Journal of the Korean Data & Information Science Society
2015, **26**(2), 419–427

http://dx.doi.org/10.7465/jkdi.2015.26.2.419
한국데이터정보과학회지

기후변화 및 식품 관련 뉴스기사의 텍스트 마이닝[†]

현윤진[1] · 김정선[2] · 정진욱[3] · 윤시몬[4] · 이문수[5]

[15]국민대학교 전문대학원 비즈니스 IT학과 · [234]한국보건사회연구원

접수 2015년 1월 8일, 수정 2015년 1월 21일, 게재확정 2015년 2월 9일

요 약

기후변화와 식품 관련 정보가 유기적인 관련이 있음에도 불구하고, 사실상 현실에서는 사용자들이 직접 그 관련성에 대한 관심을 가지고, 해당 정보에 대한 접근이 용이하다고 말하기는 어렵다.본 연구는 실제 사용자들이 직접적으로 노출되는 인터넷 포털 사이트의 뉴스 기사에 대한 빈도분석 및 연관관계 분석을 통해 기후변화 및 식품 관련 정보가 어느 정도의 연관성을 가지고 얼마나 자주 나타나고 있는지에 대해 파악하였다. 또한 추출된 기후변화 및 식품 관련 뉴스를 대상으로 기후변화 용어 사전과 식품 관련 용어 사전을 활용하여 기후변화 관련 용어와 식품 관련 용어의 총 59개의 연관관계 규칙을 도출함으로써, 특정 기후변화 관련 용어가 어떠한 식품 관련 용어와 연관관계를 갖는지 파악하여, 추후 두 용어를 패키징해 제공할 수 있는 발판을 마련하였다.

주요용어: 기후변화, 식품, 연관성, 인터넷포털사이트, 텍스트 마이닝.

1. 서 론

텍스트는 현실에서 정보를 교환하거나 표현하는 방법으로 가장 널리 사용되는 수단이다 (Witten, 2004). 최근에는 새로운 기술의 발전과 더불어 인터넷 문화가 확산됨에 따라 웹과 다양한 소셜미디어를 통해 방대한 양의 텍스트 데이터가 매일 쏟아져 나오고 있다. 따라서 많은 연구자들은 풍부한 정보를 담고 있는 텍스트 형태의 비정형 데이터에 대한 분석을 통해 의미 있는 지식을 발견하기 위해 끊임없이 노력하고 있다. 이처럼 다량의 텍스트 데이터를 분석하여 이전에는 찾을 수 없었던 새롭고 의미 있는 정보를 추출하는 과정, 즉 대용량의 방대한 텍스트로부터 구문 분석을 통해 유용한 정보를 추출하는 과정 (Hearst, 1999; Sebastiani, 2002)을 텍스트 마이닝이라고 한다. 텍스트 마이닝은 텍스트 형태의 비정형 데이터로 이루어진 정보를 사용하는 모든 분야를 아우를 정도로 그 활용 분야가 매우 다양하다. 예를 들면, 특정 기사 (article)의 원문 (source)를 파악하기 위한 연구 (Metzler 등, 2005), 특정 범죄와 다른 범죄들 간의 유사성 측정을 통해 새로운 범죄를 발견하기 위한 연구 (Fan 등, 2006), 텍스트 범주화 (categorization)를 통해 비구조적 저장소 (repository)를 구조화하기 위한 연구 (Sebastiani, 2006) 등도 텍스트 마이닝의 활용 범위에 포함된다. 앞에서도 언급하였듯이, 최근에는 특히 다양한 성향을 가지는 사용자들의 소통 수단인 소셜미디어를 통해 방대한 양의 텍스트 데이터가 공유되고 있어, 소셜미디어 데이터에 대한 텍스트 마이닝을 수행하는 기존의 데이터 분석에서는 찾을 수 없었던 새로운 유형의 지식을 찾기 위한 시도들이 활발히 이루어지고 있다 (Kim, 2012; Choi, 2012).

[†] 본 연구는 한국농촌경제연구원 (농림수산식품기술기획평가원 사업명: 생명산업기술개발)의 식품분야 기후변화 영향분석 및 영향평가 모델구축 연구과제에서 지원받아 수행된 것임.

