산업융합연구 제19권 제3호 pp. 35-40, 2021
ISSN 2635-8875 / e-ISSN 2672-0124
DOI : https://doi.org/10.22678/JIC.2021.19.3.035

빅데이터 분석을 통한 신종감염병 중요 요인 도출

서경도
금오공과대학교 산학협력단 교수

A Study on deduction of important factors for new infectious diseases through big data analysis

Kyung-Do Suh
Professor, Industry-Academic Collaboration Foundation, Kumoh National Institute of Technology

요 약 본 연구는 신종감염병에 대한 텍스트 데이터를 수집하고, 이를 분석하여 신종감염병에서 중요한 요인을 도출하고자 하였다. 이를 위해 네이버 뉴스 데이터베이스의 기사를 직접 크롤링하고, 이를 전처리 하여, 데이터 분석에 활용하였다. 또한 빅카인즈를 활용하여 추가적인 분석을 실시하였다. 우선순위 분석결과, 코로나, 전염병, 방역, 백신, 발생, 바이러스, 감염, 개발 순으로 그 중요도가 나타났다. 근접중심성 분석 결과 정부, 사망, 계획 순으로 그 중요도가 나타났으며, 빅카인즈 분석결과는 코로나 19, 질병관리 본부 등이 중요한 것으로 나타났다. 본 연구의 결과를 토대로 신종감염병에 대한 대국민 인식 제고 및 방역, 백신 및 치료제 개발 등에 정부의 정책적인 지원이 필요하다고 할 수 있다.

키워드 : 신종감염병, 코로나, Covid-19, 텍스트마이닝, 빅데이터

Abstract This study attempted to derive important factors of emerging infectious diseases by collecting and analyzing text data onto emerging infectious diseases. For this purpose, articles in the Naver News database were directly crawled, pre-processed, and used for data analysis. In addition, additional analysis was performed using Big Kinds. As a result of the priority analysis , the importance was shown in the order of corona, infectious disease, quarantine, vaccine, outbreak, virus, infection, and development. As a result of the proximity centrality analysis, the importance was shown in the order of government, death, and plan, and the analysis result of Big Kinds showed that Covid-19 and the Korea Centers for Disease Control and Prevention were important. Based on the results of this study , it can be said that the government's policy support is needed to raise public awareness of new infectious diseases, prevent disease, and develop vaccines and treatments.

Key Words : Novel infectious disease, Corona, Covid-19, Text mining, Big data

1. 서론

산업화 이후 우리 생활에는 많은 변화가 일어나고 있다. 최근에는 정보통신 기술의 발전으로 인해 우리의 생활이 많이 편리해지고 있다. 4차산업혁명 시대로 접어들면서 기술의 발전으로 인해[1-4] 인간의 생활은 더욱 더 윤택해졌지만, 많은 폐해도 일어나고 있다. 특히 물자가 풍족해지기 시작하면서 인구가 늘어나고 있으며, 늘어난 인구로 인해 다양한 폐해들이 발생하고 있다.

사람들이 지구의 환경을 압박하기 시작함으로써 생태계가 파괴되기 시작하였다. 자연의 훼손으로 인해 지구는 해양오염, 지구 온난화, 동물 및 생물 종의 파괴 등으로 인해 몸살을 앓고 있다. 특히 자연의 생태계 파괴가 심해지면 심해질수록 바이러스는 더 빠르고 쉽게 전파가 되는 양상을 보이고있다.

또한 최근에는 강아지와 고양이를 기르는 인구가 증가하면서 사람과 동물을 통해 전파되는 인수공통 감영병의 발병이 지속적으로 급증을 하고 있다.

그리고 세계적으로 평균 기온이 점차 높아지기 시작하면서 지구 온난화 현상이 빠르게 진행이 되고 있다.

