Journal of the Korean Data & Information Science Society
2013, **24**(6), 1429–1437

http://dx.doi.org/10.7465/jkdi.2013.24.6.1429
한국데이터정보과학회지

텍스트 마이닝 기법을 활용한 기후변화관련 식품분야 논문초록 분석[†]

배규용[1] · 박주현[2] · 김정선[3] · 이영섭[4]

[124]동국대학교 통계학과 · [3]한국보건사회연구원 보건정책연구본부

접수 2013년 10월 22일, 수정 2013년 11월 14일, 게재확정 2013년 11월 19일

요 약

빅 데이터 분석기법 중 비정형데이터 분석기법인 텍스트 마이닝 기법을 이용하여 기후변화 관련 식품분야 논문 초록에서 용어들의 출현빈도를 분석하였다. 이를 위하여 용어-문헌 행렬을 만들고, 용어들간의 비유사성 측도를 바탕으로 계층적 군집분석기법을 적용하여 문서들을 군집화하였다. 군집화된 문서들간의 상호 연관성과 군집별로 특정용어의 빈도를 파악하여 문서군집을 특정주제별로 분류하였다. 이러한 연구를 통하여 식품분야의 기후변화 관련 논문들의 추세와 관심주제어를 파악할 수 있었으며, 향후 기후변화 적응 및 대응 체계 로드맵 작성 시 연구 개발 기초 자료로 활용할 수 있을 것이다.

주요용어: 계층적 군집분석, 기후변화, 문서분류, 텍스트 마이닝.

1. 서 론

최근 들어 IT, 금융, 제조업 등 사회 많은 분야에서 컴퓨터의 발전과 데이터의 축적에 따라 빅 데이터 처리 및 분석 기술에 대한 관심이 증가하고 있다. 대부분의 빅 데이터 분석기법들은 통계학과 데이터마이닝 분야에서 이미 사용되던 기법들을 대용량의 데이터 처리에 맞도록 개선하여 적용하고 있다. 그러나 최근 소셜미디어 등 비정형 데이터의 증가로 인하여 기존의 정형데이터 분석 기법들을 비정형 데이터에 적용하는 기법들에 대한 연구가 활발히 진행되고 있으며, 객관적이고 실증적인 자료를 바탕으로 언어자원을 활용할 수 있는 방법론이 필요하다 (Choi와 Lee, 2011). 그 중 대표적인 기법이 텍스트 마이닝 기법이다.

텍스트 마이닝은 정형 또는 비정형 텍스트 데이터에서 자연어처리과정과 데이터마이닝 기법을 통하여 방대한 텍스트에서 의미 있는 정보를 추출해내고, 가공하는 것을 목적으로 하는 기술이다. 최근에 산업공학, 건축, 특허 등 다양한 분야에서 텍스트 마이닝 기법을 활용한 연구가 활발히 진행되어 왔다 (Cho와 Kim, 2012; Go 등, 2011; Kim과 Jeong, 2012).

한편, 환경분야에서는 전 세계적으로 지구온난화와 이상기상 발생 등 기후변화 현상이 현실화되고 있으며, 향후 상당한 기간 동안 전 지구적인 기후변화가 지속될 것으로 전망되고 있다. 유엔 산하 정부간 기후변화협의체 (IPCC)는「기후변화에 관한 제4차 보고서」(2007) 에서 지난 100년 (1906∼2005)간

[†] 본 연구는 농림축산식품부 생명산업기술개발사업 (과제번호:312028-2)에 의해 이루어진 것임.
[1] (100-715) 서울시 중구 필동로 1길 30, 동국대학교 통계학과, 석사과정.
[2] (100-715) 서울시 중구 필동로 1길 30, 동국대학교 통계학과, 교수.
[3] (122-705) 서울시 은평구 진흥로 235, 한국보건사회연구원 보건정책연구본부, 연구위원.
[4] 교신저자: (100-715) 서울시 중구 필동로 1길 30, 동국대학교 통계학과, 교수. E-mail yung@dongguk.edu

