Journal of Korea Spatial Information Society Vol.22, No.3 : 1-8, June 2014
http://dx.doi.org/10.12672/ksis.2014.22.3.001
ISSN 2287-9242(Print)
ISSN 2287-9250(Online)

트위터 사용자의 위치정보와 성향을 고려한 트윗 수집 시스템

Tweet Acquisition System by Considering Location Information and Tendency of Twitter User

최우성*·임준엽**·황병연***

Woosung Choi · Junyeob Yim · Byung-Yeon Hwang

요 약 최근 소셜 네트워크 서비스가 급격히 성장하면서, 소셜 네트워크 분석에 관련된 연구들도 많은 관심을 받고 있다. 특히 트위터는 사회적 이슈나 사건들에 대해 실시간으로 반응하기 때문에, 사회과학 분야나 정보검색 분야의 연구자들이 유용한 실험 데이터를 수집하는 데에 활용되고 있다. 그러나 정작 데이터를 수집하는 방법론에 관한 연구는 아직 미흡하다. 이에 본 논문에서는 위치 기반의 이벤트와 정치·사회적 이벤트 위주의 사용자의 성향을 고려한 트윗 수집 시스템을 제안한다. 우선 위치정보와 이벤트 관련 키워드를 포함하고 있는 트윗과 정치·사회적인 이벤트 검출에 필요한 ID들을 수집한 후, 사용자들의 성향을 분류할 ID 분석기를 설계했다. 또한 ID 분석기의 신뢰도 측정을 위해 상위 등급에 분류된 ID를 이용하여 트윗을 분석했다. 분석결과 1등급으로 분류된 ID는 88.8%의 신뢰도를 보였으며, 2등급으로 분류된 ID는 76.05%의 신뢰도를 보였다. 또한 ID 분석기는 77.5%의 신뢰도를 보였으며 소수의 ID를 사용함으로써 데이터의 수집시간을 줄였다.

키워드 : 소셜 네트워크 분석, 성향 분석, 데이터 수집, 정보 검색

Abstract While SNS services such as Twitter or Facebook are rapidly growing, research for the SNS analysis has been concerned. Especially, twitter reacts to social issues in real-time so that it is used to get useful experimental data for researchers of social science or information retrieval. However, it is still lack of research on the methodology to collect data. Therefore, this paper suggests the tweet acquisition system by considering tendency of twitter user oriented location-based event and political·social event. First the system acquires tweets including information of location and keyword about event and secure IDs for acquisition of political·social event. Then we plan ID-analyzer to classify the tendency of users. In addition for measuring reliability of ID-analyzer, it acquires and analyzes the tweet by using high-ranked ID. In analyses result, top-ranked ID shows 88.8% reliability, 2nd-ranked ID shows 76.05% and ID-analyzer shows 77.5%, it shortens collection time by using minority ID.

Keywords : Social Network Analysis, Tendency Analysis, Data Acquisition, Information Retrieval

1. 서 론

2011년에 발표된 국제전기통신연합(ITU: International Telecommunication Union)의 연간 보고서[6]에 따르면 2010년 한국의 가구당 인터넷 보급률이 96.8%로 세계 1위, 인구 100명당 무선 인터넷 접속자수가 91명으로 세계 1위를 각각 차지하였다. 또한 2011년 방송통신위원회 국정감사에 제출된 자료[7]에서는 2011년 7월 국내 스마트폰 가입자가 1,626만 명으로 전체 인구의 33.6%가 스마트폰을 사용하고 있는 것으로 나타났다. 이와 같은 기술적 추세는 사용자들에게 인터넷 접근성을 확대시켜 주었고, 그들이 인터넷을 통한 사이버 상의 새로운 커뮤니케이션 공간을 요구함에 따라 이를 제공하는 소셜 네트워크 서비스(SNS)의 급격한 성장을 가져왔다.

그러한 소셜 네트워크 서비스 중 트위터(Twitter)는 단문 텍스트를 기반으로 하는 마이크로블로그(micro-blog) 서비스이다. 트위터는 다른 소셜 네트워크 서비스와 달리 사용자 개인의 요구에 의해 타 사용자와의 일방적인 관계 형성이 가능하기 때문에, 유동적인 정

† 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).

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

보의 흐름뿐만 아니라 정보나 견해를 위한 네트워크 형성도 가능하다. 이로 인해 기존의 언론 못지않은 파급효과와 함께 새로운 형태의 미디어로 주목 받고 있으며, 기업 마케팅, 교육, 방송 등의 다양한 분야에 활용되고 있다. 특히 2011년 3월 일본지진과 같은 현실 세계의 사건에 민감하게 반응하였고[3], 튀니지의 재스민 혁명을 포함한 중동지역의 민주화 시위를 실시간으로 전 세계에 알림으로서 여러 국가들의 지지를 이끌어내는데 기여했듯이 현실 세계에 미치는 영향 또한 커지고 있다[1].