[1] (136-702) 서울시 성북구 정릉로 77, 국민대학교 전문대학원 (비지니스 IT학과), 연구원.
[2] 교신저자: (339-007) 세종특별자치시 시청대로 370, 한국보건사회연구원 (보건정책연구실), 연구위원. E-mail: kjs0416@kihasa.re.kr
[3] (339-007) 세종특별자치시 시청대로 370, 한국보건사회연구원 (보건정책연구실), 부연구위원.
[4] (339-007) 세종특별자치시 시청대로 370, 한국보건사회연구원 (보건정책연구실), 전문연구원.
[5] (136-702) 서울시 성북구 정릉로 77, 국민대학교 전문대학원 (비지니스 IT학과), 연구원.

2. 연구 방법

2.1. 텍스트 마이닝 연구자료

기존의 데이터 마이닝에서 확장된 형태의 텍스트 마이닝은 전통적인 데이터 마이닝에서 사용되는 연관관계 (association) 분석, 분류 (classification), 군집화 (clustering)뿐만 아니라, 자연어 처리, 정보 검색, 전산 언어학, 토픽추적 (topic tracking), 텍스트 범주화 등의 분야의 기술들을 종합적으로 활용한다 (Mooney와 Bunescu, 2006; Rijsbergen, 1979). 이러한 기술로 인해 기존의 데이터 마이닝 분야에서 해결해왔던 전통적인 주제뿐만 아니라, 더욱 폭넓고 다양한 주제에 대한 분석이 가능해졌다. 특히 자연어 처리 기술은 텍스트 마이닝의 성패를 좌우하는 핵심 기술이라고 할 수 있으며, 자연어 처리의 대상이 되는 텍스트는 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로 표현된다 (Stanvrianou 등, 2007). SAS enterprise miner, IBM SPSS modeler, R, linguamatics 12E 등 데이터 마이닝 소프트웨어는 텍스트 마이닝 기능을 지원하며, 대부분의 분석 도구에서 분석의 최소 단위는 각 문서가 된다. 여기서 말하는 문서란 제목, 요약, 본문, 문서전체 등 텍스트로 기술된 모든 데이터를 일컫는 폭넓은 개념을 의미한다. 각 문서는 여러 형태로 표현 가능하나, 기본적으로는 벡터공간모델 (vector space model; Albright, 2006; Salton 등, 1975)을 이용하여 표현되며, 각 문서에 사용된 용어 (term)의 빈도에 따라 해당 문서의 주제 및 특성이 요 약된다. 대부분의 경우 용어의 단순 빈도수보다는 TF-IDF (term frequency-inverse document frequency; Han과 Kamber, 2011)에 근거한 분석이 널리 활용되고 있다. TF-IDF의 개념은 어떤 문서 D에서 용어 A와 B가 동일한 빈도수로 발생하였을 때, A가 다른 문서들에서도 일반적으로 자주 발생하는 용어라면 문서 D에서 더 중요하게 사용되는 용어는 A가 아니라 B라는 인식을 전제로 한고 있다. 빈도수 기반의 분석에서 각 문서는 용어 수만큼의 차원을 갖게 되고, "(문서수)×(용어 수)"로 표현된 행렬의 각 셀에 각 문서에서 해당 용어가 나타난 빈도수를 기재함으로써 모든 문서를 행렬화할 수 있다. 하지만 문서에 포함된 용어의 수는 일반적으로 매우 방대하기 때문에, 문서간 유사성 측정을 위해 각 문서는 SVD (singular value decomposition) 등의 차원 축소 기법을 통해 저장된다 (Albright, 2006). 상용 텍스트 마이닝 도구는 이러한 이론을 기본으로 하여 파싱, 필터링, 클러스터링 등의 작업을 수행하게 되며, 이러한 작업의 결과는 이 자체로도 문서 분류, 토픽 추출 등의 작업에 사용될 수 있을 뿐 아니라, 기존의 마이닝 분석 모델인 의사결정나무, 인공신경망 등 후속 분석의 입력으로 사용될 수도 있다.

