한국빅데이터학회지
제3권 제1호, 2018, pp. 55-62

빅데이터를 이용한 독감, 폐렴 및 수족구 환자수 예측 모델 연구

The Study of Patient Prediction Models on Flu, Pneumonia and HFMD Using Big Data

우종필† · 이병욱 · 이차민 · 이지은 · 김민성 · 황재원

세종대학교 경영학과

요 약

본 연구에서는 그동안 해외에서 주로 실행되어 왔던 빅데이터를 이용한 다양한 질병(독감, 폐렴, 수족구병) 환자수 예측 모델을 개발해 보았다. 기존의 환자수 예측이 병원에서 실제 환자수를 카운팅한 수를 수집하여 발표하는 시스템이라면, 이번에 개발한 연구 모델은 실시간으로 제공되는 질병 관련 단어 및 다양한 기후 데이터를 접목하여 기계학습 방법으로 알고리즘을 만들고, 이를 기반으로 정부에서 발표하기 전 환자수를 예측하는 모델이다. 특히 유행성 질병이 빠르게 확산될 경우, 실시간으로 전파 속도를 파악할 수 있다는 점에서 그 장점이 있다. 이를 위하여 구글 플루 트렌드에서 실패한 부분을 최대한 보완하여 다양한 데이터를 활용한 예측 모델을 개발하였다.

■ 중심어 : 빅데이터, 독감, 폐렴, 수족구, 실시간 예측

Abstract

In this study, we have developed a model for predicting the number of patients (flu, pneumonia, and outbreak) using Big Data, which has been mainly performed overseas. Existing patient number system by government adopt procedures that collects the actual number and percentage of patients from several big hospital. However, prediction model in this study was developed combing a real-time collection of disease-related words and various other climate data provided in real time. Also, prediction number of patients were counted by machine learning algorithm method. The advantage of this model is that if the epidemic spreads rapidly, the propagation rate can be grasped in real time. Also, we used a variety types of data to complement the failures in Google Flu Trends.

■ Keyword : Big Data, Flu, Pneumonia, Outbreak, Real-Time Predictions

2018년 08월 02일 접수; 2018년 08월 03일 수정본 접수; 2018년 08월 31일 게재 확정
† 교신저자 yujongpil@sejong.ac.kr

Ⅰ. 서 론

구글이 웹기반 검색어 서비스인 구글 트렌드(Google trend)를 처음 오픈 한 이후, 이를 이용한 연구들이 해외에서 꾸준히 진행되어 왔다. 구글 트렌드를 이용하여 미래를 예측하는 서비스 중 특히 독감을 예측한 구글 플루 트렌드(Google Flu Trends : GFT)를 빼놓을 수 없다. GFT는 환자들이 병원을 찾기 전 그 증상을 검색한다는 점에 착안하여, 질병 관련 특정 단어의 검색량만으로도 독감 환자의 수 예측이 실시간으로 가능하며, 이는 정부에서 발표하는 시점보다 훨씬 빠르다는 연구 결과를 Nature지에 발표함으로써 세계적으로 큰 이슈를 만들어 냈다[1]. 하지만 2013년 예측이 크게 빗나가면서 현재 더 이상 서비스는 제공되지 않는 상태이다.

이렇게 구글 트렌드를 이용한 연구는 해외에서 활발히 진행되었으나, 한국의 경우 상대적으로 이에 대한 연구가 많지 않았다. 한국에서는 구글이 주된 검색 엔진이 아닌 데다, 구글의 사용자 역시 젊은 층이거나 전문직 종사자들로 편중되어 있어 일반화의 문제가 있을 수 있기 때문일 것이다.

본 연구는 이러한 점을 염두에 두고, 다음과 같은 목적을 바탕으로 연구를 진행하였다. 첫째, 구글의 플루 트렌드와 같은 예측 알고리즘이 한국에도 적용 가능한지 알아본다. 이를 위하여 구글 트렌드의 데이터를 사용하여 한국의 독감 환자수를 예측해 보는 한편, 한국에서 사용량이 많은 네이버 트렌드에서 제공하는 데이터와 함께 분석한다. 이러한 예측 모델이 정부에서 공개한 실제 독감 환자수의 증감과 일치하는지 여부를 확인한다. 둘째, 이러한 예측 알고리즘을 다양한 질병에 적용 가능한지 확인한다. 이를 위하여 독감뿐만 아니라 폐렴과 수족구의 예측 모델도 개발한다. 셋째, 구글 플루 트렌드의 예측 실패를 보완한다. 2013년 GFT의 예측 실패는 검색량만을 이용했기 때문이다. 본 연구의 모델은 질병에 영향을 미칠 수 있는 온도, 강수량 등의 기후 데이터 투입하여 모델의 예측률을 높이고자 한다.

