J. Korean Soc. Hazard Mitig.
Vol. 19, No. 4 (Aug. 2019), pp.143~149
https://doi.org/10.9798/KOSHAM.2019.19.4.143

ISSN 1738-2424(Print)
ISSN 2287-6723(Online)
www.kosham.or.kr

소방방재

빅데이터 활용을 위한 화재인자 표준화에 관한 연구

Standardization of Fire Factor for Big Data

박은석* · 민세홍**
Park, Eunseok*, and Min, Sehong**

Abstract

As a city's population becomes concentrated with economic development, approximately 46,000 fires occur yearly. The fourth industrial revolution necessitated big data analysis of fire-related statistics in the fire fighting sector and fire-related data advancement via fire pattern analysis. For big data analysis of fire statistics, this study analyzed the national fire patterns by reviewing the regional and monthly fire occurrence probability for each fire factor on the fire statistics data provided by the National Fire Agency. The data of the factors possibly causing fire effects were selected by the data evaluation criteria to accordingly assess the fire statistics data quality and build a database. Lastly, they were applied to a fire forecasting platform with standardization by data preprocessing.

Key words : Fire Forecasting Platform, Big Data Analysis, Fire Pattern, Fire Effect

요 지

경제 발전과 함께 도시의 인구가 밀집하면서 연평균 4만 6천 건에 달하는 화재가 발생하고 있으며, 현재 제 4차 산업혁명에 따라 화재소방분야에서도 화재통계 빅데이터 분석 및 화재 패턴 분석을 통한 화재 관련 데이터의 고도화가 필요하게 되었다. 본 연구에서는 화재통계 빅데이터 분석을 위해 소방청에서 제공한 화재통계 데이터에 대한 지역별·월별·화재요인별 화재발생 확률을 검토하여 우리나라의 화재패턴을 분석하였다. 이를 통해 화재영향을 줄 수 있는 인자를 데이터 평가 기준을 통해 추출하여 화재통계 데이터 품질 평가를 진행하고, 데이터베이스를 구축하였다. 마지막으로, 이러한 데이터를 전처리 과정을 통해 표준화하는 작업을 거쳐 이를 화재예측플랫폼에 적용하고자 한다.

핵심용어 : 화재예측플랫폼, 빅데이터 분석, 화재패턴, 화재영향

1. 서 론

ICT의 융합으로 이뤄지는 차세대 산업혁명인 제 4차 산업혁명에 따라 다양한 산업 분야에서 빅데이터 분석과 IoT 및 ICT 기술 등의 적용을 통해 운영시스템 및 데이터 관리시스템을 고도화하고 있다. 이에 따라 소방 분야에서도 4차 산업혁명에 따라 발전할 필요가 있다.

경제 발전과 함께 도시로의 인구가 집중되면서 연평균 4만 6천 건에 달하는 화재가 발생하고 있고, 이에 따라 인명 및 재산의 피해가 크게 발생하고 있다.

현재 소방청에서는 화재관련 정보시스템을 통해 방대한 양의 데이터를 수집하여 사전 예측 및 예방 등에 활용하고 있으나, 수집된 정형 데이터의 부정확성과 빅데이터 분석 처리의 복잡성 등으로 인하여 수집된 데이터를 통한 정확한 화재패턴의 분석은 사실상 어렵다. 또한, 소방청 국가화재정보시스템에서 제공하던 화재네비게이터는 Fig. 1과 같이 화재위험등급을 5단계로 구분지어 매우 위험, 위험, 경계, 주의 및 보통으로 화재 발생 위험도를 가시화 시키고 있으며 구(區)단위로 1년에 한번씩 매년 데이터를 업데이트 했다.

*정회원, 가천대학교 일반대학원 소방방재공학과 석사과정(E-mail: piskpask@naver.com)
Member, MS Candidate, Department Fire Protection Engineering, Graduate School, Gachon University
**교신저자, 정회원, 가천대학교 공과대학 설비・소방공학과 교수(Tel: +82-31-750-5714, Fax: +82-31-750-8746, E-mail: shmin@gachon.ac.kr)
Corresponding Author, Member, Professor, Department of Fire & Disaster Protection Engineering, Gachon University