2.2. 연관관계 분석 연구자료

데이터 마이닝은 방대한 양의 데이터로부터 유용한 정보나 패턴을 추출하는 기법으로써, 통계적 기법, 인공지능 기법 등을 통해 연관관계 분석, 분류, 군집화 등의 여러 가지 지식을 창출하는 과정 (Han과 Kamber, 2011)에 널리 활용되고 있다. 특히 연관관계 분석 (Agrawal과 Srikant, 1994)은 데이터들의 빈도수와 동시 발생 확률을 이용하여 데이터와 데이터 간의 관계를 찾고 이를 규칙으로 표현하는 분석 기법으로, 장바구니 분석, 인터넷 쇼핑몰 추천시스템, 교차판매, 매장 배치, 카탈로그 설계, 판촉전략 수립, 국가현안 및 R&D 정보 패키징 등 여러 분야 (Kim, 2008; Ahn 등, 2006; Yoon, 2005; Lee와 Kim, 2013; Hyun 등, 2013; Wang 등, 2004)에서 활용되고 있다. 이러한 연관관계 분석을 통해서 도출된 규칙들을 평가하기 위해 다양한 흥미성 척도가 고안되어 왔으며, 그 중에서도 신뢰도 (confidence), 지지도 (support) 그리고 향상도 (lift)의 세 가지 척도가 가장 일반적으로 사용되고 있다. 연관관계 분석을 통해 도출된 규칙들 중 향상도가 1 이상이면서 지지도와 신뢰도가 각각 최소지지도와 최소신뢰도 이상으로 나타날 때, 해당 규칙은 강한 연관성을 나타내는 것으로 간주된다 (Park 등, 2003). 연관규칙 (A → B)에 대해, 지지도는 전체 트랜잭션 수 대비 A와 B가 동시에 출현한 트랜잭션 수의 비율을 의미하며, 신뢰도는 A를 포함하는 트랜잭션 중 A와 B를 함께 포함하는 트랜잭션의 비율을 의미한다. 즉, 특

정 연관규칙에 대한 신뢰도는 조건에 맞는 결과가 얼마나 자주 적용될 수 있는지를 나타내고, 지지도는 연관규칙 자체가 얼마나 믿을만한 것인지를 나타낸다고 할 수 있다. 한편 향상도는 A와 B의 상관관계를 나타내는 척도로써, 그 값이 1이면 A와 B가 독립적인 관계를 나타내고, 1보다 큰 경우에는 양의 상관관계, 1보다 작은 경우에는 음의 상관관계를 나타낸다고 할 수 있다.

연관관계 분석은 그 자체로도 위와 같은 다양한 분야에 폭넓게 활용되어 왔지만, 최근에는 소셜 네트워크분석 (SNA; social network analysis), 텍스트 마이닝 등의 분야의 비약적인 발전으로 인해 그 활용범위가 더욱더 다양해졌다. 즉 분석 대상 데이터에 대한 연관관계 분석을 실시하여 관심 항목간 연관성을 도출하고, 각 항목과 이들간 연관성을 네트워크로 도식화함으로써 보다 다층적이고 심층적인 분석을 실시할 수 있게 된 것이다 (Cho와 Kim, 2011; Hyun 등, 2013). 또는 인터넷 문서, 뉴스 기사, 소셜미디어 등에 대한 분석을 통해 연관어 맵을 도출하거나 문서간 분류를 수행하는 응용의 경우 또한 기본적으로 연관관계 분석의 원리를 이용하는 것으로 파악될 수 있다.

3. 연구결과

3.1. 기후변화 및 식품 관련 뉴스 기사 분석

기후변화와 식품 관련 정보가 유기적인 관련이 있음에도 불구하고, 사실상 현실에서는 사용자들이 직접 그 관련성에 대한 관심을 가지고, 해당 정보에 대한 접근이 용이하다고 말하기는 어렵다. 대개의 경우, 사용자들은 기후변화와 관련한 정보만을 검색하거나, 식품에 관련한 정보만을 검색하여 원하는 정보를 획득하기 때문이다. 이는 사용자들의 관심도에도 중요하지만, 사용자들에게 적합한 정보를 제공해줄 수 없는 환경의 문제도 분명히 존재한다. 기후변화와 식품은 사용되는 용어의 풀 (pool) 자체가 다르기 때문에 사용자들이 기후변화 혹은 식품과 관련하여 서로 관련이 있는 특정 정보를 찾아내기까지는 무수히 많은 시행착오를 거쳐야 하는 어려움이 있다. 따라서 본 연구에서는 사용자들이 실제로 매 순간 노출되고 있는 인터넷 뉴스 기사 분석을 통해 현실에서는 얼마나 많은 기후변화 관련 뉴스 기사가 출현하고 있으며, 해당 기사에서 식품 관련 정보가 나타나는 빈도는 어느 정도인지 분석하고자 한다. 더 나아가 기후변화와 식품 관련 뉴스 기사를 대상으로 연관관계 분석을 수행하여 기후변화 관련 정보와 식품 관련 정보의 연결고리를 마련함으로써, 사용자들에게 기후변화 및 식품 관련 정보를 패키징하여 보여줄 수 있는 발판을 마련하고자 한다.