또한 트위터를 사용하는 사용자들은 트윗(Tweet)이라는 단문 메시지를 통해 개인의 견해나 정보를 공유하는데, 여기에는 개인의 일상이나 감정, 정치·사회적 이슈, 각종 사건 및 사고에 대한 정보와 함께 사용자의 동의하에 자동적으로 기록되는 GPS 좌표정보들이 포함된다. 따라서 이 정보들을 토대로 다양한 위치 기반 응용 시스템 개발이 진행되고 있다.

특히 Sakaki등[16]과 Nagarajan등[14]은 트위터 데이터를 분석하여 미리 지정한 이벤트나 지역별 이슈를 감지하는 시스템을 개발하였다. 이러한 연구를 진행하기 위해서는 트위터 사용자가 직접 경험하거나 사용자 주변에서 일어나는 사건에 대해 실시간으로 트윗을 작성할 다수의 사용자가 존재해야 하는데, 이것은 일본 내에 스마트기기를 이용하여 트위터에 접속하는 사용자가 많기 때문에 가능한 연구였다. 한국도 동일한 이유로 관련 연구를 진행하기 좋은 조건을 가지고 있다[8]. 2010년 여름에 발표된 comScore의 자료[15]를 보면 일본 전체 인구 약 1억 2천만 명 중 약 1천 6백만 명이 트위터를 사용하는데, 최근(2011년 9월 기준)의 조사[5]에 따르면, 한국은 전체 인구 약 5천만 명 중 한국인 사용자로 추정할 수 있는 트위터 계정의 수가 약 392만 개에 이른다.

Sakaki등의 연구와 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 활용된 사용자의 위치가 얼마나 잘 정의되느냐에 따라 성능의 차이가 발생할 수 있는데, Sakaki등의 연구에서는 트윗에 포함된 GPS 좌표와 프로필에 정의된 주소를 이용해 센서의 위치를 지정한 반면, Nagarajan등의 연구에서는 트위터 사용자를 국가별로 분류하였다. 국가별로 위치를 분류할 경우 국민투표나 대형사건과 같이 누구나 알고 있고, 누구나 관심을 가지는 국가적 관심사에 대해서는 추출이 가능하지만[2], 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 발생 지점을 예측할 경우 정확성이 낮아질 수 있다[9,10].

Daniel의 연구[4]에서는 Mustafaraj등의 연구[13]와 Mislove등의 연구[12]를 언급하면서 트위터를 이용한 선거 예측에 있어서 사용자의 성향을 고려해야 한다고 주장하였다. 민정식의 연구[11]에서는 트위터의 정치참여 효과를 트위터의 기능적인 특성을 반영해 분석하였다. 분석결과 트위터의 팔로어 수가 많으며, 트윗 작성 건수가 많을수록, 사회, 정치 관련 공인이나 단체의 트윗을 많이 구독할수록 정치 토의 참여, 정치의견 피력, 정치행사참여, 투표 참여의향 등 정치참여의 수준이 높았다. 이런 특징들을 고려하여 트위터 사용자 중 정치·사회적 참여수준이 높은 사용자들을 분류하고, 이러한 사용자에 대한 트윗의 수집과 분석이 선행될 수 있다면 이벤트 검출의 정확도를 높이고 수집 시간을 단축시킬 수 있다. 하지만 이러한 사용자 성향 파악에 대한 연구는 매우 부족한 실정이다.

본 논문의 의의는 다음과 같다. 첫째, ID를 인수로 한 Search API[17]와 키워드를 인수로 한 Search API를 사용하여 트윗을 수집하고 수집된 트윗을 분석함으로써 각 수집방법의 효율성을 비교하였다. 둘째, ID의 팔로워 수, 최근 트윗 사용량 그리고 키워드를 포함한 트윗량을 고려하여 ID 등급을 분류 및 분석하였다. 셋째, ID를 임의로 분류한 그룹과 ID의 등급을 분류한 그룹으로 나누고, 각 그룹으로 트윗을 수집 및 비교 분석하여 ID 분석기의 효율성을 증명하였다. 넷째, 위치기반 이벤트와 정치·사회적 이벤트 검출을 위한 사용자의 성향을 고려한 트위터 수집 시스템을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 시스템과 연관된 관련연구에 대해서 기술한다. 3장에서는 제안된 기법의 전체적인 구성과 각 부분의 기능에 대해 설명한다. 4장에서는 전체 시스템 구조를 설명한다. 5장에서는 실험 및 결과를 제시한다. 마지막으로 6장에서는 제안하는 시스템의 의미를 정리하고 향후연구 과제를 설명한다.

