트위터에서 추출한 감기 증상의 사회적 신호와 영향요인과의 상관분석

윤진영[†], 김석중^{††}, 이범석^{***}, 황병연^{****}

요 약

소셜 네트워크 서비스의 인기가 증가하면서 현재 이슈가 되는 정보를 추출하거나 증상 등을 추적하는 분석연구가 활발히 진행되고 있다. 특히 트위터에서는 트윗을 통해 이벤트 현상에 대한 반응과 사용자의 개인적인 상태 등이 잘 반영되어 트윗을 센서로 적용한 이벤트 추출이 가능하다. 최근에는 지진과 같은 재난 이벤트뿐만 아니라 사회 전반적으로 관심을 갖고 있는 질환이나 질병에 대한 사회적 신호가 질병의 확산을 조기에 감지하는데 도움을 주고 있다. 논문에서는 감기를 대상 이벤트로 지정하여 트위터로부터 수집한 정보를 감기 신호라는 사회적 신호로 간주하고 분석하였다. 추출된 감기 신호의 신뢰성을 확인하기 위해 감기 확산의 환경적 요인으로 알려진 세 가지 기후 요소와 기상청에서 제공하는 감기 기상지수와의 상관분석을 통해 변수들 사이의 상관관계를 파악한다.

A Correlation Analysis between the Social Signals of Cold Symptoms Extracted from Twitter and the Influence Factors

Jinyoung Yoon[†], Seokjung Kim^{††}, Bumsuk Lee^{***}, Byung-Yeon Hwang^{****}

ABSTRACT

With the huge success of Social Network Services, studies on social network analysis to extract the current issues or to track the symptoms of epidemic disease are being carried out actively. On Twitter, tweets reflect people's reaction to an event and users' individual status well, so it is possible to detect an event regarding a tweet as a sensory value. Recently, social signals are used to detect the spread of illness like the flu as well as the occurrence of disaster event like an earthquake in early stages. In this paper, we set up a cold as a target event and regarded tweets as Cold Signals. To evaluate the reliability of Cold Signals, we analyzed correlations between weather factors and the cold index provided by Korea Meteorological Administration.

Key words: Twitter(트위터), Cold(감기), Correlation Analysis(상관분석), Social Signal(사회적 신호)

※ 교신저자(Corresponding Author): 황병연, 주소: 경기도 부천시 원미구 지봉로 43 가톨릭대학교 컴퓨터정보공학부(420-743), 전화: 02) 2164-4363, FAX: 02) 2164-4777, E-mail: byhwang@catholic.ac.kr
접수일: 2013년 1월 30일, 수정일: 없음
완료일: 2013년 2월 16일
[†] 가톨릭대학교 컴퓨터공학과
(E-mail: my_sk_test@catholic.ac.kr)

^{††} 가톨릭대학교 컴퓨터공학과
(E-mail: typicalkorean@catholic.ac.kr)

^{***} 가톨릭대학교 컴퓨터공학과
(E-mail: bumsuk@catholic.ac.kr)

^{****} 가톨릭대학교 컴퓨터공학과
※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단 기초연구사업(No. 2011-0009407)의 연구비와 2013년 가톨릭대학교 교비연구비의 지원으로 수행되었음.

1. 서 론

최근 몇 년 동안 꾸준히 증가한 소셜 네트워크 서비스의 인기는 사용자들의 사회적 활동 범위의 확장 뿐만 아니라 정보검색이나 사회 과학 분야 연구의 발전을 이끌었다. 소셜 네트워크 서비스 연구 분야의 관심은 현재 이슈가 되는 정보를 추출하거나 실제 주변에서 일어나는 현상이나 증상을 추적하는 분석 연구로 이어지고 있다[1]. 특히 트위터는 140자로 제한된 단문 텍스트를 기반으로 하는 마이크로블로그 서비스로 사용자들은 사회적으로 이목이 집중되는 현상이나 상황에 대해 실시간으로 반응하거나 자신의 개인적인 상태 등을 트윗으로 남긴다. 작성된 트윗들은 텍스트 형태의 내용과 함께 위치정보(GPS)와 시간정보 등 다양한 정보를 포함하고 있다. 또한, 트위터 사용자는 전 세계에 걸쳐 있기 때문에 트윗에 남겨진 정보의 활용을 통한 넓은 범위의 연구 진행이 가능하다.

이러한 정보들은 이벤트를 감지하는 연구에서 유용한 센서로 작용하며 Takeshi Sakaki 등의 논문[2]에서는 사용자를 개별적인 센서로 가정하고 사용자가 작성한 트윗의 위치정보를 분석하여 지진 이벤트를 추출하였다. 지진과 관련된 트윗을 수집하기 위하여 키워드 두 가지(earthquake, shake)를 지정하였으며 키워드가 빈번하게 발생하는 지역을 실시간으로 감지하는 방법으로 이벤트에 대해 관찰하였다. 최근에는 지진과 같은 특정 현상에 대한 이벤트뿐만 아니라 사회 전반적으로 관심을 갖고 있는 질환이나 질병에 대한 연구도 활발히 이루어지고 있다. 특히 계절적 요인과 전염성 요인에 영향을 많이 받는 질병의 경우에는 사용자를 센서로 적용한 사회적신호가 질병의 확산을 조기에 감지하여 예방하는 데에 도움을 줄 수 있다.

