pISSN : 2508-3384 eISSN : 2508-3392
https://doi.org/10.22640/lxsiri.2017.47.1.267
Journal of Cadastre & Land InformatiX Vol.47 No.1 (2017)
한국국토정보공사「지적과 국토정보」제47권 제1호 2017년 6월 p. 267-276

트윗 데이터를 이용한 황사 관련 질병 유의성 분석

Significance Analysis of Yellow Dust Related Disease Using Tweet Data

정용한* · 서민송** · 유환희***
Jung, Yong-Han · Seo, Min-Song · Yoo, Hwan-Hee

Abstract

Damages have occurred in various fields such as agriculture, industry, and citizen's health due to the yellow dust. Therefore, it is urgent to take measures against it. In this regard, this study collected data of yellow dust over 11 days on a basis of Feb. 23. 2015 when yellow dust was the greatest after 2009, issue words analysis and recomposed health related tweet data. After testing the significance of yellow dust related diseases by association rule analysis with diseases, it obtained the study results as follows: As a result of significance test for the patients with rhinitis, asthma and conjunctivitis by acquiring the condition data of patients from the Health Insurance Review & Assessment Service, conjunctivitis appeared to be significant in 13 cities for 16 cities at 5% significance probability, while asthma and rhinitis showed a significance in 3 and 6 areas. As described above, it is possible to obtain information about citizens' health from SNS data, such as Tweet data and it is judged that these data will provide useful information for establishing measures of citizens' health care.

Keywords: Yellow Dust, Tweet Data, Issue Words Analysis, Association Rule Analysis, Significance Test

1. 서 론

우리나라는 매년 발생하는 황사로 인해 시민들의 호흡기 및 안질환 등 건강에 많은 피해가 발생하고 있다(Iwasaka et al 1988). 황사는 중국 북부의 황토지대에서 부유된 모래먼지가 기류를 타고 이동하여 서서히 강하하는 현상 또는 강하하는 모래먼지를 말하는 것으로 최초의 기록은 삼국사기에서 찾아볼 수 있을 정도로 오래전부터 우리나라에 발생해 왔다(손지영 외 2009; 황승식 외 2005). 황사는 미세입자로 인한 도시민의 호흡기질환, 기관지염, 천식, 안질 등의 질환을 일으킬 수 있다(김규현 2005). 이러한 피해를 효과적으로 방지하기 위해서는 직접적인 피해를 겪고 있는 시민들의 생각을 파악할 필요가 있으며, 이에 적합한 데이터가 최근 스마트폰 보급률의 증가에 따라 더욱 활성화 되고 있는 SNS 데이터이다. 특히, 트윗 데이터는 활용 가능성이 높은데

* 경상대학교 공학연구원, Engineering Research Institute, Gyeongsang National University (first jyh1315@naver.com)
** 경상대학교 도시공학과 석사과정 BK21+, Department of Urban Engineering, Gyeongsang National University (minsong-1234@hanmail.net)
***경상대학교 도시공학과 교수 BK21+, ERI, Department of Urban Engineering, Gyeongsang National University (corresponding author: hhyoo@gnu,ac,kr)

서울지역의 트윗 데이터를 대상으로 기계학습법을 사용하여 감정을 긍정과 부정으로 이분화 하여 트윗 감정의 핫스팟을 분석한 연구가 있으며(임좌상 · 김진만 2015), 트윗 데이터를 이용하여 실시간으로 지역을 탐지하는 시스템 개발에 관한 연구도 활발하게 이루어지고 있어 활용성이 매우 높다(임준엽 2015). 또한, 빅데이터 분석 방법에는 여러 가지 방법이 있다. 그 중 텍스트 마이닝은 글 속에 숨겨진 감성을 알아내는 분석기법으로 텍스트 속의 의미 있는 정보를 추출하여 다른 정보와의 연계성을 파악한 뒤, 텍스트가 가지고 있는 카테고리를 찾아내는 등의 결과를 도출할 수 있어 SNS를 분석하기에 가장 적합하다(윤홍근 2013).따라서 본 연구에서는 트윗 데이터를 활용하여 황사의 발생에 따른 시민들의 건강 피해에 대한 생각을 확인하고자 한다. 2009년 이후 서울의 미세먼지(PM10) 농도가 최대를 기록한 황사가 발생했던 2015년 2월 23일 전후로 11일 동안에 황사를 언급한 트윗 데이터를 수집하였고, 이를 R을 통해 이슈어 분석과 연관규칙 분석을 실시하여 시민들이 황사가 발생함에 따라 관심이 높아지는 질병관련 이슈어를 확인하였다. 이 이슈어 중 질병관련 이슈어인 비염, 천식, 결막염의 2015년 2월 한 달 동안 진료건수 데이터를 취득하여 황사가 발생한 넷째 주의 진료건수를 황사가 발생하지 않은 첫째 주와 둘째 주(셋째 주는 설명절이 있어서 제외됨)의 진료건수와 상관성 분석을 실시하여 3개 질병과 황사와의 유의성을 검증하였다.

