C-12

화재 발화개요 상의 명사 추출 및 분석

김은주 · 류정우

주식회사 세이프티아

화재조사데이터[1]의 발화개요는 화재 발생요인, 발화기기, 주변 및 피해상황, 조치 내력 등 화재에 대한 다양한 정보를 담고 있으나 자연어 문장으로 이루어진 비정형(unstructured data)이므로 세부 내용에 대한 분석이 어렵다. 이러한 자연어 문장을 분석하기 위해서는 문장의 구문분석(parsing)과 형태소 분석(morphological analysis)이 필요하다. 본 논문에서는 2007년 1월 1일부터 2012년 6월 30일까지 약 5년 반 동안의 화재조사데이터 상의 발화개요의 구문과 형태소를 분석하여 명사를 추출하는 발화개요 명사추출도구를 개발하고 전국, 시도권역별 상관도 분석을 수행하였다.

발화개요 상의 명사추출도구는 [Figure 1]과 같이 반자동으로 오타 및 띄어쓰기 오류를 추출하고 수정하는 오류수정모듈, 문장의 구문을 분석하는 구문분석모듈, 정규식(regular expression)을 이용하여 구문에서 주민등록번호, 차량번호 등 개인정보를 제거하는 개인정보제거모듈, KAIST의 SWRC(semantic web reaearch center)에서 개발한 한나눔(HanNanum)[2]을 이용한 형태소분석모듈과 명사추출모듈, 마지막으로 명사 별 발생 빈도인 CF(collection frequence)[3]를 계산하는 CF 계산 모듈로 구성되어 있다. CF는 정보검색(information retrieval)이나 텍스트 마이닝(text mining)에서 사용하는 문서(document)의 중요도(weight) 측정도구(measure)로 본 논문에서는 전체 발화개요에서 해당 단어가 발생한 빈도를 CF로 정의한다. 추출 결과 255,448건의 발화 개요에서 총 194,936건의 명사가 추출되었으며, 상위 20개 명사와 각각의 CF는 다음 [Table 1]과 같다. 발화개요상의 명사를 추출하고 분석을 수행한 결과 상위 단어에는 화재데이터에 자주 발생하는 “화재”, “발화” 등의 명사와 “1”, “세”, “중”과 같이 화재발생과 의미적으로 연관이 없는 명사들이 포함되어 있으므로 CF 외의 중요도 측정 도구를 이용하여 분석할 필요가 있다. 또한 화재데이터를 위한 불용어(stopword) 리스트도 필요하다.

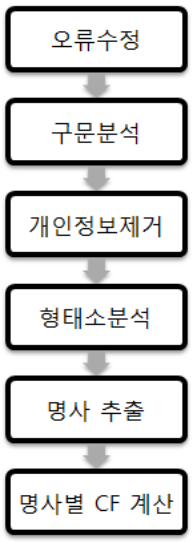

Figure 2. Noun extraction tool process at fire summary

Table 1. Top 20 CF noun in fire summary

순위	단어	CF	순위	단어	CF
1	화재	712,923	11	차량	188,238
2	것	394,015	12	연소	171,043
3	점	346,949	13	세	161,495
4	발화	336,514	14	바	155,295
5	발생	325,612	15	2	146,457
6	한	293,679	16	됨	133,269
7	등	279,925	17	부	124,836
8	1	220,027	18	내부	113,743
9	추정	205,232	19	쓰레기	113,566
10	중	193,027	20	현장	106,976

Table 2. Spearman's rank correlation coefficient(SCC) and Pearson correlation coefficient(PCC) between national and the local fire department

소방본부명	화재(건)	단어	SCC	PCC
경기전체	57,649	69,907	**0.86**	**0.94**
경기소방재난본부	41,073	52,245	0.82	0.94
서울소방재난본부	33,477	19,857	0.74	**0.69**
경남소방본부	21,529	35,786	0.76	0.93
경북소방본부	17,284	31,959	0.75	0.92
충청남도 소방안전본부	16,732	31,067	0.76	0.89
경기북부소방재난본부	16,576	33,634	0.76	0.93
부산소방본부	14,934	30,530	0.76	0.88
강원소방본부	14,093	24,884	0.74	0.93
전남소방본부	12,332	14,025	0.72	0.85
대구소방안전본부	12,042	17,360	0.72	0.87
인천소방안전본부	10,761	24,945	0.74	0.87
전북소방안전본부	9,476	19,576	0.71	0.92
대전소방본부	8,069	14,842	0.73	0.92
충북소방본부	7,958	20,524	0.72	0.91
광주소방본부	7,757	14,307	0.71	0.91
울산소방본부	7,362	20,316	0.73	0.94
제주소방본부	3,993	8,093	**0.69**	0.84

전국단위의 전체 발화개요 상의 명사를 시도권역별로 분류하고 지역별로 차이가 있는지를 분석하기 위하여 스피어만 순위 상관 계수(SCC; Spearman's Rank Correlation Coefficient)와 피어슨 상관계수(PCC; Pearson Correlation Coefficient)[4]를 이용한 전국과 시도권역별 상관도 분석을 수행하였다. 시도권역은 소방재난본부를 기준으로 분류하였고 경기소방재난본부와 경기북부소방재난본부로 나뉜 경기도만 두 소방본부의 화재건수를 합하여 "경기전체"를 추가로 계산하였다.

상관도 분석 결과 [Table 2]와 같이 경기전체와 울산이 가장 높은 양의 상관관계를 보이며, 서울소방재난본부과 제주소방재난본부에서 가장 낮은 양의 상관관계를 보인다. 특히, 다른 지역에 비하여 화재건수가 적은 제주와 달리 서울의 경우 화재건수에 비하여 단어 추출 개수가 현저하게 낮고, 상관도 역시 다른 지역에 비하여 낮으므로 추가 분석이 필요하다.

감사의 글

본 연구는 소방방재청 차세대핵심소방안전기술개발사업단에서 지원하는 2012년 차세대핵심소방안전기술개발사업으로 이루어진 결과입니다.

참고 문헌

[1] 류정우, 김은주, "의사결정트리를 이용한 지역별 화재기상요인분석", 한국화재소방학회 2012년도 추계학술발표회 초록집, pp. 263-266, 2012

[2] http://semanticweb.kaist.ac.kr/home/index.php/HanNanum

[3] Manning, C. D., Raghavan, P., Schütze, H, "Introduction to Information Retrieval", Cambridge University Press, 2008

[4] 김은주, 송원문, 송성렬, 김명원, "IPTV 서비스를 위한 효율적인 협력적 추천 기법", 정보과학회논문지: 소프트웨어 및 응용, 제 39권, 제 5호, pp. 390-398, 2012