Ⅱ. 이론적 배경

2.1 구글 트렌드 & 네이버 트렌드

구글 트렌드는 구글을 이용한 사용자들이 검색한 검색어를 그래프와 CSV(Comma Separated Value)파일로 제공하는 서비스이다. 이는 특정 국가, 지역, 도시 단위까지의 검색량을 제공하고, 검색 기간 역시 맞춤형으로 제공한다. 또한 건강, 게임, 과학, 금융, 뉴스 등 다양한 카테고리의 정보를 제공하며, 이미지, 구글 쇼핑, Youtube에서 개별 검색을 분리하여 검색할 수 있는 장점도 있다. 한꺼번에 5개까지 단어 검색이 가능하다.

구글 트렌드에서 제공하는 검색량은 실제 검색 횟수를 제공하는 것이 아니라, 검색 기간 중 가장 높은 검색량을 보이는 단어의 검색량을 100으로 설정한 후, 나머지 단어의 검색량을 0~100 scale의 기준으로 표준화한 수치를 제공한다. 이러한 데이터를 CVS 파일로 제공하기 때문에 검색어의 흐름을 알 수 있고, 이 데이터를 각종 분석에 바로 이용할 수 있다는 장점이 있다.

네이버 트렌드는 구글 트렌드와 비슷한 서비스를 제공하는데, 2016년 1월 1일부터 검색이 가능하다. 네이버 트렌드는 구글 트렌드에서 제공하는 서비스 이외에도 범위(PC 혹은 모바일), 성별(남성, 여성), 연령 등을 선택하는 기능이 있다. 이를 이용하여 해당 사용자들의 검색량을 다양하게 비교해 볼 수 있다(https://datalab.naver.com/keyword/trendSearch.naver).

2.2 구글 플루 트렌드(Google Flu Trends)

GFT는 구글 검색어를 바탕으로 독감 환자수를 예측한 웹 서비스이다. GFT는 2008년부터 25개

이상 국가에서 독감 환자의 수를 예측했는데, 현재는 더 이상 서비스를 제공하지 않고 있다. GFT는 사람들이 독감과 같은 질병에 걸렸을 경우 바로 병원에 가기보다 '독감, 고열, 기침, 두통'과 같은 증상을 검색하는 경향이 있다는 점에 착안하였다[2, 3]. 이러한 검색어가 갑자기 폭발적으로 증가하면 이는 곧 독감의 확산을 예측한다고 주장하면서 시작했다. 그리고 이러한 예측은 실제로 미국의 질병관리본부(CDC : Centers for Disease Control)보다 1~2주 정도 빠르게 독감 바이러스의 발병을 실시간으로 감지했다[1]. 이러한 결과를 바탕으로 비슷한 연구가 해외 여러 나라에서 진행되었다[4, 5].

하지만 2013년 GFT는 예측에 크게 실패하였다. 독감이 크게 유행할 것이라는 뉴스가 보도되자 독감에 걸리지 않은 사람들까지 백신을 검색하였다. 이는 GFT가 독감 환자의 수를 과다추정하게 만들었고, 결국 잘못된 예측으로 이어졌다[1]. 검색량으로만 환자수를 예측하는 것에 오류가 발생한 것이다. 결국 이러한 오류를 줄이면서 예측률을 높일 수 있는 모델을 개발하는 것이 학문적으로 또한 실무적으로도 필요한 시점이라고 할 수 있다.

Ⅲ. 연구 방법론

본 연구는 앞서 언급한 GFT 문제점을 보완하고, 독감 외에 다른 질병에 대한 예측이 가능한 한국형 모델을 만들기 위해 다음과 같이 연구를 진행하였다. 우선 구글 트렌드와 네이버 트렌드에서 다양한 검색 데이터를 수집하였다. 또한 질병에 기후가 영향을 미칠 수 있기 때문에 기상청 데이터를 활용하여 온도, 강수량 데이터를 추가하였다. 이를 이용하여 독감, 폐렴, 수족구 등 세 가지 질병에 대한 각각의 예측 모델을 개발하였다.