우리나라는 계절의 구분이 뚜렷하여 기온이 낮은 겨울과 일교차가 큰 환절기에 감기와 독감 같은 질병이 유행한다. 감기는 우리나라를 비롯한 전 세계에서 누구에게나 쉽게 발생하는 질병으로, 보통 상기도염 또는 급성 비인두염이라고도 불리나 하나의 질환이라기보다는 여러 증상들이 집합되어있는 흔한 질병이다. 감기를 발생시키는 200여 가지의 바이러스가 체내에 침투하면 일련의 면역학적 반응이 일어나 콧물, 재채기, 기침 등의 증상들을 나타나게 한다. 반면,

독감(인플루엔자)은 갑작스런 고열, 두통, 근육통과 같은 전반적인 신체 증상을 나타낸다는 점에서 감기와 매우 유사하지만 인플루엔자 바이러스에 의한 급성 호흡기 질환으로 감기의 원인균과 병의 경과가 다르기 때문에 다른 질병으로 구별된다. 독감은 매년 겨울에 소규모로 유행되고 있으며 전염성이 강하여 면역이 약한 노인이나 소아, 다른 질환을 앓고 있는 사람에게 발병되면 사망률이 증가하는 위험한 질병이다.

감기는 바이러스뿐만 아니라 온도, 일교차, 상대습도 등의 기후 요인에도 영향을 받는다. 감기가 감염되는 경로는 감기 환자로부터 재채기나 기침을 통해 적은 양의 바이러스가 옮겨지는 공기 전파와 손이나 물건의 직·간접 접촉을 통해 감염되는 접촉 전파가 있다. 하지만 감기는 독감과 다르게 질병을 예방할 수 있는 백신이 없고 약 또한 단지 증상을 완화시켜주는 용도로만 사용되고 있어 독감에 비해 훨씬 감염 속도가 빠르고 감염범위가 넓다. 따라서 감기 유행에 대한 빠른 탐지를 통한 위험성 인지가 중요하며 손을 자주 씻고 감기 환자와 접촉을 최소화하는 것이 감기를 예방하기 위한 최선의 방법이다.

독감 증상을 이벤트로 지정하여 감지하는 연구들은 이미 여러 논문에서 선행되어 왔다. Jeremy Ginsberg 등의 논문[3]에서는 구글 검색 엔진(Google Search Engine)에서 발생한 독감 질의어(query)를 분석하여 미국 지역을 대상으로 독감의 유행 흐름을 감지하는 방법을 제안하였고 Vasileios Lampos 등의 논문[4]에서는 위치정보를 가진 트윗과 영국보건국(HPA: Health Protection Agency)에서 제공하는 인플루엔자 의사환자(ILI: Influenza-Like Illness) 자료를 바탕으로 독감 유행 추적을 자동화하는 툴을 개발하였다. 특히 Vasileios Lampos 등의 논문은 독감과 관련된 키워드 설정으로 수집한 트위터의 데이터 스트림을 분석하고 독감지수로 수치화하였다는 점에서 소셜 네트워크 서비스 내의 정보를 사회적 신호로 인식한 대표적인 연구로 꼽힌다. 그러나 감기와 독감의 전염성, 유행성 측면의 차이로 인해 독감 연구와 차별화된 감기 유행 감지 방법이 고안되어야 한다.

본 논문에서는 감기의 유행을 조기에 감지하기 위해 질병을 반영하는 사회적 신호로서 트위터에 게시된 트윗들을 관찰하였다. 트위터에서는 정보 전달뿐

만 아니라 사용자 자신의 정확한 의견과 상태, 기분 등이 잘 비춰지고 있어 감기와 같은 흔한 질병에 대한 내용이 상태나 기분을 반영한 트윗에 자주 나타난다. 따라서 사용자들이 게시한 트윗들 중에서 감기와 관련된 키워드를 포함한 트윗을 추출하여 감기의 유행 정도를 측정하는 수치로 사용하였으며 이 수치를 기상청에서 제공하는 감기 기상지수와 기후 요소 데이터와 비교하였다. 감기와 관련된 트윗을 추출하기 위해서 ‘감기’라는 단어를 포함하여 재채기, 콧물 등 감기 증상을 나타내는 키워드 9개를 선정하고 서울 특별시와 부산광역시 등 9개 주요 도시를 관찰 지역으로 지정하여 트윗을 수집하였다. 해당 도시의 크기에 따라 지정된 관찰 범위가 각각 다르기 때문에 트윗과 비교 데이터들을 정규화하여 패턴을 비교하는 방법[5,6]보다는 상관관계를 분석하여 변수사이의 영향을 파악하는 방법이 더 적합하다. 상관분석은 관찰한 두 변수 사이의 원인과 결과 관계를 회귀방정식을 통해 구하는 분석방법으로 소셜 네트워크 분야에서도 상관계수를 측정하여 변수 간의 작용을 관찰하는 연구[7]가 진행되고 있다.

논문에서 검출한 사회적 신호가 감기의 유행 현상을 얼마나 정확하게 관찰할 수 있는지 판단하는 과정이 필요하며 이를 위해 트윗을 실제 유행 현상을 나타내는 2가지 요소인 감기 기상지수, 기상청 기후 관측자료와 비교하였다.