*Corresponding Author : Kyung-Do Suh (bumsoskd@hanmail.net)
Received June 1, 2021 Revised June 17, 2021
Accepted June 20, 2021 Published June 28, 2021

한 연구에 따르면, 기온의 상승과 질병발생 및 부람이 증가하고 있는 것으로 나타났다.

14세기 유럽에서 흑사병으로 인해 많은 사람들이 목숨을 잃었으며, 20세기에 가장 치명률이 높았던 전염병인 스페인 독감을 시작으로, 메르스, 현재의 코로나 까지 다양한 점염병들이 등장을 하고 있다.

메르스로 인해 전염병에 대한 인식이 일부 시작되었으나, 코로나의 등장으로 인해 전염병에 대한 중요성에 대한 인식이 높아지기 시작하였다.

대부분의 연구들은 의료인의 인식조사, 법적 관점 등에 포커스가 맞추어져 있지만 감염병 인식에 대한 빅데이터 관점에서의 연구는 부족한 실정이다. 이에 본 연구에서는 언론 기사에 나타난 빅데이터 분석을 통해 감염병에 대한 인식을 조사하는데 있다. 이를 위해 R, Jsoup 등을 통해 데이터를 수집하고 분석을 수행하였다.

2. 관련연구

2.1 감염병

감염병의 예방 및 관리에 관한 법률 제 2조에 따르면, 감염병은 제1급감염병, 제2급감염병, 제3급감염병, 제4급감염병, 기생충감염병, 세계보건기구 감시대상 감염병, 생물테러감염병, 성매개감염병, 인수(人獸)공통감염병 및 의료관련감염병을 의미한다[5].

신종감염병이란, 과거에는 존재하지 않았지만, 새롭게 질병을 일으키는 감염병을 의미한다. 즉 과거 20년 동안 사람에게서 발생이 증가한 감염병과 가까운 미래에 발생의 증가가 의심 및 예상이 되는 감염병을 통칭한다. 사스, 메르스, 코로나 등의 바이러스가 국내에서는 신종감염병으로 볼 수가 있다.

"제1급감염병"이란 생물테러감염병 또는 치명률이 높거나 집단 발생의 우려가 커서 발생 또는 유행 즉시 신고하여야 하고, 음압격리와 같은 높은 수준의 격리가 필요한 감염병으로서 다음 각 목의 감염병을 말한다. 다만, 갑작스러운 국내 유입 또는 유행이 예견되어 긴급한 예방·관리가 필요하여 질병관리청장이 보건복지부장관과 협의하여 지정하는 감염병을 포함한다.

"제2급감염병"이란 전파가능성을 고려하여 발생 또는 유행 시 24시간 이내에 신고하여야 하고, 격리가 필요한 다음 각 목의 감염병을 말한다. 다만, 갑작스러운 국내 유입 또는 유행이 예견되어 긴급한 예방·관리가 필요하여 질병관리청장이 보건복지부장관과 협의하여 지정하는 감염병을 포함한다.

"제3급감염병"이란 그 발생을 계속 감시할 필요가 있어 발생 또는 유행 시 24시간 이내에 신고하여야 하는 다음 각 목의 감염병을 말한다. 다만, 갑작스러운 국내 유입 또는 유행이 예견되어 긴급한 예방·관리가 필요하여 질병관리청장이 보건복지부장관과 협의하여 지정하는 감염병을 포함한다.

"제4급감염병"이란 제1급감염병부터 제3급감염병까지의 감염병 외에 유행 여부를 조사하기 위하여 표본감시 활동이 필요한 다음 각 목의 감염병을 말한다.

"기생충감염병"이란 기생충에 감염되어 발생하는 감염병 중 질병관리청장이 고시하는 감염병을 말한다.

"세계보건기구 감시대상 감염병"이란 세계보건기구가 국제공중보건의 비상사태에 대비하기 위하여 감시대상으로 정한 질환으로서 질병관리청장이 고시하는 감염병을 말한다.

.
.
.