3.1.1. 분석 데이터 및 환경 소개

기후변화 및 식품 관련 뉴스 기사를 분석하기에 앞서, 분석에 필요한 뉴스 기사 수집을 위해 2012년 7월 1일부터 2013년 6월 30일까지 1년 동안의 기사 중 임의로 394,303건을 크롤링하여 데이터베이스화 하였다. 하지만 모든 기사가 기후변화 및 식품 관련 뉴스를 다룬다고 볼 수 없기에, 여러 카테고리 중 '연예', '스포츠' 카테고리를 제외한 '정치', '경제', '사회', '세계', '생활/문화', 'IT/과학'의 6개 카테고리의 기사 92,440건을 분석 대상으로 선정하였다. 특히, 인터넷 뉴스 포털 사이트의 경우, 주요 언론사의 기사를 취합하여 재공급하는 특성을 가지고 있기 때문에 특정 매체의 시각에 편향되지 않은 이슈를 제공한다는 장점을 갖는다. 또한 충분히 많은 기사를 보유하고 있는 측면에서도 본 분석의 대상 데이터로 적합한 특성을 갖고 있다고 할 수 있다.

위에서 선정된 분석 데이터를 대상으로 행해지는 모든 분석 과정은 SAS enterprise miner 12.1 상에서 행해졌으며, 데이터의 저장 및 가공은 oracle DBMS 12C와 SAS enterprise guide 5.1을 활용하였다.

3.1.2. 기후변화 및 식품 관련 뉴스 기사 빈도분석

3.1.1.절에서 언급하였듯이, 인터넷 포털 사이트의 기사들 중에서도 기후변화 및 식품 관련 정보가 출현할 가능성이 높은 '정치', '경제', '사회', '세계', '생활/문화', 'IT/과학'의 6개 카테고리의 기사

92,440건을 대상으로 분석을 수행하였으며, 1차 분석 단계로, 전체 분석 대상 뉴스에서 기후변화 관련 뉴스 기사에 대한 빈도분석을 수행하였다.

우선 전체 뉴스 기사 중에서 기후변화 관련 뉴스의 빈도수를 알아보기 위해, 기후변화 용어사전을 활용하여 기후변화 관련 용어의 빈도수 및 해당 용어가 출현한 기후변화 관련 뉴스 기사를 추출하였다. 이때, 기후변화 용어사전은 관련 분야 전문가가 언론 보도용 수준의 용어를 57개 선정하여 사전 구축 후 분석 과정에 사용하였다.

SAS enterprise miner 12.1의 텍스트 마이닝 모듈에서 파싱 기능을 사용하여 기후변화 관련 용어 빈도 및 해당 용어가 출현한 뉴스 기사를 추출하였다. 이 과정을 통해 전체 뉴스 기사 중에서 start list로 적용된 기후변화 용어사전에 나온 용어만을 추출하고 해당 용어의 빈도수 및 해당 용어가 출현한 뉴스 기사를 기후변화 관련 뉴스 기사로 정의하였다. 그 결과 화면은 다음 Table 3.1과 같다.

Table 3.1의 분석 결과를 보면 처음 정의한 기후변화 용어 사전에 기후변화와 밀접한 관련을 갖지 않는 용어도 포함되어 있는 것을 알 수 있다. 이는 기후와 함께 나타났을 때 의미가 있는 용어로 해당 분석에서는 기후변화와 가장 직접적인 관련이 있는 용어 26개에 대한 결과만을 추출하였다. 그 결과의 일부가 Table 3.2에 나타나 있다.

Table 3.1 Climate change related term frequency in internet-news

term	frequency	number of publications	term	frequency	number of publications
temperature	4,105	1,531	agricultural	799	440
gas	2,342	1,242	accessibility	472	350
air	1,726	1,181	adaptation	402	346
predicted	1,169	930	flood	436	337
weather	1,275	859	food	676	301
heavy rain	1,276	786	model	395	281
trading	1,108	710	respiratory	383	246
heavy snow	1,143	705	plant	473	223
stock	987	667	fisheries	309	216
support	776	637	disaster	291	213
safety	974	636	draught	347	212
ecosystem	1,027	618	humidity	278	212
scenario	829	606	carbon dioxide	322	210
weak	782	591	extraordinary	204	180
heat wave	1,198	584	climate change	190	158
process	820	534	soil	216	148
climate	799	515	carbon	213	137