2. 관련 연구

2.1 위치정보 기반 트위터 연구

2010년에 발표된 Sakaki등의 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’과 ‘shake’가 지진 때문인지, 아니면 다른 문맥에서 사용된 단어인지를 가려냈다. 이들의 실험결과는 전체적으로 약 80% 이상의 재현율(recall), 60% 이상의 정확도(precision)를 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 인구 밀도가 높고 트위터 사용자가 많은 도심지에 가까운 형태로 이동경로가 예측되는 모습을 보였다. 이러한 실험결과는 이벤트 감지가 위치 정보가 포함된 트윗의 데이터 수에 의존적이라는 것을 보여준다. 따라서 정확한 이벤트 감지를 위해서는 위치정보를 포함하고 있는 많은 량의 트윗을 수집할 수 있는 방법이 필요하다.

2.2 정치·사회 관련 트위터 연구

Daniel의 연구에서는 Mustafaraj등의 연구를 언급하면서 소셜 미디어에는 서로 정반대의 행동을 보이는 두 그룹이 있으며, 이 두 개의 그룹을 분명하게 구분해야 한다고 주장하였다. 이 두 그룹 중 하나는 대부분의 콘텐츠를 생산해내는 소수의 유저들이며, 또 하나는 소수의 콘텐츠만을 생산하며, 리트윗을 통해 콘텐츠를 퍼트리는 대다수의 유저들을 말한다. 또한 Mislove 등의 연구를 언급하면서 트위터 사용자의 성별이나 지역적인 특성들을 고려해야 한다고 주장했다. 이것은 트위터 유저의 성향을 포함하지 않고 소셜 미디어의 콘텐츠만을 분석할 경우, 여론의 상위에 있는 몇몇의 소수의 사용자들에 의해 분석 결과가 달라질 수 있음을 말해준다. 더불어 민정식의 연구에서는 트위터의 정치참여 효과를 트위터의 기능적인 특성을 반영해 분석하였다. 이 연구는 정치·사회관련 이벤트에 관한 트위터 수집 시 트위터의 팔로어 수나, 트윗 작성 건수, 리트윗 정도를 고려한다면 좀 더 양질의 트윗을 수집하고 트윗 수집시간을 단축시킬 수 있다는 것을 말해준다.

3. 트윗 수집 시스템

본 장에서는 태풍 볼라벤과 덴빈이 발생한 2012년 8월 28일부터 31일까지 수집한 트윗을 분석하고, 지속적인 ID 수집 및 분석과 트위터 사용자의 성향을 고려하여 트윗을 수집할 수 있는 시스템을 제안한다.

3.1 트윗 수집

각 수집 방법의 효율성을 알아보기 위하여 위치정보를 공개한 사용자 ID 25,000개를 인자로 하는 Search API와 키워드(keyword)및 좌표(geocode)를 인자로 하는 Search API를 이용하여 트윗을 수집하였다. Search API는 한번의 API실행 시 최대 200개의 트윗이 수집 가능하며, 시간당 150회의 요청을 할 수 있다. 본 연구는 API실행시마다 사용자 ID를 인자로 하는 Search API는 사용자가 최근에 올린 트윗 200개를 받아오며, 키워드와 지역좌표를 이용한 Search API는 지정한 위치에서 발생한 키워드가 포함된 트윗 200개를 받아오게 설계하였다. 태풍 볼라벤과 덴빈이 발생한 8월 28일부터 31일까지 낙석, 바람, 비, 산사태, 정전, 태풍, 침수 등 총 13개의 키워드와 제주도를 포함한 전국 9개 지역의 좌표를 인자로 하여 총 145,071개의 트윗을 수집하였으며, 25,000개의 ID를 가지고 총 1,257,807개의 트윗을 수집하였다.

3.2 ID 수집

Figure 1은 ID 수집을 위한 시스템 구성도이다.

