The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.5.935
29(5), 935–947

Analysis of patterns in meteorological research and development using a text-mining algorithm

Hongju Park[a] · Habin Kim[b] · Taeyoung Park[a] · Yung-Seop Lee[b,1]

[a]Department of Applied Statistics, Yonsei University;
[b]Department of Statistics, Dongguk University

(Received May 30, 2016; Revised July 14, 2016; Accepted July 19, 2016)

Abstract

This paper considers the analysis of patterns in meteorological research and development using a text-mining algorithm as the method of analyzing unstructured data. To analyze text data, we define a list of terms related to meteorological research and development, construct times series of a term-document matrix through data preprocessing, and identify terms that have upward or downward patterns over time. The proposed methodology is applied to multi-year plans funded by Korea Meteorological Administration research and development programs from 2011 to 2015.

Keywords: text-mining, term-document matrix, unstructured data, meteorological data

1. 서론

수치형, 범주형 자료와 같은 정형 자료의 분석방법은 전통적으로 주요 관심대상이었고, 지금까지도 그 기술은 나날히 발전을 하고 있다. 하지만 최근들어서는 정형 자료 이외에도 비정형 자료에 대한 관심이 높아지고 있다. 비정형 자료는 고정된 필드에 저장되어 있지 않은 자료로서 텍스트, 이미지, 동영상 자료 등을 일컫는다 (Kang 등, 2012). 하지만 주목받은 시간이 얼마 안된 비정형 자료에 대한 분석은 아직 개척해야할 부분이 많이 남아있다.

최근 알파고와 이세돌의 바둑 대국으로 인공지능에 대한 관심이 더욱 높아지고 있다. 텍스트 마이닝이란 비정형 자료 중 텍스트에 대한 분석 방법으로 넓은 관점에서 보면 인공지능과도 관련이 있는 영역이다 (Zhang, 2007). 텍스트 마이닝의 주 기능이라고 말할 수 있는 것으로는 텍스트 자료에 대한 기초통계량 작성과 그것을 이용한 추세분석, 스코어링, 군집화 그리고 분류화 등 이라고 할 수 있다 (Srivastva와 Sahami, 2009). 예를 들면, ETS의 프로그램인 e-rater는 어떤 에세이에 대해서 텍스트 마이닝 알고리즘을 이용한 분석을 통해, 그 에세이의 문체(style)와 같은 특성을 분류하고, 그 에세이를 얼마나 잘 썼는지에 대하여 스코어링을 할 수 있다 (Attali와 Burstein, 2006).

비정형 자료에 대한 분석은 정형 자료의 분석과 시너지 효과를 발휘해 좀 더 다양한 부분에 대해서도 분

This work was funded by the Korea Meteorological Administration Research and Development Program under Grant KMIPA 2015-1110.
[1]Corresponding author: Department of Statistics, Dongguk University, 30 Pildong-ro 1-gil, Jung-gu, Seoul 04620, Korea. E-mail: yung@dongguk.edu

Table 2.1. Number of R&D projects by year and business

연도	2011	2012	2013	2014	2015	Total
기상기술개발사업	0	9	4	0	9	22
지진기술개발사업	20	24	4	7	18	73
기후기술개발사업	9	13	6	1	15	44
기상산업지원 및 활용기술개발사업	19	10	7	18	9	63
합계	48	56	21	26	51	202

석을 해볼 수 있다. 예를 들어, Goo와 Kim (2014)은 한식 음식점에 대한 주관식 설문 조사 형태의 텍스트 자료와 함께 매출액 등의 각종 수치적 자료를 합쳐서 분석하여 음식점업 사업체의 현황에 대한 좀 더 다각적인 분석을 시도한 바 있다. 또한 비정형 자료에 대한 분석 방법은 수치적 분석 방법으로 하기 어려운 표현과 분석을 수행할 수 있다. 워드클라우드나 토픽네트워크와 같은 데이터 시각화 방법은 수치적으로는 나타내기 힘든 비정형 자료에 대한 연관성을 표현한 예이다. 인터넷에 있는 방대한 양의 트위터 자료를 분석하기 위하여 Jin 등 (2013)은 텍스트 마이닝 분석기법의 일종인 동시출현단어분석(co-word analysis)을 이용하여 토픽과 관련 키워드를 직관적으로 파악가능한 네트워크로 표현하였다. 텍스트 마이닝은 텍스트 자료에 대한 추세 분석에도 용이하다. Bae 등 (2013)은 논문 초록을 텍스트 자료로 하는 텍스트 마이닝 기법으로 기후관련 연구의 트렌드와 관심주제어 등을 파악하였다. 본 논문에서는 2011년부터 2015년까지 기상청 연구개발분야의 지정과제 공고 DB로부터 텍스트 자료를 수집하고 텍스트마이닝 기법을 적용하여 각 연구개발분야에서 자주 사용되는 단어와 증가추세 및 감소추세에 있는 단어를 파악할 것이다. 이렇게 각 분야에서 자주 등장하는 단어들의 빈도분석을 통해 해당 사업분야의 키워드를 파악하고, 증가추세와 감소추세에 있는 단어들의 분석을 통해 해당사업분야에서 시간이 지남에 따라 주목을 받고 있는 단어들과 점차 관심 밖으로 멀어지고 있는 단어를 확인할 수 있을 것이다. 본 연구는 이러한 텍스트 자료분석을 통하여 실제 사업분야의 동향에 대한 설명과 예측이 가능한지에 관한 탐색적 성격을 가진다.