2.2 텍스트마이닝

빅데이터는 대량의 정형과 비정형 데이터들의 집합이라고 할 수 있다. 일반적으로 빅데이터는 정형데이터와 비정형 데이터로 분류를 하고 있다. 정형 데이터는 통계, 수치 등의 데이터들이 주를 이루고 있으며, 비정형 데이터는 정형 데이터를 제외한 모든 데이터를 의미한다. 비정형데이터 중에서 최근에 각광을 받고 있는 것이 바로 텍스트마이닝이다.

텍스트마이닝은 비정형 빅데이터 분석의 유형 중 하나이다. 비정형 데이터는 텍스트, 영상, 음성, 이미지, 로그 등의 데이터를 의미하며, 언론기사, 블로그, SNS 등에 올라온 글과 영상 등이 바로 비정형 데이터이다[1, 6].

이러한 비정형 데이터 중에서 현재 많은 연구가 진행이 되어 는 분야가 바로 텍스트마이닝 분야이다. 영상, 음성 등에 대한 분석도 중요하지만, 보편적으로 많이 사용하는 기법이 바로 텍스트 마이닝이다.

2.3 선행연구

신종 감염병에 대한 연구는 지속적으로 증가를 하고 있었으나, 2020년 코로나가 전세계적으로 급속하게 퍼지고, 관심이 높아지기 시작하면서 관련 키워드를 포함한 논문들이 지속적으로 증가를 하기 시작하였다.

대부분의 연구들은 인식조사, 대응 체계, 법 제도 등에 한정이 되어 있다.

박미화(2020)는 신종감염병인 코로나-19에 대한 간호대학생들의 인식에 대한 조사를 수행하였다. 분석결과, 신종 감염병에 대한 윤리적인 인식과 의사결정은 높은 것으로 나타났다[7]. 박혜자와 이옥철(2019)은 의료인을 대상으로 신종감염병의 윤리인식과 의사결정 수준을 파악하였으며, 분석결과 윤리적 이슈에는 민감

한 것으로 나타났고, 윤리적인 의사결정 4가지 원칙에는 충실한 것으로 나타났다[8].

송승현외(2020)는 신종감염병 위기 발생시의 대응 방안 및 체계를 검토하여 초기에 문제를 해결 할 수 있는 방안을 제시하고자 하였다[9].

전세영(2018)은 2015년 기준으로 개편된 국가 방역 체계를 거버넌스 차원에서 효과성을 분석하였다. 분석결과 전문성과 투명성은 이전보다 향상이 되었으며, 대응성과 연계성은 부족한 것으로 나타났다. 메르스 상황에서 분석을 수행한 부분이지만, 본 연구의 결과를 통해 코로나 상황에서도 거버넌스 효과성을 파악할 수 있다[10].

김대중과 최요한(2020)은 코로나-19로 인해 경제에 어떠한 영향을 미치는지에 대한 예측과 대응 방안에 대해 분석을 하였다. 분석결과 기업의 레버리지론 증가 및 부채의 증가로 인해 금융 부문의 위험이 증가되기 때문에 정부 개입을 통한 경제 회복 및 성장을 강조하였다[11].

백경희와 김자영(2020)은 신종감염병에 대한 위기대응과 보건의료 빅데이터의 수집에 대해서 법적인 관점에서의 고찰을 하였다. 분석결과 신종 감염병 위기대응의 목적에 맞는 고려가 필요하고, 보건의료와 관련된 빅데이터 수집시에 개인정보 보호에 대한 고려는 반드시 필요하다고 하였다[12]. 특히 위기대응과 개인정보보호 사이의 균형이 필요하다는 점을 강조하였다.