지구평균기온은 0.74℃ 상승하였고, 지금과 같이 인류가 화석연료에 의존하는 생활을 계속하면 21세기 말(2090∼2099)에는 지구의 평균기온이 최대 6.4℃ 추가 상승하고, 해수면은 59cm 상승할 것으로 경고하고 있다 (IPCC, 2007). 우리나라의 경우 지난 100년 (1912∼2010) 동안 기온은 1.8℃ 상승하였고, 강수량은 200mm이상 증가하여 세계 평균보다 더 빠르게 기후가 변화되고 있는 것으로 제시하고 있으며, 2020년에는 지난 40년 (1970∼2010) 평균치대비 1.8℃ 상승, 2050년에는 3.7℃상승할 것으로 전망하고 있다 (Baek 등, 2011). 따라서 현재 기후변화에 대한 연구가 각 분야에서 활발하게 진행되고 있으며, 관련 연구논문들의 발표가 증가하고 있다. 본 연구는 텍스트 마이닝 기법을 이용하여 2000년 초반부터 최근까지 기후변화 관련 논문 중 식품관련 연구성과물들에 대해 어떤 주제나 키워드가 연구되었는지를 알아보고, 이러한 주제나 키워드의 출현빈도 추세에 대해서 분석하였다. 이러한 분석을 통하여 기후변화 관련 연구성과들의 추세를 파악하고 향후 기후변화 적응 및 대응 체계 로드맵 작성 시 연구 개발 기초자료로 활용할 수 있을 것이다.

2. 자료의 수집

기후변화 (climate change)와 관련성이 높은 식품분야 논문을 수집하기 위해 www.sciencedirect.com에서 “climate change”과 “food” 두 키워드를 사용하여 2004년도부터 2012년까지 외국 학술지에 게재된 총 4500개의 영문 논문들에서 제목과 초록을 수집하였다. 이 4,500개의 논문들 중 162개의 논문은 초록이 제공되지 않아서 분석에서 제외되었으며, 나머지 4,338개의 논문 초록에서 사용된 총 용어의 수는 62,667개이었다. Figure 3.1에서 보여지는 바와 같이 각 단계별로 분석에 필요하지 않는 용어들을 제거하였고, 총 4338개의 문서와 20034개의 용어들을 얻었다. 하지만, 전체 문서수를 고려하였을 때 빈도수가 상대적으로 지나치게 낮은 용어들, 예를 들면 전체 문서에서의 출현 빈도가 26번 이하 (또는 전체문서 출현 상대빈도가 0.6%이하)인 용어들이 전체 용어의 90%를 구성하고 있고, 이러한 용어들은 대부분의 경우 잡음 (noise)으로 작용을 하여 용어-문헌 행렬에 의미 있는 소수의 잠재적인 요인들 (latent component)을 이끌어 내는데 한계점이 있다. 따라서, 본 연구를 위해 이 분야의 전문가적인 지식 (expert knowledge)을 반영해서 전체 문서에서 출현빈도가 26 이상인 용어들 중 그 용어의 중요성을 바탕으로 선택된 68개만을 고려한 용어-문헌 행렬 (term-document matrix)을 사용하였다. 최종적으로 선택된 68개의 용어들과 년도별 출현 빈도수 및 총 빈도수는 Table 2.1에 나열되어 있다. 이러한 방법은 전문가적인 지식을 반영할 뿐 아니라 용어들을 직접적으로 사용하기 때문에 그 결과를 해석하는데 용이하다는 장점을 가지고 있다. 이러한 단계를 걸쳐서, 최종적인 용어-문헌 행렬에는 총 4,338개의 문서와 68개의 용어로 구성되어 있다.