본 논문의 구성은 다음과 같다. 2절에서는 자료의 수집과정에 대해서 설명하고, 3절에서는 본 논문에서 제안된 자료 분석 방법에 대해서 소개한다. 4절에서는 제안된 텍스트 자료의 분석 방법에 따른 결과에 대해서 설명한다. 마지막으로 5절에서는 결론을 맺는다.

2. 자료의 수집

기상청 연구관리 시스템(http://rnd.kma.go.kr)에 있는 한국 기상 산업진흥원 See-At 기술 개발사업의 지정과제 중 연구기간이 2011년부터 2015년이 포함된 과제 202건에 대한 DB메타 정보 및 공시된 request for proposal(RFP) 내용에 대한 텍스트 자료를 수집하였다. 지정과제의 수는 202건이지만 각 사업이 1년부터 3년까지 진행되기 때문에 2011년부터 2017년까지 총 398개의 텍스트 자료가 수집되었다. 연도별 기상, 지진, 기후, 기상산업 지원 및 활용기술개발사업에 대한 과제 수는 Table 2.1과 같다.

각 연구의 기간별로 연구기간이 1년인 연구는 1개의 텍스트자료, 2년인 연구는 2개의 텍스트 자료, 3년인 연구는 3건의 텍스트 자료로 수집되었다. 예를 들어, 연구기간 3년인 과제가 2015년에 시작하는 경우에는, 1년차 텍스트 자료는 2015년의 자료로, 2년차 텍스트 자료는 2016년의 자료로, 3년차의 자료는 2017년의 자료로 계산된다. 각 단계에서 분석에 필요하지 않은 단어들은 배제 되었으며 사업 명칭에 포함되는 단어 또한 제외하였다. 따라서 총 398개의 문서에서 70,484개의 용어를 수집하였다. 필요하지 않은 용어를 배제하는 방법으로는 실제 분석에서 잡음으로 작용을 하는 전체용어의 0.88%를 구성하고 있는 출현빈도수가 1회 이하인 단어들을 배제하는 방식을 사용하였다. 그리고 한글과 영어를 제외한

Table 2.2. Frequency of top 10 terms by year and business

사업명	Term	2011	2012	2013	2014	2015	2016	2017	Total
기상기술개발사업	활용	.	39	43	31	36	35	32	216
	분석	.	43	35	25	40	34	31	208
	모델	.	16	36	24	29	26	30	161
	자료	.	27	29	15	36	28	21	156
	예측	.	14	42	35	22	14	20	147
	해양	.	35	36	34	11	13	13	142
	관측자료	.	22	18	20	24	28	25	137
	관측	.	13	18	20	34	30	21	136
	특성	.	26	23	15	21	16	19	120
	발생	.	23	28	22	16	15	14	118
지진기술개발사업	분석	20	49	44	44	100	94	65	416
	발생	15	21	15	25	123	99	62	360
	연구	30	57	35	46	70	62	45	345
	자료	8	23	21	32	77	71	63	295
	활용	17	32	20	37	70	61	40	277
	필요	29	39	17	27	61	56	39	268
	화산	48	29	39	37	34	34	28	249
	지진관측	19	25	22	13	64	42	38	223
	조기	9	43	24	28	43	43	29	219
	구축	18	30	19	20	47	41	34	209
기후기술개발사업	변화	42	74	94	100	116	104	92	622
	예측	14	40	59	57	74	71	60	375
	분석	48	39	37	36	70	77	38	345
	활용	31	25	45	36	89	69	45	340
	정보	34	29	35	36	60	57	51	302
	자료	18	44	57	47	51	40	36	293
	모델	14	23	44	34	36	38	31	220
	생산	15	24	16	18	38	41	38	190
	기반	11	26	20	22	36	42	30	187
기상산업지원 및 활용기술개발사업	기상정보	29	24	28	28	51	26	11	197
	관측	17	17	17	45	53	36	2	187
	서비스	33	12	29	48	34	11	0	167
	기후	19	16	21	17	29	23	4	129
	국내	24	10	9	32	31	16	2	124
	시스템	9	24	26	23	24	13	5	124
	측정	3	19	22	33	30	14	0	121
	예측	5	2	4	14	45	40	9	119
	도로	1	9	9	3	50	44	0	116
	분석	8	5	7	24	45	22	2	113

특수문자나 숫자는 전처리 과정에서 제거되었다. R에 내장된 패키지를 사용하게 되었을 때 한글 처리를 정상적으로 수행하기 위해서 수정 및 디버깅을 실시하였다.

용어가 문서에서 출현했는지의 여부보다 정확한 빈도수에 관심이 있었으므로, 용어-문서행렬을 만들 때, 한 문서에서 해당 용어가 출현 했는지만을 확인하는 0과 1이 아닌 정확한 빈도수를 입력하여 행렬을 구성하였다. 각 사업별 상위 10개 빈도 수 단어에 대한 행렬은 Table 2.2에서 볼 수 있다.

Figure 3.1. Data preprocessing procedure.