2. 연구 이론 및 방법

2.1. R 분석

본 연구에서는 트윗 데이터의 분석을 위해 프로그래밍 언어이자 소프트웨어인 R을 이용하였다. R은 R커뮤니티를 통한 기능 향상에 따른 패키지가 주기적으로 제공되고 있으며, 고유한 언어 내장 프로그램과 수백 가지 통계함수를 제공하는 등 높은 인지도를 차지하는 통계분석 프로그램이다(최경호 · 유진아 2015). 본 연구에서는 R에서 제공되는 KoNLP패키지를 이용하여 수집한 데이터에서 많이 언급된 이슈어를 확인하고 이를 wordcloud패키지를 이용하여 시각화하였다. 또한, 확인된 이슈어간의 연관성을 확인하고자 R의 arules패키지를 이용하여 연관규칙 분석을 실시하였고, arulesViz 패키지를 이용하여 시각화 하였다. 연관규칙 분석은 지지도(Support), 신뢰도(Confidence), 향상도(Lift)의 3가지 지표를 통해 의미 있는 규칙을 탐색하는 기법으로서, 본 연구에서는 두 이슈어가 하나의 트윗 내에서 얼마나 자주 나타나는가를 의미하는 지지도를 중심으로 분석하였다(유충현 · 홍성학 2015).

2.2. 트윗 데이터 분류 프로그램 개발

본 연구에서는 황사관련 트윗 데이터에서 시민들의 건강에 대한 생각을 분석하고자 R을 통해 나타난 이슈어 중 건강관련 이슈어를 포함하는 트윗 데이터를 그룹화 하였다. 이것은 연관규칙 분석을 실시할 때 건강관련 이슈어를 중심으로 트윗 데이터를 분류하여 그룹화 할 필요가 있으므로 해당 프로그램을 개발하였다. 프로그램은 Visual Basic언어기반 Visual studio2015를 이용하여 개발하였다. 개발된 프로그램은 Fig. 1과 같이 Excel 파일을 불러오는 1단계, 분류할 이슈어를 입력하는 2단계, 분류한 트윗 데이터를 구분하여 Excel 파일로 저장하는 3단계로 구성하였다.

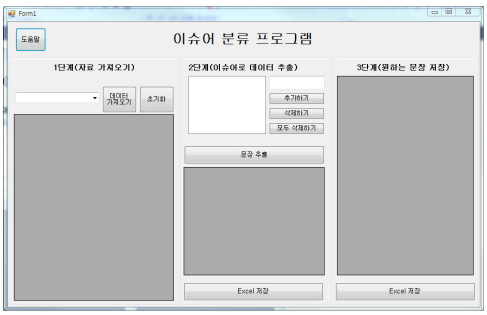

Figure 1. Sorting program

2.3. 유의성 검정

본 연구에서는 황사의 유무에 따른 시민들의 건강피해를 검정하기 위해 고급통계분석론 이론과 실습(이희연 · 노승철 2013)을 참고하여 기준일과 비교일의 개념을 활용하였다. 2015년 2월 23일 발생한 황사를 중심으로 분석하였으므로 같은 기간인 2015년 2월의 3개 질병의 진료건수를 황사가 발생한 주와 발생하지 않은 주로 나눠 황사가 발생한 주를 기준 주로 정하고 그에 대칭되는 다른 주를 비교하여 검정하였다. 유의성 검정은 대응표본 t검정을 중심으로 실시하였으나 사용되는 데이터는 대응되는 주별 7일간의 데이터로서 검정에 앞서 데이터의 정규성과 비정규성에 따라 두 가지 검정방법을 적용하였다. 먼저 Shapiro-Wilk검정을 수행하여 정규성이 검정된 데이터(p-값 〈 0.05)에 대해서는 대응표본 t검정을 실시하였다. 그 다음으로 비정규성으로 판정된 데이터는 비모수 검정인 Wilcoxon 검정을 실시하여 유의성을 검정하였으며 이러한 처리 과정은 Fig. 2와 같다(이희연 · 노승철 2013).