첫 번째 요소인 감기 기상지수[8]는 기상청에서 일교차, 최저기온, 상대습도에 따른 감기 발생 가능성을 4단계로 지수화하여 감기 발생이 증가하기 시작하는 9월부터 이듬해 4월까지 서울, 강릉, 청주 등 9개의 도시를 대상으로 제공하는 서비스이다. 기상 상황 변화에 민감한 어린이나 노약자들은 특히 감기 바이러스에 노출되기 쉽기 때문에 예방이 매우 중요하며 유익한 정보를 제공함으로써 건강한 생활 유지에 도움을 준다는 목적을 가지고 있다. 지수화된 단계는 낮음부터 매우 높음까지 4단계로 구분되고 당일과 익일의 단계를 도시별로 나타낸다. 높음과 매우 높음 단계로 예보된 기간에는 감기에 걸릴 가능성이 증가하며, 이는 감기와 관련된 키워드를 포함하는 트윗 수의 증가를 예상할 수 있다.

두 번째는 기상청에서 제공하는 기후 관측 자료이다. 기후는 면역력을 감소시켜 감기를 발생시키거나 증상을 악화시킬 수 있는 요인이며 그 중 일교차, 최

저기온, 상대습도는 기후 요소 중 가장 많은 영향을 미치는 요소이다. 장경화의 논문[9]에서는 국내 인플루엔자 의사환자 데이터와 기후요인과의 연관성을 분석하여 독감 발병에 대한 기후요소들의 영향력을 연구하였다. 감기는 독감과 원인균이 다르지만 발병 기후 조건이 비슷하므로 감기 신호의 신뢰성을 판단하는 과정에서 기후가 중요한 요인으로 작용할 것이라고 예상된다. 따라서 일 단위로 측정된 세 가지 기후 데이터와 트윗 데이터의 상관관계를 분석하여 감기 증상과의 관련성을 찾는다. 이처럼 두 가지 비교 데이터 집합과 트윗 데이터의 상관분석을 통해 소셜 네트워크 서비스로부터 검출된 사회적 신호가 얼마나 대상 이벤트를 감지하는 데에 유용한 신호로 작용할 수 있는지 증명할 수 있다.

논문의 구성은 다음과 같다. 2장에서는 관련연구로 본 연구의 바탕이 된 독감 감지 시스템에 관한 연구들을 소개하고 3장에서는 연구에서 사용된 데이터 세트의 수집과 정제 과정에 대해 설명한다. 4장에서는 분석 결과를 제시하고 마지막으로 5장에서 연구의 결론과 의미를 정리하며 향후 연구계획을 소개한다.

2. 관련연구

대상 이벤트를 감지하기 위한 연구는 이미 여러 분야에서 진행되고 있으며 독감을 이벤트로 지정하여 유행을 추적하는 연구들도 선행되었다. 초기 연구 시기에 발표된 Philip M. Polgreen 등의 연구[10]와 Jeremy Ginsberg 등의 연구는 검색 엔진의 질의어 분석으로 질병의 활동을 감지하여 독감 관련 질의어 검색과 실제 독감의 발병 현상사이의 관련성을 찾는다는 공통점을 가지고 있다. Jeremy Ginsberg 등의 연구에서는 구글 검색 엔진에서 발생한 검색 질의어 중 독감 관련 키워드를 관찰하여 미국 각 지역에 대해 주별 독감 활동을 탐지하는 방법을 제시하였으며 실제 인플루엔자 의사환자 비율과 검색 질의어의 비교에서 유사한 패턴을 가진다는 결과를 얻었다.

그림 1은 관찰한 질의어와 IILI 데이터에 대한 수치를 나타낸 그래프로 각각 붉은 색과 검은 색으로 표시된 데이터의 패턴 비교를 통해 연관성을 나타내고 있다. 독감과 관련된 질의어는 다른 계절에 비해 겨울철에 증가하는 패턴을 보이고 있으며, 인플루엔자

그림 1. 독감 관련 질의어 검색과 ILI 데이터 사이의 패턴 분석

의사환자 데이터 또한 겨울에 해당하는 연말과 연초 기간의 수치가 가장 높게 나타나 두 데이터 사이의 패턴 유사성을 확인할 수 있다. 이 연구에서는 검색 엔진에서 발생한 질의어를 특정 대상 이벤트 관찰을 위한 신호로 사용하였다는 특징을 가지고 있으나 단순히 독감에 대한 키워드를 검색했다는 사실로 사용자에게 독감이 발병되었다고 판단하기에는 부족하다. 사용자가 검색 엔진에서 독감에 대한 검색을 하는 경우는 실제로 독감에 걸려 질병의 증상과 치료에 대한 정보를 얻기 위한 것과 발병과는 상관없이 예방과 원인 정보를 얻기 위한 것으로 나누어진다. 따라서 독감에 대한 이벤트 관찰 연구의 방식이 검색 엔진을 바탕으로 한 연구에서 소셜 네트워크 서비스를 사회적 신호로 활용한 연구로 발전되고 있다.