분석에 앞서 각각의 문서에서 용어의 빈도수를 측정하고 그 빈도수를 전체 문서 중 몇 개의 문서에서 출현하는지 여부를 보정하는 방법이 적용되는데, 일반적으로 논문 초록은 전체 논문에서 가장 핵심적인 내용들을 간략한 형태로 제시하기 때문에 중요한 용어가 여러 번 나타나는 경우가 드물다. 따라서 초록에서 사용된 용어들은 각각의 문서에서 각 용어의 빈도수가 아니라 각각의 문서에 출현했는지 여부만을 반영하는 형태로 자료를 구성하여 분석에 사용하였다. Table 2.1에서와 같이 특정 용어의 연도별 추세를 보기 위해서 각각의 자료들을 1년 단위로 용어의 총 출현 빈도를 이용해서 분석을 하였다.

Table 2.1 Frequencies of 68 selected terms by publication year

term	2004	2005	2006	2007	2008	2009	2010	2011	2012	total
production	198	196	198	204	179	201	202	216	208	1,802
model	140	177	198	205	148	189	165	194	195	1,611
water	149	142	129	127	136	109	125	136	141	1,194
environmental	223	223	167	216	205	183	156	194	160	1727
increase	188	216	204	225	217	249	244	245	267	2,055
effects	151	169	158	194	183	154	224	194	233	1,660
temperature	139	102	99	125	134	104	141	151	161	1,156
global	99	95	156	156	150	165	187	183	187	1,378
impacts	111	113	130	153	150	170	197	194	207	1,425
level	129	131	127	129	129	138	128	144	152	1,207
related	173	139	171	160	172	153	141	150	160	1,419
systems	117	108	106	129	108	127	149	138	155	1,137
conditions	102	177	123	157	112	106	96	97	159	1,129
agricultural	68	79	85	92	75	114	93	124	112	842
management	83	81	107	102	78	116	159	115	114	955
adaptation	46	71	72	57	78	94	113	114	139	784
analysis	146	139	141	122	137	127	132	144	150	1,238
potential	108	123	117	132	118	125	149	140	160	1172
significant	135	138	140	121	147	125	139	114	135	1,194
growth	144	102	75	107	86	95	97	100	97	903
policies	54	60	71	73	81	107	104	112	105	767
plant	99	81	84	76	86	73	83	83	88	753
regions	127	121	154	169	150	134	152	142	160	1,309
scenarios	37	40	63	73	60	75	76	88	92	604
population	81	72	81	87	89	76	87	98	77	748
ecosystem	84	80	97	112	94	87	88	75	88	805
indicate	127	128	120	132	129	118	124	106	104	1,088
reduction	43	85	113	137	123	59	58	62	155	835
scale	96	78	103	85	78	69	97	105	108	819
economic	60	70	82	78	73	95	104	103	110	775
organic	72	87	111	89	108	63	72	53	70	725
processes	92	84	123	83	98	96	93	93	80	842
assessment	80	110	99	97	109	106	101	106	119	927
fartors	103	101	100	98	96	82	99	102	73	854
approach	74	58	76	65	82	93	91	93	101	733
current	69	82	91	94	98	100	125	110	103	872
research	58	72	56	83	86	77	95	91	101	719
decrease	74	86	79	92	79	65	76	76	82	709
response	87	97	86	101	101	97	121	112	107	909
predicted	58	67	54	74	65	59	88	78	85	628
resources	77	60	56	77	62	66	70	93	95	656
haman	60	66	82	78	91	54	70	63	80	644
seasonal	88	61	70	69	81	62	67	51	56	605
local	79	57	72	67	71	61	79	80	83	649
influence	84	81	83	91	87	87	76	76	75	740
risk	41	37	36	46	51	59	71	63	83	487
countries	37	34	62	60	47	60	61	81	65	507
sustainable	53	45	63	56	62	88	62	77	63	569
distribution	57	67	76	57	78	53	69	68	69	594
patterns	71	80	86	71	89	72	76	59	60	664
natural	73	86	62	77	72	78	73	62	76	659
ecological	62	58	69	63	62	59	64	67	59	563
strategies	48	50	52	50	58	74	75	76	80	563
interactions	64	54	75	59	85	45	67	76	81	606
abundance	61	54	63	44	61	39	38	32	38	430
consumption	47	45	46	50	49	54	53	63	49	456
limited	66	63	65	70	85	75	68	81	68	641
structure	65	59	64	66	66	62	49	49	56	536
concentrations	67	72	58	53	52	50	54	51	48	505
spatial	54	58	67	49	58	59	53	56	53	507
information	64	48	66	62	68	57	86	56	77	584
trends	49	62	53	59	69	42	56	51	58	499
dynamics	51	56	60	56	72	51	68	53	55	522
annual	53	68	53	60	48	54	50	48	57	491
alternative	45	48	62	70	54	62	73	61	62	537
characteristics	73	88	69	70	65	42	60	46	47	560
mitigation	16	24	26	52	40	54	59	73	66	410
integrated	63	48	62	53	61	50	55	68	60	520