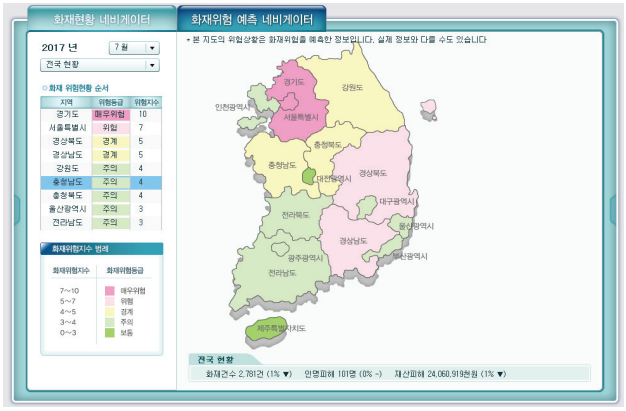

Fig. 1. Fire Navigator Previously Provided by NFDS

Yoon and Shin (2018)에서 확인한 결과, 미국에서는 범죄 발생현황과 범죄 관련 영향인자를 통해 범죄 예측시스템을 도입하여 범죄발생건수를 약 20% 이상 감소시키는 효과를 달성하였다.

실시간 화재발생 예측 플랫폼을 구축하기 위해서는 기존 소방청 화재통계 데이터를 통한 국내 화재 패턴 분석이 필요하며, 이를 고도화하기 위해서는 화재에 영향을 미치는 인자가 무엇인지 분석해야 한다.

이렇게 분석한 화재영향인자를 화재통계 데이터에 적용하기 위해 데이터 평가 기준을 통해 데이터를 추출하고 전처리 과정을 거쳐 표준화를 진행하였다.

2. 소방청 통계분석을 통한 화재 패턴 분석

2.1 소방청 데이터 분석

소방청(NFA, 2017)에서 제공한 화재통계 데이터는 2008년부터 2017년까지 10년간의 데이터로 91가지의 변수로 구성되어있다. 이러한 변수들 중 화재사고 발생 지역 예측 및 화재 피해 규모 예측과 최적화에 적용할 수 있는 발생한 화재에 대한 시간 정보, 화재 유형, 대응 장비를 기준으로 25가지의 변수를 추출하였다. 추출한 25가지의 변수는 화재 발생 날짜(월, 일, 시, 분, 요일), 날씨(풍속, 풍향, 온도, 습도), 화재발생지역(시도별, 시군구별, 읍/면/동별), 출동소요시간, 화재발생층, 화재유형, 화재진압시간, 사망자, 부상자, 총 인명피해, 총 재산피해, 발화요인, 발화장소, 소방동원명수, 동원장비, 펌프 물탱크 수로 구성되어있다.

2.2 통계분석 범위 선정

소방청에서 제공한 화재통계에 기재된 자료를 기반으로 지난 10년간(2008~2017년까지) 화재발생건수를 분석한 내용으로 시도별 범위는 서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기도, 충청북도, 충청남도, 전라북도, 전라남도, 경상북도, 경상남도, 제주로 전국 16개의 시도를 분석하였고, 세종특별자치시의 경우, 신도시로 인한 데이터 부족 및 행정구역의 변화 등으로 대전에 포함시켜 분석을 진행하였다.

또한, 발화요인별 범위는 전기적 요소, 기계적 요소, 가스누출, 화학적 요인, 교통사고, 부주의, 자연적인 요인, 미상, 방화, 방화의심 등 11개의 발화요인으로 한정하였다. 마지막으로 계절의 분류는 봄은 3월에서 5월, 여름은 6월에서 8월, 가을은 9월에서 11월, 겨울은 12월에서 2월로 분류하였다.

2.3 화재 패턴 분석 결과

시도별 화재 패턴 분석 결과, 각 발화요인은 지역별, 월별로 차이를 보이지만 부주의, 전기적 요인, 미상, 기계적 요인, 방화의심, 기타, 교통사고, 방화, 화학적 요인, 자연적인 요인, 가스누출(폭발) 순으로 발생 비율이 높았다. 11개의 발화요인 중 부주의, 전기적 요인이 약 70% 이상으로 대부분을 차지한다. 또한, 부주의는 Table 1과 같이 11개 발화요인 중 가장 큰 비율을 차지하며, 서울, 부산을 제외한 나머지 14개 지역에서 3월(봄)에 최댓값을 가진다.

국가화재정보시스템(NFDS, 2017)을 통해 부주의 화재는 12개의 세부 발화요인으로 분류되고 담배꽁초, 음식물 조리 중, 불씨/불꽃/화원 방치, 불장난, 용접/절단/연마, 기타, 논/임야 태우기, 가연물 근접방치, 빨래 삶기, 유류 취급 중, 폭죽놀이 순으로 발생 비율이 높았다. 또한, Fig. 2와 같이 12개 세부 발화요인 중 담배꽁초, 음식물 조리 중이 약 46% 이상으로 부주의 화재에 절반을 차지하였다.