트위터와 같은 서비스에서는 사용자가 직접 자신의 상태에 대한 정보를 트윗이라는 단문 텍스트로 남긴다는 점에서 질의어 검색보다 대상 이벤트와의 직접적인 연관성을 갖는다고 볼 수 있다. Vasileios Lamos 등은 영국의 49개 도심지에서 발생한 위치 정보를 가진 트윗들을 통해 인플루엔자 유행을 추적하는 자동화된 툴을 구현하였다. 그들은 트윗과 인플루엔자 의사환자 비율을 데이터로 사용하기 위해 2009년 6월부터 2010년 3월까지 40주에 걸쳐 하루 평균 대략 200,000개의 지리적 위치정보를 포함하는 트윗을 수집하고 영국 보건국으로부터 북부, 중부, 남부로 나눈 세 지역에 해당하는 인플루엔자 의사환자 데이터를 모았다. 트위터의 검색(Search) API를 통해 독감 증상을 나타내는 키워드를 지정하여 수집된 관련 트윗들을 RSS(Really Simple Syndication) 형태로 탐색하는 과정을 거쳐 RSS 피드를 읽어와 사용하기 쉬운 형태로 변경해주는 오픈 소스인

ROME 자바 API를 이용하여 파싱하였다. 이 연구는 트위터를 통해 독감에 대한 사회적 신호를 감지하였다는 점에서 본 논문의 연구방법과 유사한 점을 보이고 있다. 이 밖에도 일반적인 이벤트들의 개념을 확장하여 소셜 웹을 관찰해 독감 증상을 감지하는 방법을 제안하고 소셜 웹상의 체계적이지 않은 많은 양의 텍스트를 통해 이벤트를 검출하는 다양한 연구[11]가 진행되어 왔다.

관련연구에서 소개한 논문들은 대상 이벤트를 독감으로 지정하였으며 질의어나 트윗 분석을 통해 얻은 데이터들을 인플루엔자 의사환자 데이터와 비교하였다. 하지만 인플루엔자 의사환자 데이터는 수집과 분석과정에 소요되는 시간 때문에 데이터의 관찰 시간이 1-2주 정도 지연된다는 단점을 가지고 있다. 감기는 독감과 유사하지만 감염과 예방방법에서 차이를 보이는 질병으로 전염성이 더 강하고 백신 또한 존재하지 않아 이른 단계의 감지가 필수적인 질병이다. 따라서 키워드를 감기로 설정하는 실험 방법과 비교 대상이 달라져야 한다. 본 논문에서는 감기에 대한 사회적 신호를 감지하기 위해 트윗을 분석하였으며 비교 대상을 질병 확산에 영향을 미치는 기후 요소와 감기 기상지수로 확장시켰다.

3. 사회적 신호 검출과 데이터 정제

3.1 사회적 신호 검출

감기에 대한 사회적 신호 검출은 트위터로부터 감기와 관련된 트윗을 수집하여 분석하는 것으로 시작된다. 감기는 기온이 낮은 겨울철과 일교차가 큰 환절기에 더 많이 발생하기 때문에 수집기간을 2011년 11월 28일부터 2012년 5월 2일까지로 설정하여 관찰

하였다. 트위터에는 특정한 조건을 부여해 해당 질의에 상응하는 트윗을 수집하는 여러 방법이 존재하는데, 검색 API와 스트리밍(Streaming) API를 이용하는 것이 대표적인 방법이다. 본 논문에서는 API를 선택하기 위해 다음과 같은 조건을 고려하였다.

- 1) 검색할 키워드를 모두 적용할 수 있는가?
- 2) 검색할 지역을 정확히 적용할 수 있는가?

트윗을 검색하기 위해서는 여러 조건을 이용할 수 있다. 키워드, 지역, 사용자 등의 조건을 부여하여 트윗을 수집할 수 있는데 1)의 경우에는 검색 API나 스트리밍 API가 모두 키워드에 해당하는 조건 사용을 허용하기 때문에 실험에 제약이 없다. 하지만 2)처럼 지역을 조건으로 사용하는 경우에는 조건 지정 방식에서 두 API가 차이를 보인다. 검색 API는 중심좌표와 반경에 대한 값을 지정하여 질의문을 원의 범위로 작성하는 반면 스트리밍 API의 경우에는 좌측 하단점과 우측 상단점에 해당하는 두 개의 좌표값으로 질의문을 작성하여 최종적으로 사각형을 범위로 한다. 그러나 두 API 모두 도시의 경계를 범위로 지정할 수 없기 때문에 정확하게 해당 도시에서 발생하는 트윗만을 수집하는 것은 불가능하다. 게다가 기상청에서 감기 기상지수를 제공한 9개의 도시는 서로 지리적 규모가 달라 논문에서는 도시의 경계를 가급적 벗어나지 않으면서 가능한 많은 지역을 포함할 수 있는 검색 API를 선택했다. 아래 표 1은 실험에 적용한 지역의 중심좌표와 반경을 나타낸다.