3. 자료 분석 방법

3.1. 용어-문헌 행렬 생성

수집된 논문 초록을 이용하여 텍스트 마이닝에 필요한 용어-문헌 행렬 (term-document matrix)을 구성하기 위해 R 프로그램 3.0.1 버전의 tm 패키지 (Feinerer 등, 2008)를 사용하였다. 이 패키지는 텍스트 마이닝에 필요한 함수들을 포함한 패키지로써, Figure 3.1의 Step 2-4에서 불필요한 용어 (stop word)를 제거하거나 같은 어근을 갖는 용어를 찾는 과정 (stemming)을 수행하는데 이용하였다. 각 단계 별 필요한 함수들의 이용방법은 패키지 사용설명서 (Feinerer, 2013)에 자세히 정리되어 있다.

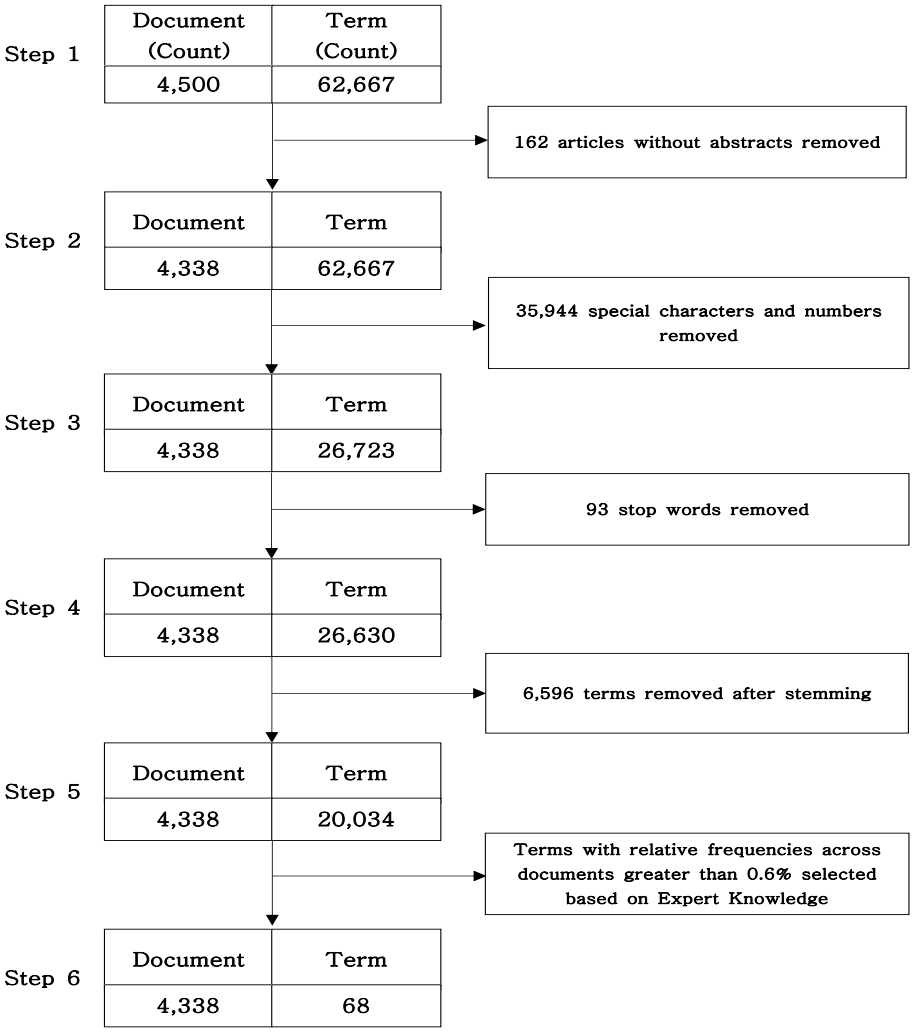

Figure 3.1 Flow of establishing a term-document matrix

3.2. 용어들 간의 계층적 군집분석

군집분석에 대한 많은 연구가 진행되고 있으며 (Lim과 Lim, 2012; Yeo, 2011), 본 연구에서는 특정 주제에 관련된 용어들이 어떻게 군집을 이루는 지를 확인하기 위해 계층적 군집분석 방법 (hierarchical clustering)을 사용하였다. 계층적 군집방법은 자율 군집 (unsupervised clustering)의 한 방법으로써 수집된 논문들을 분류하는데 알려져 있는 체계가 없는 상황에서 논문에 나타난 용어들의 비유사성 (dissimilarity)을 바탕으로 계층에 따라 어떻게 용어 들이 군집을 구성하는지를 보여준다. 계층적 군집분석 기법중 하나인 Ward 방법을 사용하여 용어들 간의 비유사성은 개별대상간의 거리로 측정을 하였으며, 군집의 수는 실루엣 (silhouette) 방법 (Rousseeuw, 1987)과 현장 전문가의 지식을 활용하여 집단 간의 유사성 (homogeneity)을 가장 크게 하는 군집의 수를 찾았다.