3.1 환자 데이터 수집

실제 환자 수를 알아보기 위해 보건의료 빅데이터 개방시스템(http://opendata.hira.or.kr/home.do)에서 독감, 폐렴, 수족구 환자 수에 대한 월별 데이터를 수집하였다. 데이터 수집 기간은 2016년 1월부터 2017년 8월까지로 한정하였다. 데이터 수집 당시 2017년 8월까지의 데이터만 공개된 상황이었기 때문이다.

3.2 빅데이터 수집

구글 트렌드와 네이버 트렌드에서 질병과 관련된 다양한 검색어 데이터(예 : '독감', '독감증상', '기침', '고열', '두통', '폐렴', '수족구')를 수집하였다. 데이터 수집 기간은 2016년 1월부터 2018년 4월 현재까지이다. 한편 동일한 기간 기상청에서 공개한 각종 기상 자료 (예 : 평균기온, 최저기온, 최고기온, 일강수량)를 수집하였다. 이를 앞서 수집한 빅데이터와 하나의 데이터 세트로 만들어 분석하였다.

3.3 분석 과정

분석 과정은 2016년 1월부터 2017년 8월까지 정부기관에서 발표한 자료를 가지고 기존의 빅데이터를 학습데이터로 활용하여 알고리즘을 개발한 후, 완성된 알고리즘으로 2017년 9월부터 2018년 4월 현재까지의 환자 수를 예측하는 과정을 따랐다.

3.4 분석 방법

데이터는 SPSS, 엑셀 및 IBM Modeler에 분석되었으며, Random forest, 인공신경망, LSVM 등 7가지 다양한 기법을 사용하여 예측 환자 수 오차를 최소화하기 위해 노력하였다. 또한 기관에서 발표하는 월별 환자 수 데이터를 일별 환자 수로 전환하여, 일별로 환자 수를 예측하는 수준으로 예측력을 끌어 올렸다.

Ⅳ. 연구 결과

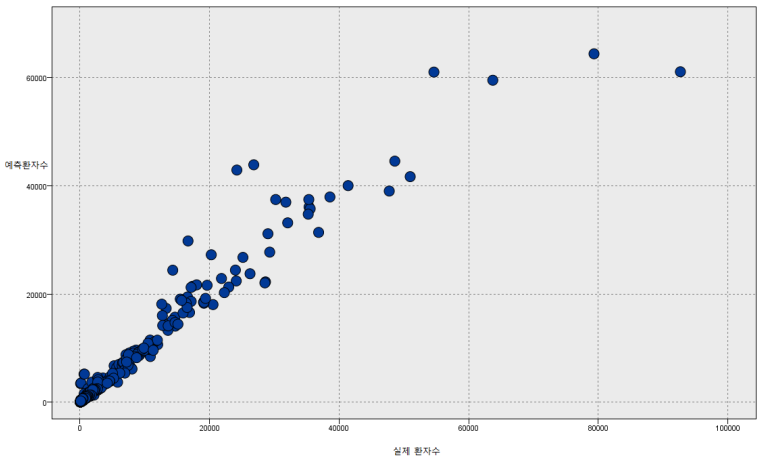
연구 결과는 각 질병(독감, 폐렴, 수족구)에 따라 개별적으로 분석되어 각기 다른 결과가 도출되었다. 각 질병의 연구 모델의 결과 및 예측 환자 수 그래프는 각 모델별로 제공하고자 한다.

4.1 독감환자 모델

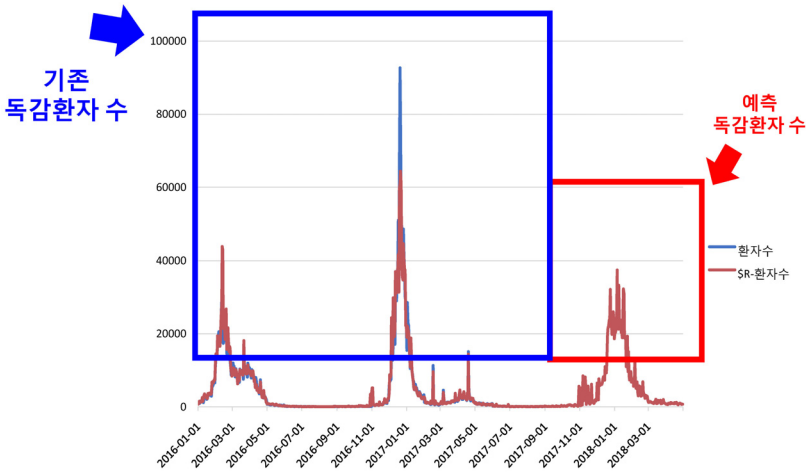
독감환자 모델의 경우, 독립변수는 다양한 검색어와 기후관련 데이터이며, 종속변수는 기존 환자 수로 지정한 후, 앞서 언급한 7가지 기법으로 분석했다. 분석 결과 기존 환자 수와 예측 환자 수 사이에 상관은 .8~.9로 매우 높은 상관관계를 보여줘 기존 데이터를 통해 완성된 알고리즘으로 환자 수 예측이 가능하다는 결과를 보여주었다. 이는 <그림 1>에서 확인할 수 있다.