실험에서는 표 1에서 지정한 9개 지역을 중심으로 감기 관련 트윗을 수집하기 위해 질의에 포함할 키워드 리스트를 생성하였다. 수집에 사용한 키워드는 감기, 기침, 오한, 콧물, 재채기 등 질병 이름과 증상을

나타내는 단어목록으로 구성되었다. 지정한 지역과 키워드 리스트를 조건으로 만족하는 트윗들은 작성된 시간과 위치 등 모든 정보를 포함하여 JSON 데이터 형식으로 저장된다. 수집된 트윗의 정보와 내용은 구현된 JSON 파서를 거쳐 트윗_ID, 사용자_ID, 트윗 내용 등 실험에 필요한 정보로 이루어진 문자열로 생성하였다. 트윗을 저장하는 데이터 형식으로 지정 가능한 JSON은 XML보다 기능이 적지만 파싱이 빠르고 간단하다는 장점을 가지고 있어 많은 데이터가 발생하고 이를 처리해야 하는 트위터 연구에서 많이 사용된다[12].

3.2 비교 데이터 수집

논문에서는 2011년 11월부터 기상청에서 제공하는 감기 기상지수가 실제 감기의 유행을 잘 추적하는지 관찰하고자 하였다. 하지만, 감기 기상지수는 트윗을 수집한 9개의 지역에 대해 기후 정보를 바탕으로 감기가 걸릴 가능성을 지수화한 것이기 때문에 기상청이 예측한 가능성의 정확도는 입증된 바가 없다. 따라서 트윗에서 추출한 사회적 신호로 파악된 감기 환자 추정치와 감기 기상지수의 그래프 형태와 상관관계를 분석함으로써 얼마나 신뢰성을 가진 정보인지를 판단할 수 있다. 위험도에 따라 4단계인 매우 높음, 높음, 보통, 낮음으로 구분되는 감기 기상지수는 오후 3시 이전에는 당일과 다음날에 대한 예보를, 이후에는 다음날과 모레에 대한 예보 서비스를 제공한다. 실험을 위해 2011년 11월 이후의 감기 기상지수 데이터 중 트윗 수집 종료기간인 2012년 5월 2일까지의 데이터를 사용하였다.

기후는 독감뿐 아니라 감기 바이러스 확산과 면역력에도 영향을 미치는 요인으로 Brian P. Hanley 등의 연구[13]에서는 인플루엔자 바이러스의 전파에 대한 겨울철 온도와 상대습도와의 연관성을 증명하였다. 이 연구에서는 바이러스 생존과 전달에 유리한 조건으로 0℃~20℃ 사이의 온도와 50%이하의 낮은 상대습도를 꼽았다. 감기도 이 두 가지의 기후 요인에 영향을 받을 것으로 예상되며, 감기의 원인 중 하나로 알려진 일교차를 포함한 세 요인의 데이터를 기상청 기후 관측 자료로부터 얻었다.

3.3 데이터 정제

트윗 수집을 위한 조건으로 사용된 위치는 지역의

표 1. 데이터 수집에 이용한 중심좌표와 반경

지역	중심좌표(위도,경도)	반경
서울	37.568889,126.976667	50km
강릉	37.750000,128.900000	51km
청주	36.641994,127.488803	12km
대전	36.350000,127.385000	23km
전주	35.816667,127.150000	14.4km
광주	35.166667,126.916667	22.5km
대구	35.871250,128.601262	29.7km
부산	35.179562,129.075641	26.6km
제주	33.366667,126.533333	43km

크기와 인구수에 따라 서로 다른 범위를 가지고 있고, 비교를 위해 사용된 데이터들도 각각 나타내는 수치적 기준이 서로 다르기 때문에 모든 데이터 값을 다음과 같은 간단한 식으로 정규화하는 과정이 필요하다.

$$v' = \frac{(v - \text{Min}(V))}{(\text{Max}(V) - \text{Min}(V))}$$

정규화된 v' 는 각각의 데이터 집합 V 의 최대값과 최소값을 이용하여 단일 데이터 값 v 를 미리 정한 범위인 0.0과 1.0 사이로 표시한 값이다. 감기와 관련하여 발생한 트윗의 수와 기후 요소 데이터는 그 값들이 수치로 표현되어 있어 정규화 과정을 통한 범위 변경이 가능하였으나 감기 기상지수는 단계화되어 나타내어졌기 때문에 낮은 단계부터 1에서 4까지의 수치 값으로 변환하였다.

4. 실험 및 결과

본 장에서는 트위터에서 발생한 트윗이 대상 이벤트를 반영하는 사회적 신호로써 얼마나 정확하게 작용하는지 확인하기 위하여 감기 관련 수치 데이터들과 비교하는 실험을 진행하였다. API를 통해 수집된 트윗은 기본적으로 트위터 측에서 정한 항목들에 대한 정보를 모두 포함하고 있다. 하지만 사회적 신호 역할을 하기 위해 필요한 정보들은 대상 이벤트와 사용목적에 따라 다르게 정해진다. 실험에서는 사회적 신호 역할을 하는데 사용되는 정보만을 추출하기 위해 데이터 정제 과정을 거쳐 표 2의 항목들로 분류하고 “트윗_id#.#사용자_id#.#트윗내용#.#키워드#.#작

성시간#.#위치” 형태의 문자열로 생성하였다.