3.3. 년도별 군집의 상대 출현빈도

년도별로 논문에서 주로 다뤄지는 분야에 변화가 있었는지를 확인하기 위해 용어들 간의 계층적 군집분석의 결과로 얻는 군집들을 바탕으로 각 년도별로 전체 논문들에서 각 주제들 즉, 군집이 얼마나 많이 다루어 졌는지를 측정하기 위해 식 3.1과 같이 년도별 평균 상대 빈도 $\bar{X}_{y,h}$를 정의하였다.

$$\overline{x}_{y,h} = \sum_{i=1}^{N_y} \frac{\left(\sum\limits_{j \in S_h} F_{y,i,j}/n_h \right)}{N_y} \tag{3.1}$$

여기서 N_y는 y년도의 총 문서의 수, $F_{y,i,j}$는 y년도 i번째 문서에 나타난 j번째 용어의 빈도수, S_h는 h번째 군집에 포함되는 용어들의 지수 집합 (index set), 그리고 n_h는 h번째 군집에 포함되는 용어의 수를 의미한다.

3.4. 문서의 분류

특정 문서가 어느 분야의 논문인지를 분류하기 위해 용어들의 계층적 군집분석을 바탕으로 하는 주제별 상대 비중을 계산하였다. 주제별 상대 비중은 각각의 군집 (즉, 주제 분야)에 포함 된 용어들이 한 문서에 상대적으로 얼마나 자주 출현하는 지를 보여준다. 어떤 문서의 군집 h, $h = 1, 2, ..., H$에 대한 상대 비중은 식 3.2와 같이 표현된다.

$$Pr(C_i = h) = \frac{\sum\limits_{j \in S_h} F_{i,j}/n_h}{\sum\limits_{h=1}^{H} \left(\sum_{j \in S_h} F_{i,j}/n_h \right)} \tag{3.2}$$

여기서 C_i는 i번째 문서가 H개의 군집들 중 어떤 군집에 포함되는지를 보여주는 지시 변수, $F_{i,j}$는 i번째 문서에서 j번째 용어의 빈도수, S_h는 h번째 군집에 포함되는 용어들의 지수 집합, 그리고 n_h는 h번째 군집에 포함되는 용어의 수를 의미한다.

