JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 22, No. 5: 722~727, May 2018

SPSS를 이용한 대기질과 기상인자와의 미세먼지 상관관계 분석

차진욱[1] · 김장영[1*]

Analysis of fine dust correlation between air quality and meteorological factors using SPSS

Jinwook Cha[1] · Jangyoung Kim[1*]

[1*]Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

요 약

현재까지 미세먼지에 대한 연구는 예측, 분석, 측정 등으로 나눠지는데, 주로 대기환경 분야에서 이루어져 왔다. 미세먼지는 대기질 인자와 기상인자 그리고 배출 등 여러가지 원인으로 인해 발생한다. 각 요소들이 미세먼지에 얼마나 많은 영향을 끼치는지 상관관계를 분석하는 것이 우선이라고 판단하였고, 이를 실험하였다. 이 상관 분석에는 기상청과 에어코리아를 통해 확보한 대기질 인자와 기상인자 데이터를 이용, IBM사의 SPSS라는 Tool을 사용하여 이루어졌다. 그 결과 각 대기질 인자와 기상인자들이 미세먼지 수치에 미치는 영향정도와 상관관계를 좀 더 명확하게 알 수 있었다. 본 논문에서는 미세먼지 수치와 영향요소 및 상관관계의 정확한 분석을 위해 상관분석 및 피어슨 상관계수로 결과를 나타낸다.

ABSTRACT

Until now, the study of fine dust has been divided into prediction, analysis and measurement, mainly in the field of atmospheric environment. Fine dust is caused by various causes such as atmospheric quality factor, meteorological factor and emission. It was determined that it was a priority to analyze the correlation of how much each element affects fine dust, and it was experimented. This correlation analysis was done using IBM SPSS tool using air quality factor and meteorological factor data obtained from Korea Meteorological Administration and Air Korea. As a result, the influence of air quality factors and meteorological factors on the fine dust level was more clearly understood. In this paper, we present experimental results as correlation analysis and pearson coefficient for more precise analysis between PM10 values and affected factors.

키워드 : 미세먼지, SPSS, 상관 분석, 알고리즘

Keyword : PM10, SPSS, Correlation analysis, Algorithm

Received 21 February 2018, **Revised** 22 March 2018, **Accepted** 5 April 2018

*** Corresponding Author** Jangyoung Kim (E-mail: jykim77@suwon.ac.kr, Tel: +82-31-229-8345)
Department of Computer Science, The University of Suwon, Hwaseong 18323, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2018.22.5.722 pISSN:2234-4772

Ⅰ. 서 론

미세먼지는 수많은 대기오염 물질을 포함한다. 자동차, 공장 등의 산업 활동에서 생성되기도 하지만, 일상생활에서도 미세먼지는 발생한다. 이와 더불어 국내에서 생산되는 먼지뿐만 아니라 중국발 미세먼지 및 황사도 미세먼지 수치에 굉장히 많은 영향을 끼친다[1]. 미세먼지는 인체에 많은 영향을 끼치는데, 미세먼지의 농도가 증가할수록 심장혈관과 호흡기의 질환으로 인한 사망률이 증가한다는 연구가 있다[2]. 뿐만 아니라 암과 전신 질환에도 깊은 관련이 있다는 연구 결과도 발표된 바가 있어[3], 미세먼지에 대한 경각심과 연구가 필요하다. 미세먼지에 관한 연구는 대부분 대기환경 분야에서 이루어져 왔으며 미세먼지가 인체에 끼치는 영향을 의학 분야에서 연구하기도 하였다. 정보통신 분야에서 미세먼지에 대한 연구는 거의 이루어지지 않았다. 현재 기상청에서 쓰는 미세먼지 예측 모델[4]에도, 정보통신 분야에서 많이 쓰이는 알고리즘들이 각각의 요소로 들어가 있는 이 시점에, 본 논문은 그동안 축적된 데이터를 이용하여, 미세먼지 수치와 각각의 인자들과의 상관관계를 밝혀보려 한다. 이를 위해, 대기, 기상 등에 대한 데이터를 확보하였다. 확보한 데이터는 SPSS를 이용하여 상관관계 분석을 시도하였다. SPSS는 분석을 위한 계획, 데이터 수집, 분석, 보고의 전과정을 지원 할 수 있는 프로그램이다.