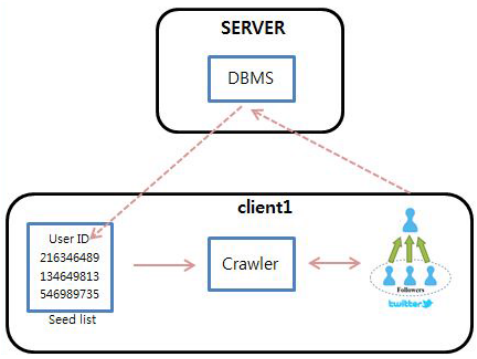

Figure 1. ID Gathering System

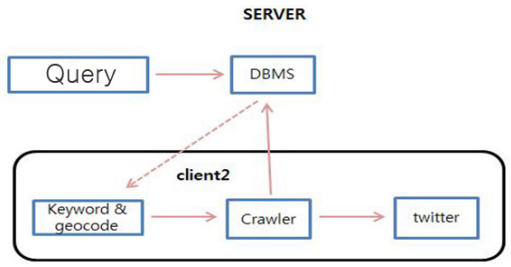

Figure 2. Gathering System using Keyword and Local Information

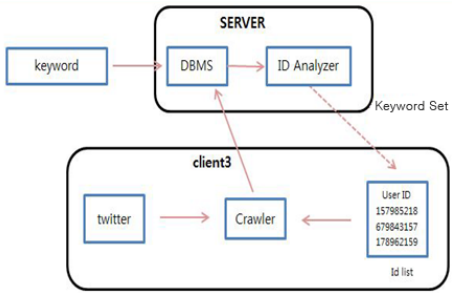

Figure 3. Tweet Gathering System using ID

2011년 새롭게 변경된 트위터 API는 트위터 데이터에 대한 애플리케이션의 시간당 접근 횟수를 제한하기 때문에 1시간마다 지속적으로 ID 수집을 수행한다. ID 수집은 DB에서 저장되어 있는 유저 ID 리스트를 입력받아 이 리스트를 시드로하여 시드에 입력된 사용자의 팔로워 정보를 JSON 형태로 받아온다. 이렇게 받아온 ID들은 파싱을 거쳐 다시 DB에 저장되며, 향후 수집의 시드 데이터로 쓰이게 된다.

3.3 키워드와 지역좌표를 이용한 트윗 수집

Figure 2는 키워드와 지역좌표를 이용한 트윗 수집 시스템이다. 이 시스템의 경우, 구현 방식이 두 가지가 있다. 첫째는 Streaming API[18]를 사용하는 것이고, 또 한 가지는 Search API를 사용하는 것이다. Streaming API를 사용할 경우, 데이터에 대한 애플리케이션의 횟수 제한은 없지만, 과거의 데이터는 수집이 불가능하다. 반대로 Search API를 사용할 경우에는 애플리케이션의 접근 횟수제한을 받지만, 과거의 데이터를 수집할 수 있다. 시스템의 목적이 실시간 이벤트 검출이라면 Streaming API가 용이하며, 과거의 지나간 이벤트에 대한 트윗 분석이 목적이라면 Search API를 사용해야한다. 본 논문에서는 두 가지 방식을 모두 사용한다. 과거의 데이터는 Search API를 이용하여 수집하며, 현재의 데이터는 Streaming API를 사용하여 데이터를 수집 후 통합한다.

질의를 받으면 서버에선 클라이언트에게 키워드와 지역좌표를 넘겨준다. 일정 좌표에서 일정 거리 이내의 트윗만을 받아올 수 있기 때문에, 본 논문에서는 총 9개의 중심지 좌표와 수집거리를 넘긴다. 크롤러는 키워드와 지역좌표를 인수로 사용해 트윗을 JSON형태로 수집하며 수집된 트윗은 파싱을 거쳐 DB에 저장된다.

3.4 ID를 이용한 트윗 수집

Figure 3은 ID를 이용한 트윗 수집 시스템이다. ID를 이용한 트윗 수집 시스템은 주로 정치·사회 분야의 트윗 수집에 사용된다. ID를 이용한 트윗 수집은 수집된 ID 중 팔로워의 수가 많고 키워드를 포함한 트윗의 수와 최근 트위터의 사용량이 많은 ID를 이용하여 트윗을 수집한다. 이는 광고·홍보의 성격을 띠는 트윗을 걸러내기 위한 작업으로 데이터의 신뢰성을 높이기 위함이다. ID를 입력받은 크롤러는 ID를 인자로 하여 트윗을 수집한다. 수집 알고리즘은 인자만 달라질 뿐, 키워드와 지역좌표를 사용하는 Search API방식과 동일하다.