4. 자료 분석 결과

총 68개의 용어들 중에서 계층적 군집방법을 통해 Table 4.1과 같이 4개의 군집을 형성하였다. 각 군집에 포함되는 용어의 수는 계층적 군집방법에서 용어들 간의 비유사성을 바탕으로 군집을 형성하기 때문에 동일하지 않다. 여기서 '비유사성'의 의미는 임의의 두 용어가 함께 한 문서에 나타나면 유사성

이 높은 것이고 각각 다른 문서에 나타나면 비유사성이 높은 것이다. 따라서 같은 군집 안에 있는 용어들은 다른 군집에 있는 용어들에 비해서 한 문서에 함께 나타날 가능성이 크다는 것을 의미한다. 한편, 일반적으로 특정 주제에서 자주 언급되는 용어들은 한 논문의 초록에서 같이 나타날 가능성이 크므로, 각 군집에 포함되어 있는 핵심 용어들을 바탕으로 그 군집이 나타내는 주제를 추정할 수 있다. 이를 위해 Table 4.1에서와 같이 농식품 분야의 전문가의 의견을 바탕으로 각 군집에 주제어를 부여하였다. Figure 4.1에서는 기후변화와 관련하여 논문에서 어떤 주제들이 주로 다루어졌는지를 보기 위해 군집의 수 $H = 4$일 때 식 3.1을 이용하여 년도 별로 주제들의 상대 출현빈도를 살펴보았다. 주제가 "Assessment of climate-change prediction model"인 군집 2는 지난 9년 동안 전체 논문의 약 20%이상에서 꾸준하게 다뤄졌으며 그 비중이 매년 증가하는데, 2004년도에는 논문의 23%에서 언급되었지만 2012년도에는 전체 논문의 약 32%가 언급되어 평균적으로 매년 약 1%씩 증가하는 추세를 보여주고 있다. "Research and polices in preparation for climate change"가 주제인 군집 1 또한 지속적으로 증가하는 추세를 보여주고 있다. 이에 반해, "Spatial and temporal trends of climate-change effects"과 "Growth of human population and environment"가 주제인 군집 3과 4는 지난 9년 동안 큰 변화없이 전체의 13∼15%의 논문에서 다뤄졌음을 볼 수 있다.

Table 4.1 Clustering of 68 selected terms

cluster	term	Topic phrase
1	consumption, countries, policies, economic, reduction, mitigation, information, risk, research, adaptation, strategies, approach, integrated, systems, agricultural, management, resources, sustainable	Research and polices in preparation for climate change
2	level, significant, indicate, related, analysis, decrease, conditions, water, temperature, current, scenarios, assessment, increase, impacts, potential, model, effects, global, production, environmental	Assessment of climate-change prediction model
3	regions, local, patterns, scale, spatial, trends, seasonal, annual, distribution, abundance, factors, influence, characteristics, organic, concentrations	Spatial and Temporal trends of climate-change effects
4	ecosystem, ecological, interactions, dynamics, structure, processes, growth, population, plant, response, predicted, limited, alternative, human, natural	Growth of human population and environment

