검색량 정보를 통한 코로나 확산 파악

최승일*
*공주대학교 산업시스템공학과
e-mail:sichoi@kongju.ac.kr

Detection of the Spread of Corona via Search Volume Information

Seungil Choi*
*Dept. of Industrial & Systems Engineering, Kongju National University

요 약

본 연구에서는 검색량 정보를 통해 코로나 확산 정도를 미리 파악할 수 있는지를 확인하고자 하였다. 한국의 경우 검색량의 7~8일 이동평균과 일별 신규 확진자 수의 상관관계가 높게 나타나, 검색량으로 코로나 확산 정도를 어느 정도 파악할 수 있었다. 하지만 미국에서는 이러한 상관관계가 명확하게 나타나지 않아 보다 정교하고 보편적인 예측 모델을 만들기 위한 추가 연구가 필요하다.

1. 서론

전염병의 확산과정을 이해하고 확산을 효과적으로 차단하기 위한 방역 대책을 수립하기 위해 바이러스 확산 정도를 거의 실시간으로 파악하는 것은 매우 중요하다. 하지만 의료기관을 통해 수집하는 전염병 통계에는 무증상 또는 경증 감염으로 의료기관을 방문하지 않는 경우가 반영되지 않는다. 바이러스의 확산 과정을 설명하기 위한 다양한 이론들이 제시되어 있으나, 바이러스의 실제 전파 속도를 빠르게 파악하는 것에는 한계를 보이고 있다[1,2].

전염병의 확산 과정을 실시간으로 빠르게 파악하기 위해 많이 활용되는 방법은 검색량, SNS 데이터 등을 분석하는 것으로 구글, 트위터 등의 자료를 분석한 기존 연구들이 있다[3-5]. 이번 코로나 사태는 전파 속도가 빠르고 백신과 치료제의 부재로 장기화되면서 사회 전반에 미치는 영향이 매우 크다. 코로나 바이러스의 확산 정도를 보다 빠르게 파악하는데 검색량을 활용한 방법이 효과적인지를 확인하기 위해 한국은 네이버 검색량, 미국은 구글 검색량을 수집하여 분석한다. 분석기간은 한국에서 확진자가 처음으로 나온 2020년 1월 20일부터 9월 13일까지이며, 네이버는 '코로나', 구글은 'Corona'와 'Covid'를 키워드로 선정하여 검색량 자료를 수집하였다. 검색량과 일별 신규 확진자 수의 상관관계 분석을 통해 코로나 확산 정도를 검색량으로 파악하기 위해 필요한 적정 통계량을 찾고자 한다.

2. 코로나 확산과 검색량의 연관성

2.1 확진자 수 통계

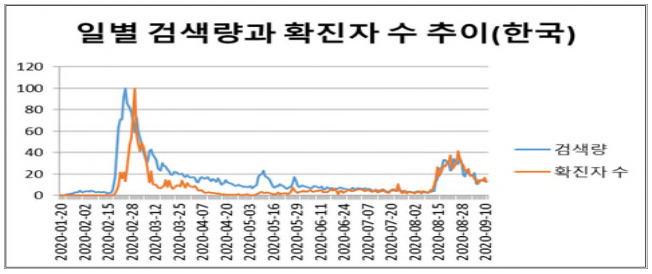
코로나19의 경우 무증상 감염을 통해 확산되는 경우가 많아 코로나 검사 결과 양성으로 나오는 확진자를 최대한 찾아내고 접촉자를 추적하여 감염 경로를 차단하는 것이 매우 중요하다. 먼저 현재까지의 코로나19의 확산 속도를 확인하기 위해 확진자 수에 대한 통계를 정리하였다. 한국의 경우 질병관리청(KDCA: Korea Disease Control and Prevention Agency)에서 발표한 자료를 기반으로 일별 신규 확진자 수를 파악하였다. [그림 1]을 보면 신규 확진자 수는 3월 1일 1,062명(2월 29일 16:00 ~ 3월 1일 24:00 집계)으로 최대를 기록하였고, 8월 26일 441명으로 2차 피크를 보였다.

[그림 1] 일별 신규 확진자 수(한국, 질병관리청 기준)

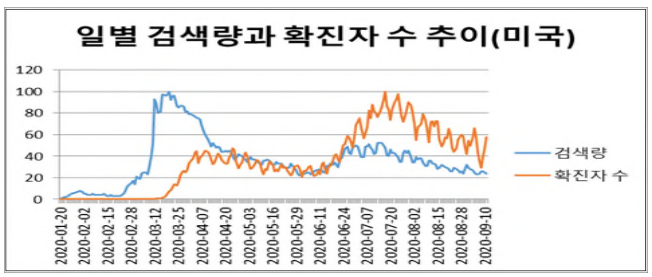
미국의 경우 질병통제예방센터(CDC: Centers for Disease Control and Prevention Agency)에서 발표한 자료를 기반으로 일별 신규 확진자 수를 파악하였다. 일별 신규 확진자 수는 7월 17일 76,091명으로 최대가 되는데, 이는 구글 트렌드에서 'Corona'와 'Covid'의 검색량이 최대였던 3월과는 많은 시차를 보인다.