이러한 알고리즘을 가지고 예측 데이터로 분석했을 시 예측되는 환자 수 결과는 <그림 2>와 같다.

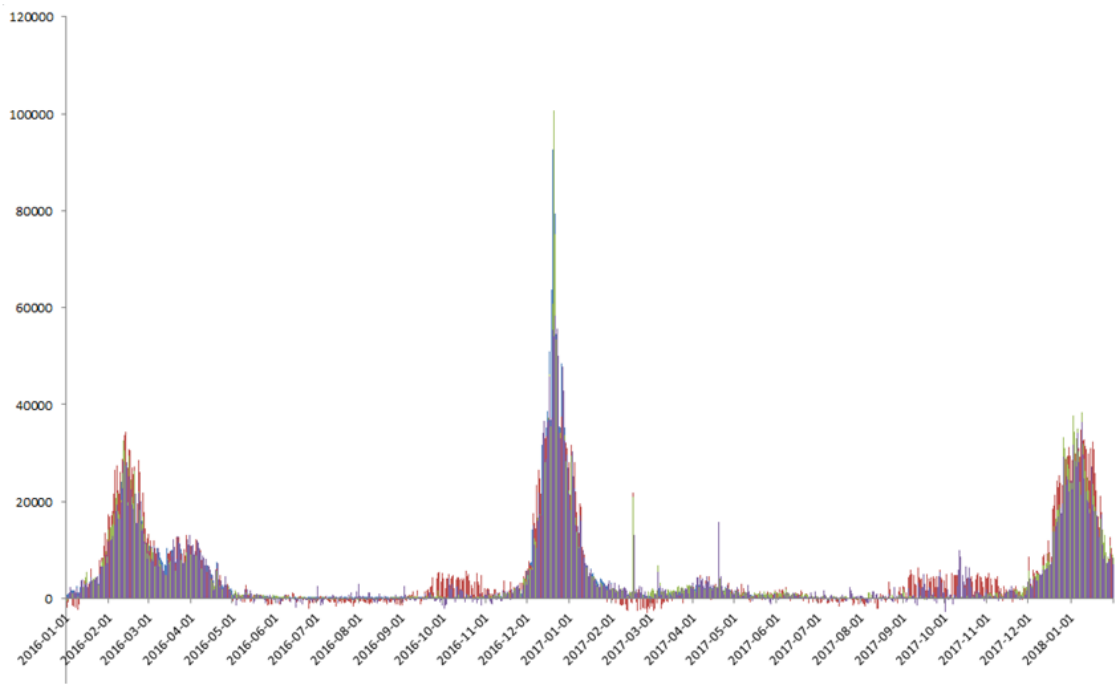
마지막으로, 하나의 기법에서 발생할 수 있는 환자 수의 오류를 줄이기 위해 다양한 기법에서 제공한 환자 수를 한꺼번에 보여준 모델은 다음 <그림 3>과 같다.