Ⅱ. 관련연구 및 연구내용

2.1. 데이터 수집 및 가공

대기질 관측데이터는 에어코리아(www.airkorea.or.kr)에서 확보하였다. 에어코리아는 한국환경공단에서 운영하는 전국 실시간 대기오염도 공개 홈페이지이다. 이곳에서 국립 환경과학원의 최종 확정자료를 excel파일로 다운로드 하였다. 대기질 최종 확정 데이터는 2014년도부터 제공되었기 때문에 2014년부터 2017년 6월까지의 서울지역 대기질 데이터를 확보하였다.

기상 데이터는 기상청(www.kma.go.kr)에서 제공하는 기상자료개방포털(https://data.kma.go.kr)에서 기상관측 데이터를 얻을 수 있었다. 정확한 실험 결과를 얻기 위해 대기질 관측데이터와 마찬가지로 2014년부터 2017년 6월까지의 서울지역 기상관측 데이터를 확보하였다.

획득한 두 데이터를 하나의 excel 파일로 시간과 장소에 따라 하나의 파일로 융합하는 가공을 하여 실험의 input data로 활용하였다.

2.2. 기존연구

서울시 8개의 지점(구의, 구로, 도봉, 동대문, 신사, 종로, 강서, 북한산)에서 초미세먼지를 채취하여 분석하고 황산이온, 질산이온, 암모늄이온과의 관계를 분석한 연구가 있다. 황산이온, 질산이온, 암모늄이온의 농도분포 분석을 통하여 고농도 초미세먼지 발생 원인을 해석하고, 이 중 암모늄이온 농도의 분포가 계절적 특성보다 고농도 미세먼지 발생과 더욱 밀접한 관계라는 것을 밝혔다[5].

또한 초미세먼지의 농도는 기상인자(풍속, 강수량, 일사량 등)에 영향을 받고, 대기물일인 이산화질소, 이산화황, 오존 등에도 영향을 받는다. 우리나라의 자동차 수나 오염원으로 인한 초미세먼지 외에도 중국으로부터 유입되는 초미세먼지까지 고려하기 위해 풍향과 풍속까지 고려하여 공간자기상관 행렬에 기초한 공간패널모형을 소개하는 연구가 있다 [6].

2.3. 배경지식

상관분석은 확률론과 통계학에서 두 변수 간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법이다. 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계라 한다. 본 논문에서는 대기질 인자 변수인 O3, NO2, CO3, SO2, PM2.5와 기상인자 변수인 일평균 기온, 일강수량, 평균풍속, 평균 상대습도, 합계일조시간, 합계일사, 평균 지면온도, 안개 계속시간을 분석 하여 상관계수를 구하고, 상관분석을 하려 한다. 상관계수는 0.0 ~ 0.2 사이라면 상관관계가 거의 없다. 0.2 ~ 0.4 사이면 상관관계가 낮은편이고 0.4 ~ 0.6 이면 상관관계가 있다고 말한다. 0.6 ~ 0.8 사이라면 상관관계가 높다고 말할 수 있고, 0.8 ~ 1.0이면 상관관계가 매우 높다고 한다. 단, 계수가 음수로 나오면 절대값을 취하면 된다.

피어슨 상관계수(Pearson correlation coefficient)는 두 변수간의 관련성을 구하기 위해 보편적으로 이용된다. X와 Y라는 변수가 있을 때 계수 r은 X와 Y가 함께

변하는 정도 / X와 Y가 따로 변하는 정도이다. r은 X 와 Y 가 완전히 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일하면 -1을 가진다.