3. 자료의 분석 방법

데이터베이스화 된 자료를 R 프로그램의 KoNLP패키지 (Jeon, 2013)와 tm패키지 (Feinerer 등, 2008)를 이용하여 용어-문서행렬을 구성하였다. KoNLP패키지는 입력된 한글 문서 벡터를 형태소 단위로 인식하고 적절한 품사를 부여해주는데 사용되며, tm패키지는 문서 벡터를 Corpus로 구성하고 용어-문서 행렬을 만들거나 Corpus를 이용한 텍스트 전처리를 수행하는데 사용된다. KoNLP패키지의 함수를 다루는 법은 KoNLP 사용설명서 (Jeon, 2015)를 참고할 수 있고, tm패키지에서 필요한 함수들을 다루는 방법은 tm 사용설명서 (Feinerer, 2013)를 참고할 수 있다. 이러한 패키지들을 이용하여 각 사업 분야별로 용어-문서행렬을 만드는 전처리 과정은 Figure 3.1과 같다.

실제 분석에서는 절대도수가 아닌 상대도수를 이용하여 분석하였다. 그 이유는 각 사업과 연도별로 문서 수와 총 단어 수가 다르기 때문에 단순히 절대도수로 분석을 하는 것은 문서내에 그 단어가 얼마나 자주 쓰였는지를 표현하는데 한계가 있기 때문이다. h사업의 y년도 j번째 단어의 상대도수 $x_{h,y,j}$를 계산하는 방식은 다음과 같다.

$$x_{h,y,j} = \frac{\sum_{i=1}^{n_h} F_{h,y,i,j}}{T_{h,y}},$$

여기서 $T_{h,y}$는 y년도의 h사업의 총 단어의 수, $F_{h,y,i,j}$는 y년도 h사업의 i번째 문서에 나타난 j번째 용어의 빈도수, 그리고 n_h는 h사업에 포함되는 총 문서의 수를 의미한다.

이렇게 계산된 상대도수를 이용하여 연도별 상대도수 행렬을 생성하였다. 그리고 각 사업별로 상대도수

Table 4.1. Terms with high frequency, and upward or downward patterns over time

분류	사업명칭	단어
상대도수가 높은 단어	기상기술개발사업	모델, 예측, 해양, 분석, 자료, 활용
	지진기술개발사업	발생, 연구, 필요, 분석, 자료, 활용
	기후기술개발사업	변화, 예측, 정보, 분석, 자료, 활용
	기상산업 지원 및 활용기술개발사업	관측, 기상정보, 서비스, 국내, 기후, 시스템
증가추세에 있는 단어	기상기술개발사업	관측자료, 구름, 미세, 빅데이터
	지진기술개발사업	기준, 활동, 백두산, 자료
	기후기술개발사업	가뭄, 예측, 우리나라
	기상산업지원 및 활용기술개발사업	검증, 관련, 실시간, 연계
감소추세에 있는 단어	기상기술개발사업	동해, 방사능, 해역
	지진기술개발사업	재난, 전조, 화산재, 확산
	기후기술개발사업	관리, 몬순, 수문기상, 수자원
	기상산업지원 및 활용기술개발사업	기상장비, 서비스, 선진국, 수입

가 높은 6개의 단어에 대해 연도별 변화를 관찰하였다. 그리고 연도를 설명변수로 하고 상대도수를 반응변수로 하여 단순회귀분석을 실시하여 회귀계수가 0보다 큰 단어는 증가추세에 있는 단어로 0보다 작은 단어는 감소추세에 있는 단어로 분류하였고, 그 중 회귀선의 기울기의 절댓값이 큰 단어들에 대해서 집중적으로 살펴보았다.

4. 자료의 분석 결과

4개의 사업분야를 구분하여 분석을 실시하였다. 총 202개의 사업에 대한 398개의 문서에 있어서 각 사업분야 별 문서 수는 기상기술개발사업 60개, 지진기술개발사업 128개, 기후기술개발사업 106개, 그리고 기상산업지원 및 활용기술개발사업 104개이다. 각 사업별로 상대도수가 높은 단어, 증가추세에 있는 단어 그리고 감소추세에 있는 단어는 Table 4.1과 같다. Figure 4.1은 시각적으로 각 사업별 주요단어를 한 눈에 알아보기 편하게 하기 위해 생성한 워드클라우드이다.

사업별 빈도가 높은 단어의 연도별 상대도수 그래프는 Figure 4.2에서 볼 수 있다. 공통적으로 파악할 수 있는 것은 사업별 명칭에 들어가는 단어들은 대부분 상대도수가 높다는 점이었다. 하지만 위에서 언급한대로 의미있는 분석을 위해서 사업명칭에 포함이 되는 단어는 불용어로 분석에서 제외하였다. 분석, 자료, 활용은 기상산업지원 및 활용기술개발사업분야를 제외한 3개의 사업에서 높은 상대도수를 갖는 것을 확인 할 수 있다. 활용의 경우는 사실 4가지 사업분야에서 모두 높은 상대도수를 기록하였지만 기상산업지원 및 활용기술개발사업분야에서는 사업명칭에 포함되는 단어라서 분석에서 제외되었다. 기상산업지원 및 활용기술개발사업분야를 제외한 3개의 사업에서 분석과 자료라는 단어가 높은 상대도수를 가진다는 것은 다른 3개의 사업분야와는 달리 기상산업지원 및 활용기술개발사업분야는 자료를 바탕으로 분석을 하는 이론적인 부분보다 다른 부분에 중점을 둔다는 것을 짐작해 볼 수 있다.