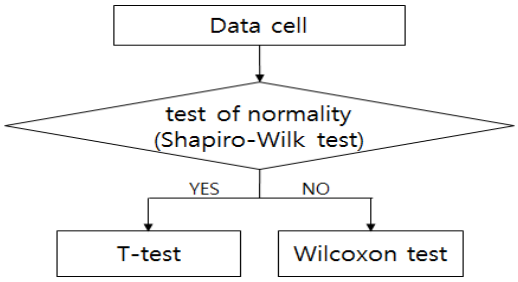

Figure 2. Diagram significance test

3. 결과 분석

3.1. 트윗 데이터 수집 및 처리

본 연구에서는 2009년 이후 서울시 미세먼지의 농도가 최대치를 기록하여 최악의 황사로 기록된 2015년 2월 23일을 중심으로 18일부터 28일까지 총 11일간의 '황사' 키워드가 포함되어 있는 트윗 데이터를 Pulse-Ks/w를 통해 수집하였다. Pulse-K는 코난테크놀로지(주)가 개발한 소셜미디어 분석 및 모니터링 서비스이다. Fig. 3은 총 11일간의 수집된 황사관련 트윗 데이터의 수를 날짜별로 나타낸 것으로서 황사관련 총 데이터의 양은 16,497개이며, 황사가 최대로 발생한 2월 23일에는 트윗 데이터 수가 8,751개로서 가장 많이 수집되었다. Fig. 4는 그 기간 동안 기상청에서 발표한 미세먼지의 농도를 나타내는 그래프로서 최대 황사가 발생한 2월 23일에 가장 농도가 높게 나타나고 있다.

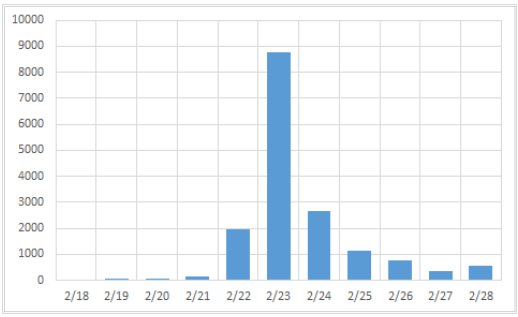

Figure 3. Number of Tweet data

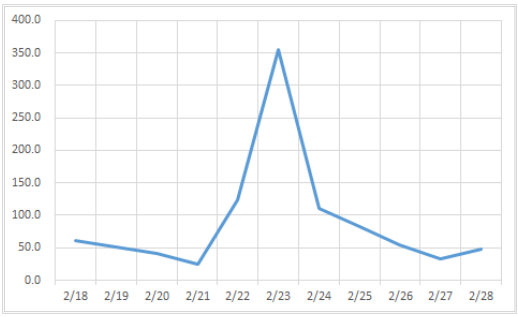

Figure 4. Average concentration of PM10

수집된 트윗 데이터는 사용자에 의해 자유롭게 작성된 SNS 데이터이므로 정확한 분석을 위해 나라인포테크(주)에서 개발하여 무료로 이용 가능한 한국어 맞춤법/문법 검사기를 이용하여 띄어쓰기와 오타 등을 수정하였다.

3.2. R을 이용한 트윗 데이터 분석

수집된 전체 트윗 데이터를 R을 통해 워드클라우드로 표현하였는데(Fig. 5). '황사'를 검색어로 수집한 트윗 데이터이므로 황사와 관련이 있는 '마스크'가 가장 크게 가운데에 위치하고 있었다. 이것은 시민들이 황사에 대비하여 마스크를 가장 관심 있게 생각하고 있음을 알 수 있었으며, 그 다음으로 '오늘', '미세먼지' 등이 나타나고 있어서 황사발생 시점을 의미하는 오늘과 황사 관련 미세먼지 농도 등이 많이 언급된 것을 확인할 수 있었다. 또한, 시민들의 건강에 대한 생각을 더욱 자세하게 확인하고자 건강관련 이슈어를 중심으로 트윗 데이터 그룹을 구성하였다. 건강 관련 트윗 데이터는 전체 트윗 데이터 중 17.0%를 차지하고 있다.

Figure 5. Wordcloud of tweet data

본 연구에서는 황사에 따른 시민들의 건강관리에 대한 자세한 분석을 위해 황사관련 이슈어 그룹 중에서 건강관련 이슈어를 Table 1과 같이 정리하였으며 '재채기, 몸살, 기침, 감기, 천식, 호흡기질환' 등의 시민들의 건강에 직접적인 피해를 줄 수 있는 이슈어를 확인할 수 있었다.

이러한 이슈어들 중 시민들이 황사가 발생했을 때 어떤 질병에 관심이 있는지에 대한 분석을 하기 위하여 건강과 관련된 트윗 데이터의 이슈어들 간의 연관성을 분석하기 위하여 빅데이터 분석툴인 R을 이용하여 연관규칙을 분석하였다.