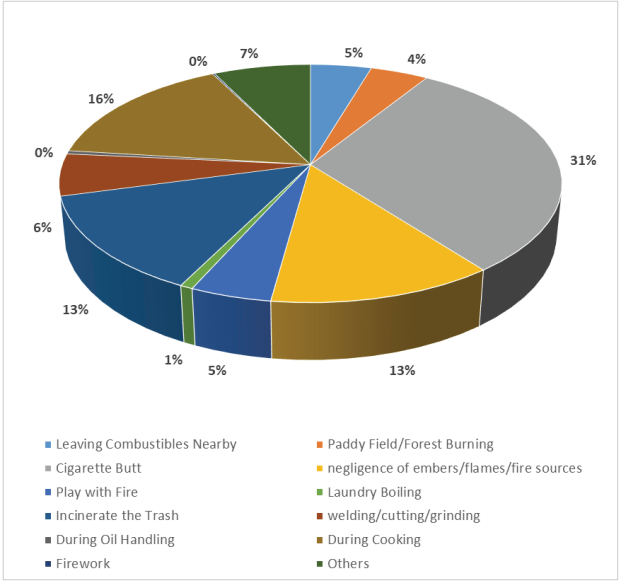

Fig. 2. 10-year Carelessness Fire Detail Factor Ratio

전기적 요인은 Table 2와 같이 전국 16개 지역에서 7월(여름)에 최댓값을 가지며, 이는 여름철 증가하는 전기 사용에 따라 발생 빈도가 증가하는 것으로 판단된다. 대부분 지역이 부주의, 전기적 요인이 큰 비중을 갖지만, 인천, 경기도, 충청북도, 경상북도 지역의 경우, 각 기역의 기계적 요인의

Table 1. Carelessness Fire Pattern Analysis

Season	Winter			Spring			Summer			Autumn		
City	12	1	2	3	4	5	6	7	8	9	10	11
Seoul	48.0	47.4	53.0	56.5	56.7	57.3	52.9	42.3	45.0	54.6	57.4	52.6
Busan	55.3	55.5	55.2	55.2	55.5	54.8	50.5	41.5	43.9	51.8	55.0	52.1
Daegu	49.5	50.1	52.8	55.5	51.9	52.7	48.2	35.3	36.2	46.2	50.7	47.8
Incheon	39.4	39.9	45.0	53.9	47.3	44.7	43.5	26.0	30.7	41.3	44.1	39.2
Gwangju	50.0	51.4	63.6	71.3	62.8	65.4	61.0	47.2	45.0	55.3	56.8	56.3
Daejeon	51.3	53.6	59.7	67.5	63.9	62.9	62.1	48.1	49.1	57.3	59.3	54.8
Ulsan	56.9	60.9	57.5	62.4	60.0	58.9	51.7	44.4	46.0	50.2	54.0	53.6
Gyeonggi	36.7	38.0	45.5	56.9	53.3	47.3	42.5	26.6	28.3	39.2	39.7	38.5
Gangwon	49.8	53.6	54.8	64.4	60.8	59.3	51.7	33.2	34.8	46.5	48.9	46.2
Chungbuk	37.1	41.0	42.7	50.9	46.6	41.8	39.2	22.5	25.4	33.7	39.2	38.4
Chungnam	38.8	42.7	55.6	65.5	59.6	51.6	50.6	26.9	28.7	45.2	42.0	44.0
Jeonbuk	41.1	39.3	47.6	58.6	49.1	44.6	42.5	22.6	26.2	36.1	40.4	45.3
Jeonnam	54.7	55.1	62.0	65.8	59.6	57.1	54.2	30.6	35.4	50.9	52.2	51.8
Kyungbuk	44.3	47.7	50.7	53.1	48.4	46.3	40.8	25.7	26.7	33.1	36.8	42.3
Kyungnam	52.1	53.1	59.5	59.7	55.9	54.1	51.8	33.6	37.6	44.1	48.6	51.0
Jeju	51.4	56.9	47.0	59.1	55.1	57.5	53.6	30.9	36.0	43.9	51.0	48.1