〈그림 1〉 Random forest 기법의 실제 환자 수와 예측 환자 수 상관관계



〈그림 2〉 Random forest 기법을 이용한 2017년 9월 이후의 예측 환자 수



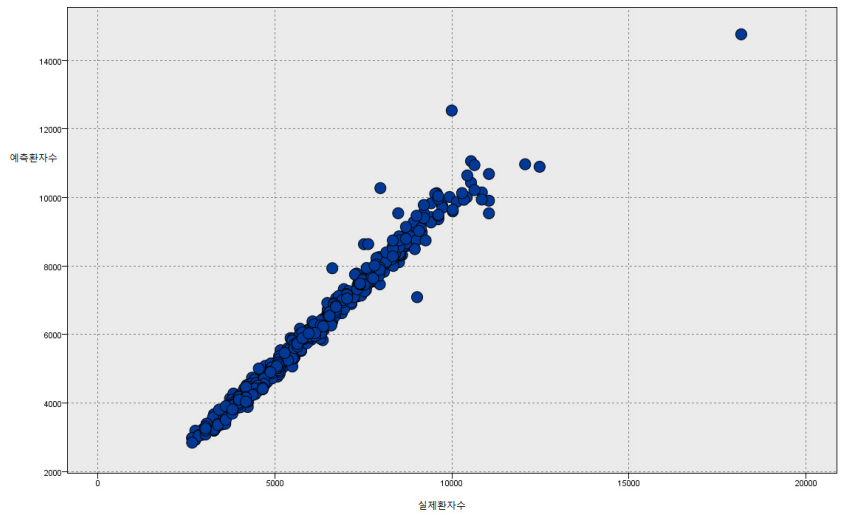
〈그림 3〉 다양한 기법에서 제시한 예측 환자 수

4.2 폐렴환자 모델

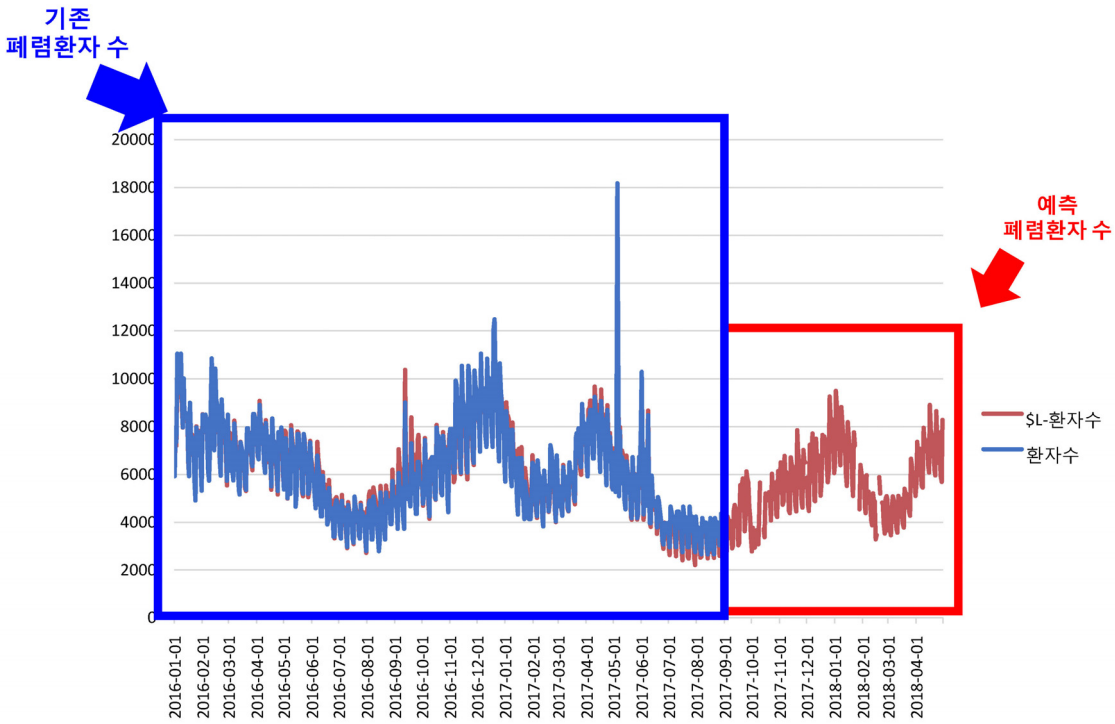
폐렴환자의 경우 역시, 기존 환자 수와 예측 환자 수 사이에 상관은 .9 이상으로 기존 데이터로 환자 수 예측이 가능하다는 결과를 보여주었다.