Table 3.2 Frequency of selected terminologies directly related to climate change in internet-news

term ID	term	publication number	date of publication	term frequency
750	water	298802	20120701	1
9076	ecosystem	334915	20120701	2
22183	air	64168	20120701	1
23941	draught	183444	20120701	1
32206	soil	334915	20120701	1
240	environment	190911	20120702	1
1962	flood	225698	20120702	1
9076	ecosystem	338991	20120702	1
23941	draught	30847	20120702	1
240	environment	325688	20120703	1
1962	flood	13934	20120703	1
5188	heavy rain	351456	20120703	1
13839	climate	54555	20120703	1
15543	disaster	224776	20120703	1
22163	temperature	239877	20120703	3
22175	heat wave	170300	20120703	1
22183	air	198413	20120703	1
32969	weather	372637	20120703	1
32969	weather	170300	20120703	1
104890	carbon dioxide	54555	20120703	3
22163	temperature	105596	20120704	3
22175	heat wave	265245	20120704	2
92550	humidity	17037	20120711	2

Table 3.2의 결과를 바탕으로 기후변화 관련 뉴스 기사를 추출한 결과 전체 분석 대상 뉴스 기사 92,440건 중에서 약 16%인 14,336건의 뉴스 기사가 기후변화 관련 뉴스 기사로 추출되었다. 이렇게

추출된 뉴스 기사를 대상으로 식품 관련 용어 사전을 활용하여 2차 빈도분석을 수행하였다. 이때, 활용되는 식품 관련 용어 사전 역시 기후변화 관련 용어 사전과 마찬가지로 관련 분야 전문가가 언론 보도용 수준에서 51개의 용어를 선정하여 사전 구축 후 분석에 활용하였다.

2차 빈도분석 단계에서는 1차 분석 단계에서 정의된 기후변화 관련 뉴스를 대상 데이터로 하여 식품 관련 용어 사전을 start list로 적용한 후, 기후변화 용어와 함께 출현한 식품 관련 용어를 추출한 후, 해당 용어의 빈도수 및 해당 용어가 출현한 뉴스 기사, 즉 기후변화 및 식품 관련 용어가 동시에 출현한 뉴스 기사를 추출하였다. Table 3.3은 분석 결과의 일부를 보여주고 있다.

Table 3.3 Food related term frequency in climate change related internet-news

term ID	term	publication number	date of publication	term frequency
1997	disease	20	20121203	1
3788	food	331	20121207	1
5821	food	512	20130417	2
5821	food	517	20120930	1
5821	food	520	20130619	6
7625	consumption	756	20130405	1
9173	pollution	930	20130228	1
9173	pollution	956	20120826	1
7625	consumption	1,084	20121130	4
10417	eat out	1,084	20121130	1
10378	distribution	1,084	20121130	1
1997	disease	1,432	20120922	1
9173	pollution	1,469	20130409	2
13727	virus	1,469	20130409	2
13707	groceries	1,469	20130409	1
10378	distribution	1,537	20121231	5
14390	food	1,537	20121231	1
13707	groceries	1,537	20121231	2
14390	food	1,735	20121029	1
7625	consumption	1,865	20130415	1
13707	groceries	1,865	20130415	1
10378	distribution	1,865	20130415	5
13707	groceries	2,206	20130313	1
16874	restaurant	2,206	20130313	2
16874	restaurant	2,310	20130502	1
13757	processed food	2,310	20130502	2
13707	groceries	2,325	20130328	1

Table 3.3의 결과를 바탕으로 기후변화 관련 뉴스 기사에서 식품 관련 용어가 출현한 뉴스 기사를 추출한 결과, 총 14,336건의 기후변화 관련 뉴스 기사 중 약 27%인 3,887건의 뉴스 기사에서 식품 관련 용어가 함께 출현했다는 것을 알 수 있었다. 즉, 전체 뉴스 기사 92,440건 중에서 약 4% 정도의 뉴스 기사에서 기후변화와 식품 관련 정보가 함께 나타남을 알 수 있었다. Table 3.4는 빈도분석 결과를 월별 및 분기별로 나누어 나타낸 것이다.