Table 2. Electrical Factor Fire Pattern Analysis

Season	Winter			Spring			Summer			Autumn		
City	12	1	2	3	4	5	6	7	8	9	10	11
Seoul	29.0	30.2	25.9	21.5	21.2	22.4	27.2	37.7	34.4	24.1	20.8	24.4
Busan	22.0	19.9	21.7	21.5	22.7	21.3	25.5	35.0	33.0	23.2	21.8	23.5
Daegu	22.5	22.6	21.1	19.4	21.5	19.4	23.0	34.5	32.1	23.0	21.0	20.9
Incheon	30.3	30.6	27.1	20.8	23.8	26.1	28.5	43.8	35.6	26.4	25.4	29.5
Gwangju	28.1	28.0	19.9	14.5	18.5	17.5	21.3	34.2	32.4	23.0	20.0	23.1
Daejeon	26.0	26.5	21.7	17.4	18.3	19.5	21.1	34.5	31.8	26.1	21.3	22.7
Ulsan	15.4	15.1	16.2	14.6	14.2	16.4	20.9	26.7	25.6	21.3	17.0	16.4
Gyeonggi	27.3	27.6	24.2	17.4	18.0	21.0	23.4	37.6	33.4	25.2	22.7	25.8
Gangwon	20.2	19.5	20.6	14.7	15.8	15.3	19.6	33.3	31.8	23.1	20.5	22.7
Chungbuk	23.0	24.9	24.1	20.5	20.8	23.1	21.6	37.4	30.0	23.2	21.4	22.9
Chungnam	25.8	24.6	18.8	12.8	15.5	19.6	20.2	33.6	30.7	22.1	20.1	19.9
Jeonbuk	22.2	25.5	22.8	17.7	20.0	20.1	19.6	37.3	29.6	20.6	18.9	18.9
Jeonnam	19.5	19.6	17.8	15.8	18.0	20.4	22.5	34.5	30.0	21.0	18.7	20.7
Kyungbuk	21.6	21.2	20.5	19.4	19.8	19.0	22.7	31.3	29.6	26.3	22.9	22.0
Kyungnam	18.7	18.2	15.7	15.4	17.1	19.0	18.8	32.8	27.0	22.8	20.3	19.2
Jeju	18.7	19.6	21.6	16.1	17.1	17.7	23.0	32.0	33.8	18.4	21.0	19.5

Table 3. Mechanical Factor Fire Pattern Analysis

Season	Winter			Spring			Summer			Autumn		
City	12	1	2	3	4	5	6	7	8	9	10	11
Seoul	6.4	5.6	4.8	4.8	5.1	4.2	5.1	5.6	6.3	5.7	5.0	6.0
Busan	7.2	8.7	7.0	7.0	6.5	7.4	7.7	8.3	7.1	7.5	7.0	7.6
Daegu	10.6	10.0	10.0	8.5	10.2	10.1	9.8	12.1	12.0	12.0	10.1	12.2
Incheon	14.9	13.9	12.1	9.8	10.3	9.4	10.1	12.1	14.4	12.7	12.2	12.1
Gwangju	6.8	6.3	5.1	3.8	4.8	4.8	5.0	6.4	8.5	5.6	7.6	4.2
Daejeon	8.1	6.3	3.4	4.0	4.6	3.6	4.5	5.4	6.2	3.6	5.7	7.8
Ulsan	6.1	5.1	4.8	3.7	4.8	6.1	6.1	8.7	8.7	6.1	5.9	6.4
Gyeonggi	15.6	14.8	11.8	8.8	10.1	10.8	11.1	13.7	14.6	13.8	15.2	15.1
Gangwon	11.3	10.1	9.6	7.8	7.1	8.0	8.7	12.6	13.1	13.4	12.1	13.0
Chungbuk	16.5	14.6	12.2	9.0	9.9	12.0	15.0	15.5	15.0	16.0	16.1	15.4
Chungnam	13.0	13.2	8.8	7.4	7.6	8.7	9.1	13.4	13.8	11.0	14.6	12.7
Jeonbuk	13.7	11.9	8.7	6.5	8.1	9.9	12.9	14.9	13.4	12.6	13.9	11.6
Jeonnam	9.2	9.5	6.7	6.3	6.8	8.0	8.1	12.5	12.6	9.9	10.8	9.9
Kyungbuk	12.9	11.7	10.7	9.6	9.6	10.9	13.1	15.9	17.5	15.9	15.6	13.9
Kyungnam	18.7	18.2	15.7	15.4	17.1	19.0	18.8	32.8	27.0	22.8	20.3	19.2
Jeju	18.7	19.6	21.6	16.1	17.1	17.7	23.0	32.0	33.8	18.4	21.0	19.5

비율은 Table 3과 같이 순서대로 12.00%, 12.96%, 13.93%, 13.11%로 전국 평균에 비해 다소 높게 나타나는 것으로 볼 때, 지역의 특성이 발화요인과 관계가 있음을 확인할 수 있었다.