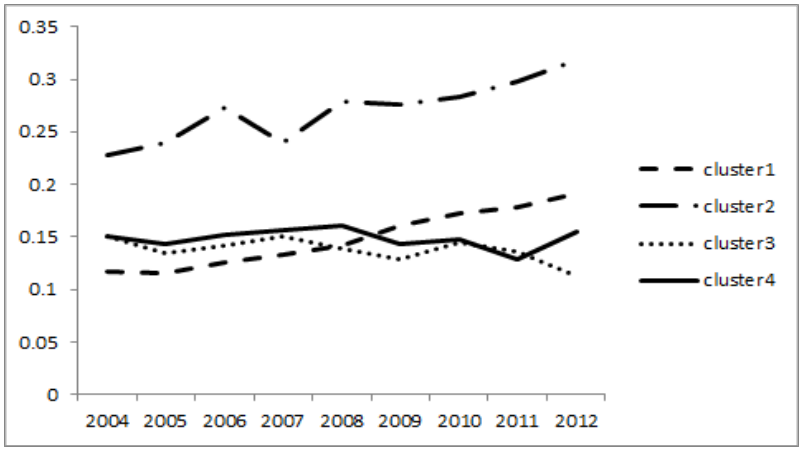

Figure 4.1 Trend of relative frequencies of clusters over past 9 years

Table 4.2는 용어들의 계층적 군집 결과를 바탕으로 군집의 수 $H = 4$일 때, 식 3.2를 적용하여 각 문서들이 어떤 주제들을 주로 다루었는지에 대해 처음 5개의 문서와 마지막 문서 5개의 결과를 보여 준다. 문서1은 군집 1부터 군집 4까지의 주제들을 모두 다루고 있지만, 주로 군집 1과 군집 2의 주제 (각각 38.0%와 29.3%)가 다른 두 군집보다는 조금 더 높은 비중을 가지고 언급되었음을 알 수 있다. 이에 반해 문서 4336은 주로 군집2의 주제를 주로 다루었고 (65.2%), 군집 1과 군집 3의 주제는 전혀 다루지 않았음을 알 수 있다. 각각의 문서를 주어진 4개의 군집 중 하나로 분류하고자 하는 경우, 상대 비중이 가장 큰 군집으로 지정하여 분류를 하였다. 예를 들어, Table 4.2에 보여진 것 같이 문서 1은 군집 1로 분류되지만, 문서 2와 문서 3의 경우는 군집 2에 포함된다.

Table 4.2 Classification of documents into 4 clusters

	Cluster1	Cluster2	Cluster3	Cluster4	selected cluster
Document1	0.380	0.293	0.196	0.130	Cluster 1
Document2	0.217	0.470	0.104	0.209	Cluster 2
Document3	0.308	0.415	0.185	0.092	Cluster 2
Document4	0.217	0.391	0.261	0.130	Cluster 2
Document5	0.308	0.138	0.369	0.185	Cluster 3
⋮	⋮	⋮	⋮	⋮	⋮
Document4334	0.577	0.115	0.000	0.308	Cluster 1
Document4335	0.294	0.706	0.000	0.000	Cluster 2
Document4336	0.000	0.652	0.000	0.348	Cluster 2
Document4337	0.250	0.450	0.150	0.150	Cluster 2
Document4338	0.000	0.462	0.385	0.154	Cluster 2

5. 결론 및 토의

본 연구는 기후 변화에 맞추어 영문 학술지에 게제된 논문을 중심으로 데이터 마이닝의 기법중 하나인 텍스트 마이닝을 적용하였다. 이를 통해, 지난 9년간 학술지에서 다뤄진 주제들의 특성을 파악하였고, 용어의 계층적 군집방법을 통해 찾아낸 4개의 주제어들에 각각의 문서를 분류하는 방법을 제시하였다. 이러한 결과물은 새로운 연구를 시작하기에 앞서 기존의 유사한 논문들을 찾아내는데 유용하게 사용될 수 있을 뿐 아니라, 특정 연구 주제가 최근 학술지에서 주로 다뤄지는지 여부를 확인 할 수 있다. 본 연구의 중요성과 연구와 관련된 제언은 다음과 같이 정리될 수 있다. 첫째, 최근 대두되는 주요 화제인 “기후 변화”에 관련된 연구 논문들에 대해 처음으로 텍스트 마이닝 방법을 적용하였다는 것이다. 하지만, 본 연구에서는 영문 학술지에 게재된 논문들만을 대상으로 하였기 때문에 영어가 아닌 다른 언어, 특히 한국어로 게재된 논문들이 포함되지 않았다는 제한이 있다. 이는 아직 국내에서 기후 변화와 식품에 관련된 충분한 양의 논문이 없기 때문이기도 하지만 같은 의미를 다른 두 언어들로 표현할 경우, 예를 들면, ‘climate’과 ‘기후’, 두 용어를 하나의 의미로 인식할 수 있는 소프트웨어의 부재이기도 하다. 따라서 보다 포괄적인 연구를 위해 이러한 문제를 해결할 수 있는 텍스트 마이닝 소프트웨어의 개발이 필요하다. 둘째, 본 연구는 전문가의 의견을 반영하여 각 년도 별로 관련성이 높은 100개의 논문들을 찾고 분석하였다. 이는 전문가의 의견을 반영하는 장점이 있으나 전문가에 따라 다른 논문 자료가 선택될 수 있다는 선택적 편의 (selection bias)가 발생할 수 있다. 향후 연구에서는 기후 변화에 대한 문헌 검색 프로그램이나 논문 데이터 베이스를 통해 관련성이 있는 모든 논문을 분석 대상으로 하여 이러한 전문가에 의한 편의를 최소한으로 줄이는 노력이 필요할 것이다.

