

# GAM 모델 기반의 미세먼지 취약계층 대상 새로운 위험지수 개발 및 활동 추천시스템을 통한 생활밀착형 스마트시티 연구

권재선\*, 김지연\*\*, 유현수\*\*, 최지원\*\*

\*이화여자대학교 건축도시시스템공학과

\*\*이화여자대학교 통계학과

jaesun99@ewhain.net, jiyeoneye@ewhain.net, hsu607@ewhain.net, jiii111@ewhain.net

## A Smart city study through development of new risk index based on GAM model and activity recommendation system for the vulnerable class of fine dust

Jae-Sun Kwon<sup>1</sup>, Ji-Yeon Kim<sup>2</sup>, Hyun-Su Yu<sup>2</sup>, Ji-Won Choi<sup>2</sup><sup>1</sup>Dept. of Architectural and Urban Systems Engineering, Ewha Womans University<sup>2</sup>Dept. of Statistics, Ewha Womans University

### 요 약

최근 미세먼지는 중대한 건강위험요소로 고려되고 있고, 미세먼지 취약계층은 이에 대한 적극적 대응이 필요하다. 그러나 현재의 대기환경지수는 세분화 되어있지 않아 본 논문에서는 위해성 평가와 GAM 모형을 기반으로 건강취약계층 대상을 위한 미세먼지 위험지수를 새롭게 개발하였다. 또한, 이에 따라 실내 및 실외활동을 추천하는 시스템을 구현함으로써 생활밀착형 스마트시티로 발돋움하도록 한다.

### 1. 서론

초미세먼지는 환경보건 관련 가장 큰 위협 요소로 꼽히며 이에 대한 시민들의 우려는 커지고 있다. 기저 질환 환자, 노약자 등은 초미세먼지에 대한 장기 노출의 건강 영향이 더욱 크며 사망률과도 유의미한 관계를 보인다. 도시 미세먼지 취약계층은 적극적 대응이 필요한 대상으로, 정부도 그동안 다양한 사업을 수행했으나[1] 교통 및 대규모 배출원 등에 대한 규제에 비중을 두고 있으며, 생활권 공간 단위에서의 구체적인 서비스는 적게 사용된다. 나아가 현재 통합대기환경지수(Comprehensive air -quality index, CAI)는 해외와 비교해 완화된 수준이므로 시민들의 실제 감각을 반영하지 못하고 있다. 이에 스마트시티인 경기도 시흥시에서 수집되는 미세먼지 데이터를 활용하여 사회적 약자를 지원할 수 있는 서비스를 기획하였다.

### 2. 선행 연구

#### 2.1 위험도 및 위험성 평가

위험성 평가란 위험상황에 놓였을 때 개인이나 집단의 건강에 손상이 가는지 추정하는 과학적인 과정으로, 건강 위해성 평가는 건강 위해도를 추정하는 과정이다. 위험도(Risk)란 독성물질에 노출된 개인 및 집단이 유해물질에 대해 특정한 농도에서 유해한 결과가 유발될 확률이나 가능성을 말하며, OECD의 정의에서는 다음과 같이 나타낸다[2].

$$\text{위험도(Risk)} = \text{노출량(Exposure)} \times \text{유해성(Hazard)}$$

#### 2.2 GAM 모형 기반 지수

일반화 부가모형(Generalized additive model: GAM)은 계절적 변동, 기상요인, 요일 등의 영향으로 인한 교란작용을 제거하기 위하여 대기오염과 사망에

많은 영향을 미치는 기상요인을 평활화하여 변수를 보정할 수 있다. 그렇기 때문에 대기 오염이 건강에 미치는 영향을 분석하기 적합한 모형이다[3].

#### 2.3 추천 알고리즘 현황

추천 시스템은 크게 콘텐츠 기반 필터링(Content-based Filtering, CBF)과 협업 필터링(Collaborative Filtering, CF)으로 구분할 수 있다. CBF 에는 콘텐츠의 특징을 이용해 사용자에게 아이템을 추천해주는 키워드 기반 추천 시스템(Item-based Collaborative Filtering, IBCF) 등이 있으며, 협업 필터링에는 유사도를 특정 아이템과 가장 유사한 아이템을 추천하는 item 기반 추천 시스템 등이 있다.