정제된 데이터들을 위치와 키워드 별로 군집화시키는 과정에서 지역에 따라 키워드별 발생빈도 차이를 발견하였다. 감기 이벤트를 잘 반영하는 트윗을 수집하기 위해 우리는 앞서 키워드 9개를 지정하였으며 이 키워드 리스트는 직접적으로 병명을 나타내는 단어인 감기, 고열과 간접적인 단어인 증상 관련 단어들로 이루어져 있다. 모든 지역에서 감기라는 단어를 포함한 트윗의 수가 절대적인 수치로 나타낼 것이라고 예상하였으나 그림 2에서 보듯 지역별로 키워드가 발생한 빈도의 차이가 확연하게 나타났다. 강릉에서는 고열을 제외한 거의 모든 키워드가 비슷한 비율로 발생하였지만 전주의 경우에는 감기 키워드 비율이 70%에 가까웠다. 또한, 고열은 감기를 의미하지만 현대에는 잘 쓰지 않는 단어 중 하나로 모든 지역의 트윗에서 거의 사용되지 않은 점을 확인할 수 있었다.

다음은 정규화 과정을 거친 트윗 데이터를 패턴 해석을 통해 분석한 결과이다. 트윗을 비롯한 기후 요소, 감기 기상지수 데이터는 수집 기준 단위가 1일로 지정된 데이터이며 그 값들은 시간이 지남에 따라 곡선의 그래프와 유사한 형태를 보인다. 그림 3은 서울에서 발생한 트윗들로부터 추출한 감기 신호를 나타낸 그래프이다. 12월과 2월 초 그리고 4월 말에 감기 신호가 크게 증가 했으며 감기 기상지수와 기후 요소도 이러한 그래프 형태로 나타나 패턴을 비교할 수 있었다.

패턴을 분석하여 두 데이터 집합 사이의 연관성을 파악하는 과정은 정확한 연관관계 정도를 나타내기

표 2. 수집된 트윗에서 추출한 정보 샘플

트윗_id	사용자_id	트윗 내용	키워드	작성 시간	위치
#0235521	M39884	@sy_style05 #친목당 감기가 아직도ㅎㅎ	감기	Mon, 02 Apr 2012	강릉
#9806464	M29866	목이따끔따끔 감기가오려나	감기	Sun, 01 Apr 2012	강릉
#3215490	M97145	기침 피로워..	기침	Thu, 29 Mar 2012	서울
#4887552	M38411	혹 몸살감긴가?아 아픔당	몸살	Sun, 01 Apr 2012	제주
#0855168	M03631	재채기가 10연속 속사포로 나왔다.... 정신을 못 차리겠네	재채기	Thu, 05 Apr 2012	청주
#2828929	M29819	열37.8도래..ㅋㅋ 링겔맞고왔는데도 온몸이쭈신다ㅠ	열	Thu, 12 Apr 2012	광주
#3377280	M68842	아 코를 풀어도 풀어도 콧물이 나오네 ㅠㅠ	콧물	Wed, 04 Apr 2012	부산
#1090177	M02047	@EarthLittleBoy 몸이 으슬으슬하니..조심해야겠어요ㅎㅎ	으슬	Tue, 10 Apr 2012	서울

그림 2. 각 지역 트윗들의 키워드별 발생빈도(단위:%)

그림 3. 트윗에서 추출한 감기 신호

어렵기 때문에 두 변수 사이의 원인과 결과 관계를 설명할 수 있는 상관관계 분석법을 사용하였다. 상관 분석으로 두 변수가 함께 변화하는 직선적 관계를 예측하고 인과관계를 분석하여 한 변수의 변화에 따른 다른 변수의 변화를 예측할 수 있다. 또한 상관계수를 통해 변수 사이의 관계정도를 측정하여 정상관계 또는 부상관관계를 구분한다. 상관분석 실험의 첫 번째 단계는 산점도를 그려 두 변수의 대략적 관계를 알아보는 것이다. 상관계수는 두 변수 간의 직선적 관계의 유무정도를 파악하는 값이기 때문에 산점도를 통한 관계 분석이 선행되어야 한다. 그림 4는 서울에서 12월 한 달간 발생한 트윗과 상대습도, 최저기온 사이의 산점도를 나타낸 그림이다. 산점도 분석 결과로 추정회귀방정식과 결정계수(R^2)를 얻었으

며 상대습도와 최저기온의 선형관계에서 모두 음의 기울기를 갖는 부상관관계를 보이고 있는 것을 확인할 수 있었다.

산점도 분석 후에는 상관계수를 구하여 두 변수 사이의 연관 정도를 수치적으로 나타낸다. 상관계수는 두 변수의 표준편차의 곱에 대한 공분산의 비율로 정의하며 상관계수의 부호가 (+)이면 양의 상관관계, (-)이면 음의 상관관계라고 한다. 상관계수 r 의 범위는 $-1 \leq r \leq 1$ 이고 -1 또는 1 에 가까울수록 상관이 높고 0 에 가까울수록 상관이 낮다. 표 3은 감기 신호와 세 가지 기후요소, 감기 기상지수 사이의 상관정도를 보여준다. 감기 신호는 감기 관련 키워드를 포함하는 트윗으로부터 추출했기 때문에 발생한 트윗량의 증가는 감기 환자수가 증가했다는 것을 말하고,

그림 4. 감기 신호와 상대습도, 최저기온 사이의 산점도

각각의 변수들과의 상관계수를 통해 상관관계를 분석할 수 있다. 일교차는 하루의 최고기온과 최저기온의 차이를 나타내며 일교차가 커질수록 면역력이 약해져 감기 바이러스에 감염될 확률이 높아진다. 이러한 사실을 기반으로 했을 때 일교차의 증가에 따라 트윗량도 증가하는 양의 상관관계를 보여야 하는데, 실험 결과에서 상관계수의 부호가 (+)인 것으로 두 데이터 사이의 연관성을 확인하였다.

