소셜 빅데이터 분석을 활용한 사전탐색적 불량식품 판별법 개발 후보 도출

An Exploratory Study on Extraction of Food Fraud Candidates for Developing Detection Methods Using Social Big Data Analysis

장현후[1], 김주란[1], 라하나[2], 류소영[2], 정소영[2], 조천호[3], 권기성[3], 전향숙[2*]

Hyunhoo Jang[1], Jooran Kim[1], Ha Na Ra[2], So Young Ryu[2], So-young Jeong[2], Cheon Ho Jo[3], Kisung Kwon[3], Hyang Sook Chun[2*]

[1]㈜버즈메트릭스, [2]중앙대학교 식품공학부, [3]식품의약품안전평가원

[1]Buzz Metrix,
[2]School of Food Science and Technology, Chung-Ang University
[3]National Institute of Food and Drug Safety Evaluation

1. 서론

식품의 생산, 제조, 유통, 판매 등 전 단계에서 발생할 수 있는 모든 법을 위반한 제품을 '불량식품'으로 정의하고 있으며, 식품의약품안전처는 불량식품의 유형을 17개로 분류하였다. 불량식품은 안전한 식품을 섭취할 수 있는 소비자의 권리를 침해하는 요인으로 대두되며 불량식품 근절을 위한 다양한 정책이 시행되고 있다. 그러나 식품 공급망이 세계화되고 복잡해짐에 따라 국제규모의 위·변조 행위가 발생할 수 있다(전 등, 2014). 경제적 이익을 취하기 위해 의도적으로 원료를 바꾸거나 부정물질을 식품에 첨가하는 수법이 점점 정교해지고 급증하고 있으나 사법권역이 달라 법적인 대응이 쉽지 않은 실정이다(장 등, 2017). 이와 같이 불량식품의 발생은 사후처리에 어려움이 있으며, 감시·단속과 같은 방법으로 불량식품 발생에 대응하기에는 한계가 있을 것으로 판단된다(장 등, 2017). 따라서 불량식품을 사전에 차단할 수 있는 과학적 체계가 필요하며, 이를 위해 소비자가 생각하는 불량식품에 대한 인식과 관심도를 파악하는 것은 매우 중요한 기초자료가 될 것이다.

빅데이터는 대량의 정형 또는 비정형 데이터를 포함한다. 트위터, 페이스북, 블로그, 카페 등의 소셜 네트워크를 기반으로 생성된 빅데이터는 개인의 생각과 화제를 게재 할 때, 외부적인 요인에 의해 영향을 받지 않기 때문에 문서의 형태가 다양하고, 트렌드를 반영할 수 있는 장점이 있어 기존의 설문조사방법의 한계를 극복하는데 사용될 수 있다(Roh, 2017). 빅데이터의 주요 특성은 방대한 자료의 양(volume), 문서 입출력의 속도(velocity), 수집 자료의 다양성(variety)으로 구분할 수 있으며, 수집된 빅데이터의 분석을 통해 새로운 가치(value)를 창출하기 위해 사회안전(Lee, 2017), 경제(Park 등, 2017), 교육(김 및 김, 2017; 박, 2017) 등 다양한 분야에서 활용되고 있다(Raghupathi & Raghupathi, 2014). 불량식품 판별법 개발 연구의 궁극적인 목적은 소비자의 권리를 보장할 수

* Correspondence to: Hyang Sook Chun, Ph.D.
School of Food Science and Technology, Chung-Ang University
#4726 Seodong-daero, Anseong-si, Gyeonggi-do 17546, South Korea
Tel: +82-31-670-3290 (office) Fax: +82-31-676-2102 E-mail: hschun@cau.ac.kr