위와 같은 요인별 화재 발생 확률을 통해 대부분의 화재가 부주의와 전기적인 이유에서 발생한다는 것을 확인할 수 있었으며, 이를 근거로 화재에 영향을 주는 인자를 선정하였다. 예를 들어, 부주의의 세부 발생요인과 관련하여 담배꽁초에 대한 화재와 음식물 조리중 발생하는 화재에 대한 영향인자를 선정하기 위해 공공 데이터 포털에서 금연구역 및 흡연율에 대한 데이터와 노령화 및 스트레스 인지율 등 인간의 삶의 질에 관한 데이터를 확보하였다.

3. 화재영향인자 선정

3.1 데이터 평가 모델 구축

먼저 양질의 화재영향인자를 선정하기 위해 평가 기준을 제시하였다. 즉, 문헌 자료 및 소방청 화재통계 데이터를 통해 보편적이고 객관적으로 적용될 수 있다고 판단된 평가 항목을 모아 재구성하여 평가기준을 작성하였으며, 평가 기준의 검증을 위해 다수의 화재영향인자를 적용하였다. 그 결과, 현실적으로 의미가 없거나 평가하기 어려운 항목은 수정하거나 삭제 및 개선하여 최종 평가 기준을 수립하였다.

최종 선정된 평가 기준을 사용하여 Lee (2014)에서 언급되어진 바와 같이 데이터 평가 모델을 구축하였다. Fig. 3과 같이 평가 모델의 적용과정을 단계별로 정리하면 아래와 같다.

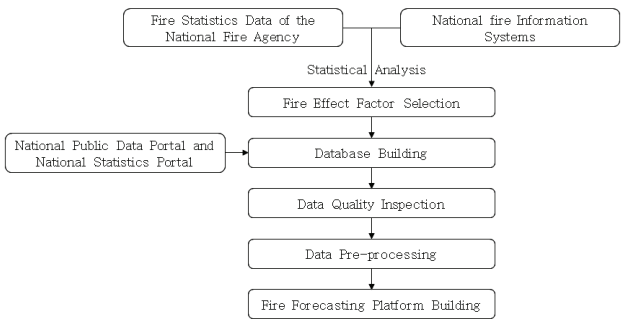

Fig. 3. Application of Step-by-Step Evaluation Model Flow-chart

(1) 소방청 및 국가화재정보시스템을 통해 화재 발생요인을 분석하여 화재와 연관성이 있는 화재영향인자를 선정한다.
(2) 국가공공데이터포털 및 국가통계포털 등 여러 공공데이터포털을 이용하여 화재영향인자를 제공하는 데이터베이스를 구축한다.
(3) 데이터 평가 기준을 통해 선정된 화재영향인자에 대한

평가를 진행한다. 평가의 범주는 연관성, 호환성, 지속성, 생산성으로 구분하여 평가하며, 평가 기준에 따라 객관적 데이터 품질 검사를 진행한다.

(4) 데이터 평가에 의해 최종 선정된 양질의 화재영향인자를 데이터 전처리 과정을 거쳐 화재예측플랫폼에 적용한다.

3.2 데이터 평가 기준 선정

화재영향인자 데이터를 선정하기 위해 총 4가지의 기준을 확립하였고, 4가지 선정기준은 데이터의 연관성, 지속성, 호환성, 생산성으로 구성되어 있다.

첫 번째로 데이터의 연관성의 경우, 화재와 화재영향인자의 연관성을 분석하는 단계로서 화재통계 데이터의 화재원인 분석을 통해 판단하였다.

데이터의 지속성은 데이터가 Open API를 통해 지속적인 자동 업데이트가 가능한 것을 기준으로 판단하였다. 여기서 Open API란 직접 응용 프로그램과 서비스를 개발할 수 있도록 정부나 유관기관에서 공개된 Application Programming Interface (API)를 말한다.

다음으로 데이터의 호환성은 소방청에서 제공하고 있는 화재통계 데이터의 범주는 기본적으로 읍/면/동 단위까지의 데이터로 기준으로 데이터의 호환성을 분석하였다.

마지막으로 데이터의 생산성은 소방청에서 제공하고 있는 화재통계 데이터는 2008년부터 2017년까지 총 10년치 데이터로서 Artificial Intelligence (AI) 반복학습을 위해서는 최소 5년 이상의 데이터가 필요하다.