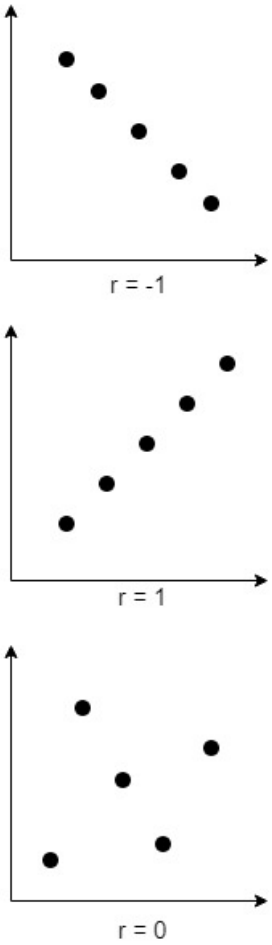

Fig. 1 Correlation coefficient graph

결정계수 (Coefficient of determination) 는 r^2 로 계산하며 이것은 X 로부터 Y 를 예측할 수 있는 정도를 의미한다. 통상적으로 r이 -1.0 ~ -0.7 사이이면, 강한 음적 선형관계, -0.7~ -0.3 사이이면, 뚜렷한 음적 선형관계, -0.3~ -0.1 사이이면, 약한 음적 선형관계, -0.1~0.1 사이이면, 거의 무시해도 되는 선형관계, 0.1~0.3 사이이면, 약한 양적 선형관계, 0.3~0.7 사이이면, 뚜렷한 양적 선형관계, 0.7~1.0 사이이면, 강한 양적 선형관계로 해석한다[7]. 그림1은 상관관계에 대한 배경지식을 돕고자한 그림으로, r은 계수를 의미한다.

Ⅲ. 제안 기법

3.1. 제안 모델

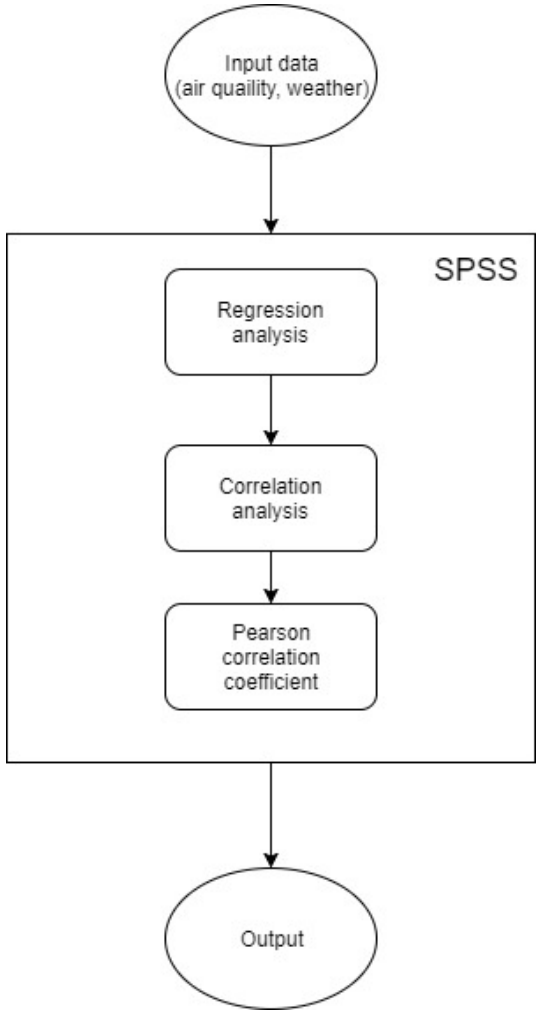

Fig. 2 Proposed model flow chart

본 논문에서는 SPSS를 이용하여, 대기질 데이터와 기상 데이터를 가지고 분석을 실시하여 상관계수를 구하고, 피어슨 상관계수를 구하는 상관분석을 하려 한다.

그림2는 제안모델의 알고리즘 순서도를 설명하고 있다. 이 모델은 대기질 인자 변수인 O3, NO2, CO3, SO2, PM2.5 와 기상인자 변수인 일평균 기온, 일강수량, 평균풍속, 평균 상대습도, 합계일조시간, 합계 일사, 평균 지면온도, 안개 계속시간을 input data로 사용한다.