Table 3.4 Climate change and food related term frequency in climate change related internet-news listed by month and season

data collection period	number of climate change related publications	climate change and food related term frequency	probability of climate change related publication containing food related terms
July, 2012	1,221	293	24.00
August, 2012	1,768	299	16.91
September, 2012	1,185	239	20.17
October, 2012	1,001	257	25.60
November, 2012	904	231	25.55
December, 2012	1,266	275	21.72
January, 2013	1,379	327	23.71
February, 2013	1,192	278	23.32
March, 2013	1,039	279	26.85
April, 2013	1,272	314	24.69
May, 2013	1,248	312	25.00
June, 2013	1,185	283	23.88
Spring (Mar, Apr, May)	3,559	905	25.43
Summer (Jun, Jul, Aug)	4,174	875	20.96
Autumn (Sep, Oct, Nov)	3,093	727	23.50
Winter (Dec, Jan, Feb)	3,837	880	22.93

3.2. 기후변화 및 식품 관련 키워드 연관관계 분석

앞의 과정을 통해 전체 뉴스 기사에서 기후변화 및 식품 관련 뉴스 기사의 빈도분석을 수행함으로써, 실제 사용자들이 노출되는 인터넷 언론보도에서 기후변화와 식품 관련 정보가 얼마나 자주 동시에 언급되고 있는지에 대해 파악할 수 있었다. 하지만 앞선 분석에서는 해당 용어의 동시 출현 뉴스 기사의 빈도수만을 파악하는데 그쳤기 때문에, 특정 기후변화 관련 용어가 과연 어떠한 식품 관련 용어와 연관관계를 가지고 함께 나타나는지는 파악하기 어렵다는 한계가 존재한다. 따라서 본 연구에서는 3.1.2.절에서 추출된 기후변화 및 식품 관련 뉴스를 대상으로 기후변화 용어 사전과 식품 관련 용어 사전을 활용하여 기후변화 관련 용어와 식품 관련 용어의 연관관계 규칙을 도출함으로써, 특정 기후변화 관련 용어가 어떠한 식품 관련 용어와 연관관계를 갖는지 파악하여, 추후 두 용어를 패키징해 제공할 수 있는 발판을 마련하고자 한다. 우선 분석의 품질을 향상시키고 분석 시간을 단축시키기 위한 정제작업을 수행하였으며, 이 과정에서 기후변화 및 식품 관련 용어를 제외한 다른 용어들을 제거하기 위해 기후변화 용어 사전과 식품 관련 용어 사전을 사용하였다. Table 3.5는 그 결과를 보여주고 있다. 물, 환경, 음식, 유통, 식품, 질환, 식당, 소비, 공기, 기온 등 빈도순 상위 10개 용어 외에 곰팡이, 이산화탄소, 화학물질, 독소, 박테리아, 해충, 기생충, 장염 등의 식품안전에 유해한 물질들에 대한 용어들을 볼 수 있다.

Table 3.5 Data cleaning result of climate change and food related news

term	frequency	number of publications	term	frequency	number of publications
water	3,044	1,440	food	187	105
environment	2,153	1,249	cooking	160	104
food	2,425	915	heavy snow	152	101
distribution	1,041	531	mold	237	97
groceries	1,246	527	draught	137	82
disease	1,321	525	infectious disease	108	82
restaurant	1,055	480	eating habits	108	79
consumption	841	466	flood	83	73
air	558	370	heavy rain	96	70
temperature	426	242	eat out	96	64
pollution	358	213	soil	88	62
restaurant	342	213	processed food	97	56
infection	421	197	climate change	66	54
germ	437	188	carbon dioxide	75	48
virus	504	187	chemical substance	63	47
climate	303	166	toxin	76	46
ecosystem	313	160	bacteria	86	44
food item	222	151	disaster	66	40
heat wave	309	140	food poisoning	94	36
weather	178	125	harmful insects	82	36
process	174	120	carbon	61	32
humidity	164	113	parasite	90	30
food	295	107	enteritis	50	30

이렇게 추출된 기후변화 및 식품 관련 용어의 집합에 대해 연관관계 분석을 수행하여 총 59개의 연관규칙을 도출하였으며, 그 결과의 일부가 Table 3.6에 나타나 있다. 이때, 분석에 사용된 대표적인 연관성 척도인 신뢰도와 지지도에 대한 간략한 수정 정의는 다음과 같다.

$$\text{support}(A \grave{a}\ B) = \frac{(\text{기후변화 용어 } A\text{와 식품 관련 용어 } B\text{가 동시에 명시된 문서의 수})}{(\text{전체 문서의 수})}$$

$$\text{confidence}(A \grave{a}\ B) = \frac{(\text{기후변화 용어 } A\text{와 식품 관련 용어 } B\text{가 동시에 명시된 문서의 수})}{(\text{기후변화 용어 } A\text{가 명시된 문서의 수})}$$