Table 1. Health related issue words

Health related Issues
mask, health, sneeze, aftereffects, cold, hospital, asthma, bronchitis, aches, respiratory, disease, rhinitis, breathing, virus, preventive, neck care, death, cough, rhinitis, stress, pharmacy, patients, asthma, sputum, bronchitis, conjunctivitis …

이를 위하여 전체 트윗 데이터를 대상으로 분석된 이슈어들 중 건강관련 이슈어가 포함되어 있는 트윗 데이터만을 별도로 분류하여 소위 건강관련 트윗 데이터 그룹으로 재구성하였다. 건강관련 트윗 데이터그룹을 재구성하기 위하여 본 연구에서는 Visual Studio 2015를 이용하여 Visual Basic 언어를 기반으로 하는 프로그램을 개발하였고(Fig. 1), 이를 통해 건강관련 트윗 데이터그룹을 재구성하였다. 이 그룹을 대상으로 황사와 어떤 질병이 연관성이 높은 지를 분석하기 위해 R을 이용하여 연관규칙 분석을 지지도 10%로 실시하였다(Fig. 6).

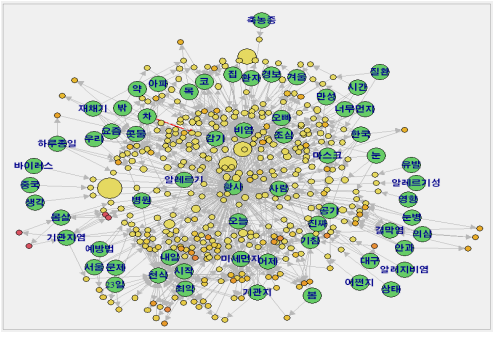

Figure 6. The association rule graph for the health group

황사관련 트윗 데이터에 나타난 황사와 질병의 연관성을 Fig. 6에서 분석하면, 황사가 중심에 위치하고 '감기, 비염, 알레르기, 천식, 결막염, 기관지염' 등에 연관성을 보이고 있었다. 관련 질병 중 천식은 '기관지염, 기침'과 비염은 '감기, 마스크', 결막염은 '눈병, 안과' 등과 연관성이 높게 나타났다. 이와 같이 황사가 발생함에 따라 시민들이 관심을 가지게 되는 질병의 종류에 대해 트윗 데이터를 분석함으로서 시민들의 건강에 대한 걱정과 생각을 파악할 수 있었다. 그러나 황사발생에 관련하

여 시민들의 걱정이 실제로 질병으로 발생하여 환자가 발생하는 것인지에 대해서는 확인할 필요가 있다. 즉, 단순한 시민들의 걱정인지 아니면 실제로 관련 질병이 발생하여 환자수가 증가하는지에 대한 분석을 실시하여 확인할 필요가 있다.

따라서 본 연구에서는 황사관련 트윗 데이터에서 이슈어 분석과 연관규칙분석으로 도출된 질병과 건강보험심사평가원에서 취득 가능한 환자 실태자료를 종합하여 비염, 천식, 결막염 환자에 대한 분석을 실시하였다. 건강보험심사평가원에서 제공받은 병원진료 환자수에 대한 자료는 2015년 2월의 전국 비염, 천식, 결막염 환자 진료수이다. 따라서 황사가 최대로 발생한 2월 23일이 포함된 넷째 주를 기준주로 하고 첫째 주와 둘째 주의 환자수와 비교분석 하므로써 상관성이 유의함을 검증하였다. 셋째 주는 설날이 포함되어 휴무가 있어서 본 연구에서는 제외하였으며, 주단위로 검증한 것은 병원의 환자진료수의 패턴이 요일별로 비슷한 변화를 보이므로 주 단위를 하나의 데이터 셀으로 정의하여 주단위로 상관성이 유의함을 검증하였다.

3.3. 황사 발생에 따른 질병별 상관성 분석

황사발생에 따른 비염, 천식, 결막염 환자의 진료건수와 황사가 발생하지 않은 평시의 환자 진료건수를 주단위로 비교하여 상관성이 유의함을 검증하여 황사발생이 관련 질병에 연관성이 있는지를 분석하였다. 그동안 황사가 발생함에 따른 전국의 지역별 황사농도는 다소 차이가 있었으며 최대 황사가 발생한 2015년 2월 23일 기상청에서 발표한 전국의 미세먼지 농도도 지역별로 최대 594㎍/㎥ 에서 최소 212㎍/㎥로 분포하였다.