2.2 검색량과의 상관관계 분석

일별 검색량과 확진자 수의 상관관계를 분석하여 검색량 변화를 통해 확진자 수의 변화를 미리 파악할 수 있는지 분석해 보았다. 한국의 경우 일별 네이버 검색량(검색키워드:코로나)과 질병관리청(KDCA)의 확진자 수 통계를 사용하였고, 미국의 경우 일별 구글 검색량(검색키워드:Corona+Covid)과 질병통제예방센터(CDC)의 확진자 수 통계를 사용하였다. 네이버 검색량과 구글 검색량은 최대 검색량을 100으로 했을 때 이에 대한 상대적 수치로 계산되는데, 상관관계 분석을 위한 사전 작업으로 확진자 수 데이터를 변환하는 과정이 필요하다. 검색량 자료와 마찬가지로 확진자 수를 최댓값에 대한 상대적 수치로 나타내기 위해, 개별 국가의 최대 확진자 수를 100으로 하고 이에 대한 상대적 수치를 구한다. 한국의 경우 변환된 시계열 자료의 그래프인 [그림 2]에서 확진자 수의 1차 피크와 2차 피크가 검색량 피크와 거의 일치하고 검색량과 확진자 수가 밀접하게 관련되어 있음을 쉽게 확인할 수 있다. 반면 미국의 경우 [그림 3]을 보면 이러한 연관성이 잘 관찰되지 않는다.



[그림 2] 일별 네이버 검색량과 확진자 수 추이(한국)



[그림 3] 일별 구글 검색량과 확진자 수 추이(미국)

전염병의 경우 주변 사람들에게서 염려스러운 상황이나 의심 증상이 나타나면 검색량이 증가한다. 그리고 증상이 명확히 발현되면 병원을 방문하게 되어 검색량 변화가 환자 수 변화를 선행하는 경향이 있다. 이러한 경향을 고려하여 검색량과 확진자 수의 시차를 설명하기 위해 일정 기간 동안의 검색량 이동평균을 계산하여 확진자 수와의 상관계수를 구하면 <표 1>의 결과를 얻는다. 한국의 경우 검색량의 7~8일 이동평균과 확진자 수의 상관계수가 크게 나오는데 미국의 경우는 상관계수가 대체로 작게 나온다.

<표 1> 검색량의 n일 이동평균과 확진자 수의 상관계수

	n=1	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9
한국	0.716	0.745	0.770	0.790	0.806	0.820	0.826	0.827	0.822
미국	0.235	0.248	0.257	0.264	0.269	0.274	0.279	0.286	0.293

3. 결론

본 연구에서는 검색량의 이동평균을 통해 확진자 수의 변화를 미리 파악할 수 있음을 보였다. 코로나19의 잠복기간인 1~14일(평균 4~7일)을 고려하면 검색량과 확진자 수와의 이러한 시차가 설명된다. 이번 연구에서는 빅데이터인 검색량을 활용하여 확진자 수를 예측할 수 있는 가능성을 확인하였다. 하지만 이를 다른 국가로 일반화하기에는 어려움이 있다는 것을 미국의 사례에서 볼 수 있었고, 보다 정교하고 보편적인 예측 모델을 만들기 위한 추가 연구가 필요하다.

참고문헌

[1] R. Pastor-Satorras and A. Vespignani, "Epidemic Spreading in Scale-Free Networks", Physical Review Letters, 86, pp. 3200-3203, 2001.

[2] A. L. Lloyd and R. M. May, "How Viruses Spread Among Computers and People", Science, 292, pp. 1316-1317, 2001.

[3] J. Ginsberg et al., "Detecting Influenza Epidemics Using Search Engine Query Data", Nature, 457, pp. 1012-1014, 2009.

[4] A. Culotta, "Towards Detecting Influenza Epidemics by Analyzing Twitter Messages", Proceedings of the First Workshop on Social Media Analytics, pp. 115-122, 2010.

[5] A. F. Dugas, "Google Flu Trends: Correlation with Emergency Influenza Rates and Crowding Metrics", Clinical Infectious Diseases, 54(4), pp. 463-469, 2012.