이러한 선정기준을 통하여 데이터 평가가 이루어지며 평가결과는 우수/양호/보통/미흡 총 4단계로 나누어지며 각각 3/2/1/0점으로 평가된다. 평가된 점수의 합이 9점 이상일 시에는 이를 화재영향인자로 선정하고, 점수 미달 시 인자에서 제외하였다. 여기서 평가 점수 결과의 9점 이상만을 인자로 사용한 이유는 질이 낮은 데이터가 화재예측플랫폼에 포함된다면 질이 높은 데이터로만 구축된 플랫폼에 비해 예측 결과의 신뢰성과 정확성을 떨어뜨릴 수 있기 때문이다.

3.3 화재영향인자 데이터베이스 구축

앞서 진행된 소방청 화재통계 데이터의 패턴을 분석한 결과, 부주의, 전기적 요인의 화재가 전체 화재 건수 중 70%를 차지하는 것과 Min et al. (2018)에서 언급된 초기 화재영향인자 분석을 통해 화재 영향 인자를 선정하였다. 부주의, 전기적 요인의 화재와 연관이 있는 데이터베이스를 구축하기 위해 각종 공공데이터포털을 분석하였으며, 전기사용량, 가스 사용량, E-지방지표, 건축물 노후도, 노령화 인구, 행정구역별 토지이용률, 가구 및 소득분포, 전국 금연구역 현황의 총 8가지의 데이터베이스를 구축하였다. 특히, 부주의 화재는 인간의 심리적・환경적 원인에 따라 발생하는 화재로 이에 대한 통계는 따로 존재하지 않아 E-지방지표, 노령화 인구, 가구 및 소득분포, 전국 금연구역 현황 통계를 포함하였다.

각각의 데이터베이스는 여러 가지의 데이터 요소로 구성되어 있으며, 다양한 공공데이터 포털에서 제공하고 있다. 각 데이터베이스마다 세부적인 데이터 구성과 출처는 다음과 같다.

(1) 전기 사용량은 대지위치와 사용량으로 구성되어 있으며, 건축 데이터 민간 개방시스템에서 제공하고 읍/면/동의 범주로 이루어져 있다.

(2) 가스 사용량은 대지위치와 사용량으로 구성되어 있으며, 건축 데이터 민간 개방시스템에서 제공하고 읍/면/동의 범주로 이루어져 있다.

(3) E-지방지표는 전입인구, 전출인구, 흡연율, 음주율, 스트레스 인지율, EQ-5D 지표, 인구 십만명당 자살율로 구성되어 있으며, 국가통계포털에서 제공하고 시/군/구의 범주로 이루어져 있다.

(4) 건축물 노후도는 건물명, 건축물 구조, 주요용도명, 건물높이, 지상층수, 지하층수, 건물 연령으로 구성되어 있으며, 국가공간정보포털에서 제공하고 읍/면/동의 범주로 이루어져 있다.

(5) 노령화 인구는 60세부터 100세 이상까지를 5세 단위로 구분하고 총 노령인구로 구성되어 있으며, 국가통계포털에서 제공하고 읍/면/동의 범주로 이루어져 있다.

(6) 행정구역별 토지이용률은 전, 답, 과수원, 목장용지, 임야, 광천지, 염전, 대, 공장용지, 학교용지, 주차장, 주유소 용지, 창고용지, 도로, 철도 용지, 제방, 하천, 구거, 유지, 양어장, 수도용지, 공원, 체육 용지, 유원지, 종교용지, 사적지, 묘지, 잡종지로 구성되어 있으며, 국가통계포털에서 제공하고 시/군/구의 범주로 이루어져 있다.

(7) 가구 및 소득분포는 가구분포, 가구원 수, 가구주 연령, 경상소득, 자산, 부채, 순자산액으로 구성되어 있으며, 국가통계포털에서 제공하고 시/도의 범주로 이루어져 있다.

(8) 전국 금연구역 현황은 소재지 지번주소, 금연구역 구분, 제공기관명으로 구성되어 있으며, 공공데이터 포털에서 제공하고 읍/면/동의 범주로 이루어져 있다.

3.4 데이터 평가 결과 분석

Table 4는 구축한 데이터베이스를 평가한 결과로 전국 금연구역 현황을 제외한 6가지의 데이터베이스가 평가 기준을 만족하여 화재예측플랫폼에 적용하기로 하였다. 전국 금연구역 현황의 경우, 데이터가 지역별로 모든 구의 금연구역이 명시되지 않아 기존의 화재통계 데이터와의 호환성이 미흡하고 데이터의 생산성이 2년으로 데이터가 적어 총 6점으로 평가기준에 미달하여 화재예측플랫폼에 적용하지 않았다.