비염, 천식, 결막염의 건수는 2월 한달 간 비염의 경우 130,206건, 천식의 경우 1,960,289건, 결막염의 경우 130,206건으로 나타났다. 또한, 병명 각각에 따라 지역별 환자수도 차이가 있을 것으로 판단하여 확인해본 결과 서울이 비염 247,917건, 천식 77,128건, 결막염 101,882건으로 가장 많이 발생하였으며 전국을 Table 2와 같이 권역별로 나누고 최대 황사가 발생한 2월 넷째 주를 기준으로 첫째 주와 둘째 주를 비교하였다. 그리고 권역별에 따른 환자 진료건수에 대해 상관성이 유의한지를 검증하였다. 유의성 검정은 먼저 Shapiro-Wilk검정을 수행하여 정규성이 검정된 데이터에 대해서는 대응표본 t검정을 실시하고, 비정규성으로 판정된 데이터는 비모수 검정인 Wilcoxon 검정을 실시하여 유의성을 검정하였다.

Table 2. Section for Area

Region	Area
Metropolitan	Seoul, Incheon, Gyeonggi-do
Gangwon	Gangwon-do
Chungcheong	Daejeon, Chungcheongbuk-do Chungcheongnam-do
Jeonla	Gwangju, Jeonlabuk-do, Jeonlanam-do
Gyeongsang	Daegu, Ulsan, Busan, Gyeongsangbuk-do, Gyeongsangnam-do
Jeju	Jeju Island

Table 2와 같이 6개 권역 16개 지역에 대해 비염, 천식, 결막염에 대한 Shapiro-Wilk검정을 수행하여 정규성을 검정한 결과 비염에서 1건, 천식에서 1건, 결막염에서 6건의 총 8건의 정규성을 만족하지 않은 검정 결과가 나왔다. 따라서 정규성이 만족되는 데이터는 대응표본-t검정을 실시하였고, 만족되지 않은 나머지 8건은 비모수 검정의 Wilcoxon 검정을 실시하여 유의성을 확인하였다.

Table 3은 수도권 지역의 대응표본 t-검정의 결과를 나타낸 표이다. 비염은 인천지역의 첫째 주에 비해 넷째 주에서 0.2%가 많았던 것을 제외하면 모든 지역에서 황사가 발생한 주 보다 진료건수가 작은 것 으로 나타났으며, 경기지역의 둘째 주에서는 작은 것이 유의수준 5%에서 통계적으로 유의한 결과로 나타났다. 천식의 경우 증가율이 증가와 감소가 반복되고 있으며 p-값에서 통계

Table 3. Matching sample t-test result of Metropolitan region

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Seoul	First week	- 2.3%	0.237	1.2%	0.800	18.0%	0.002*
	Second week	- 5.9%	0.088	- 4.2%	0.450	21.5%	0.009*
Incheon	First week	0.2%	0.937	5.8%	0.332	15.5%	0.005*
	Second week	- 6.8%	0.148	- 0.8%	0.909	17.2%	0.026*
Gyeonggi	First week	- 1.6%	0.504	6.3%	0.092	16.3%	0.004*
	Second week	- 9.6%	0.044*	- 0.9%	0.851	18.6%	0.016*

* Statistically significant at 5% level of significance

Table 4. Matching sample t-test result of Gangwon region

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Gangwon	First week	10.9%	0.025*	10.2%	0.090	9.9%	0.010*
	Second week	1.2%	0.812	1.5%	0.765		

* Statistically significant at 5% level of significance

Table 5. Matching sample t-test result of Chungcheong region

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Daejeon	First week	1.8%	0.512	15.2%	0.054	15.4%	0.007*
	Second week	- 4.9%	0.202	- 1.8%	0.757	13.2%	0.036*
Chung-buk	First week	8.3%	0.079	10.9%	0.047*		
	Second week	- 0.4%	0.920	6.3%	0.231		
Chung-nam	First week	6.4%	0.151	2.7%	0.539	14.8%	0.018*
	Second week	- 1.3%	0.809	- 4.9%	0.500		

* Statistically significant at 5% level of significance

적 유의성을 나타내지 못하였다. 결막염에서는 모든 지역의 진료건수가 증가한 것으로 나타났으며 둘째 주에 비해 넷째 주가 최대 21.5%가 많았고 또한, 모든 지역이 유의수준 5%에서 통계적으로 유의한 결과를 나타내었다.

Table 4는 강원권 지역의 대응표본 t-검정의 결과이며 강원지역의 둘째 주와 같이 정규성이 만족되지 않은 데이터는 검정이 불가하여 비워두었다. 첫째 주에 비해 넷째주가 3개 질병 모두 9.9~10.9%의 증가율을 나타내었으나, 비염과 천식에서만 유의수준 5%에서의 통계적으로 유의한 결과 값을 나타내어 첫째 주가 둘째 주에 비해 영향력이 높게 나타난 것을 확인할 수 있었다.