기상기술개발사업에서 사업 명칭에 포함되는 단어를 제외한 단어 중 상대도수가 높게 나온 단어로는 모델, 예측, 해양, 분석, 자료, 활용이 있었다. 해양을 제외한 모델, 예측, 분석, 자료, 활용은 모델이라는 단어와 잘 어울리는 단어로 볼 수 있는데, 이 점으로 미루어 보아 기상기술개발사업에서는 모델이 차지하고 비중이 높다는 것을 짐작해 볼 수 있다. 지진기술개발사업에서 사업명칭을 제외한 단어로는 발생, 연구, 필요, 분석, 자료, 활용이 상대도수가 높게 나왔다. 이러한 단어들은 아직 우리나라에 지진이 거의 일어나지 않았음을 암시해 준다고 생각해 볼 수도 있다. 만약 우리나라가 실질적으로 지진이 발생하는 나라라고 했다면, 진도나 피해와 같은 지진에 대한 구체적인 단어가 좀 더 많이 나왔을 것이다. 기후

Figure 4.1. Word cloud of terms by business.

기술개발사업에서 많이 나온 단어는 변화, 예측, 분석, 활용, 정보, 자료이다. 변화나 정보와 같은 단어들은 지진기술개발사업에서는 자주 출현하지 않은 단어로 지진기술개발사업에서 나온 단어들 보다는 좀 더 우리 생활에 맞닿아 있는 단어라고 생각해 볼 수 있다. 우리나라에서는 지진이 자주 일어나지 않아서 기후기술개발사업에서 다루는 기후라는 주제가 지진기술개발사업에서 다루는 지진이라는 주제보다 좀 더 현실에서 직접적으로 느낄 수 있는 요소가 많기 때문에 기후기술개발사업에서 좀 더 현장감이 있는 단어를 사용한다고 추측해 볼 수 있다. 마지막으로 기상산업지원 및 활용기술개발사업 분야에 있어서는 관측, 기상정보, 서비스, 국내, 기후, 시스템이 상대도수가 높게 나온 단어이다. 사업명칭을 보면 기상정보나 서비스같은 단어의 경우에 기상산업지원 및 활용기술개발사업분야 이외에 다른 사업분야에서는 등장하기 힘든 단어로 사업의 특징을 잘 보여준다고 할 수 있다.

그 다음으로는 회귀선의 기울기가 양수인 증가추세에 있는 단어를 살펴보았다. 상대적으로 큰 값의 회귀계수를 가지고 있는 단어들 중 기울기에 대한 유의성을 나타내는 p-값이 0.05보다 낮은 단어들에 대해

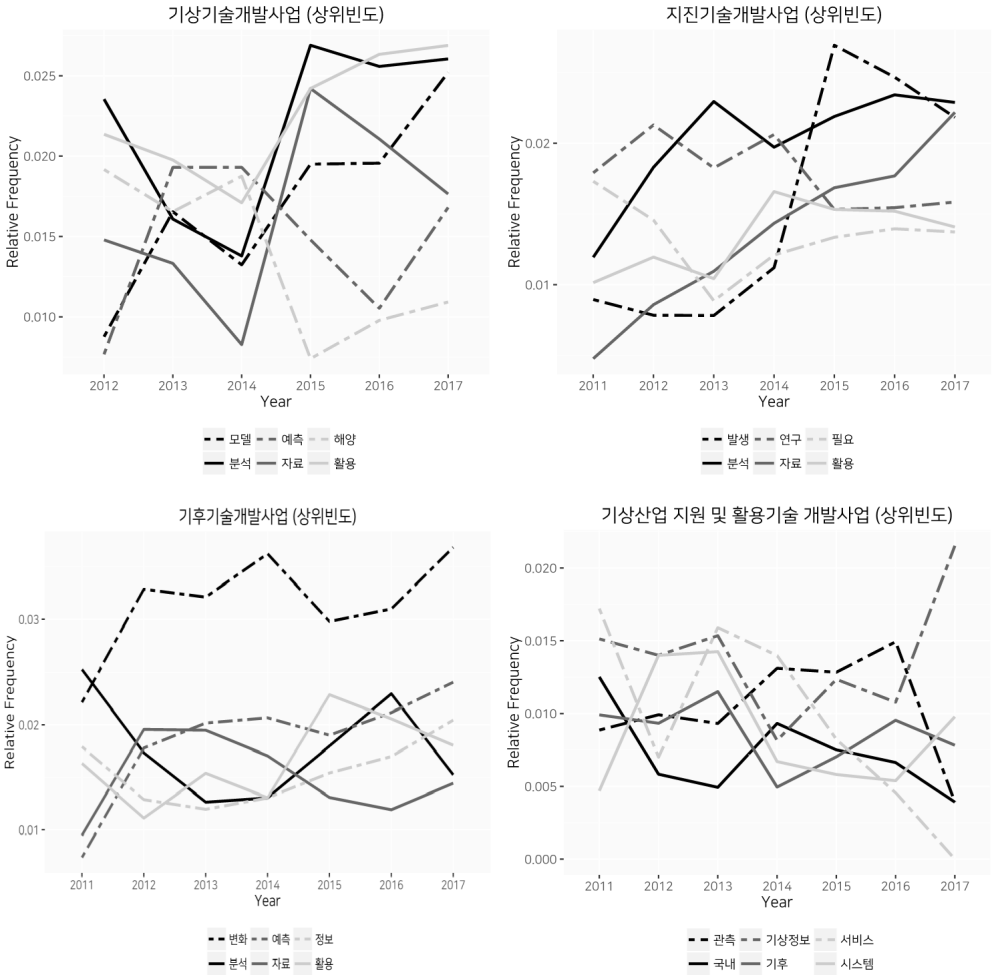

Figure 4.2. Time series plots of terms with high frequency by business.