Table 4. Data Evaluation Result

Data evaluation criteria	Correlation	Persistence	Compatibility	Productivity	Sum
Electricity Consumption	3	3	3	2	11
Gas Consumption	3	3	3	2	11
E-Local index	2	2	3	2	9
Building aging	3	3	3	3	12
Aging Population	2	3	3	3	11
Land Use Rate	3	2	3	3	11
Household and Income Distribution	3	2	3	2	10
Non-smoking Area Status	3	0	3	0	6

이와 달리, 가장 높은 점수를 받은 건축물 노후도 데이터는 노후도가 높은 건축물일수록 강화된 소방법이 적용되지 않았기 때문에 화재와의 연관성이 매우 높았으며, Open API를 통해 지속적으로 데이터 업데이트가 가능하며 데이터가 10년치 이상이 축적되어 가장 높은 점수인 12점으로 평가되었다.

이렇게 데이터 평가 기준에 의해 선정된 데이터베이스를 화재예측플랫폼에 적용하기 위해서는 데이터 전처리를 진행하여 표준화를 시켜야 한다.

4. 선정된 데이터의 표준화를 위한 전처리

4.1 데이터 전처리의 필요성 및 과정

화재예측플랫폼에 적용하는데 선정된 소방청 화재통계 데이터와 화재영향인자 데이터를 적용하기 위한 데이터 전처리 과정은 데이터를 처리하고 분석하기 전에 얻으려고 하는 정보를 결정하고, 데이터를 선택하고 정제하는 단계이다. 실제 화재영향인자 데이터베이스는 통일성이 없으며 결측값이 처리되지 않아 불완전하고 코드 데이터와 같은 불필요한 데이터를 포함하고 있다. 이러한 데이터베이스에서 불필요한 데이터를 제거하고 필요한 데이터는 일반화를 통해 데이터의 폼을 맞춘 후, 결측값을 채워주거나 삭제하는 작업을 데이터 정제라 할 수 있다.

여기서 결측값이란 어떠한 데이터베이스에서 변수 항목은 존재하지만, 데이터의 값이 다른 값과 기울기에 비해 오차범위가 큰 이상치를 말한다. 데이터 정제 과정에서 결측값을 처리하는 방법은 제거하거나 수작업으로 채우는 방법 등 다양한 방법이 존재한다. 이번 데이터 정제 작업에는 결측값을 제거하는 방식을 통해 데이터를 정제하였다. 소방청에서 제공한 화재통계 데이터의 경우, 이러한 결측값 데이터가 많이 존재하였는데 대표적으로 출동소요시간이 0초일 때 화재진압시간이 0초라는 데이터는 결측값으로 판단하여 제거하였으며, 소방동원명수/동원장비소계의 평균 비율은 2.5인데 일부 데이터에서 비율이 500이라는 큰 숫자가 나오는 경우는 이상치로 판단하여 제거하였다.

그 다음으로는 데이터의 통합을 위해 여러 곳에 나누어 저장된 데이터를 결합하여 분석하기 위해 데이터의 속성 추가 작업 및 일반화를 진행하였다. 속성 추가 작업이란 데이터의 폼을 맞추기 위해서 기존의 없는 내용을 추가하여 통일성을 부여하는 작업으로서 화재영향인자 데이터베이스에서는 대표적으로 데이터 생성 년도의 속성을 추가하여 작업을 진행하였다. 다음으로 데이터 일반화를 위해 데이터의 형태를 데이터의 생성 년도, 시도별, 시군구별, 읍면동별, 각 데이터베이스 당 세부내용 순으로 데이터를 구성하였다. 또한, 코드 데이터 등 화재영향인자 데이터와 상관없는 데이터를 제거하였다. 대표적인 예를 들면, 건축물 노후도 데이터베이스는 GIS 건물통합식별번호, 고유번호, 법정동 코드, 특수지구구분코드, 대장종류코드 등을 제거하였다.