Table 5는 충청권 지역의 대응표본 t-검정의 결과이며 충북과 충남의 둘째 주의 데이터는 정규성을 만족하지 못하여 검정이 불가하므로 해당 칸을 비워두었다. 비염에서는 큰 변화와 유의성 또한 나타나지 않았고, 천식에서는 대전지역의 첫째 주에 비해 넷째 주가 15.2% 많았으나 유의한 결과를 나타내진 못하였고, 충북지역의 첫째 주에 비해 넷째 주가 10.9%가 많았으며 유의수준 5%에서 유의한 결과를 나타내었다. 또한, 결막염에서는 분석한 데이터 내에서 모든 지역이 13.2~15.4% 진료건수가 많았으며 유의수준 5%에서의 유의성을 나타내었다.

Table 6. Matching sample t-test result of Jeonla region

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Gwangju	First week	2.6%	0.455	3.3%	0.537	16.2%	0.013*
	Second week	- 5.7%	0.269	- 4.8%	0.533	16.3%	0.048*
Jeon-buk	First week	9.2%	0.133	13.7%	0.079	9.8%	0.037*
	Second week	0.9%	0.849	0.4%	0.957	11.6%	0.048*
Jeon-nam	First week	9.4%	0.038*	10.2%	0.231	10.4%	0.022*
	Second week			- 1.3%	0.841	9.5%	0.155

* Statistically significant at 5% level of significance

Table 7. Matching sample t-test result of Gyeongsang region

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Daegu	First week	4.1%	0.251	4.3%	0.225	16.3%	0.003*
	Second week	- 3.6%	0.338	1.9%	0.668	16.9%	0.010*
Ulsan	First week	9.1%	0.153	9.8%	0.187	20.3%	0.005*
	Second week	- 2.2%	0.745	2.7%	0.666	23.8%	0.011*
Busan	First week	4.1%	0.339	11.7%	0.067	12.9%	0.019*
	Second week	- 5.7%	0.103	2.1%	0.637	12.1%	0.031*
Gyeong-buk	First week	14.6%	0.015*	14.4%	0.019*		
	Second week	3.8%	0.363	8.3%	0.163		
Gyeong-nam	First week	9.7%	0.043*	9.7%	0.128	19.0%	0.004*
	Second week	- 0.4%	0.901	- 0.8%	0.882	21.1%	0.009*

* Statistically significant at 5% level of significance

Table 6은 전라권 지역의 대응표본 t-검정의 결과를 나타낸 것이다. 전남지역 둘째 주에서 비염은 정규성을 만족하지 못하여 비워두었다. 비염에서 전남의 첫째 주에 비해 넷째 주가 9.4%가 많았으며 유의수준 5%에서 통계적으로 유의한 결과를 나타내었고, 천식에서는 전북의 첫째 주와 전남의 첫째 주에 비해 넷째 주가 10% 많았지만 유의성은 나타내지 못하였다. 결막염에서는 전남의 둘째 주를 제외한 모든 지역에서 9.8~16.3% 많았던 것을 보이며 유의수준 5%에서 유의한 결과를 나타내었다.

Table 7은 경상권의 대응표본 t-검정의 결과를 나타낸 것이며 경북지역의 결막염은 정규성을 만족하지 못하여 비워두었다. 비염에서는 경북지역의 첫째 주에 비해 넷째 주가 14.6%가 많았던 것을 보였으며 유의수준 5%에서 유의한 결과를 확인하였고, 경남지역의 첫째 주에 비해 넷째주가 9.7%더 많았으며 유의성을 나타내었다. 천식에서는 경북지역의 첫째 주에 비해 넷째 주가 14.4%더 많았으며 유의성을 만족시켰고, 결막염은 나타난 데이터 내에서 모든 지역이 최대 23.8%를 보이며 유의수준 5%에서 통계적으로 유의한 결과를 보였다.

Table 8은 제주권의 대응표본 t-검정의 결과 값을 나타낸 것이다. 제주지역에서는 천식의 둘째 주의 데이터가 정규성을 만족하지 못하여 비워두었다. 제주지역의 모든 질병의 진료건수가 많았으며, 특히, 첫째 주에 비해 넷째 주가 비염의 경우 43.7% 많았고 결막염은 첫째 주에 비해 넷째주가 57.7%더 많았다. 또한, 둘째 주에

Table 8. Matching sample t-test result of Jeju region

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Jeju	First week	43.7%	0.006*	13.2%	0.044*	57.7%	0.001*
	Second week	22.3%	0.046*			48.6%	0.006*