	SO2	SolarRadiation	WindSpeed	Fog	Precipitation	Humidity	PM2.5	O3	Temperature	NO2	Sunshine	CO	GroundTemperature
SO2	1.000	-.025	-.222	.045	-.022	.108	-.229	-.126	.020	-.308	-.020	-.116	.024
SolarRadiation	-.025	1.000	-.069	-.059	.052	.120	.098	-.331	.096	-.220	-.755	.064	-.203
WindSpeed	-.222	-.069	1.000	-.028	-.094	-.107	.104	-.229	.020	.255	.062	.075	.094
Fog	.045	-.059	-.028	1.000	.042	-.121	.001	.022	.035	.009	.042	-.033	-.011
Precipitation	-.022	.052	-.094	.042	1.000	-.221	.128	-.024	-.002	-.050	.060	.023	-.010
Humidity	.108	.120	-.107	-.121	-.221	1.000	-.143	.158	-.089	.209	.212	-.150	-.055
PM2.5	-.229	.098	.104	.001	.128	-.143	1.000	-.444	.142	-.047	.007	-.535	-.158
O3	-.126	-.331	-.229	.022	-.024	.158	-.444	1.000	.040	.356	.179	.127	-.144
Temperature	.020	.096	.020	.035	-.002	-.089	.142	.040	1.000	-.069	.118	-.174	-.954
NO2	-.308	-.220	.255	.009	-.050	.209	-.047	.356	-.069	1.000	.169	-.468	.005
Sunshine	-.020	-.755	.062	.042	.060	.212	.007	.179	.118	.169	1.000	-.107	-.060
CO	-.116	.064	.075	-.033	.023	-.150	-.535	.127	-.174	-.468	-.107	1.000	.264
GroundTemperature	.024	-.203	.094	-.011	-.010	-.055	-.158	-.144	-.954	.005	-.060	.264	1.000

Fig. 3 Correlation coefficient

input data는 14개의 인자, 그리고 각 인자당 1096일 치의 데이터(2014.1.1 ~ 2017.6.31.)로, 총 15334개의 변수를 이용하였다. 많은 수의 변수를 일일이 계산하기 위하여 SPSS를 이용, 연산하였고 이를 통하여 분석을 한다. 상관관계는 상호적인 관계이기 때문에, 두 변수간의 인과관계를 파악하려면 상관관계가 필요하다. 그러나 상관관계가 높은 변수라고 반드시 인과관계가 있지는 않기 때문에 독립변수와 종속변수의 인과적 관계를 파악하는 분석도 함께 실시하여 실험에 정확도를 높이고자 하였다. 그 후 상관분석을 통하여, 각 변수 사이의 상관계수를 구하였다. 그 후 가장 대중적으로 쓰이는 피어슨 상관계수를 이용하여 각각의 변수를 매칭하여 두 변수의 관계가 선형적인지 비선형적인지를 판별하고, 선형적이라면 어느 정도의 선형관계를 가지고 있는지를 각각 알아보았다.

Ⅳ. 실험 결과

4.1. 상관분석

그림3은 각 변수들이 어느 정도의 상관관계가 있는지 나타내는 지수인 상관계수 이다. 미세먼지인 PM10을 제외한 다른 변수들 간에, 어느 정도의 상관관계가 있는지 분석한 결과이다. SO2에 가장 높은 상관관계를 가지고 있는 변수인 NO2는 수치가 0.308이다. 가장 낮은 상관관계를 가지고 있는 변수는 수치 0.020으로 일사량과 기온이다. 일사량과 가장 높은 상관관계를 가지고 있는 변수는 일조량이며, 그 수치는 0.765이다. 반면 가장 낮은 상관관계를 가진 변수는 SO2이고, 0.025이다. 그러나 상관계수가 0.8 이상일 경우 상관관계가 매우 높고, 적어도 0.6은 넘어야 상관관계가 높다고 말할 수 있는데, 그림2를 보면 0.6을 넘는 수치는 손에 꼽힌다. 0.4는 넘어야 상관관계가 존재하는데 이 또한 적다. 상관계수가 미세먼지와 별개로 서로에 대한 상관관계는 예상보다 높지 않은 것으로 나타났다.