3.5 ID 분석기

ID 분석기는 크게 3가지 요소를 고려하여 ID의 등급을 분류한다. 첫 번째로 사용자를 팔로우하고 있는 팔로워의 수이다. 트위터 서비스가 크게 성공하면서 트위터를 이용한 광고·홍보 성격의 트윗의 양이 늘어나고 있다. 이러한 트윗들은 키워드만을 고려했을 시 정상적인 트윗에서 분류해내기 어렵다. 트위터 사용자의 팔로워의 수가 많다는 것은 이 사용자가 올리는 트윗의 정보가 신뢰도가 높고, 유용하다는 것을 간접적으로 나타내며, 이러한 사용자는 광고·홍보성 트윗을 올릴 확률이 적다. 또한 민정식의 연구의 분석결과에서 알 수 있듯이 팔로워의 수가 많으면 정치·사회적인 트윗을 올릴 확률이 높다. Table 1은 팔로워 수를 고려한 ID의 등급 기준이다. 각각의 ID는 팔로워의 수에 따라 총 5개의 등급으로 분류되며, 각 등급별로 0.5~3점의 점수를 부여받는다.

두 번째로는 사용자의 최근 트윗 양이다. 최근 트윗 양은 현재 그 사용자가 트위터를 사용하는 빈도수를 보여준다. 만약 트위터 사용자가 팔로워 수가 많더라도 최근 트윗을 사용하지 않았다면, 원하는 데이터를

Table 1. ID Grading Criteria Considering the Number of Followers

Grade	Criteria	Score
1	More than 15,000 followers	3
2	Less than 15,000 and more than 10,000 followers	2.5
3	Less than 10,000 and more than 5,000 followers	2
4	Less than 5,000 and more than 1,000 followers	1.5
5	Less than 1,000 followers	0.5

Table 2. ID Grading Criteria Considering the Amount of Recent Tweet Consumption

Grade	Criteria	Score
1	Amount of more than 150 recent tweets	3
2	Recent tweets in the amount of less than 150, more than 100	2.5
3	Recent tweets in the amount of less than 100, more than 50	2
4	Recent tweets in the amount of less than 50, more than 25	1.5
5	Recent tweets in the amount of 25 or more, less than 10 cases	1
6	Recent tweets in the amount of less than 10, more than one	0.5

보유하고 있지 않을 확률이 높다. 또한 사용자가 트위터를 사용하고 있다는 것을 보장해준다. 본 연구에서는 최근 2개월간의 트윗 사용량을 고려하였다. Table 2는 최근 트윗 사용량을 고려한 ID 등급 기준이다. 각각의 ID는 최근 트윗 사용량에 따라 총 6개의 등급으로 분류되며, 각 등급별로 0.5～3점을 부여받는다.

세 번째로 키워드를 포함한 트윗의 수이다. 팔로워의 수가 많고 최근 트윗량이 많다 하더라도 그 사용자가 정치·사회에 관심을 가지고 있다는 것은 보장되지 않는다. 트윗은 사용자의 감정이나 일상생활을 포함하며, 트윗의 주제가 매우 광범위하다. 키워드를 많이 포함하고 있다는 것은 사용자가 그 분야에 관해 관심이 많다는 것을 보여주는 지표이므로 키워드를 포함한 트윗량을 고려한다. Table 3은 키워드를 포함한 트윗량을 고려한 ID 등급 기준이다. 각각의 ID는 키워드를 포함하고 있는 트윗량에 따라 총 4개의 등급으로 분류되며, 각 등급별로 0.5～3점의 점수를 부여받는다.

마지막으로 이렇게 3가지 요소에 대하여 얻은 등급

Table 3. ID Grading Criteria Considering the Amount of Tweets Containing Keyword

Grade	Criteria	Score
1	Tweet amount includes more than 50 keywords	3
2	Tweet amount includes less than 50 keywords, 25 and over	2.5
3	Tweet amount includes less than 25 keywords, more than 10	1
4	Tweet amount includes less than 10 keywords	0.5

Table 4. Final Grading Classification of ID based on the Score of each Element

Grade	Criteria
1	More than 8 points
2	More than 6.5 and Less than 8 points
3	More than 5 and Less than 6.5 points
4	More than 2.5 and Less than 5 points
5	Less than 2.5 points

의 점수의 총 합을 구하여 그 값에 따라서 ID의 최종 등급을 결정한다. Table 4는 ID의 각 요소별 획득 점수에 따른 최종등급 분류 기준을 나타낸 것이다. 각 ID는 획득 점수에 따라 총 5개의 등급으로 분류된다. 각 요소별 등급과 최종 등급은 ID의 새로운 트윗 데이터를 수집하여 DB에 저장하면서 갱신하게 된다.