* Statistically significant at 5% level of significance

Table 9. Test results of Wilcoxon Signed-Rank

Area	Compare week	Rhinitis		Asthma		Conjunctivitis	
		Growth rate	P-value	Growth rate	P-value	Growth rate	P-value
Gangwon	First week					13.3%	0.027*
Chung-buk	First week					10.9%	0.028*
	Second week					12.0%	0.063
Chung-nam	Second week					20.3%	0.018*
Jeon-nam	Second week	2.1%	0.735				
Gyeong-buk	First week					8.6%	0.018*
	Second week					8.6%	0.090
Jeju	Second week			2.0%	0.866		

* Statistically significant at 5% level of significance

비해 넷째 주가 48.6% 더 많이 나타났으며, 모든 데이터 내에서 유의수준 5%에서 통계적으로 유의성을 나타내어 황사의 발생에 가장 큰 3개 질병의 진료건수의 영향을 받는 것으로 나타났다. 정규성이 만족된 데이터의 대응표본 t-검정에서 3개 질병 중에서는 결막염이 거의 모든 지역에서 유의하게 보이며 영향이 높은 것으로 나타났고, 지역 중에서는 제주에서 3개 질병 모두에서 유의한 결과가 나타났고 유의수준 5%에서 유의한 것으로 나타났다. 또한 정규성 검정에서 제외된 데이터셀에 대해서는 비모수 검정인 Wilcoxon 검정을 실시하여 유의성을 검정하였다.

Table 9는 비정규성을 나타낸 데이터셀 만을 대상으로 Wilcoxon 검정을 실시하여 나타낸 결과로서 정규성을 갖는 데이터셀은 앞서 분석이 이뤄졌으므로 Table 9에서는 빈칸으로 두었다. 결과를 보면 비염과 천식에서 각 1건의 데이터에서는 유의성을 나타내지 못하였으나, 결막염에서는 충북지역의 둘째 주와 경북지역의 둘째 주를 제외하고는 모든 지역에서 유의수준 5%에서 통계적으로 유의한 결과를 나타냈다.

4. 결 론

본 연구에서는 2009년 이후 황사발생으로 인해 서울시의 미세먼지 농도가 최대치를 기록하였던 2015년 2월 23일을 중심으로 11일간의 트윗 데이터 분석하여 다음과 같은 결론을 도출하였다.

첫째, 2015년 2월 18일부터 28일까지 총 11일의 황사 관련 트윗 데이터를 수집하고 R프로그램을 통해 이슈어 분석을 실시하고 Visual Studio 2015프로그램을 이용하여 Visual Basic 언어를 기반으로 그룹을 분류하는 프로그램을 개발하였다. 개발된 분류 프로그램을 이용하여 건강관련 트윗데이터를 재구성 하였으며 전체 트윗 데이터중 건강관련 이슈어는 17.0%로 나타났다. 이것을 대상으로 연관규칙을 분석한 결과 천식은 '기관지염, 기침'과 비염은 '감기, 마스크', 결막염은 '눈병, 안과' 등과 연관성이 높게 나타났다.

둘째, 황사관련 트윗데이터에서 도출된 질병과 건강보험심사평가원에서 취득한 환자실태자료를 종합하여 비염, 천식, 결막염 환자에 대한 유의성 검정을 실시하였으며 정규성인 경우 Shapiro-Wilk검정을, 비정규성인 경우는 Wilcoxon 검정을 실시하여 유의성 여부를 판단하였다. 그 결과 유의확률 5%에서 결막염은 16개 시・도 중 13개 지역에서 유의하게 나타났으며, 비염은 6개의 지역에서, 천식은 3지역에서 유의한 것으로 나타났다.

이상과 같이 트윗데이터와 같은 SNS데이터는 시민들의 건강에 대한 정보를 취득할 수 있는 가능성을 보여주었다. 향 후 분석 결과의 신뢰성을 높이기 위해서 자료 취득 기간의 다양화와 의학적인 전문성이 융합된 연구가 추가적으로 수행되어야 할 것으로 판단된다.

참고문헌

References

김규현. 2005. 황사기간과 비황사기간의 대구지역 PM10 및 중금속 오염도 특성평가. 석사학위논문. 경북대학교 산업대학원.

Kim KH. 2005. *Assessment of PM10 and heavy metals compare with Yellow-Sand period and non Yellow-Sand period in the Daegu area*[Thesis]. Kyungpook National University Graduate School of Industry.

손지영, 조용성, 김윤신, 이종태, 김연정. 2009.도시 대기오염의 위해 평가에 있어서 황사효과 분석 – 서울시 총사망 및 원인별 사망률에 미치는 영향. 한국환경보건학회지. 35(4): 249-258.

Soon JY, Cho YS, Kim YS, Lee JT, Kim YJ. 2009. An Analysis of Air Pollution Effect in Urban Area Related to Asian Dust on All-cause and Cause-specific Mortality in Seoul, Korea, *Journal of Environmental Health Science*. 35(4): 249-258.