References

Baek, H., Cho, C., Kwon, W., Kim, S., Cho, J. and Kim, Y. (2011). Development strategy for new climate change scenarios based on RCP. *Journal of Climate Change Research*, **2**, 55-68.

Cho, S. and Kim, S. (2012). Finding meaningful pattern of key words in IIE transactions using text mining. *Journal of the Korean Institute of Industrial Engineers*, **38**, 67-73.

Choi, K. and Lee, Y. (2011). The deduction of objective linguistic information using statistical methods - The grouping of the possibility of interdisciplinary research. *Journal of the Korean Data & Information Science Society*, **22**, 49-55.

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

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

Go, G., Jung, W., Shin, Y., Park, S. and Jang, D. (2011). A study on development of patent information retrieval using text mining, *Journal of the Korea Academia-Industrial Cooperation Society*, **12**, 3677-3688.

Kim, J. and Jeong, C. (2012). Analysis of trend in construction using text mining method. *Journal of The Korean Digital Architecture·Interior Association*, **12**, 53-60.

Lim, J. and Lim, D. (2012). Comparison of clustering methods of microarray gene expression data. *Journal of the Korean Data & Information Science Society*, **23**, 39-51.

Rousseeuw, P. J. (1987). Silhouettes : Graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 54-65.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M. and Miller, H. L. (2007). *Climate change 2007*, Cambridge University Press, Cambridge, United Kingdom, 996.

Yeo, I. (2011). Clustering analysis of Korea's meteorological data. *Journal of the Korean Data & Information Science Society*, **22**, 941-949.

Journal of the Korean Data & Information Science Society
2013, **24**(6), 1429–1437

http://dx.doi.org/10.7465/jkdi.2013.24.6.1429
한국데이터정보과학회지

Analysis of the abstracts of research articles in food related to climate change using a text-mining algorithm[†]

Kyu Yong Bae[1] · Ju-Hyun Park[2] · Jeong Seon Kim[3] · Yung-Seop Lee[4]

[124]Department of Statistics, Dongguk University-Seoul
[3]Health Policy Research Department, Korea Institute for Health and Social Affairs

Received 22 October 2013, revised 14 November 2013, accepted 19 November 2013

Abstract

Research articles in food related to climate change were analyzed by implementing a text-mining algorithm, which is one of nonstructural data analysis tools in big data analysis with a focus on frequencies of terms appearing in the abstracts. As a first step, a term-document matrix was established, followed by implementing a hierarchical clustering algorithm based on dissimilarities among the selected terms and expertise in the field to classify the documents under consideration into a few labeled groups. Through this research, we were able to find out important topics appearing in the field of food related to climate change and their trends over past years. It is expected that the results of the article can be utilized for future research to make systematic responses and adaptation to climate change.

Keywords: Climate change, document classification, hierarchical clustering, text-mining.

[†] This research was supported by Bio-industry Technology Development Program, Ministry of Agriculture, Food and Rural Affairs (Project No. 312028-2).
[1] Graduate student, Department of Statistics, Dongguk University-Seoul, Seoul 100-715, Korea.
[2] Professor, Department of Statistics, Dongguk University- Seoul, Seoul 100-715, Korea.
[3] Research fellow, Health Policy Research Department, Korea Institute for Health and Social Affairs, Seoul 122-705, Korea.
[4] Corresponding author: Professor, Department of Statistics, Dongguk University-Seoul, Seoul 100-715, Korea. Email: yung@dongguk.edu