감기 신호와 비교한 네 가지 변수 중 가장 상관관계가 높은 변수는 상대습도이다. 최저기온과 상대습도는 일교차보다 감기에 더 영향을 미치는 요인으로 나타났으며 최저기온과 상대습도가 낮아지면 감기

에 걸리는 환자는 반대로 증가하는 음의 상관관계를 확인할 수 있었다. 상대습도는 대체적으로 강수량 등에 의해 달라져 변화의 폭이 매우 넓다는 특성을 가지고 있지만 겨울철은 대체로 상대습도가 낮고 건조하기 때문에 감기에 대한 상관도가 가장 높은 것으로 보인다. 이러한 결과는 인플루엔자 바이러스의 확산과 전파에 대한 연구[13]에서 도출된 바가 있다. 이 연구에서는 기온과 상대습도가 낮은 20℃ 이하, 50% 이하의 조건에서 인플루엔자 바이러스의 전달률이 가장 높게 나타났다. 이는 유사 질병인 감기의 발병기후 조건과 매우 비슷하다는 점에서 감기 신호가 대상 이벤트에 대한 사회적 신호로 잘 검출되어 그 유용성이 충분함을 보여준다.

기상청이 제공하는 감기 기상지수는 여러 기후 요인들의 상호작용을 감안하여 감기에 대한 위험성을 알리는 지수이며 실험을 위해 단계적으로 나타내어진 데이터를 1부터 4라는 숫자 값으로 수치화하였다. 그림 5에서와 같이 감기 기상지수가 높아진 2011년 12월 말과 1월 초, 2월 초·중순 기간에는 특히 최저기온과 상대습도가 매우 낮은 것으로 확인되었다. 또한 감기 신호와 양의 상관관계를 나타내고 그래프의 증감 시점 패턴이 트윗에서 추출한 감기 신호와 유사한 패턴을 띄고 있다.

마지막으로, 표 4에서는 2012년 3월과 4월 감기 신호와 기후 요인, 감기 기상지수 사이의 상관계수를 지역별로 비교 분석하였다. 감기 기상지수와 가장 큰 상관관계를 나타낸 지역은 대구였지만 서울과 대전 지역에서는 감기 기상지수가 높게 나타나도 트윗이 감소하는 음의 상관관계를 보였다. 이는 분석한 기간이 환절기이므로 겨울철 보다 기후요소에 대한 영향을 덜 받는 것으로 해석된다. 전체적으로 트윗이 감기 신호로써 잘 반영된 지역은 일교차와 감기 기상지수에 정상관계를 갖고 최저기온과 상대습도에는 부상관계를 보이는 부산, 제주 지역이다. 또한 최

표 3. 감기 신호와 각 변수들 사이의 상관계수

상관계수 r	Tweets	일교차	최저기온	상대습도	감기 기상지수
Tweets	1				
일교차	0.08629	1			
최저기온	-0.26763	-0.56994	1		
상대습도	-0.46707	-0.45578	0.71202	1	
감기 기상지수	0.08492	0.51002	-0.76524	-0.56285	1

그림 5. 감기 신호와 감기 기상지수

표 4. 각 지역별 상관계수

상관계수 r	서울	강릉	청주	대전	전주	광주	대구	부산	제주
트윗수	1	1	1	1	1	1	1	1	1
감기 기상지수	-0.0166	0.1033	0.0919	-0.0571	0.0531	0.0429	0.3499	0.2527	0.1035
최저기온	-0.0271	-0.0598	0.1011	-0.3371	-0.3452	-0.4555	-0.4404	-0.5079	-0.1476
일교차	-0.0549	-0.0543	0.1313	-0.1657	-0.2225	-0.1620	-0.0112	0.3338	0.0204
상대습도	0.1048	0.0009	-0.0817	0.2862	0.1245	0.0389	-0.1623	-0.2538	-0.0504

저기온은 청주를 제외한 모든 지역에서 상관계수가 음의 부호를 갖고 특히 부산 지역에서는 그 정도가 중위 부상관으로 높게 나타나 최저기온에 감기 신호가 민감하게 작용하는 것으로 확인되었다.

5. 결론 및 향후 연구계획

본 논문에서는 감기의 유행을 초기에 감지하기 위해 감기를 대상 이벤트로 지정하여 트위터로부터 수집한 정보를 사회적 신호로 간주하고 분석하였다. 감기는 바이러스의 공기 전파와 접촉 전파 형태로 감염되나 독감과 구별되는 질병으로 이른 시기의 탐지를 통한 위험성 인지와 예방이 중요하다. 이에 실험에서는 기상청에서 제공하는 감기 기상지수와 세 가지 기후 요소 데이터와의 비교 분석을 통해 검출한 사회적 신호의 신뢰성을 확인하였다. 실험 결과에 따르면, 기후 요소 중 최저기온과 상대습도의 그래프가 감기 신호 그래프 추이와 매우 비슷한 패턴을 가지고 있었으며 기온이 낮고 상대습도가 낮은 날씨에 감기 신호가 증가하는 부상관관계 결과를 보였다. 하지만 3가지 기후 요소는 단독적일 때 보다 모든 요소가 다 고려될 때 더 감기 신호와 유사했으며 이는 모든

요소들의 상호작용을 통해 얻어진 감기 기상지수와 의 패턴 분석에서 두드러지게 나타났다.