서 살펴보았다. 이런 조건을 만족시키는 단어가 시간에 따른 유의미한 증가추세를 가진다고 해석할 수 있기 때문이다. 회귀선의 기울기가 양수이면서 p-값이 0.05이하인 단어의 비중은 사업별로, 기상기술개발사업 7.31%, 지진기술개발사업 11.56%, 기후기술개발사업 7.18%, 기상산업지원 및 활용기술개발사업 2.62%이다. 이러한 조건을 만족시키는 단어들 중 해석에 있어서 의미가 있어보이는 3–4개의 단어를 사업별로 선별하였다. 예외적으로 지진기술개발사업의 백두산의 경우는 p-값이 0.05보다 작지 않지만 분석대상에 추가하였다. 여기서 선별한 단어들의 그래프는 Figure 4.3에서 회귀선의 기울기의 값과 p-값은 Table 4.2에서 볼 수 있다.

기상기술개발사업에서 증가추세에 있는 단어는 관측자료, 구름, 미세, 빅데이터로 총 4개의 단어가 있었다. 그리고 지진기술개발사업에서 유의한 증가추세에 있는 단어들은 기준, 백두산, 자료, 활동으로 총 4개의 단어가 있었다. 다음으로 기후기술개발 사업에서 유의한 증가추세에 있는 단어로는 가뭄, 예측, 우리나라로 총 3개의 단어가 있었다. 마지막으로 기상산업 지원 및 활용기술개발 사업에 있어서 유의한

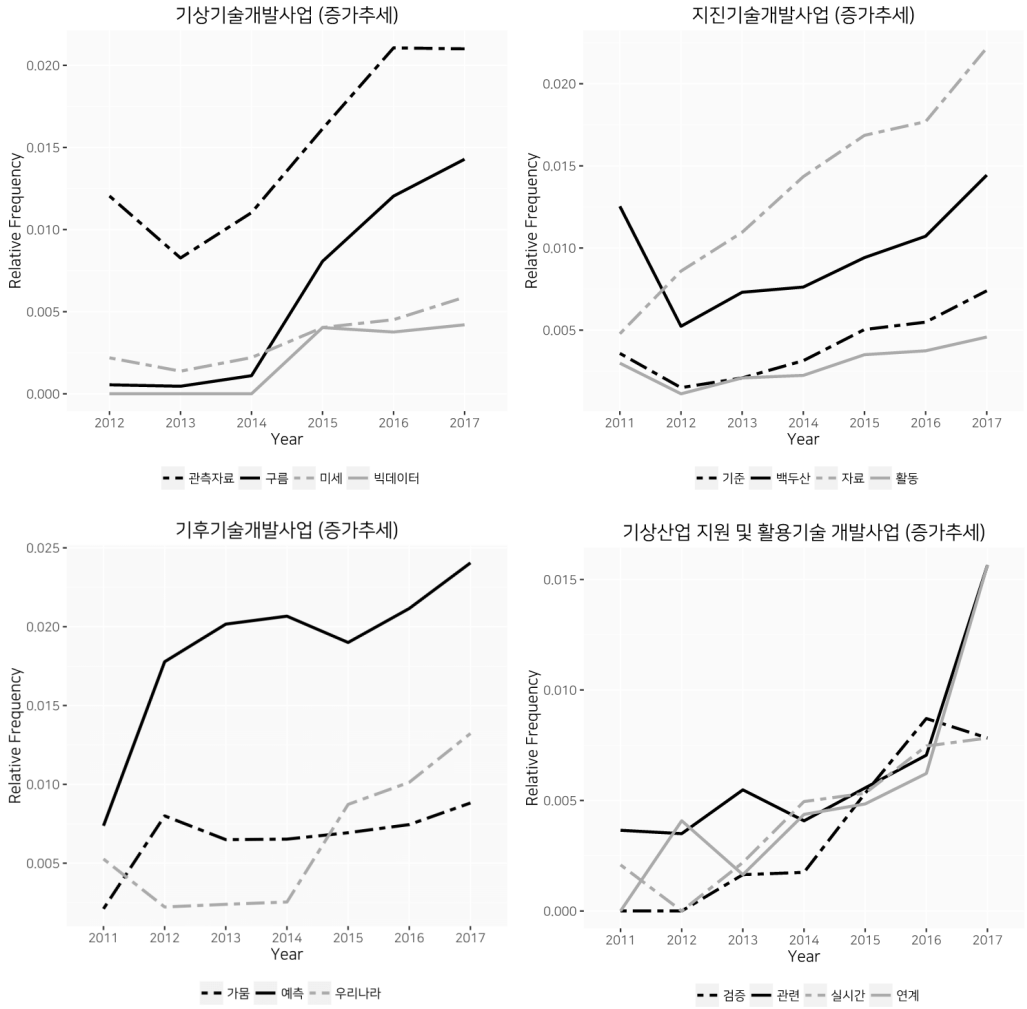

Figure 4.3. Time series plots of terms with upward patterns over time by business.