그러나 추천시스템의 성능을 저해하는 초기 평가 문제(Early Rate), 데이터의 희소성(Sparsity) 등 여러 문제점이 존재한다. 윤수진 등(2005), 최준희 등(2017)은 위의 문제를 해결하고자 KNN, 베이지안 분류기 등을 적용하거나[4] 다중 데이터 기반의 추천시스템을 시도해보았다[5]. 본 연구에서도 이러한 문제 등을 보완할 수 있는 알고리즘을 비교 분석해보고자 한다.

### 3. 본론

#### 3.1 사회적 약자를 위한 미세먼지 지수 개발

##### 3.1.1 데이터 설명

분석에 사용한 데이터는 일별 기상, 대기오염물질, 사망자수, 입원환자 수 데이터를 사용했으며, 기간은 2018, 2019 년도, 지역은 시흥시로 한정하였다. 국내외 연구결과에서 대기 오염이 호흡기 및 심혈관계 질환 발생 및 입원에 유의한 영향이 있다고 보고되고 있기 때문에[3] 입원환자 수와 사망자수의 경우 주상병코드가 호흡기계, 심혈관계인 데이터만



데이터에 활동 특성' 컬럼을 생성하여 실내 활동(1), 실내 장소(2), 실외 장소(3)으로 분류하였고, 개발된 미세먼지 지수에 따라 아주 나쁨 단계일 시, 실내활동만을 추천하고, 보통 단계일 시 실외장소와 실내장소 중에서 추천하는 등의 알고리즘을 설계하였다. 추천시스템 모델은 콘텐츠기반 필터링(CBF), 아이템 기반 협업 필터링(BCF) 두가지로 생성하였다. 본 연구의 모델은 텍스트 데이터가 존재하여, TF-IDF 를 이용한 임베딩으로 특징을 아이템의 추출한다. 이러한 방식으로 인한 단순성 문제를 해결하기 위해 카테고리 유사도, 평점개수, 평균평점, 구매자 타겟 유사도 행렬에 가중치를 조절하는 방식을 활용하여, 여러 정보를 고려하는 추천시스템을 구현하였다. IBCF 는 데이터 희소성 문제를 해결하기 위해 랜덤으로 30 명의 사용자 리뷰 데이터를 추출하고, 이들의 평균 평점을 이용해 데이터를 재구성하였다. 재구성한 데이터를 바탕으로 사용자 기반 협업 필터링 추천시스템을 구현하였다.

4. 실험 및 평가

4.1 시흥시 데이터 설명 및 실험 방식

데이터는 KT 에서 제공된 스마트시티인 시흥시의 미세먼지 데이터이다. 시간별 미세먼지 데이터로 초단위까지 나와있으나, 실험의 편의성을 위해 1~4 월 데이터 중 매일 오전 9 시, 오후 3 시, 오후 9 시 대기오염물질 수치를 임의로 선정해서 300 개를 추출하여 기존 CAI 에서의 오염 등급 단계와 새로운 지수에서의 오염 등급 단계를 도출하였다.

4.2 추천시스템 평가성능 지표

연구에 활용한 데이터의 유형에 맞춰 분류 정확도도 모델을 평가한다. 추천시스템에서의 정확도는 상위 k 개의 추천에 대한 평가지표로 precision@k, Mean Average Precision@k(MAP@k) 등이 있다. 이때, MAP@k 는 precision@k 에서 k 를 점점 늘려가며 얻게 되는 precision score 를 평균 낸 것으로, 이번 연구의 평가지표이다.

4.3 결과 분석

#	CO	NO2	O3	PM10	PM2.5	SO2	CAI GRADE	CAI	NEW GRADE	NEW INDEX
1	0.3	0.009	0.036	17	7	0.002	보통	55	중음	63.894737
2	0.4	0.02	0.029	28	13	0.003	중음	48	중음	75.252525
3	0.5	0.042	0.004	27	10	0.002	보통	70	보통	124.727794
4	0.4	0.02	0.026	30	17	0.003	중음	50	중음	75.252525
5	1	0.044	0.003	44	34	0.005	보통	73	보통	127.535817
6	0.5	0.028	0.012	44	26	0.003	보통	69	보통	105.071633
7	0.5	0.014	0.027	47	46	0.004	보통	85	보통	136
8	0.4	0.019	0.02	24	8	0.003	보통	51	중음	70.30303
9	0.4	0.025	0.012	17	14	0.004	보통	56	중음	100
10	0.5	0.018	0.016	78	64	0.003	나쁨	109	나쁨	164
...	...	...	...	...	...	...	...	...	...	...
295	0.4	0.004	0.083	57	27	0.002	보통	94	조금 나쁨	153.512821
296	0.4	0.022	0.106	82	34	0.006	나쁨	139	나쁨	182.410256
297	0.4	0.011	0.039	10	2	0.002	보통	58	중음	71.631579
298	0.5	0.013	0.04	13	4	0.002	보통	58	중음	74.210526
299	0.5	0.014	0.029	9	5	0.002	중음	48	중음	48.333333
300	0.3	0.01	0.037	25	19	0.002	보통	56	중음	66.473684