4.2 데이터 표준화를 통한 화재예측플랫폼 적용

데이터 전처리를 통해 소방청 화재통계 데이터와 화재영향인자 데이터의 표준화 작업을 진행하였다. 이후 소방청 화재통계 데이터와 화재영향인자 데이터를 화재예측플랫폼에 코드화 작업을 진행 후 화재예측플랫폼에 적용하여 Fig. 4와 같이 구축하였다. 또한, 두 데이터 사이의 빅데이터 상관성 분석을 통해 세부 데이터 항목에 대한 분석을 진행할 예정이며, Lee et al. (2010)에서 제시한 상관계수 지표를 기준으로 적용하여 각 데이터 인자들의 차이를 분석할 것이다.

Fig. 4. An Example of Building a Fire Forecasting Platform

이렇게 구축된 화재예측플랫폼을 통해 화재발생확률을 예측하고, 각 읍/면/동별 관련 화재발생패턴, 전기, 가스 사용량, 건축물 노후도 현황, 노령화 인구 현황 등을 실시간으로 확인을 통해 예방하여 전국 화재에 대한 피해를 감소시킬 수 있다. 화재예측플랫폼은 연구 초기 단계로 일(日)단위로 각 시/군/구, 읍/면/동별 화재발생을 예측하며 신뢰성의 문제는 빅데이터 분석의 특성상 화재통계 데이터가 증가함에 따라 해결될 것이다.

5. 결 론

소방청 화재통계 데이터 분석을 통하여 전국의 화재 패턴 분석한 결과, 부주의와 전기적 요인이 화재발생건수 중 많은 부분을 차지하는 것을 확인할 수 있었고 분석한 결과에 의해 화재영향인자 데이터베이스를 구축하였다. 이후 구축 데이터베이스를 전처리 과정을 거쳐 데이터를 표준화시켜 화재예측플랫폼에 적용시켰다.

본 연구의 결과, 다음과 같은 결론을 얻을 수 있었다.

(1) 소방청에서 제공한 2008년부터 2017년까지 화재통계 데이터를 분석하여 전국의 지역별・월별・발화요인별 화재패턴 분석을 진행하였고 이것을 통해 부주의 화재와 전기적 요인의 화재가 높은 비중을 차지한다는 결과를 도출하였다.

(2) 데이터 평가 모델을 구축하여 데이터의 연관성, 호환성, 지속성, 생산성을 기준으로 범주를 구성하여 평가기준에 따라 객관적으로 데이터 품질을 평가하여 화재영향인자를 선정하였다.

(3) 데이터 평가 기준을 통해 국가통계포털, 공공데이터포털, 건축 데이터 민간 개방시스템 등에서 전기, 가스 사용량, E-지방지표, 건축물 노후도, 노령화 인구, 토지이용률, 가구 및 소득분포, 금연구역 현황에 대한 데이터베이스를 구축하였다.

(4) 구축한 데이터베이스의 데이터 폼을 통일하기 위해서 이상치 제거를 통한 결측값을 해결하였고, 데이터 속성 추가 작업 및 일반화를 진행하였다.

(5) 표준화된 데이터를 지능형 위험분석, 피해예측 기반의 재난(화재)상황 대응 플랫폼 기술개발과제의 AI 화재예측플랫폼에 적용하여 구축하였고 이를 소방업무에 적용한다면 화재 예방에 중요한 역할을 담당할 것으로 기대된다.

감사의 글

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 – 재난안전플랫폼기술개발사업의 지원을 받아 수행된 연구임(NO. NRF-2017M3D7A1071832)

References

Lee, H.R., Moon, S.K., Kim, S.K., Kim, K.H., and Kim, J.J. (2010). A correlation analysis between the capability of construction firms and efficiency of construction company using DEA. *Journal of The Architectural Institute of Korea, Structure & Construction*, Vol. 26, No. 5, pp. 125-132.

Lee, J.S. (2014). A study on the data mining preprocessing tool for efficient database marketing. *Journal of Digital Convergence*, Vol. 12, No. 11, pp. 257-264.

Min, S.H., Lee, J.M., Park, E.S., Lim, S.B., Kim, J.B., and Choi, D.H. (2018). A study on perfromance improvement of fire prediction program through fire influencing factor. *Proceedings of Annual Conference*, The Korean Society of Disaster Information, pp. 37-38.

National Fire Agency (NFA). (2017). *Number of fires in 10 years (2017)*.

National Fire Data System (NFDS). (2017). *10-year carelessness fire detail factor ratio (2017)*.

Yoon, S.Y., and Shin, S.H. (2018). Current and future of crime prediction: Improvement of Korean crime prediction system. *Korean Association of Public Safety and Criminal Justice Review*, Vol. 27, No. 3, pp. 243-272.

Received	May 9, 2019
Revised	May 14, 2019
Accepted	May 28, 2019