증가추세에 있는 단어는 총 4개로 검증, 관련, 실시간, 연계가 있었다.

여기서 흥미롭게 볼 수 있는 부분은 증가추세 있는 단어 중 일부분은 사회적으로 이슈가 되었던, 혹은 지금까지도 이슈가 되고 있는 문제들이라는 것이다. 기상기술개발사업의 미세나 빅데이터 같은 단어의 경우에는 최근들어서 관심이 급증하고 있는 분야임을 알 수 있다. 특히 미세먼지를 뜻하는 미세의 경우는 올해 2016년의 기상관련 이슈 중 가장 큰 이슈가 되는 문제이다. 그런데 이는 2012년 이후 꾸준히 상승세를 가졌고, 2016년에 와서 큰 이슈가 된걸로 미루어 짐작해 보면 학계에서 다루는 빈도가 점차 증가하는 경우에는 실제로 언젠가 크게 사회적으로 이슈가 될 수 있는 가능성을 내포하고 있다고 해석할 수 있다.

백두산의 경우 p-값이 0.05보다 큼에도 불구하고 증가추세에 있는 단어로 선별하였다. 왜냐하면 2011년에서 2012년 사이에 빈도수가 급격하게 감소하였지만 2012년부터 2017년까지는 서서히 증가하는 추세

Table 4.2. Summary of regression results for terms with upward patterns over time by business

사업명	Term	Coefficient	p-값
기상기술개발사업	관측자료	2.52e−3	0.0104
	구름	3.15e−3	0.0020
	미세	8.48e−4	0.0043
	빅데이터	1.03e−3	0.0096
지진기술개발사업	기준	7.99e−3	0.0098
	백두산*	6.71e−4	0.1542
	자료	2.72e−3	0.0000
	활동	4.08e−4	0.0245
기후기술개발사업	가뭄	3.16e−3	0.0411
	예측	1.99e−3	0.0142
	우리나라	1.64e−3	0.0138
기상산업지원 및 활용기술개발사업	검증	1.59e−3	0.0008
	관련	1.54e−3	0.0189
	실시간	1.26e−3	0.0014
	연계	1.94e−3	0.0098

*는 단어 선별 기준을 충족시키지 않는 예외의 경우.

Table 4.3. Summary of regression results for terms with downward patterns over time by business

사업명	Term	Coefficient	p-값
기상기술개발사업	동해	−9.03e−4	0.0097
	방사능	−1.37e−3	0.0172
	해역	−1.33e−3	0.0101
지진기술개발사업	재난	−4.90e−4	0.0017
	전조	−4.66e−4	0.0153
	화산재	−9.38e−4	0.0058
	확산	−8.74e−4	0.0042
기후기술개발사업	관리	−6.90e−4	0.0140
	몬순	−6.19e−4	0.0118
	수문기상	−9.48e−4	0.0010
	수자원	−7.57e−4	0.0063
기상산업지원 및 활용기술개발사업	기상장비	−1.49e−3	0.0077
	서비스	−2.29e−3	0.0197
	선진국	−2.24e−4	0.0032
	수입	−5.38e−4	0.0073

를 가지는데 이것이 현상적으로 유의미한 해석이 가능하기 때문이다. 이것은 2011년에 일본 대지진과 쓰나미 사건으로 백두산 관련 이슈가 사회적으로 관심을 받았다가 2012년 이후에 기억속에 잊혀졌었지만, 학계에서는 그 이후로도 계속해서 관심을 발전시켜 나갔다는 것을 의미한다. 2016년 쿠마모토에서 대지진이 일어난 이후로 다시 백두산 관련 이슈가 사회적으로 관심을 받고 있다는 사실은 학계에서 지속적으로 관심을 증가시켰던 부분에 있어서는 크게 이슈가 될 가능성을 내포하고 있다는 주장할 수 있는 근거를 뒷받침 해준다고 볼 수 있다. 이러한 측면에서 기후기술개발사업의 가뭄의경우 가까운 미래에 가뭄이 사회적으로 크게 이슈가 될 가능성을 내포하고 있다고 조심스레 예측해 볼 수 있다.

그 다음으로는 회귀선의 기울기가 음수인 감소추세에 있는 단어를 살펴보았다. 비교적 절대값이 큰 회귀계수를 가지고 있는 단어들 중 기울기에 대한 유의성을 나타내는 p-값이 0.05보다 낮은 단어들에 대해