임준엽. 2015. 트위터를 이용한 실시간 이벤트 지역 탐지 시스템. 석사학위논문. 가톨릭대학교.

Yim JY. 2015. *Twitter Based Realtime Event-Location Detector*[Thesis]. Catholic Naional University.

임좌상, 김진만. 2015. 한국어 트위터 감정의 핫스팟 분석. 멀티미디어학회논문지. 18(2): 233-243.

Lim JS, Kim JM. 2015. Hotspot Analysis of Korean Twitter Sentiments. *korea Multimedia Society*. 18(2): 233-243.

유충현, 홍성학. 2015. R을 활용한 데이터 시각화. 교보문고. p. 676-679.

Yoo CH, Hong SH. 2015. *R Visualization*. Kyobobook, p. 676-679.

윤홍근. 2013. 문화산업에서 빅데이터의 황용방안에 관한 연구. 글로벌문화콘텐츠. 10: 157-180.

Yoon HG. 2013. Research on the Application Methods of Big Data within the Cultural Industry. *AAGCC*. 10: 157-180.

이희연, 노승철. 2013. 고급통계분석론 이론과 실습. 문우사. p. 138-166.

Lee HY, Rho SC. 2013. Advanced Atatistical Theory Theory and Practice. MoonWoo. p. 138-166.

하병국. 2015. 데이터 분석 방법론을 이용한 트윗 핫스팟 선정에 관한 연구. 박사학위논문. 광운대학교 경영대학원.

Ha BG. 2015. *Study on geotagged SNS data analysis methodology to select the tweets hotspots*[Thesis]. Kwangwoon Naional University.

황승식, 조수헌, 권호장. 2005. 2002년 봄 서울 지역에 발생한 심한 황사가 일별 사망에 미치는 영향. 예방의학회지. 38(2): 197-202.

Hwang SS, Kwon HJ, Cho SH. 2005. Effects of the Severe Asian Dust Events on Daily Mortality during the Spring of 2002, in Seoul, Korea. *Journal of Preventive Medicine and Public*

Health. 38(2): 197-202.

Pulse-K. 2015. 트윗 데이터 [인터넷]. [http://www.pulsek.com/]. 2015년 3월 20일 검색.

Pulse-K. 2015. Tweet Data[Internet]]. [http://www.pulsek.com/]. Last accessed 20 March 2015.

건강보험심사평가원. 2016. 환자실태자료. [인터넷]. [http://www.hira.or.kr/]. 2016년 3월 10일.

Health Insurance Review and Evaluation Center. 2016. Patient Status Data. [Internet]. [http://www.hira.or.kr/]. Last accessed 10 March 2016.

Iwasaka, Y, Yamato, M, Imasu, R. 1988. Transport of Asian dust(KOSA) particles; importance of weak KOSA events on the geochemical cycle of soil particles. *Tellus*. 40(5): 495-503.

2017년 5월 01일 원고접수(Received)
2017년 6월 07일 1차심사(1st Reviewed)
2017년 6월 19일 2차심사(2nd Reviewed)
2017년 6월 20일 게재확정(Accepted)

초 록

우리나라는 황사로 인해 농업 및 산업분야, 시민건강 등 다양한 분야에 걸쳐 피해가 발생되고 있으며 이에 대한 대책 마련이 시급한 실정이다. 이에 본 연구에서는 2009년 이후 최대 황사가 나타났던 2015년 2월 23일을 기준으로 전후 11일간의 황사 관련 트윗 데이터를 수집하고, 이슈어 분석, 건강과 관련된 트윗 데이터 그룹 재구성, 질병과의 연관규칙 분석 등을 걸쳐 황사발생과 관련 질병의 유의성을 검정한 결과 다음과 같은 결론을 얻었다. 황사관련 트윗 데이터로부터 도출된 질병과 건강보험심사평가원에서 취득한 환자실태 자료를 종합하여 비염, 천식, 결막염 환자에 대한 유의성 검정을 실시한 결과, 유의확률 5%에서 결막염은 16개 시 · 도 중 13개 지역에서 유의하게 나타났으며, 비염은 6개 지역에서, 천식은 3개 지역에서 질병 발생에 유의한 것으로 나타났다. 이상과 같이 트윗 데이터와 같은 SNS데이터로 부터 시민들의 건강에 대한 정보를 취득할 수 있었으며, 이를 활용한 시민건강 관리 대책을 수립하는데 유용한 정보를 제공해 줄 수 있을 것으로 판단된다.

주요어 : 황사, 트윗 데이터, 이슈어 분석, 연관규칙 분석, 유의성 검정