연관관계가 있는 키워드 중에서 지지도 척도가 높은 순으로 상위 10위를 살펴보면 세균과 물, 음식과 물, 오염과 환경, 식품과 물, 식당과 물, 질환과 물, 유통과 환경, 음식점과 물, 감염과 질환 그리고 감염과 물이 연관성이 높게 나타났고, 신뢰도가 높은 순으로 상위 10위의 키워드를 살펴보면 음식과 물, 식품과 물, 음식과 환경, 유통과 환경 그리고 질환과 물이 연관성이 높은 것으로 나타났다. 또한 기후변화

와 식품 관련 뉴스 기사들 중 연관성이 높은 키워드로서 물의 빈도수가 가장 높게 나타나서, 특히 물의 중요도를 보여주고 있다.

Table 3.6 Association rules between climate change and food related terms

confidence	support	lift	rule	
40.64	2.24	6.99	virus	infection
38.58	2.24	6.99	infection	virus
43.15	2.51	2.78	infection	disease
16.19	2.51	2.78	disease	infection
39.36	2.18	2.54	germ	disease
14.10	2.18	2.54	disease	germ
34.27	2.16	2.42	eatery	restaurant
15.21	2.16	2.42	restaurant	eatery
30.99	2.21	2.00	temperature	disease
14.29	2.21	2.00	disease	temperature
29.19	3.19	1.88	air	disease
20.57	3.19	1.88	disease	air
55.40	3.48	1.50	pollution	environment
61.70	3.42	1.45	germ	water
38.33	16.30	1.42	water	food
60.33	16.30	1.42	food	water
36.00	5.58	1.33	disease	food
20.66	5.58	1.33	food	disease
35.95	3.93	1.33	air	food
14.54	3.93	1.33	food	air
44.63	7.00	1.21	distribution	environment
18.98	7.00	1.21	environment	distribution
32.64	5.08	1.21	groceries	food
18.80	5.08	1.21	food	groceries
50.47	7.85	1.19	groceries	water
18.47	7.85	1.19	water	groceries
48.75	6.91	1.15	restaurant	water
16.25	6.91	1.15	water	restaurant
42.27	5.82	1.15	consumption	environment

4. 결론

본 연구는 실제 사용자들이 직접적으로 노출되는 인터넷 포털 사이트의 뉴스 기사에 대한 빈도분석 및 연관관계 분석을 통해 기후변화 및 식품 관련 정보가 어느 정도의 연관성을 가지고 얼마나 자주 나타나고 있는지에 대해 파악하였다. 하지만 사용자들에게 기후변화 및 식품 관련 정보를 패키징하여 제공해 주기 위해서는 빈도분석 및 연관관계 분석을 수행함에 있어, 분석 결과에 큰 영향을 미치는 기후변화 및 식품 관련 용어 사전의 질과 양을 향상시키기 위한 노력이 반드시 필요하다. 따라서 향후 연구에서는 더욱더 풍부한 양질의 용어 사전을 구축이 필요할 뿐 아니라, 추가적인 패키징 방법론 개발에 대한 노력이 필요할 것으로 판단된다. 향후 뉴스 기사 분석과 더불어 트위터, 블러그, 페이스북 등 SNS 빅데이터 분석을 활용하여 사용자의 자발적인 의사 표현까지 수집한 후 사용자들의 행태를 파악하고 기후변화와 식품 관련 소비자 행태 개선방안을 제시할 수 있는 도구를 개발할 수 있겠다.

References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, **1215**, 487-499.

Ahn, H. C., Han, I. K. and Kim, K. J. (2006). The product recommender system combining association rules and classification models: The case of G Internet shopping mall. *Information System Review*, **8**, 181-201.

Albright, R. (2006). *Taming text with the SVD*, SAS Institute Inc., Cary, NC.

Cho, I. D. and Kim, N. K. (2011). Recommending core and connecting keywords of research area using social network and data mining techniques. *Journal of Intelligence and Information Systems*, **17**, 127-138.

Choi, K. S. (2012). Hybrid big data analysis techniques and case studies in the SNS-era. In *2012 Big Data Search and Analysis Techniques Insight*, Korea Database Agency, Seoul.