3.6 전체 시스템 구조

Figure 4는 시스템의 전체적인 구조도이다. 질의가 들어오면, 클라이언트 2와 클라이언트 3을 이용하여 트윗을 수집한다. 이때, 태풍이나 지진 같은 위치기반의 이벤트의 성질을 띠는 경우 클라이언트 2를 이용한다. 클라이언트 2는 키워드와 지역 좌표를 이용하여 트윗을 수집한다. 정치·사회의 성격을 띠는 이벤트의 경우 클라이언트 3을 이용하여 트윗을 수집하게 된다. 이때, 클라이언트 3은 ID 분석기를 통해 분류된 ID 그룹을 인자로 받고 이 ID 그룹을 이용하여 데이터를 수집한다. 데이터 수집을 하지 않을시, 클라이언트 2는 유휴상태이며, 클라이언트 1은 ID를 수집하여 서버에 저장한다. 이때 저장된 ID는 다시 클라이언트 3의 인자로 넘겨지며 클라이언트 3은 넘겨받은 ID를 이용하여 트윗을 수집한다. 클라이언트 3이 트윗을 수집하는 이유는 새로 수집된 ID에 대해 ID 분석기에서

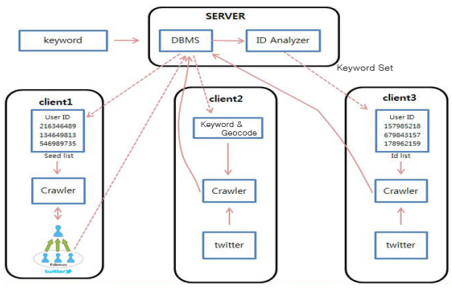

Figure 4. Overall System Structure

사용할 트윗 데이터를 수집하기 위해서이다.

4. 실험 및 결과

본 실험에서는 10,850개의 ID를 팔로워 수, 키워드를 포함하고 있는 트윗량 그리고 최근 트위터 사용량을 고려하여 ID 등급을 분류하였다. 그리고 분류된 ID를 크게 5개의 그룹으로 나누어 각 ID 그룹으로 트윗을 수집하고 분석하였다. 또한 수집시스템의 신뢰도를 분석하였다.

4.1 ID 등급 분류

첫 번째로 10,850개의 ID를 팔로워 수를 고려하여 5개의 등급으로 분류하였다. Table 5는 팔로워 수로 분류한 ID의 등급 정보를 나타낸 것이다. 등급이 낮아질수록 ID의 개수는 증가하지만 ID당 평균 트윗량은 감소하는 모습을 보였다.

두 번째로 10,850개의 ID를 ID의 최근 트윗 사용량을 고려하여 6개의 등급으로 분류하였다. Table 6은 최근 트윗 사용량을 고려하여 분류한 ID의 등급 정보를 나타낸 것이다. 10,850개의 ID에 대하여 분류하였으나, 실제 등급이 분류된 ID의 개수는 총 4,407개였다. 6,443개의 ID가 제외된 원인은 최근 2개월 동안 트윗 사용량이 없었기 때문으로 분석된다. 등급 분류에서 제외된 6,443개의 ID는 최근 트윗 사용량의 등급 점수에서 0점을 획득한다.

세 번째로 10,850개의 ID에 대하여 키워드를 포함한 트윗량을 고려해 4개의 등급으로 분류하였다. Table 7은 키워드를 포함한 트윗량을 고려하여 분류한 ID의 등급 정보를 나타낸 것이다. 등급이 내려갈수록 ID 당 평균 키워드를 포함한 트윗량이 크게 감소하는데 이것은 상위 소수의 유저들이 키워드를 포함하고 있는 트윗의 대부분을 작성하기 때문으로 분석된다.

마지막으로 3가지 요소를 고려하여 분류한 등급에 따른 ID의 점수에 따라서 5개의 등급으로 ID의 최종 등급을 분류하였다. Table 8은 3가지 요소를 모두 고려하여 분류한 최종 ID 등급 정보를 나타낸 것이다. 등급이 낮아질수록 ID의 개수는 증가하지만 ID당 평균 트윗 양은 감소함을 알 수 있다.