	PM10	Temperature	Precipitation	WindSpeed	Humidity	Sunshine	SolarRadiation	GroundTemperature	Fog	PM2.5	O3	NO2	CO	SO2
PM10	1.000	-.168	-.176	-.078	-.124	.048	.022	-.167	.126	.764	.036	.409	.491	.507
Temperature	-.168	1.000	.138	-.164	.357	.062	.403	.980	-.064	-.082	.570	-.280	-.467	-.365
Precipitation	-.176	.138	1.000	.120	.413	-.351	-.263	.103	.011	-.185	-.041	-.166	-.150	-.205
WindSpeed	-.078	-.164	.120	1.000	-.021	.020	.034	-.122	.033	-.262	.250	-.478	-.339	-.078
Humidity	-.124	.357	.413	-.021	1.000	-.589	-.418	.286	.120	.026	-.041	-.166	-.057	-.284
Sunshine	.048	.062	-.351	.020	-.589	1.000	.830	.160	-.091	-.043	.349	-.042	-.160	.104
SolarRadiation	.022	.403	-.263	.034	-.418	.830	1.000	.499	-.086	-.059	.619	-.153	-.349	-.014
GroundTemperature	-.167	.980	.103	-.122	.286	.160	.499	1.000	-.069	-.100	.643	-.333	-.530	-.371
Fog	.126	-.064	.011	.033	.120	-.091	-.086	-.069	1.000	.010	-.074	-.001	.050	-.029
PM2.5	.764	-.082	-.185	-.262	.026	-.043	-.059	-.100	.010	1.000	.039	.553	.678	.593
O3	.036	.570	-.041	.250	-.041	.349	.619	.643	-.074	.039	1.000	-.471	-.507	-.147
NO2	.409	-.280	-.166	-.478	-.166	-.042	-.153	-.333	-.001	.553	-.471	1.000	.811	.623
CO	.491	-.467	-.150	-.339	-.057	-.160	-.349	-.530	.050	.678	-.507	.811	1.000	.643
SO2	.507	-.365	-.205	-.078	-.284	.104	-.014	-.371	-.029	.593	-.147	.623	.643	1.000

Fig. 4 Pearson correlation coefficient

4.2. 피어슨 상관계수

그림4는 피어슨 상관계수(Pearson correlation coefficient)로 각 변수간의 관련성을 나타내었다. PM2.5는 피어슨 상관계수가 0.764로 PM10과 가장 강한 선형관계를 가진다. Sunshine과 Solar Radiation, O3, Wind Speed는 거의 무시해도 되는 선형관계를 가지고 있는 것으로 확인되었다. Temperature, Precipitation, Humidity, Ground Temperature는 약한 음적 선형관계를 갖는다. Fog는 약한 선형관계를 가지며, NO2와 CO, S02는 뚜렷한 선형관계를 갖는 것으로 확인되었다.

그림5는 실험결과로써 복잡한 표로 되어있는 그림3의 이해를 돕고자 결정적인 수치들을 그래프화 하여 가독성을 돕고자한 그림이다.

수량, 평균풍속, 평균 상대습도, 평균 지면온도, 안개계속시간은 음적 선형 관계를 갖는 것을 알 수 있었다. 그 중에서도 합계일조시간과 합계 일사, O3, 풍속은 선형관계라 보기 힘들다는 것도 발견하였다.

미세먼지 배출 등의 데이터는 따로 존재하지 않아 미세먼지에 영향을 끼치는 모든 인자를 변수로 두고 실험하지 못한 것이 아쉬운 부분이다. 표 1은 그림 4에 대한 이해를 돕기 위한 차트로써 그림4에서 꼭 확인해야 하는 부분만 따로 다시 기재하였다.

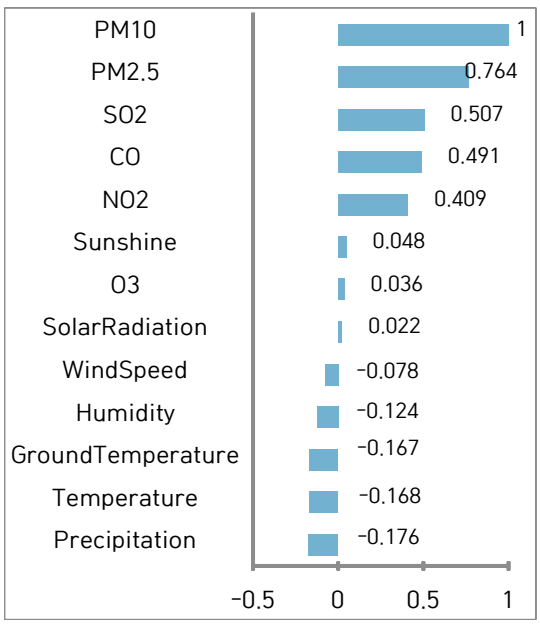

Fig. 5 Pearson correlation graph

Table. 1 Pearson factor and numerical value

factor	numerical value
PM10	1
PM2.5	0.764
SO2	0.507
CO	0.491
NO2	0.409
Sunshine	0.048
O3	0.036
Solar Radiation	0.022
Wind Speed	-0.078
Humidity	-0.124
Ground Temperature	-0.167
Temperature	-0.168
Precipitation	-0.176