Fan, W., Wallace, W., Rich, S. and Zhang, Z. (2006). Tapping the ower of text mining. *Communications of the ACM*, **49**, 76-82.

Han, J. and Kamber, M. (2011). *Data mining: Concepts and techniques*, 3rd Ed., Morgan Kaufmann Publishers, San Francisco, US.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computation Linguistics, 3-10.

Hyun, Y., Han, H., Choi, H., Park, J., Lee, K., Kwahk, K. and Kim, N. (2013). Methodology using text analysis for packaging R&D information services on pending national issues. *Journal of Information Applications & Management*, **20**, 231-257.

Kim, I. H. (2012). Values and application strategy of big data. In *2012 Big Data Search and Analysis Techniques Insight*, Korea Database Agency, Seoul.

Kim, N. (2008). Effect of market basket size on the accuracy of association aule measures. *Asia Pacific Journal of Information Systems*, **18**, 95-114.

Lee, Y. J. and Kim, K. J. (2013). Product recommender systems using multi-model ensemble techniques. *Journal of Intelligence and Information Systems*, **19**, 39-54.

Metzler, D., Bernstein, Y., Crofit, W. B., Moffat, A. and Zobel, J. (2005). Similarity measures for tracking information flow. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 517-524.

Mooney, R. J. and Bunescu, R. (2006). Mining knowledge from text using information extraction. *ACM SIGKDD Explorations Newsletter*, **7**, 3-10.

Park, W. C., Seung, H. W. and Yong, H. S. (2003). *Data mining: Concepts and techniques*, Free Academy, Seoul.

Rijsbergen, C. J. V. (1979). *Information retrieval*, 2nd Ed., Butterworth, London,

Salton, G., Wong, A. and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**, 613-620.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**, 1-47.

Sebastiani, F. (2006). Classification of text, automatic. In *The encyclopedia of language and linguistics 14*, 2nd Ed., Elsevier Science Pub., Amsterdam, 457-462.

Stanvrianou, A., Andritsos, P. and Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *ACM Sigmod Record*, **36**, 23-34.

Wang, W. F., Chung, Y. L., Hus, M. H. and Keh, H. C. (2004). A personalized recommender system for the cosmetic business. *Expert Systems with Applications*, **26**, 427-434.

Witten, I. H. (2004). Text mining. In *Practical Handbook of Internet Computing*, edited by M. P. Singh, CRC Press, New York.

Yoon, S. J. (2005). A study of churn prediction model for department store customers using data mining technique. *Asia Marketing Journal*, **6**, 45-72.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 419–427

http://dx.doi.org/10.7465/jkdi.2015.26.2.419
한국데이터정보과학회지

Text mining on internet-news regarding climate change and food[†]

Yoonjin Hyun[1] · Jeong Seon Kim[2] · Jin-Wook Jeong[3] · Simon Yun[4] · Moon-Soo Lee[5]

[15]Graduate School of Business IT, Kookmin University
[234]Korea Institute for Health and Social Affairs

Received 8 January 2015, revised 21 January 2015, accepted 9 February 2015

Abstract

Despite of correlation between climate changes and food-related information, it is still not easy for many users to get access to the information with interest. This study investigated how much climate change and food-related information are correlated with each other and how often they are exposed through frequency and correlation analysis on news articles on the internet portals. Through analysis on the frequency of climate change and food-related news articles, this study was able to figure out how often they are exposed at the same time by the internet news portals. In addition, a total of 59 correlation rules regarding the climate change and food-related vocabularies were derived from these news articles using the climate change and food-related glossaries. Then, a correlation between certain climate change-related and food-related words was analyzed in order to package the related words.

Keywords: Climate change, food, internet portals, relation, text mining.

[†] This research was supported by Korea Institute of Planning & Evaluation for Technology in Food, Agriculture, Forestry & Fisheries.
[1] Researcher, Graduate School of Business IT, Kookmin University, Seoul 100-715, Korea.
[2] Corresponding author: Senior researcher, Health Policy Research Department, Korea Institute for Health and Social Affairs, Sejong 339-007, Korea. E-mail: kjs0416@kihasa.re.kr
[3] Senior researcher, Health Policy Research Department, Korea Institute for Health and Social Affairs, Sejong 339-007, Korea.
[4] Researcher, Health Policy Research Department, Korea Institute for Health and Social Affairs, Sejong 339-007, Korea.
[5] Researcher, Graduate School of Business IT, Kookmin University, Seoul 100-715, Korea.