4.3 수족구 환자 모델

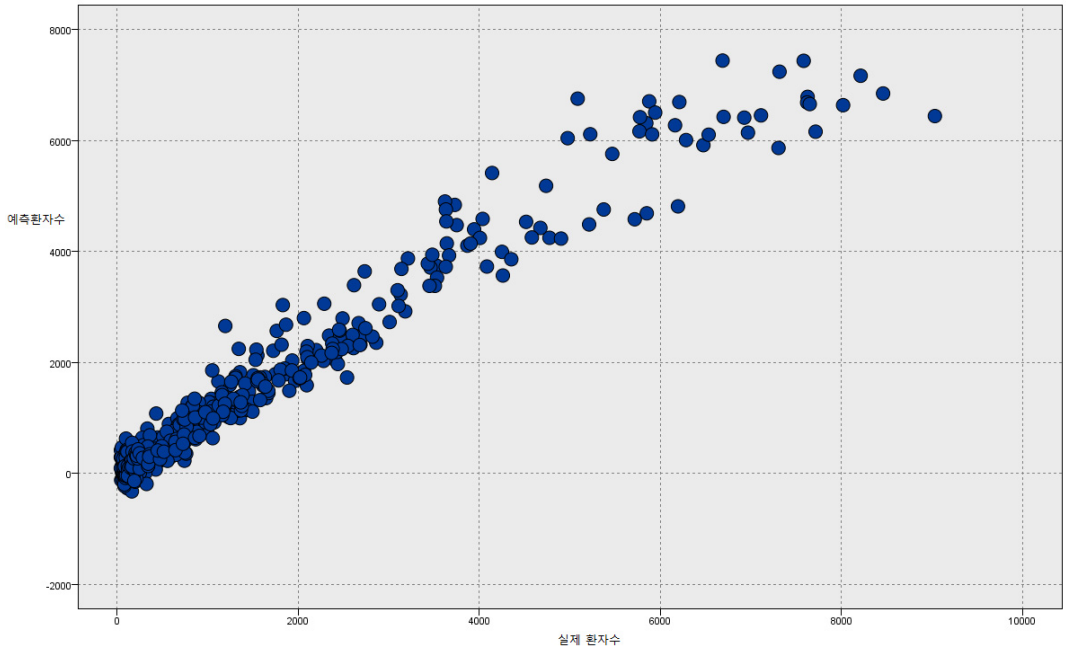
수족구의 경우 역시, 기존 환자 수와 예측 환자 수 사이에 상관은 .8~.9 이상으로 기존 데이터로 환자 수 예측이 가능하다는 결과를 보여주었다.



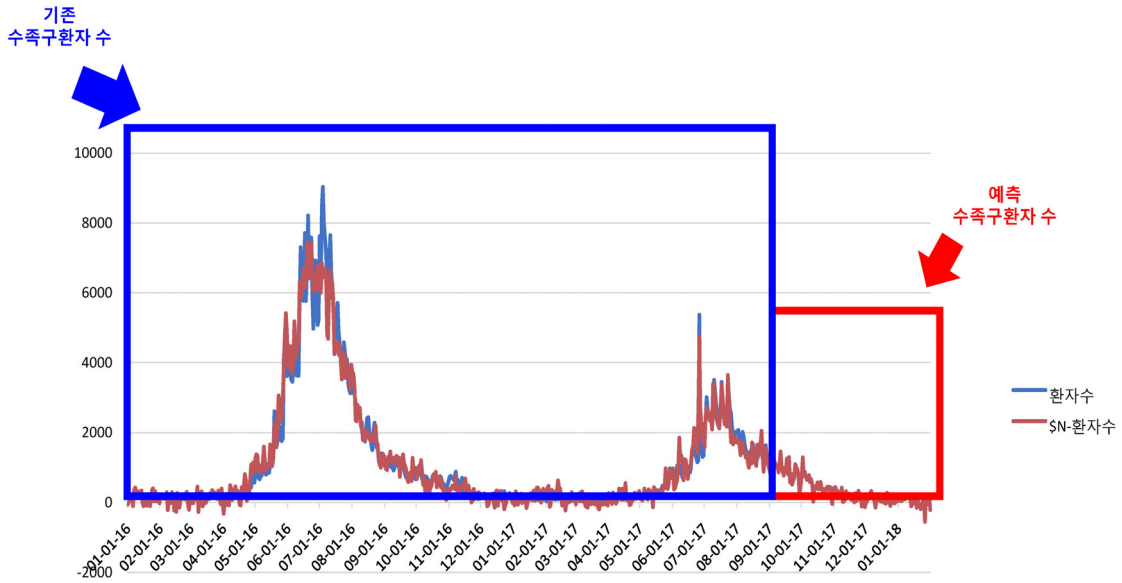
〈그림 4〉 Random forest 기법의 실제 환자 수와 예측 환자 수 상관관계