서경도외(2020a)는 미래감염병 발생 시에 신속하고 적합하게 위기관리 체계를 수행하고 이를 통해 피해 경감을 최소화할 수 있도록 중앙 정부의 보건의료 정책을 위한 신종감염병의 감시 체계 방안을 모색하였다[13]. 또한 서경도외(2020b)는 신종감염병에 대한 예방 및 관리 체계에 대해 형사 정책적 관점에서 신종감염병의 예방 및 관리에 관한 법률을 중심으로 살펴보고, 이를 통해 문제점과 한계점를 검토하고, 주요 국가 제도를 비교 및 분석을 통해, 신종감염병 예방 및 관리를 위해 효과적 및 실효적인 형사정책적 관점의 대응방안을 제시하였다[14].

3. 연구방법 및 절차

본 연구에서는 Jsoup를 사용하여 네이버 뉴스를 크롤링하였다[15]. 수집된 텍스트 데이터들은 전처리 과정을 거쳐, R을 활용하여 분석을 수행하였다. 또한 추가적으로 빅카인즈에서 “감염병”의 키워드를 통해 검색을 하고 분석을 수행하였다. 분석 절차 프로세스는 다음과 같다.

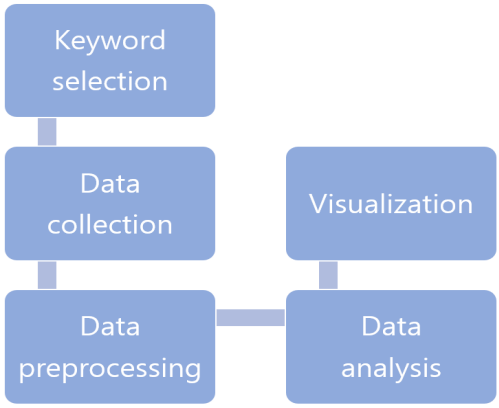

Fig. 1. Research Porcess

데이터 수집 기간은 2020년 1월 1일 ~ 2020년 12월 31일까지로 설정하였으며, 네이버 뉴스 데이터베이스에 있는 기사만을 자료 분석에 활용하였다.

4. 분석결과

4.1 우선순위 도출 결과

R을 활용하여 우선순위를 도출한 결과는 다음과 같다. 먼저 코로나가 가장 많이 언급이 되어 1위로 나타났다. 2020년 1월부터 전세계적으로 코로나가 전파되기 시작하면서 코로나를 키워드로 하는 기사 원고가 증가를 한 것이 원인으로 보인다.

2위는 전염병으로 나타났다. 코로나를 처음에는 대수롭지 않게 생각했는데, 확산 속도, 사망 등 전세계적으로 셧다운을 일으키고 있기 때문에 그 이전에 일어난 스페인 독감, 흑사병, 메르스 등에 대해 관심이 증가하면서 이에 대한 기사가 많이 생성되어서 이러한 결과가 나온 것으로 보인다.

3위는 방역으로 나타났다. 한국은 k-방역을 내세우면서 방역을 초기에 잘 한 국가로 인식이 되었다. 이에 전세계적으로 각국이 앞다투어 빗장을 걸고, 국내외 이동을 차단하는 시도가 일어났지만, 국내의 경우는 어느 정도 자유롭게 이동을 할 수 있게 하면서도 방역에 성공하였기 때문에 이러한 기사들이 주로 생성된 것으로 보인다.

4위는 백신으로 나타났다. 코로나의 확산세가 꺽이지 않고 있다. 각종 변이들이 발생을 하게 되면서 백신에 대한 관심이 증가하고 있다. 2021년부터 전세계적으로 백신 접종이 시작되었다. 화이자, 모더나, 얀센, AZ 등의 백신 외에도 다양한 백신이 현재 개발중이거나 임상중이다. 특히 변이 바이러스에 대한 효과가 있는지 여부가 중요한 이슈로 떠올라서 백신에 대한 기사가 많이 생성이 된 것으로 보인다.

5위는 발생으로 나타났다. 코로나의 경우 발생 지역에 대한 논란이 지속적으로 있어 왔다. 현재까지 발생 원인을 파악하기 위해 노력을 하고 있지만, 아직까지 정확하게 밝혀진 바는 없다고 할 수 있다. 또한 변이 바이러스 발생이 전세계적으로 지속되고 있기 때문에 이러한 결과가 나온 것으로 보인다.