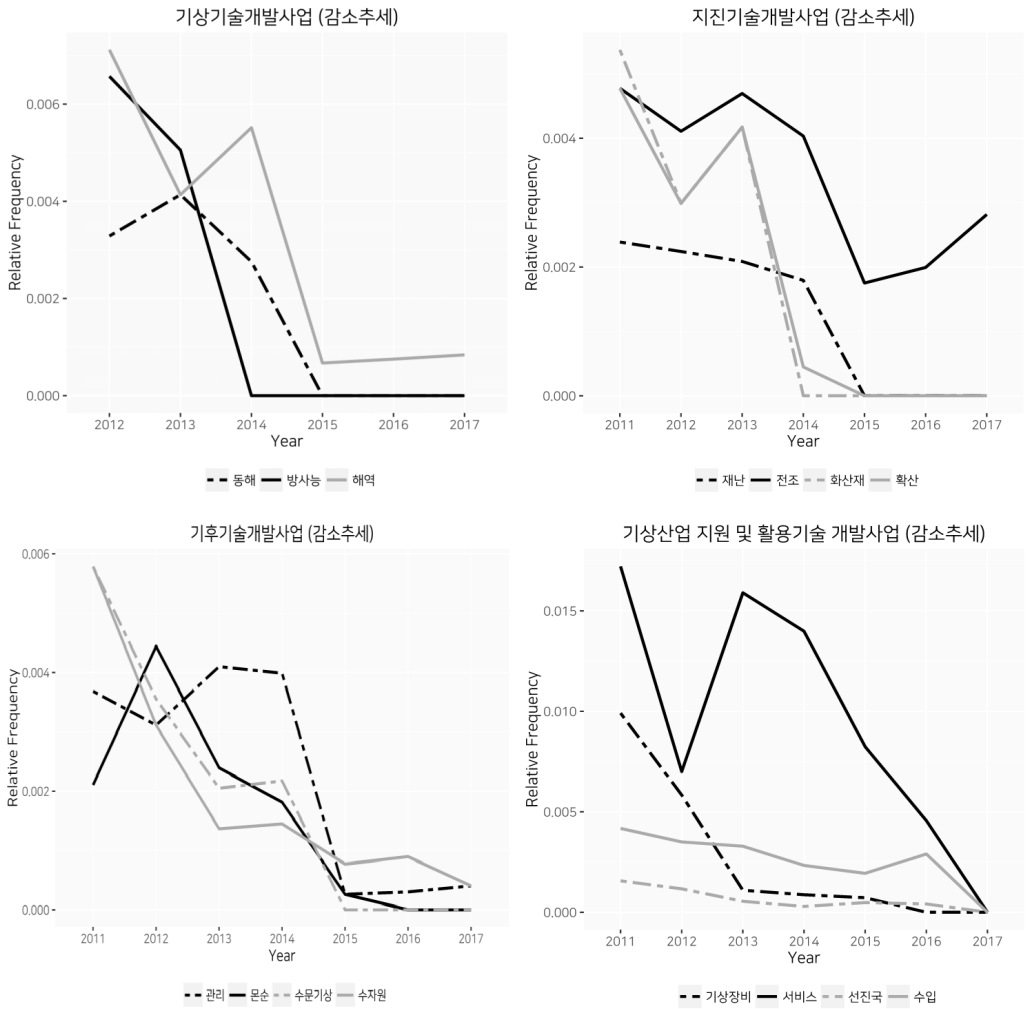

Figure 4.4. Time series plots of terms with downward patterns over time by business.

서 살펴보았다. 이런 조건을 만족시키는 단어가 시간에 따른 유의미한 감소추세를 가진다고 해석할 수 있기 때문이다. 회귀선의 기울기가 음수이면서 p-값이 0.05이하인 단어의 비중은 사업별로, 기상기술개발사업 1.88%, 지진기술개발사업 2.13%, 기후기술개발사업 1.73%, 기상산업지원 및 활용기술개발사업 2.14%이다. 이런 조건을 만족시키는 단어들 중 해석에 있어서 의미가 있어보이는 3–4개의 단어를 사업별로 선별하였다. 여기서 선별한 단어들의 그래프는 Figure 4.4에서 회귀선의 기울기의 값과 p-값은 Table 4.3에서 볼 수 있다.

먼저 기상기술개발사업에서 감소추세를 가지는 단어는 동해, 방사능, 해역으로 3개이다. 그리고 지진기술개발사업에서 유의한 감소추세가 있는 단어들은 재난, 전조, 화산재, 확산으로 총 4개의 단어가 있었다. 다음으로 기후기술개발 사업에서 유의한 감소추세에 있는 단어로는 관리, 몬순, 수문기상, 수자원으로 총 4개의 단어가 있었고, 마지막으로 기상산업 지원 및 활용기술개발 사업에 있어서 유의한 감소추세

를 가지는 단어는 총 4개로 기상장비, 서비스, 선진국, 수입이 있었다.

단어가 감소추세를 가지는 경우 중 하나는 몇 개의 연구에서 그 단어의 비중이 높여 사용하다가, 그 연구들의 연구기간이 종료된 경우 갑작스런 감소추세를 가지며 상대도수가 0이 되는 경우이다. 기상기술개발사업에서 동해, 방사능, 지진기술개발사업에서 재난, 화산재, 확산, 기후기술개발사업에서 수문기상 같은 단어들이 이러한 경우라고 추측해 볼 수 있다. 감소추세에 있는 단어도 역시 사회적 이슈의 영향을 받는 경우도 있다. 기후기술개발사업의 몬순의 경우가 그러하다. 몬순은 엘니뇨랑 관계가 깊은 단어인데 엘니뇨는 지금도 문제이기는 하지만 비교적으로 과거에 전 세계적으로 이슈였다. 하지만 최근에 많은 주목을 받지 못하고 우리의 기억에서 점차 사라져 가고 있는데 이것은 몬순의 단어가 감소추세를 갖는다는 사실과 일맥상통한다. 감소추세를 갖는 단어들 중 우리가 주의해야 할 부분은 실제로 우리 사회에 중요한 부분을 차지하고 있으면서도 연구에 대한 관심부족으로 그 분야에 대한 연구활동이 감소하는 경우를 제시해 준다는 점이다. 기상기술개발사업의 방사능의 경우 일본의 후쿠시마원전사고가 2011년에 일어난 후 관심을 받다가 최근부터는 연구에 큰 집중을 받지 못했다는 것을 추측해 볼 수 있다. 하지만 방사능의 경우 지금 사회적으로도 아직까지 큰 이슈이고 우리 삶에 치명적인 영향을 줄 수 있는 부분으로서 지속적인 관심이 필요하다는 사실을 추세분석으로 환기해 볼 수 있다.

