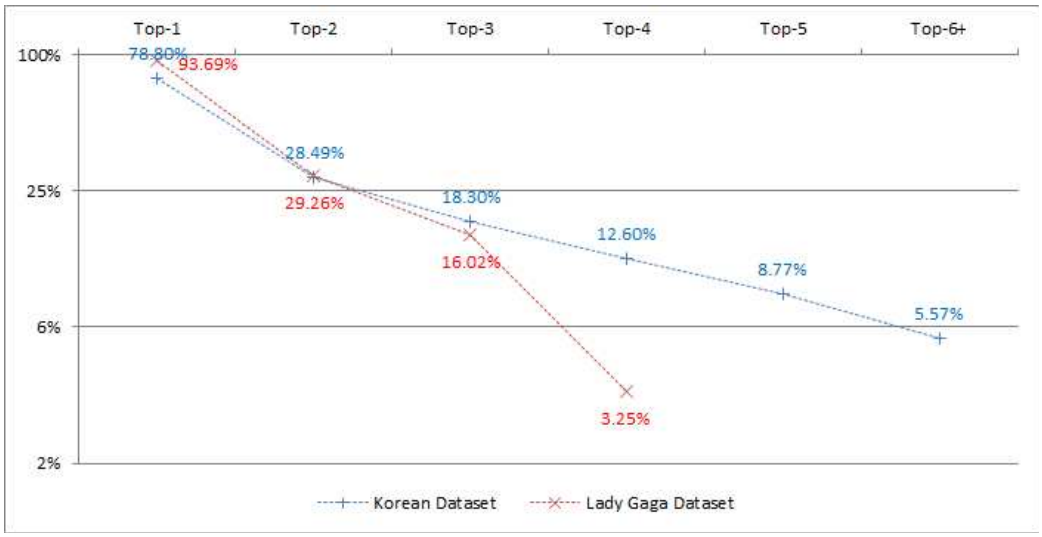

그림 9. 각 그룹별 신뢰도 분석

없을 때, 해당 사용자를 가장 많은 계정이 속한 Top-1 그룹으로 가정하고 트윗을 분석하면 그림7(49.73%)과 그림 9(78.80%)의 결과에 의해 39.1%의 신뢰도를 가진 L2 정보를 확보할 수 있다는 점을 알 수 있었다. 같은 방법으로 레이디가가 데이터셋을 분석하면 84.9%의 신뢰도를 가진 L2 정보를 확보할 수 있음을 알 수 있었다. 이러한 비교를 통해 프로필에 위치정보가 없는 국내 사용자는 트윗의 위치정보만으로 해당 사용자의 정확한 위치정보를 판단하는데 어려움이 있지만, 해외 사용자의 경우는 비교적 믿을 만한 위치정보를 가져올 수 있다는 점을 유추할 수 있었다.

5. 결론 및 향후 연구계획

본 논문에서는 트위터에서 사용자가 제공한 프로필 위치정보와 트윗에 포함된 트윗 작성위치의 관계 분석을 통해 프로필 위치의 신뢰성을 분석하였다. 이러한 분석은 보편화되고 있는 위치정보 기반의 응용[14,15]에 소셜 정보를 적용할 때, 결과물의 신뢰성 확보를 위해 필수적이다. 실험 결과에 따르면, 전체적으로 국내 사용자보다는 해외 사용자의 위치정보가 더 높은 신뢰성을 가지는 것으로 나타났다. 국내 사용자의 50%는 프로필의 위치에서 가장 많은 트윗을 작성하지만 동시에 약 30%의 사용자는 프로필 위치에서 작성한 트윗이 전혀 없는 것으로 나타났다. 반면 해외 사용자의 경우에는 약 90%의 사용자가 프로필 위치에서 가장 많은 트윗을 작성하는 것으로 나타났다. 또한 각 Top-k 그룹별 트윗 작성위치의 수 k를 분석한 결과 k가 증가하면 트윗 작성위치의 수도 점점 늘어남을 알 수 있다. 이를 통해 다양한 지역에서 트윗을 작성하는 사용자는 이벤트의 위치정보를 판단하기 위한 데이터에서 배제하는 것이 더 나은 결과를 가져올 것으로 예측할 수 있었다. 이러한 점들을 종합적으로 고려할 때, 본 논문의 결과는 트위터에서 추출된 이벤트의 정확성을 판단할 수 있는 기초적인 정보를 제공할 뿐만 아니라, 추출된 이벤트 위치의 정확성을 향상시키는데 도움이 될 것으로 기대한다. 특히, 국가적 수준의 이벤트 보다는 교통사고나 화재 같은 지역적 수준의 이벤트를 추출할 때 더 도움이 될 것이다. 향후 연구에서는 본 논문의 결과를 실제 이벤트 감지 시스템에 적용하여 어느 정도의 성능을 가져올 수 있는지 확인해야한다.