바이러스는 6위로 나타났다. 코로나도 바이러스의 일종으로 보통 기사에서는 코로나 바이러스 혹은 COVID-19, 코로나 19 등으로 기사 원고가 작성이 되고 있다. 그러나 대부분 언론이나 사람들은 코로나로 부르고 있기 때문에 바이러스는 6위로 나타났다.

7위는 감염으로 나타났다. 감염은 기생충과 세균, 바이러스 등이 몸속으로 들어가 빠르게 증식하고 이를 통해 많은 질환을 유발한다. 현재도 코로나 19의 전파 및 감염으로 인한 증상이 지속되고 있다. 이에 7위로 나타났다.

8위는 개발로 나타났다. 이는 4위로 나타난 백신과 유사하다고 할 수 있다. 즉, 백신의 개발에 대한 필요성이 2020년부터 제기되었고, 현재 화이자, 모더나, 얀센이 미국의 FDA 승인을 받았으며, 많은 제약 및 바이오 기업들이 현재도 코로나 치료제 및 백신 개발에 힘을 쏟고 있어서 이러한 결과가 나타난 것으로 보인다.

9위는 접종으로 나타났다. 국내에서는 2021년 들어서 백신의 접종이 시작되었지만, 일부 국가에서는 2020년 하반기 부터 시작되었기 때문에 백신 접종에 대한 기사가 주를 이루고 있어서 9위로 나타났다. 확산은 10위로 나타났다. 코로나가 잠시 주춤했다가 지금 다시 전세계적으로 확산이 되고 있다. 대만의 경우는 코로나 방역을 잘한 국가 중의 하나였으나, 현재는 환자가 급증하고 있어 이러한 결과가 도출된 것으로 보인다.

Table 1. Priority derivation result

	rev	Freq
1	corona	882
2	Epidemic	828
3	quarantine	763
4	vaccine	709
5	Occur	547
6	virus	543
7	infection	436
8	Development	290
9	inoculation	280
10	diffusion	280
11	diffusion	261
12	government	260
13	Response	250
14	USA	250
15	situation	249
16	People	236
17	area	233
18	disease	219
19	Farm	199
20	Disinfection	197

이외에도 중국, 정부, 대응 미국, 상황, 사람, 지역, 질병, 농장, 소독 등이 중요한 요인으로 도출이 되었다.

4.2 근접 중심성 결과

근접중심성은 각 노드 간의 거리를 기준으로 중심성을 측정하는 방법이다. 직접 연결된 노드뿐만이 아니라 간접 연결된 모든 노드들 간의 거리를 합산해서 중심성을 측정한다. 그리고 각 노드가 네트워크 구조 내에서 얼마나 중심에 위치하고 있는지를 나타내는 중심성 지표이다[16].

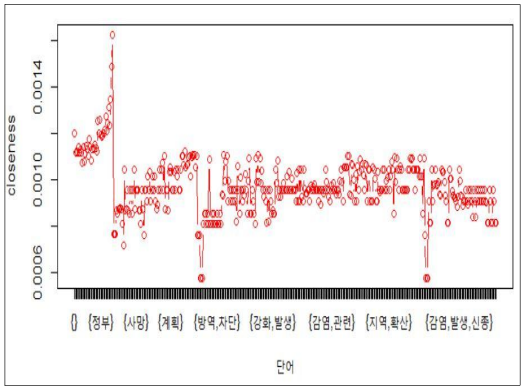

Fig. 2 Closeness Centrality

근접중심성을 분석한 결과, 정부, 사망, 계획 순으로 그 중요도가 나타났으며, 방역·차단, 강화·발생 감염·관련, 지역·확산, 감염·발생·신종 등 순으로 그 중요도가 나타났다.