Ⅴ. 결론 및 향후과제

먼저 미세먼지를 제외한 나머지 변수들 간의 상관관계는 그리 높지 않은 것으로 발견되었다. 지면온도와 일평균기온, 일조량과 일사량 등 아주 밀접한 관계가 성립되어 있는 변수들이 강한 상관관계가 있는 정도이다. 그러나 각각 정도의 차이는 있지만 PM10과 O3, NO2, CO, SO2, PM2.5, 합계일조시간 합계 일사는 양적 선형관계를, 기상인자 변수인 일평균 기온, 일강수량, 평균풍속, 평균 상대습도, 평균 지면온도, 안개 계속시간은 음적 선형 관계를 갖는 것을 알 수 있었다. 그 중에서도 합계일조시간과 합계 일사, O3, 풍속은 선형관계라 보기 힘들다는 것도 발견하였다.

미세먼지 배출 등의 데이터는 따로 존재하지 않아 미세먼지에 영향을 끼치는 모든 인자를 변수로 두고 실험하지 못한 것이 아쉬운 부분이다.

향후 과제로는 더 많은 데이터를 확보하여, 미세먼지에 조금이라도 영향을 끼치는 인자 데이터를 최대한 많이 확보하여 최대한의 인자들과의 상관관계를 밝혀 미세먼지 수치 예측 등 미세먼지에 관한 많은 연구에 기여할 예정이다 [8-9].

ACKNOWLEDGEMENT

The paper was supported by the research grant of the University of Suwon in 2017.

REFERENCES

[1] D. H. Shin, and Y. M. Noh, "Aerosol Optical Properties and Separation of Asian Dust using AERONET Sun/Sky Radiometer Measurement at the Asian Dust Source Region," *Korean Journal of Remote Sensing*, vol. 32, no.3, pp.245-251, June 2016.

[2] J. S. Oh, S. H. Park, M. K. Kwak, C. H. Pyo, K. H. Park, H. B. Kim, S. Y. Shin, and H. J. Choi, "Ambient Particulate Matter and Emergency Department Visit for Chronic Obstructive Pulmonary Disease," *Journal of The Korean Society of Emergency Medicine*, vol. 28,no. 1, pp. 32-39, Jan. 2017.

[3] H. J. Bae, "Effect of Short-term Exposure to PM10 and PM2.5 on Mortality in Seoul," *Journal of Korea Society of Environmental Health,* vol.40, no.5, pp. 346-354, May 2014.

[4] Y. S. Koo, H. Y. Yun, H. Y. Kwon, and S. H. Yu, "A Develope of PM10 Forecasting System," *Journal of Korean Society for Atmospheric Environment*, vol. 26, no. 6, pp. 666-682, Nov. 2010.

[5] Y. H. Seo, "Characterization of high concentration PM2.5 by nitrate and ammonium ions of PM2.5 in Seoul," *Journal of Korea Society of Environmental Administration*, vol. 21, no.1, pp. 1-7, Mar. 2015.

[6] J. H. Lee, Y. M. Kim, and Y. K. Kim "Spatial panel analysis for PM2.4 concentrations in Korea," *Journal of the Korean Data & Information Science Society*, vol. 28, no. 3, pp. 473-481 Mar. 2017.

[7] G. Kader, and C. Franklin "The Evolution of Pearson's Correlation Coefficient," *Journal of National Council of Teachers of Mathematics*, vol. 102, no. 4, pp. 292-299, Nov. 2008.

[8] N. Arora, M. Martolia, and A. Ashok "A Comparative study of the Image Registration Process on the Multimodal Medical Images", *Asia-pacific Journal of Convergent Research Interchange*, HSST, vol.3, no.1, pp. 1-17, Mar. 2017.

[9] J. Cha, and J. Kim, "Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 22, no. 4, pp. 595-601, April 2018.

차진욱(Jinwook Cha)

2017년 3월: 수원대학교 컴퓨터학과 석사 재학

※관심분야 : Big data, Networks

김장영(Jangyoung Kim)

2005년 2월: 연세대학교 컴퓨터과학 공학사
2010년 5월: Pennsylvania State Univ, 공학석사
2013년 7월: State University of New York 공학박사
2013년 8월: University of South Carolina 조교수
2014년 3월: 수원대학교 컴퓨터학부 조교수
※관심분야 : Big data, Cloud computing, Networks