Table 5. ID Grading Information Considering the number of Followers

Grade	The number of IDs	The average amount of tweets per ID
1	144	181
2	130	179
3	227	174
4	1,524	166

Table 6. ID Grading Information Considering the Amount of Recent Tweet Consumption

Grade	The number of IDs
1	333
2	359
3	450
4	431
5	657
6	2,237

Table 7. ID Grading Information Considering the Amount of Tweets Containing Keyword

Grade	The number of IDs	The average amount of tweets containing the keyword ID
1	57	76
2	109	32
3	289	13
4	10,394	1.8

Table 8. Final ID Grading Information Considering 3 Elements

Grade	The number of IDs	The average amount of tweets per ID
1	9	199
2	71	187
3	265	177
4	1,974	153
5	8,510	95

4.2 ID 분석기의 효율성 검증

ID 분석기의 효율성을 검증하기 위하여 ID 분석을 통해 분류된 ID를 크게 5개의 그룹으로 나누고 각 ID 그룹으로 트윗을 수집하고 분석하였다. 첫 번째 ID 그룹은 임의로 분류한 3,000개의 ID이다. 두 번째 ID 그룹은 최근 트윗 사용량을 고려하여 분류한 총 6개의 등급 중 3등급 이상의 ID이다. 세 번째 ID 그룹은 팔로워 수를 고려하여 분류한 총 5개의 등급 중 3등급 이상의 ID이다. 네 번째 ID 그룹은 키워드를 포함하고 있는 트윗 수를 고려하여 분류한 총 4개의 등급 중 3등급 이상의 ID 이다. 마지막으로 다섯 번째 ID 그룹은 최근 트윗 사용량, 팔로워 수 그리고 키워드를 포함하고 있는 트윗 수의 등급 점수의 합계로 분류한 총 5개의 최종 등급 중 2등급 이상의 ID이다. 각 실험은 첫 번째 ID 그룹과 각각의 ID 그룹을 이용하여 트윗을 수집, 분석하여 그 결과를 비교하였다. 그 결과 임의로 분류한 ID 그룹에 비하여 키워드를 포함하고 있는 트윗 수가 약 1,500개가 적었지만 ID 당 키워드를 포함한 평균 트윗 수는 약 27배가 많았으며 수집시간 역시 약 1/27로 단축하였다.

4.3 ID 분석기의 신뢰도 분석

ID 분석기의 신뢰도를 분석하기 위해 4.2절에서 사용된 3가지 요소를 모두 고려한 ID 그룹으로 수집한 트윗을 분석하였다. 수집된 트윗을 ID별로 나누어 ID 분석기로 분류된 ID가 이벤트 검출에 적합한 트윗을 보유하고 있는지를 검사하였다. Table 9는 ID의 등급별 신뢰도이다. 1등급으로 분류된 ID 9개중 8개의 ID가 이벤트 검출에 적합한 트윗을 보유하고 있었으며, 나머지 1개의 ID는 홍보성 트윗이 다수를 이루어 이벤트 검출에 적합하지 않았다. 또한 2등급으로 분류된 ID는 총 71개 중 54개가 이벤트 검출에 적합한 트윗을 보유하고 있었다. 이벤트 검출에 적합하지 않다고 판단된 17개의 ID 중 3개의 ID는 정치인들의 ID이었으며, 나머지 14개의 ID는 유용한 생활 정보 등을 주로 작성하는 ID이었다. Figure 5는 각 등급별 ID 그룹의 신뢰도와 시스템 신뢰도를 그래프로 나타낸 것이다. 1등급으로 분류된 ID 그룹의 신뢰도는 약 88.8%를 나타내었으며, 2등급으로 분류된 ID 그룹의 신뢰도는 약 76%를 나타내었다. 따라서 위의 두 등급의 합산을 본 논문에서 제시하는 ID분석기에 대한 신뢰도로 보았을 때, 이 시스템의 신뢰도는 약 77.5%이다.

Table 9. Reliability of ID Grades

Grade	Number of valid ID	Number of Full ID	Reliability (%)
1	8	9	88.8
2	54	71	76.05