4.3 빅카인즈 활용 결과

빅카인즈에서 뉴스검색을 동일한 기간(2020년 1월 1일~2020년 12월 31일)으로 설정을 하고 분석을 수행하였다.

Fig. 3 Wordcloud analysis

분석결과, 코로나 19가 가장 중요한 요인으로 도출이 되었으며, 질병관리본부, 보건복지부, 신종 코로나바이러스 감염증 등이 중요한 요인으로 도출이 되었다.

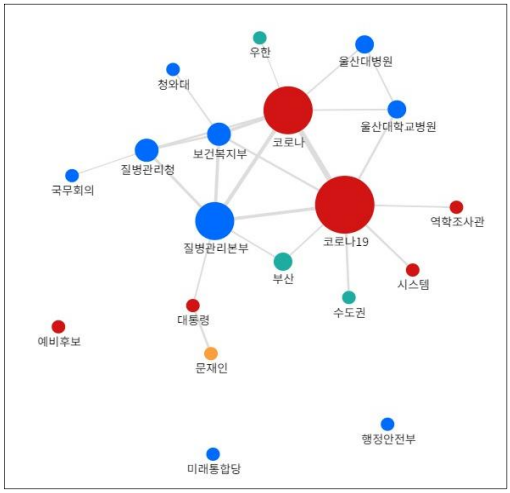

Fig. 4 Network Analysis

네트워크 분석결과는 다음과 같다. 코로나 19, 코로나, 질병관리본부가 중요한 요인으로 나타났다. 코로나 19, 질병관리본부, 보건복지부 등이 서로 관련이 높은 것으로 나타났다.

5. 결론

본 연구는 신종감염병에 대한 동향을 파악하고자 네이버 뉴스기사를 크롤링하여 텍스트 데이터를 수집하고 이를 전처리하고 분석을 수행하였다.

우선순위 분석결과, 코로나, 전염병, 방역, 백신, 발생, 바이러스, 감염, 개발 순으로 그 중요도가 나타났다. 근접중심성 분석 결과 정부, 사망, 계획 순으로 그 중요도가 나타났으며, 빅카인즈 분석결과는 코로나 19, 질병관리 본부 등이 중요한 것으로 나타났다.

본 연구는 신종 감염병을 키워드로 뉴스 기사에 나타난 중요한 키워드가 무엇인지를 파악하였다는 점에서 기존 연구와 차별성을 가진다. 아직까지 텍스트 마이닝 기법을 활용하여 신종 감염병에 대한 동향 및 중요도를 파악한 연구는 부족한 실정이기 때문이다.

또한 본 연구의 결과는 정책적으로 시사점을 가지고 있다. 방역, 백신, 감염, 개발 등에 대한 관심이 높다는 것이 입증되었다. 이에 정책적으로 코로나 방역을 위한 노력을 경주해야 하며, 백신과 치료제 등에 대한 개발을 정책적 우선순위로 하여 정부 차원에서 지원을 해서 우리나라 뿐만이 아니라 전세계 시장에서 코로나와 관련한 백신과 치료제 시장을 확보할 수 있을 것으로 보인다. 이를 위해서는 정책적으로 다양한 정책을 지원해야만 한다.

본 연구의 한계점으로는 특정 기간동안의 신종감염병에 대한 키워드와 네이버 뉴스 DB 기사만을 대상으로 하였으며, 다양한 분석 방법을 활용하지 못했다는 한계점이 있다. 향후 연구에서는 다양한 데이터의 수집이 필요하며, 텍스트마이닝 외에도 네트워크 분석 기법 등의 분석이 필요할 것으로 보인다.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5B5A07073840).

REFERENCES

[1] E. M. Park & J. H. Seo. (2020). Analysis of Research Trends in Technology Innovation: Focus on SCOPUS DB. *Journal of Convergence for Information Technology, 10(8),* 120-126.