〈그림 5〉 Random forest 기법을 이용한 2017년 9월 이후의 예측 환자 수



〈그림 6〉 신경망 기법의 기법의 실제 환자 수와 예측 환자 수 상관관계



〈그림 7〉 신경망 기법을 이용한 2017년 9월 이후의 예측 환자 수

Ⅴ. 결론 및 시사점

본 연구는 해외에서 시도되었던 플루 트렌드 형태의 모델이 한국에서도 개발 가능한지 여부 및 독감 이외의 질병도 예측 가능한지 여부를 시도한 것이다. 특히 데이터 수집에 있어서 한국에서 사용자가 많지 않은 구글 트렌드 데이터 이외에 추가로 네이버 트렌드에서 제공하는 검색 데이터도 사용하였다. 또한 검색량 중심의 예측 모델은 GFT에서처럼 오차가 발생할 수 있어, 검색어 데이터 외에 유행병에 영향을 미칠 수 있는 다양한 기후 데이터를 추가하여 분석했다. 분석 결과, 거의 모든 질병 데이터에서 실제 환자 수 데이터와 예측 데이터 간 상관이 .8~.9 이상으로 매우 높게 나와 예측력 높은 알고리즘을 개발했고, 이러한 알고리즘을 바탕으로 실시간 환자 수를 예측 할 수 있었다.

결론적으로 본 연구를 통해 한국에서도 검색 데이터를 제대로 활용한다면, 얼마든지 유행 질병 확산 예측에 도움이 됨을 알 수 있었다. 여기에 기존 제약회사의 데이터나 정부 기관의 다른 데이터가 추가된다면 훨씬 높은 예측력을 가진 모델 개발이 가능할 것이다.

마지막으로 현재 검색 서비스를 제공하는 구글, 네이버 등의 검색 엔진이 검색어의 검색 지역 정보까지 제공한다면, 질병이 어디서 발생하여 어느 지역으로 확산되고 있는지, 그 확산 속도까지 예측 가능할 것으로 예상된다.

참 고 문 헌

[1] Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature,* 457(7232), 1012.

[2] Lazer, D. M., R. Kennedy, G. King, and A. Vespignani, “The parable of google flu : Traps in big data analysis”, *Science Magazine(AAAS)*, 2014.

[3] Lazer D., R. Kennedy, G. King, A. Vespignani, “The parable of Google flu : traps in big data analysis”, *Science*, Vol.343, No.6176, pp.1203-1205, 2014.

[4] Kelly, H. and K. Grant, “Interim analysis of pandemic influenza (H1N1) 2009 in Australia: surveillance trends, age of infection and effectiveness of seasonal vaccination”, *EuroSurveill*, Vol.14, (31) : pii=19288, 2009. Available: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19288.

[5] Wilson, N., K. Mason, M. Tobias, M. Peacey, QS Huang et al., “Interpreting ‘‘GFT’’ Data for Pandemic H1N1: The New Zealand Experience”, *EuroSurveill*, Vol.14(44), pii=19386, 2009. Available : http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19386.

저 자 소 개

우 종 필(Jong-Pil Yu)
- 2007년~현재 : 세종대학교 경영학과 교수, 빅데이터 MBA 주임교수
- 관심분야 : 마케팅, 유통, 빅데이터

이 병 욱(Byung-Uk Lee)
- 2014년~현재: 세종대학교 경영학과 (학사)
- 관심분야: 데이터 마이닝, 빅데이터

이 차 민(Cha-min Lee)
- 2014년~현재 : 세종대학교 경영학과 (학사)
- 관심분야 : 빅데이터, 머신러닝, R Programming, 인공지능, 자율주행

이 지 은(Ji-Eun Lee)
- 2016년~현재 : 세종대학교 경영학과 (학사)
- 관심분야 : 데이터 마이닝, 빅데이터

김 민 성(Min-sung Kim)
- 2013년~현재 : 세종대학교 경영학과 (학사)
- 2017년~현재 : 세종대학교 경영학과 빅데이터팀 연구원
- 관심분야 : 데이터 마이닝, 모델링, 빅데이터

황 재 원(Jae-won Hwang)
- 2017년~현재 : 세종대학교 경영대학 대우교수
- 관심분야 : 마케팅, 소비자 행동, 마케팅 커뮤니케이션