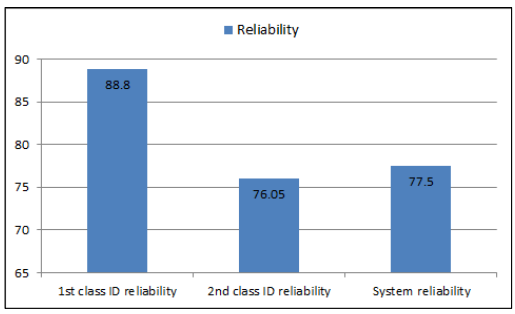

Figure 5. Each ID Group and System Reliability

5. 결론 및 향후 연구

본 논문에서는 위치기반의 재난 및 질병에 관한 이벤트뿐만 아니라, 정치·사회적 성향의 이벤트 감지를 위한 트위터 수집 시스템을 제안하였다. 정치·사회적인 측면에서 개인적인 성향 파악은 결코 무시 할 수 있는 부분이 아니며, 사용자의 성향 파악을 위해서는 지속적인 ID 수집 및 관리가 필수적이다. 본 시스템은 ID 분석기를 통하여 사용자의 팔로워 수, 트윗 사용빈도 그리고 키워드를 포함하고 있는 트윗 양을 고려하여 ID의 등급을 분류한다. 또한 분류한 ID 등급으로 정치·사회적 이벤트 감지에 적합한 ID를 추출하고, 데이터를 수집한다. 추가로 위치 기반의 이벤트 감지를 위한 데이터 수집도 지원하며, 지속적인 ID 수집 및 분석을 통하여 데이터 수집 및 이벤트 감지에 충분한 양의 ID를 확보한다. ID의 등급 분류를 함으로써 고려해야 되는 ID의 양을 줄여 데이터 수집 기간을 줄이고 이벤트 검출 및 사용자 성향 파악에 필요한 충분한 데이터를 수집할 수 있다.

ID 분석기의 신뢰도 측정을 위해 ID의 등급 분류 및 분류된 ID로 수집된 트윗의 분석 결과 77.5%의 정확도를 보였으나, 2등급으로 분류된 ID 중 정치적 이벤트 검출에 적합하지 않다고 판단되는 17개의 ID가 발견되었다. 이 결과는 팔로워 수와 키워드를 포함한 수, 그리고 최근 트위터 사용량을 고려하는 것만으로는 정확한 사용자의 성향을 판단하기에는 부족하다는 것을 보여준다.

향후 연구로는 좀 더 정확한 사용자 성향 파악을

위한 자연어 처리 기술과 감정 분석 기술에 관한 연구와 수집 데이터를 이용한 이벤트 검출에 대한 연구를 수행할 것이다.

References

[1] Cheblb, N. K; Sohall, R. M. 2011, The Reasons Social Media Contributed to the 2011 Egyptian Revolution, Journal of Business Research and Management, 2(3):139-162.

[2] Cho, A. R; Kang, Y. O. 2010, The Design and Implemenation of Festival Information Website using the GeoRSS Function, Journal of Korea Spatial Information Society, 18(1):89-99.

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

[4] Daniel. G.-A. 2012, A Balanced Survey on Election Prediction using Twitter Data, Department of Computer Science, University of Oviedo May 1.

[5] Hwang, K. S. 2011, Twitter Users Increased and the Concentration of Top 1% has Intensified, The Kyunghyang Shinmun.

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

[7] Korea Communications Commission, 2011, 2011 Responses to National Audit Written Interrogatories.

[8] Kwon, O. J; Kim, J. H; Li, K. J. 2010, A Spatial Data Stream Processing System for Spatial Context Analysis in Real-time, Journal of Korea Spatial Information Society, 18(1):69-76.

[9] Lee, B. S; Hwang, B. Y. 2012, A Study of the Correlation between the Spatial Attributes on Twitter, Paper presented at the 28th Conference of Data Engineering Workshop on Spatio Temporal data Integration and Retrieval, 337-340.

[10] Lee, B. S; Kim S. J; Choi, W. S; Jang, K. H; Yoon, J, Y; Hwang, B. Y. 2011, Analyzing the Credibility of the Location Information Provided by Twitter Users, Paper presented at the 28th Conference of Korea Multimedia Society, 1-3.

[11] Min, J. S. 2012, Study on Twitter users political participation, Journal of Korea Regional Communication Research Association, 12(2):274-303.

[12] Mislove, A; Lehmann, S; Ahn, Y. Y; Onnela, J. P; Rosenquist. J. N. 2011, Understanding the Demographics of Twitter Users, Paper presented at the 5th Conference of AAAI on Weblogs and Social Media (ICWSM'11), 554-557.

[13] Mustafaraj, E; Finn, S; Whitlock C; Metaxas, P. T. 2011, Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail, Paper presented at Conference of Social Com / PASSAT, 103-110.

[14] Nagarajan, M; Gomadam, K; Sheth, A. P; Ranabahu, A; Mutharaju, R; Jadhav, A. 2009, Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences, Paper presented at the 10th Conference of LNCS on Web Information Systems Engineering, 539-553.

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

[16] Sakaki, T; Okzaki, M; Matsuo, Y. 2010, Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, Paper presented at the 19th Conference of World Wide Web, 851-860.

[17] Twitter Search API, 2013, https://dev.twitter.com/docs/api/1/get/search

[18] Twitter Streaming API, 2013, https://dev.twitter.com/docs/streaming-apis

논문접수 : 2014.2.12
수 정 일 : 2014.6.5
심사완료 : 2014.6.9