코로나 19 뉴스데이터 분석 및 시각화

허태성O, 황인용*
O인하공업전문대학 컴퓨터정보공학과,
*인하공업전문대학 컴퓨터정보공학과
e-mail: tshur@inhatc.ac.kr, hiyong5978@gmail.com

Covid 19 news data analysis

Hur Tai-seongO, Hwang In Yong*
ODept. of Computer Science Engineering, Inha Technical College,
*Dept. of Computer Science Engineering, Inha Technical College

● 요 약 ●

본 논문에서는 2020년 1월부터 2020년 8월까지 8개월간의 유통되었던 코로나 19와 관련된 뉴스 데이터를 이용하여 기간 및 지역별 단어의 빈도수를 구하고, 그 결과를 활용해 코로나 19와의 상관관계를 분석하고, 시각화하였다. 뉴스데이터는 한국언론진흥재단에서 운영하는 뉴스 빅데이터 시스템인 '빅카인즈'에서 수집된 데이터를 이용하였다. 본 논문에서 웹서비스를 활용해 시각화하였으며 지역과 기간을 선택하면 분석한 결과를 불러와 전체 지역대비 선택한 지역의 뉴스 빈도수, 선택한 지역의 주요 키워드, 주요 키워드의 지역별 일자별 변화 등을 보여주고 있다. 이러한 시각화를 통해 이전에 발생되었던 사건에 대해 주요 키워드와 코로나 19의 상관관계를 쉽게 파악을 할 수 있다.

키워드: 코로나-19(Covid-19), 언론 보도 분석(Analysis of media coverage), 데이터 시각화(Data Visualize)

I. Introduction

1918년 스페인 독감이라는 인플루엔자 바이러스가 발병하여 세계적으로 적게는 2천만명, 많게는 8천만명 정도가 독감으로 사망한 것으로 기록되어 있다. 또한 2012년에는 호흡기 감염증인 메르스가 사우디아라비아에서 발생되었고, 이후 중동지역에서 지속적으로 발생하고 있다. 2020년에는 세계적으로 코로나19 시대를 경험하고 있으며, 현재 우리나라에서는 2020년 1월 20일 한 중국인 여성이 신종 코로나 바이러스 감염자로 확진되면서 대한민국에 첫 번째 감염자가 되었다.

본 논문에서는 한국언론진흥대단에서 운영하는 뉴스 빅데이터 시스템인 '빅카인즈'에서 수집된 코로나 19 뉴스 데이터를 활용하여 지역 및 기간별 주요 단어를 분석하여 나온 결과 활용하여 이를 시각화 할 것이다. 본 논문에서는 csv 형태로 제공된 데이터를 사용하여 이전에 발생한 사건에 대해서만 확인이 가능하지만, 추후 API 서비스와 연동 시 최근 발생한 사건에 대해 주요 키워드와 코로나 19 상관관계를 예측함으로 감염 예방에 도움을 줄 수 있다.

II. Preliminaries

1. Related works

1.1 국내 동향

현재 정부에서 제공되는 코로나 19 데이터는 코로나19에 대한 감염 정보를 제공하거나 확산 추세 예측에 사용되고 있다. 사용자들의 휴대전화를 이용하여 위치 데이터를 취합하고 지도 서비스를 이용하여 이동 현황을 분석하고 있다. 하지만 개인정보 보호를 위해 개인별 위치와 개인 식별 정보를 차단하고 있다. 하지만 이 경우 대한민국에서는 확진자 경로를 정부에서 제공하고 있는 홈페이지에 표시되고 있다. 홈페이지의 경우 텍스트로 표현되고 있으며 체인점의 경우 장소만 제공되어 따로 지도에 검색해야 하는 번거로움이 있다.

III. The Proposed Scheme

본 논문에서는 데이터는 Python 프로그램을 사용하여 정리 분석하여 가공하고 있으며, 데이터 시각화는 chart.js와 d3.js 라이브러리를 사용하여 데이터 시각화를 했다.