DOI: 10.22156/CS4SMB.2020.10.08.120

[2] S. T. Park & Y. K. Kim. (2019). A study on deriving an optimal route for foreign tourists through the analysis of big data. *Journal of Convergence for Information Technology, 9(10),* 56-63. DOI: 10.22156/CS4SMB.2019.9.10.056

[3] S. T. Park, D. Y. Kim & G. Li. (2020). An analysis of environmental big data through the establishment of emotional classification system model based on machine learning: focus on multimedia contents for portal applications. *Multimedia Tools and Applications,* 1-19. DOI: 10.1007/s11042-020-08818-5

[4] S. T. Park, J. R. Jung & C. Liu. (2020). A study on policy measure for knowledge-based management in ICT companies: focused on appropriability mechanisms. *Information Technology and Management, 21(1),* 1-13. DOI: 10.1007/s10799-019-00298-w

[5] Korea Ministry of Government Legislation, https://www.law.go.kr

[6] E. M. Park & J. H. Seo. (2019). A Study on Leadership Typology in Sports Leaders Based on Big Data Analysis. *Journal of the Korea Convergence Society, 10(7),* 191-198. DOI: 10.15207/JKCS.2019.10.7.191

[7] M. Park. (2020). Awareness about pandemic influenza, Ethical Awareness, and Ethical Decision-making among Nursing Students in the situation of COVID-19 pandemic. *Journal of Digital Convergence, 18(10),* 335-344. DOI: 10.14400/JDC.2020.18.10.335

[8] H. J. Park & O. C. Lee. (2019). Ethical Awareness and Decision-making of Healthcare Providers in Response to Pandemic Influenza-Focused on Middle East Respiratory Symptom Coronavirus. *Crisisonomy, 15(1),* 19-29.

[9] S. H. Song, J. K. Choi & S. R. Kim. (2020). Response System for Emerging Infectious Disease Crisis-Focusing on the Organization and the Operation of an Initial Response Task Force. *Crisisonomy, 16(5),* 1-16.

[10] S. Y. Jeon. (2020). An Analysis of the Effectiveness of the National Prevention System: Response to the MERS Outbreak in 2018. *Journal of the Korean Society of Hazard Mitigation, 20,* 39-50. DOI: 10.9798/KOSHAM.2020.20.3.39

[11] D. J. Kim & Y. H. Choi. (2020). Economic impact of the COVID-19 and policy challenges. *Future Growth Research, 6(1),* 163-174.

[12] K. H. Baek & J. Y. Kim. (2021). A Legal Study on Infectious Disease Crisis Response and Healthcare Big Data Collection. *Chosun law journal, 28(1),* 3-31.

[13] K. D. Suh & P. A. Choi. (2020). A Study on Healthcare Policy Response to Risks of Future Infectious Diseases: Focused on Infectious Disease Surveillance Systems. *Journal of Industrial Convergence, 18(3),* 109-116. DOI: 10.22678/JIC.2020.18.3.109

[14] K. D. Suh, J. I. Choi & P. A. Choi. (2020). Research on criminal policy measures for the prevention and management of infectious diseases: Focusing on Mers. *Journal of Industrial Convergence, 18(6),* 9-17. DOI: 10.22678/JIC.2020.18.6.009

[15] S. T. Park & C. Liu. (2020). A study on topic models using LDA and Word2Vec in travel route recommendation: focus on convergence travel and tours reviews. *Personal and Ubiquitous Computing,* 1-17. DOI: 10.1007/s00779-020-01476-2

[16] K. S. Noh, J. H. Kim, S. T. Park & B. S. Kim. (2020). Big data analysis using R, Wowpass.

서 경 도(Kyung-Do Suh) [종신회원]

· August, 2012 : Kumoh National Institute of Technology, Department of Management (Ph.D.)
· March, 2019 ~ Present : Kumoh National Institute of Technology Industry-Academic Collaboration Foundation

· Research Interests : Infectious, Management
· E-Mail : bumsoskd@hanmail.net