논문의 감기 신호는 대상 이벤트를 감지하기 위해 검색엔진의 질의어를 분석하는 대신 소셜 네트워크 서비스에서 나타난 정보를 바탕으로 검출한 사회적 신호이다. 트위터를 이용하는 점에서 이전 연구들과 신호 추출 방법이 유사하지만 대상 이벤트에 따른 다양한 키워드 설정과 관련 데이터들과의 분석을 통해 감기에 대한 이른 감지를 가능하게 했다. 이번 연구에서는 대상 이벤트 관련 트윗들을 검색 API로 수집하였기 때문에 지난 트윗에 대한 사회적 신호 검출 연구가 이루어졌으나 향후에는 스트리밍 API를 이용한 실시간 데이터 수집으로 계속적인 분석을 통한 경보 시스템 구축이 가능해질 것으로 보인다. 더 나아가 실시간으로 수집되는 데이터에 예측 알고리즘을 추가하여 감기 발생가능성을 예보하는 연구로 확장할 수 있을 것이다.

참 고 문 헌

- [1] 손영우, "Social Networks과 Twitter 서비스에 관한 고찰," 멀티미디어학회논문지, 제14권, 제4

- 호, pp. 546-553, 2011.
- [2] 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.
- [3] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, "Detecting Influenza Epidemics using Search Engine Query Data," *Nature*, Vol. 457, pp. 1012-1014, 2009.
- [4] V. Lampos, T.D. Bie, and N. Cristianini, "Flu Detector - Tracking Epidemics on Twitter," *Proc. of the European Conf on Machine Learning and Principles and Practice on Knowledge Discovery in Database*, pp. 599-602, 2010.
- [5] Bumsuk Lee, Jinyoung Yoon, Seokjung Kim, and Byung-Yeon Hwang, "Detecting Social Signals of Flu Symptoms," *Proc. of the 8th IEEE Int'l Conf on Collaborative Computing: Networking, Applications and Worksharing*, pp. 544-545, 2012.
- [6] 이범석, 김석중, 윤진영, 황병연, "온라인 소셜 트렌드의 감지 및 시각화," 제30회 한국멀티미디어학회 추계학술대회 논문집, 제15권, 제2호, pp. 95, 2012.
- [7] A. Abbasi, J. Altmann, and L. Hossain, "Identifying the Effects of Co-authorship Networks on the Performance of Scholars: A Correlation and Regression Analysis of Performance Measures and Social Network Analysis Measures," *Journal of Informetrics*, Vol. 5, No. 4, pp. 594 - 607, 2011.
- [8] 한국 기상청, http://www.kma.go.kr/weather/lifenindustry/health_05.jsp, 2011.
- [9] 장경화, "인플루엔자 의사환자와 기후요인과의 연관성 - 2005년~2008년의 인플루엔자 표본감시 자료를 중심으로," 고려대학교 석사논문, 2009.
- [10] P.M. Polgreen, Y. Chen, D.M. Pennock, F.D. Nelson, and R.A. Weinstein, "Using Internet Searches for Influenza Surveillance," *Clinical Infectious Diseases*, Vol. 47, No. 11, pp. 1443-1448, 2008.
- [11] V. Lampos and N. Cristianini, "Nowcasting Events from the Social Web with Statistical Learning," *ACM Transactions on Intelligent Systems and Technology*, Vol. 3, No. 4, Article 60, pp. 1-22, 2011.
- [12] G. Wang, "Improving Data Transmission in Web Applications via the Translation between Xml and Json," *Proc. of the Third Int'l Conf on Communications and Mobile Computing*, pp. 182 - 185, 2011.
- [13] B.P. Hanley and B. Borup, "Aerosol Influenza Transmission Risk Contours: A Study of Humid Tropics versus Winter Temperature Zone," *Virology Journal*, Vol. 7, Article 98, pp. 1-18, 2010.

윤진영

2011년 가톨릭대학교 컴퓨터공학과 학사
2011년~현재 가톨릭대학교 컴퓨터공학과 석사과정
관심분야 : 데이터베이스, 정보검색, 소셜 네트워크 분석, 데이터마이닝, 시맨틱 웹

김석중

2011년 가톨릭대학교 컴퓨터공학과 학사
2013년 가톨릭대학교 컴퓨터공학과 석사
관심분야 : 소셜 네트워크 분석, 데이터마이닝, 오피니언 마이닝, 데이터베이스, 정보검색

이범석

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 : 데이터베이스, 소셜 네트워크 분석, 웹데이터 마이닝, 정보검색

황병연

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

1994년~현재 가톨릭대학교 컴퓨터정보공학부 교수
1999년 University of Minnesota Visiting Scholar
2007년 California State University, Sacramento Visiting Scholar
관심분야 : 소셜 네트워크 분석, XML 데이터베이스, 데이터마이닝, 정보검색, 지리정보시스템