5. 결론 및 토의

본 연구는 텍스트 마이닝 방법을 이용하여 기상청 연구개발분야의 사업별 주요단어와 추세에 관하여 분석하여 보았다. 이를 통해, 현재 각 사업별로 중요시 되는 주제에 대해서 알아볼 수 있었고, 점차 시간에 따라 상승하거나 감소하는 추세를 가진 단어들이 갖는 의미에 대해서도 생각해 볼 수 있었다. 본 연구의 중요성과 연구와 관련된 제언은 다음과 같이 정리할 수 있다. 첫째, 지속적으로 상대도수가 높은 단어들은 각 사업별로 어떤 부분에 대해서 중점을 두고 있는지 파악할 수 있게 해준다는 것이다. 그리고 둘째로는, 각 사업에서 다루는 빈도가 상승추세를 갖는 단어들은 아직까지는 사회적으로 이슈가 되지 않았더라도 이슈가 될 가능성을 내포하고 있을 수 있다는 것이다. 마지막으로 감소추세를 가지는 단어를 통해서 실제로 우리 사회에 있어 중요한 부분인데도 불구하고 우리의 관심에서 멀어진 경우를 파악할 수 있다는 점이다. 이러한 부분들을 미리 감지하여 대비한다면 연구의 방향성을 바로잡고 사회적인 충격을 좀 더 유연하게 대처할 수 있을 것이다. 향후 연구에서는 지정 과제 뿐만 아니라 자유 과제에 대해서도 분석 대상으로 하여, 좀 더 넓은 관점에서 사업별 추세와 패턴에 대해서 분석하여 사업의 청사진을 제시하는 방향으로 발전시켜 나갈 수 있을 것이다.

References

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2, *The Journal of Technology, Learning, and Assessment*, **4**, Available from: http://www.jtla.org.

Bae, K. Y., Park, J. H., Kim, J. S., and Lee, Y. S. (2013). Analysis of the abstracts of research article in food related to climate change using a text-mining algorithm, *Journal of the Korean Data and Information Science Society*, **24**, 1429–1437.

Feinerer, I. (2013). Introduction to the tm package text mining in R, http://CRAN.R-project.org/doc/Rnews/

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text mining infrastructure in R, *Journal of Statistical Software*, **25**, 1–54.

Goo, J. and Kim, K. (2014). Text mining for Korean: characteristics and application to 2011 Korean Economic Census Data, *Korean Journal of Applied Statistics*, **27**, 1207–1217.

Jeon, H. (2013). KoNLP: Korean NLP package, *R package version 0.76*, **8**.

Jeon, H. (2015). Package KoNLP, Available from: https://cran.r-project.org/web/packages/KoNLP/KoNLP.pdf.

Jin, S. A., Heo, G. E., Jeong, Y. K., and Song, M. (2013). Topic-network based topic shift detection on twitter, *Korea Society for Information Management*, **30**, 285–302.

Kang, M. M., Kim, S. R., and Park, S. M. (2012). Analysis and utilization of big data, *Korea Information Science Society review*, **30**, 25–32.

Srivastava, A. N. and Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*, CRC Press.

Zhang, B. T. (2007). Next-generation machine learning technologies, *Communications of the Korea Information Science Society*, **3**, 96–107.

텍스트 마이닝 알고리즘을 이용한 기상청 연구개발분야 과제의 추세 분석

박홍주[a] · 김하빈[b] · 박태영[a] · 이영섭[b,1]

[a]연세대학교 응용통계학과, [b]동국대학교 통계학과

(2016년 5월 30일 접수, 2016년 7월 14일 수정, 2016년 7월 19일 채택)

요 약

이 연구에서는 비정형 자료 분석 기법 중 하나인 텍스트 마이닝 기법으로 기상청 연구개발분야 과제의 동향에 대하여 분석하였다. 이를 위하여 용어사전을 구축하고, 전처리를 하여 용어-문서 행렬을 만들었다. 이것을 이용해 연도별 용어 빈도수를 측정하고, 자주 나타나는 단어들에 대해서는 상대도수의 변화에 대해서 관찰하였다. 그리고 회귀분석을 사용하여 증가추세와 감소추세를 가지는 용어들을 파악하였다. 이러한 분석으로 기상청 최근 연구개발 분야의 트렌드를 파악하였다. 이와 같은 연구는 향후 기상청 연구개발에 관한 기초 자료로 사용될 수 있으며, 연구개발의 방향성과 청사진을 제시하는데 이용될 수 있을 것이다.

주요용어: 텍스트 마이닝, 용어-문서 행렬, 비정형 자료, 기상 자료.

이 연구는 기상청 기상기술개발사업(KIMPA 2015-1110)의 지원으로 수행되었습니다.

[1]교신저자: (04620) 서울시 중구 필동로 1길 30, 동국대학교 통계학과. E-mail: yung@dongguk.edu