개발환경은 다음과 같다. Python 3.7 버전을 사용하였으며, 뉴스 데이터를 가져오기 위해 공공 데이터 포털에 있는 "한국언론진흥재단_뉴스빅데이터_메타데이터_코로나"를 이용하였다.

Table 1. 분석 및 시각화에 사용된 패키지

구분	종류	내용
데이터 분석을 위한 패키지	pandas	csv 데이터를 분석하기 위한 패키지
	pymysql	분석된 결과를 DB에 저장하기 위한 패키지
시각화를 하기 위한 패키지	php	DB에 저장된 결과를 불러오기 위한 패키지
	chart.js	막대, 선등의 그래프를 보여주기 위한 패키지
	d3.js	지도, 네트워크 등 복잡한 그래프를 시각화 하기 위한 패키지

제공된 뉴스 데이터는 csv 형태로 되어 있으며 이를 데이터를 가공하여 기간 및 지역별로 구분하였다. 이렇게 구분된 데이터를 이용하여 시각화 하였다.

뉴스 데이터들을 기간별로 구분하였고, 구분된 데이터를 다시 지역별로 구분하여 단어의 빈도수를 구하여 본 논문의 자료로 사용하였다.

가공된 자료를 기간 및 일자별로 구분하여 구분된 데이터는 아래와 같다.

Fig. 1. 가공된 데이터

분석된 결과를 활용하여 2020년 3월 서울로 검색시 "콜센터"라는 단어가 상위에 있음을 알 수 있는데, 해당 기간에 "콜센터 집단 감염사태"로 인해 뉴스의 비율이 급증하였으며, 해당 사건으로 인해 코로나 확진자도 많이 늘어난 것을 확인할 수 있다. 뿐만 아니라 인근 지역인 경기도에서도 뉴스가 급증한 것을 알 수 있다.

다음은 분석된 결과를 지도와 차트를 통해 시각화하여 나타낸 것이다.

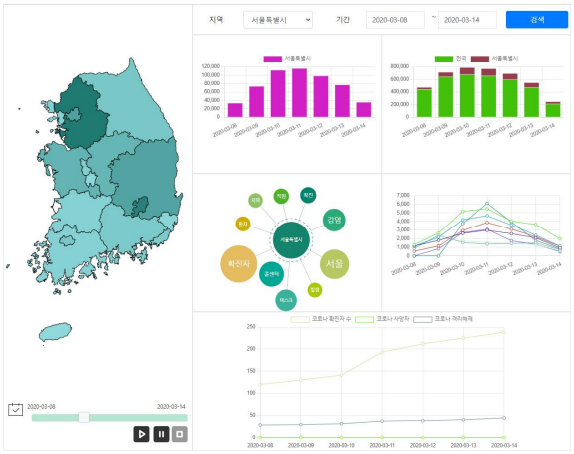

Fig. 2. 기간 및 지역별 뉴스 변화량

위의 그림처럼 데이터를 시각화하면 기간 및 지역별 주요 단어를 쉽게 알 수 있으며 코로나 19와의 상관관계를 알 수 있다.

IV. Conclusions

본 논문에서는 기간 및 지역별로 뉴스의 단어 빈도수를 구분하여 시각화하였으며, 코로나 19 확진자와 관련된 데이터를 같이 시각화하였다. 이를 통해 특정 기간에 발생한 사건에 대해 주요 키워드와 코로나 19의 상관관계를 알 수 있으며, 추후 API 서비스와 연동 시 최근 발생한 사건에 대해 코로나 19 상관관계를 예측함으로 감염 예방에 도움을 줄 수 있다.

REFERENCES

[1] https://www.data.go.kr/data/15069309/fileData.do
[2] https://github.com/jooeungen/coronaboard_kr
[3] https://github.com/d3/d3/wiki
[4] https://www.chartjs.org/