식품안전 관련 트위터 정보의 연관 관계 분석 및 시각화

소현수0, 강승식, 오세욱 국민대학교 컴퓨터공학부, 국민대학교 식품영양학과

hyss4412@naver.com, sskang@kookmin.ac.kr, swoh@kookmin.ac.kr

Coocurrence Relation Analysis and Visualization in Tweet for Food Safety Domain

Hyun-Su So⁰, Seung-Shik Kang, Se-Wook Oh School of Computer Science & Foods and Nutrition, Kookmin University

요 약

식품안전 사고가 발생했을 때 뉴스, 인터넷 기사를 통해 정보를 인지하기 전에 그 음식을 섭취하는 경 우가 발생하는 문제점 최소화하기 위하여 실시간 트윗 분석으로 현재 발생한 식품안전 키워드와 어느 지 역에서 발생했는지를 신속하게 파악하고. 키워드 연관관계 분석 프로그램을 활용하여 정확한 정보를 추출 한다. 이와 더불어, SNS 등 다양한 정보 소스로부터 추출한 정보를 간단명료하게 파악하기 위해서 워드 클라우드 등 데이터 시각화 기법을 활용하여 시각화로 정보를 제공한다. 이 기법은 식품안전 뿐만 아니라 최근 발생한 콜레라 감염 발생과 같은 문제를 해결하기 위한 방법으로 활용될 수 있을 것이다.

주제어: 식품 안전, 트위터, 트윗, 실시간 분석, 데이터 시각화, 워드 클라우드, 키워드 연관관계

1. 서론

SNS는 뉴스보다 빠르게 정보를 전파할 수 있다. 미국 에서 이륙하려던 비행기가 활주로를 벗어나는 사고가 발 생했을 때, 승객들은 트위터를 활용해서 구조 요청을 했 고, 빠르게 구조가 이뤄졌다. 이처럼, SNS는 실시간 정 보를 빠르게 전파할 수 있는 매개체이다. 본 논문은 SNS 중 트윗 정보를 분석하여 실시간으로 식품안전 사고를 인지하고 최대한 빠른 시간내에 조치할 수 있도록 하거 나 식품안전 관련 지식정보 자료로서 온톨로지를 구축하 는데 도움이 되는 시스템을 구축하고자 한다[1,2]...

2. 실시간 트위터 분석

트위터 정보를 실시간으로 분석하는 프로그램은 Python에서 제공하는 twython library를 이용하여 구현 하였는데 이 프로그램은 실시간으로 올라오는 SNS의 게 시물을 활용하여 식품 안전정보를 보다 빠르게 사용자에 게 전달하기 위한 것이다[3.4]. 이 프로그램은 2가지 방 법으로 정보를 얻을 수 있다. 첫째, 특정 키워드가 포함 된 게시물을 검색하여 관련 정보를 사용자에게 제공한 다. 예를 들어, '식중독'이라는 키워드를 사용할 경우 에 트위터에 식중독이 포함된 트윗 게시물을 실시간으로 전송받는다. 이를 통해, 실시간으로 식중독이 발생하고 있다는 것을 파악할 수 있으며 추가로 게시물 내용을 분 석하여 발원지와 발생의 원인 등을 분석한다.

둘째, 특정 GPS 좌표의 주변에서 작성된 게시물 정보 를 사용자에게 제공한다. 예를 들어, 국민대학교의 GPS 좌표를 입력해 두면 지정한 영역에서 작성된 게시물을

받을 수 있다. 이를 통해, 실시간으로 해당 지역에서 작 성된 게시물들을 분석하여 식품안전과 관련된 정보가 발 견되면 이를 사용자에게 실시간으로 제공한다.

Date: Sat Apr 30 22:11:05 +0000 2016 이, 이번에는 매실청을 조금 만들어들까 해. 여름에 챙겨 마시면 줄거든. 그런 의미에서 매실차를 추천할 계. 매실에는 카테킨산이 용부해서 장 안의 좋지 않은 균을 억계하고 장을 지키기에 좋아. <u>식중독이나 배탈</u> 에도 좋다고 하니까 더운 날에는 챙겨 마셔. Date: Sat Apr 30 21:41:50 +0000 2016 봄나물 캠먹어서.. 식중독같아..

Date: Sat Apr 30 21:19:12 +0000 2016 대전동구, Sa인 미만 급식취약시설 식중독 예방 위생점점

Date: Sat Apr 30 21:04:52 +0000 2016 RT @ryujing68: 식중독 예방 뮤지컬 보러 5386명 성남시청 찾아 ihansung https://t.co/MZtLojaKGe

Date: Sat Apr 3D 19:3B:46 +0000 2016 《Casin.P.FE (^...식중도에 검렬...지도오...) 응! 처음에는 아무반응없었는데 점점가면감수록 평평울었 어....(닷맥은지 볼을긁적인다) =..캡틴의 기도가 이뤄질까나?! 이뤄졌으면 좋겠다! 캡틴이 설탕복음밥을 못먹을수있도록 도와구세요..!

Date: Sat Apr 30 19:18:09 +0000 2016 RT @ha_nosh: 1. 한국에 주로 알려진 일본의 공해병은 이타이이타이병이나 미나마타병이지 싫다. 오늘 소개 하고 싶은 사건은 카네마유소 암말히 말하자면 공해병은 아니다. <u>식중독 사건이다</u>. 그러나 화학물질에 의 한 사건이고, 가슴기 살균게 사건과..