참 고 문 헌

[1] A. Chowdhury, Global Pulse, http:// blog.twitter.com/2011/06/global-pulse.html, 2011.

[2] N. K. Cheblb and R. M. Sohall, “The Reasons Social Media Contributed to the 2011 Egyptian Revolution,” *Int'l Journal of Business Research and Management*, Vol. 2, Issue 3, pp. 139-162, 2011.

[3] 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.

[4] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav, “Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences,” *Proc. of the 10th Int'l Conf. on Web Information Systems Engineering, LNCS*, Vol. 5802, pp. 539-553, 2009.

[5] J. Russell, Japan Overtakes Indonesia as Biggest Twitter User in Asia, http://www.asiancorrespondent.com, 2011.

[6] 황경상, 트위터 사용자 늘고 ‘상위 1%’ 집중도 심화, http://wkh.kr/uWi4Pj, 경향신문, 2011.

[7] ITU, ITU Measuring the Information Society 2011, http://www.itu.int/ITU-D/ict/, International Telecommunication Union, 2011.

[8] 방송통신위원회, 2011년도 국정감사 서면질의 답변서, http://www.kcc.go.kr, 방송통신위원회, 2011.

[9] 이범석, 김석중, 최우성, 장경훈, 윤진영, 황병연, “트위터에서 사용자 위치와 트윗 작성위치의 관계분석,” 제28회 한국멀티미디어학회 추계학술대회논문집, 제14권, 제2호, pp. 1-3, 2011.

[10] R. Lee and K. Sumiya, “Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection,”

Proc. of the 2nd ACM SIGSPATIAL Int'l Workshop on Location Based Social Networks, pp. 1-10, 2010.

[11] Bumsuk Lee and Byung-Yeon Hwang, "A Study of the Correlation between the Spatial Attributes on Twitter," *Proc. of the 28th IEEE Int'l Conf. on Data Engineering Workshop on Spatio Temporal data Integration and Retrieval*, 2012.

[12] Version 1.0 야후! 주소 ↔ 좌표 변환 API, http://kr.open.gugi.yahoo.com/document/geocooder.php, 2011.

[13] List of sovereign states and dependent territories by population density, http://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_population_density, 2012.

[14] 윤혜진, 창병모, "위치 인식을 이용한 음식점 추천 시스템의 설계 및 구현", 멀티미디어학회논문지, 제14권, 제1호, pp. 112-120, 2011.

[15] 이중기, 김창수, "스마트폰 앱기반 재난정보 서비스 및 검색기능 구현", 멀티미디어학회논문지, 제15권, 제2호, pp. 273-290, 2012.

이 범 석

2004년 가톨릭대학교 분자생물학과, 국사학과 학사
2006년 가톨릭대학교 컴퓨터공학과 석사
2010년 가톨릭대학교 컴퓨터공학과 박사

2010년-2011년 University of Alberta 박사후연구원
2011년-현재 가톨릭대학교 컴퓨터공학과 연구원
관심분야 : 소셜네트워크분석, 웹데이터마이닝, 정보검색, 데이터베이스

김 석 중

2011년 가톨릭대학교 컴퓨터공학과 학사
2011년-현재 가톨릭대학교 컴퓨터공학과 석사과정
관심분야 : 소셜네트워크분석, 데이터마이닝, 오피니언마이닝, 데이터베이스, 정보검색

황 병 연

1986년 서울대학교 컴퓨터공학과 학사
1989년 한국과학기술원 전산학과 석사
1994년 한국과학기술원 전산학과 박사

1994년-현재 가톨릭대학교 컴퓨터정보공학부 교수
1999년 University of Minnesota 방문교수
2007년 California State University 방문교수
관심분야 : 소셜네트워크분석, XML 데이터베이스, 데이터마이닝, 정보검색, 지리정보시스템