(그림 2) 새로운 지수로 구분한 데이터 셋

(그림 2)의 형태와 같은 테스트 데이터를 구성하여, 사회적 약자를 위한 미세먼지 위험 지수 모델과 추천시스템에 입력하였다. 미세먼지 수치(PM2.5)가 34 인 날에 심혈관 질환이 있는 user a 는 미세먼지 위험 지수가 127.535 로 보통 단계이고, 추천 활동은 실내장소+실외장소 데이터의 모델에서 결과를 도출할 수 있다.

위의 표를 통해 이전에 보통으로 묶었던 단계가 좋음, 보통, 조금 나쁨 등으로 세분화되었다는 사실을 관찰할 수 있다. 낮은 단계의 세분화는 특히 약자에게 의미 있는 방향으로 발전했다고 할 수 있다.

추천시스템의 경우, 나쁨 단계와 보통 단계일때의 MAP@10 성능을 확인한 결과, 아래의 <표 2>와 같이 IBCF 모델의 성능이 더 좋았으며 이는 정확도 수치가 0.099, 0.108[7],0.027, 0.045, 0.055 [8]등의 연구와 비교했을 때 2~3 배인 수치이기에 합리적인 모델이라고 생각한다.

<표 2> 추천시스템 MAP@10 성능

	실내활동		실내장소+실외장소	
	CB	IBCF	CB	IBCF
MAP@10	0.277	0.309	0.054	0.223

5. 결론

본 논문에서는 사회적 약자를 위한 미세먼지 위험지수 모델을 개발하고, 이를 바탕으로 사용자의 개인적인 취향을 함께 반영해 활동 추천 시스템을 구현하였다. 미세먼지 취약자를 중심으로 한 위해도 연구이며 기존의 CAI 보다 WHO 및 해외의 기준에 알맞게 세분화되어 있으므로 약자 뿐 아니라 모두에게 확장 가능하다는 점에서 의미가 있다. 추천시스템은 산출한 미세먼지 위험 지수를 바탕으로, 실내 또는 실외 활동을 추천해 사용자가 실용적으로 이용할 수 있도록 하였다. 또한, 추천 시스템의 단순성 및 희소성 문제를 해결하기 위해 다양한 방식의 추천 시스템을 구현하고자 하였다. 구현한 추천 시스템은 전국으로 확대하는 효과를 기대할 수 있다.

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링프로젝트 결과물입니다.

참고문헌

[1] 황인창, "고령화와 초미세먼지 건강영향." 정책리포트 -.287 1-22, 2019.

[2] 홍지환, "초미세먼지(PM2.5) 농도와 위해도 지수간 시차분석을 통한 상관관계에 관한 연구." 국내박사학위논문 한세대학교 대학원, 2020.

[3] 신재은, "새로운 통합대기환경지수 개발과 호흡기 및 심혈관계 질환 사망에 미치는 영향." 環境政策 28.3, 2020, 113-157.

[4] 윤수진(Sujin Yun), and 윤희병(Heebyung Yoon). "개인화 추천시스템의 성능 향상 적용 알고리즘 분석." 한국지능시스템학회 학술발표 논문집 15.1, 2005, 181-184.

[5] 최준희(Jun-hee Choi), 박찬수(Chan-soo Park), 황태규(Tae-gyu Hwang), 윤태욱(Tae-uk Yun), and 김성권(Sung Kwon Kim). "초기 사용자 문제를 위한 다중 데이터 기반의 추천시스템." 한국정보과학회 학술발표논문집 2017.12, 2017, 847-849.

[6] Geneva: World Health Organization "WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.", 2021.

[7] 조용민, 남기환 "협업 필터링 및 하이브리드 필터링을 이용한 동종 브랜드 판매 매장간 취급 SKU 추천 시스템" 한국지능정보시스템학회, 2017, 77-110.

[8] 이승진, "음악 추천을 위한 가사정보 및 음악신호 기반 특성 탐색 연구", 서울대학교 융합과학기술대학원 공학석사 학위 논문, 2019.