그림 1 . 특정 키워드가 포함된 게시물 검색 '식중독' 관련 트윗 정보 예제

Tweet from @nole_twt Date: Sun May 01 01:18:33 +0800 2016 290RT @erid4244: 한국대기절 연구하시는 분이 마세먼지주의보 안 내려도 5e마이크로그램 이상이면 야외운동 자제하고 꼭 마스크쓰라고 당부했다. 요몇년 사이 정말 심각할 정도로 미세먼지 내 중금속성분 비율이 높아졌다고

그림 2 . 특정 GPS좌표의 주변에서 작성된 게시물 검색 서울특별시 광화문 반경 20Km 정보

Tweet from gtwonc_co_kr Dete: Sun May 01 01:18:21 +0000 2016 104[한국언론의 세대교체 ◆브레이크뉴스◆] <u>인원시교육청, 2019년 학생 감영병 때방관리 단당가 연수 실시</u> https://t.co/www.adbhlAp

그림 3 . 특정 GPS좌표의 주변에서 작성된 게시물 검색 인천광역시 주안역 반경 20Km 정보

이 프로그램을 활용하여 얻은 정보는 텍스트 형태가 아니라 그 내용을 시각화하여 사용자에게 한눈에 보여주

¹⁾ 이 논문은 2015년 정부(미래창조과학부)의 재원으로 국가과 학기술연구회 융합연구단 사업(No. CRC-15-05-ETRI)의 지원 을 받아 수행된 연구임.

기 위하여 한국어 형태소 분석기²⁾를 활용하여 키워드를 추출하고 그 키워드들을 R 프로그래밍 라이브러리를 이 용하여 시각화하였다.



그림 4 . 트위터 정보에 대한 워드 클라우드 생성

3. 키워드 연관관계 분석

연관관계 분석은 문서에 나오는 단어끼리 어떤 연관관계를 맺고 있는지 파악하는 방법이다. 예를 들어, 삼국지 문서를 분석하면 유비의 측근들은 유비의 주변에 나타나고 조조의 측근들은 조조의 주변에 나타나게 된다. 연관관계 분석을 분석하기 위해서 프로그램은 텍스트 파일로 문서가 들어오면 한국어 형태소 분석기를 활용하여한 문장씩 형태소 분석을 시작한다. 한 문장의 명사들은 각각을 Key와 Value 쌍의 형태로 만들어서 한 문장에 함께 출현했음을 표시한다. 마지막으로 모든 분석이 끝났으면 Key를 내림차순 정렬하여 출현 빈도가 높은 Key순으로 어떤 연관관계가 있는지를 D3를 통해 시각적으로보여준다.

한국, 지카바이러스 안전지대 아니다…국가별유행확률

日本 10160604 1028 | 中間 20160694 1028

SK680 000 000 3457000

한국, 지카바이러스 안전지대 아니다...국가별유행확률

우리나라에 지카 바이러스 유행 가능성은 15~30%

브라질 등 중남미에서 유행 중인 지카 바이러스 감염증 의 확산이 심상치 않다.

최근 일본 연구진이 전 세계 189개국을 대상으로, 국가 별 지카 바이러스 유입 및 지역 감염(유행) 위험을 추정 하는 새로운 통계모델을 개발하고 일부 추정 결과를 발 표했다.

(생략)

http://nownews.seoul.co.kr/news/newsView.php?id=20160604601015

그림 5 . 문서를 텍스트 파일로 저장

연관관계 분석은 문서에서 많이 출현한 단어와 그 단어들과 관련이 있는 단어들을 시각적으로 확인할 수 있어서 단어의 빈도 정보만 활용하는 'wordcolud' 방법보다 더 자세하게 문서의 내용을 파악할 수 있다.

그림 6 . 연관관계 어휘 추출

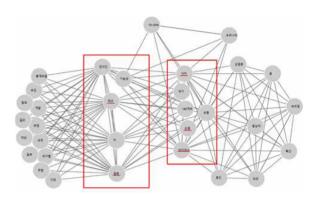


그림 7 . D3.js로 시각화 빈도수가 많이 포함된 단어를 기준으로 연관관계

4. 결론

본 연구에서는 트위터의 실시간 정보를 키워드와 위치기반으로 제공받고 빈도수로만 게시물을 파악하는 것이 아닌 키워드 연관관계를 분석하여 내용을 보다 정확하게 파악할 수 있다. 그리고 R과 D3를 이용하여 한눈에 볼수 있게 시각화하였다. 연구 결과로 개발된 연관관계 분석 프로그램은 문서의 특징을 파악하는데 활용하거나 콜레라 발생 및 신종 인플루엔자와 같은 전염 바이러스의 발생을 인지하고 확산을 방지하는 목적으로 활용될 수있을 것이다.

참고문헌

- [1] S. Blanchemanche and A. Rona-Tas, "An Ontology of Scientific Uncertainty: Methodological Lessons from Analyzing Expressions of Uncertainty in Food Risk Assessment", ISA/ESA Amsterdam conference, 2013.
- [2] N. Meenachi and M. Baba, "A Survey on usage of Ontology in Different Domains", International Journal of Applied Information Systems(IJAIS), Vol.4, No.2, 2012.
- [3] A. Beulens, Y. Li, M.R. Kramer and J. Vorst, "Possibilities for Applying Data Mining for Early Warning in Food Supply Networks", Workshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment, 2006.
- [4] Y. Maeda, N. Kurita and S. Ikeda, "An Early Warning Support System for Food Safety Risks", New Frontiers in Artificial Intelligence, LNCS Vol. 4012, pp. 446-457, 2005.

²⁾ http://cafe.daum.net/nlpk/