Table 1. 불량식품 17개 유형 및 수집 키워드

불량식품 유형	수집 키워드	최종수집 문서 건(%)	빅데이터 상세분석
부패/변질된 위해 우려 식품	상한, 변질, 썩은, 곰팡이, 벌레, 구더기, 부패, 불량	23,277(23.0)	–
유독/유해한 물질이 함유된 위해 우려 식품	유해, 농약, 자연독, 독성, 항생제, 발암물질, 발암성, 방부제, 보존제, 색소, 동물용의약품	22,763(22.5)	–
사용이 금지된 물질 함유 식품	마약류, GMO, 유전자조작, 비식용, 불법첨가제, 멜라민, 색소, 사용금지, 미승인	10,873(10.8)	○
불법도축, 병든 고기나 그 원료로 만든 식품	기생충, 불법도축, 밀도축, 병든, 밀도살, 불법도살, 불량고기	5,432(5.4)	○
유독/유해물질이 함유된 식품용기/포장	포장, 용기, 가소제, 중금속용출, 증발잔류물	9,807(9.7)	–
유해물질 기준, 규격 부적합 식품	기준치 초과, 부적합, 이물혼입, 위반	8,279(8.2)	○
비위생적으로 제조/조리/재사용한 식품	재사용, 기준치, 위생불량, 비위생	3,646(3.6)	–
수입 신고를 하지 않고 반입된 식품	밀수입, 밀반입, 수입신고	1,149(1.1)	–
제품의 성분(함량), 품질, 가격을 속인 식품	가짜, 짝퉁, 중량미달, 불법증량, 함량미달, 거짓, 카라멜, 불량품질, 불량성분	7,075(7.0)	○
성분, 영양가, 신고사항 등 허위표시식품	허위표시, 미표시	2,260(2.2)	○
병원성 미생물 등에 오염된 위해우려 식품	식중독, 대장균, 장염, 바이러스, 살모넬라	42,825(42.4)	–
성분, 규격에 맞지 않는 식품	규격미달, 기준미달, 부적합 성분	363(0.4)	○
무허가/무신고 식품	무허가, 무신고	668(0.7)	–
원산지를 속인 식품	원산지 허위, 원산지 속인, 원산지 미표시, 원산지 미표기	1,642(1.6)	○
유통기한 위/변조 식품	유통기한	5,063(5.0)	–
질병치료나 의약품으로 오인/혼동 광고 식품	허위과장, 과대광고, 소비자혼동	2,545(2.5)	–
어린이 현혹 저가/저품질 정서저해식품	어린이, 정서저해, 미끼	1,486(1.5)	–

있는 기준을 마련하고, 식품의 안전성을 확보하기 위함이다. 이를 위해 소비자가 남긴 블로그나 카페의 빅데이터를 분석하는 것은 불량식품에 대한 소비자의 의견이 반영될 수 있어 현실적인 정책수립 및 대응이 가능하다.

이에 본 고에서는 불량식품 유형에 따른 키워드를 기반으로 빅데이터를 수집하여 발생빈도가 높은 식품에 대한 소비자의 다양한 의견을 워드 클라우드로 분석하였다. 또한 식품 전문가로부터 판별법이 필요한 불량식품에 대한 의견을 수집하였다. 이를 통해 도출된 주요식품을 대상으로 전문가 FGD(Focus group discussion)를 실시하여 불

량식품 판별법 개발을 위한 사전탐색적 후보 품목의 우선순위를 도출하였다.

2. 연구방법

2.1. 불량식품 관련 소셜 빅데이터 수집 및 분석방법

2014년 1월부터 2016년 12월까지 생성된 국내 온라인 포털 사이트(네이버, 다음)의 블로그와 카페에 게재된 문서를 수집하였다. ㈜Buzz metrix에서 자체 개발한 크롤러(Crawler)를 이용해 원문 데이터를 검색하고, 식품의약품안전처의 불량식품 17개 유형에 따른 키워드를 기반으로 총 1,815,435건의 데이터를 수집하였다. 수집된 자료 중 불량식품과 연관성이 없는 문서는 기계적으로 1차 필터링 하였으며, 시스템으로 확인이 어려운 부분은 인적 필터링을 통해 제거하여 149,153건을 최종 분석 데이터로 설정하였다(Table 1). 이 중 미국 FDA(Food and Drug Administration)의 정의에 따라 소비자를 의도적으로 속여 경제적인 이득을 취하기 위해 식품이나 첨가물을 대체, 첨가 및 변조하는 등의 행위(Spink & Moyer, 2011)에 해당하는 불량식품 유형 7개에 대해 소셜 빅데이터를 상세분석하여 키워드 발현량이 높은 식품을 10위까지 도출하였으며, 소비자가 우려하는 식품의 버즈 발생량에 따른 의견을 워드 클라우드로 시각화하였다.

2.2. 사전탐색적 불량식품 품목 도출 및 개발 필요 판별법 협의를 위한 FGD

빅데이터 분석 시 키워드 발생빈도가 높고, 불량식품 유형에 중복되어 나타나는 식품을 토대로 사전탐색적 불량식품 판별법 개발이 요구되는 대상 품목을 발굴하고, 우선순위를 도출하기 위해 FGD를 실시하였다(Fig. 1). 2017년 9월, 8인의 전문가를 대상으로 불량식품 판별법 개발 여부를 검토하여 이미 판별법이 개발된 불량식품은 대상에서 제외하였다. 불량식품의 판별법이 개발되어 있으나 불량식품의 유형이 달라 추가적인 판별법이 요구되는 식

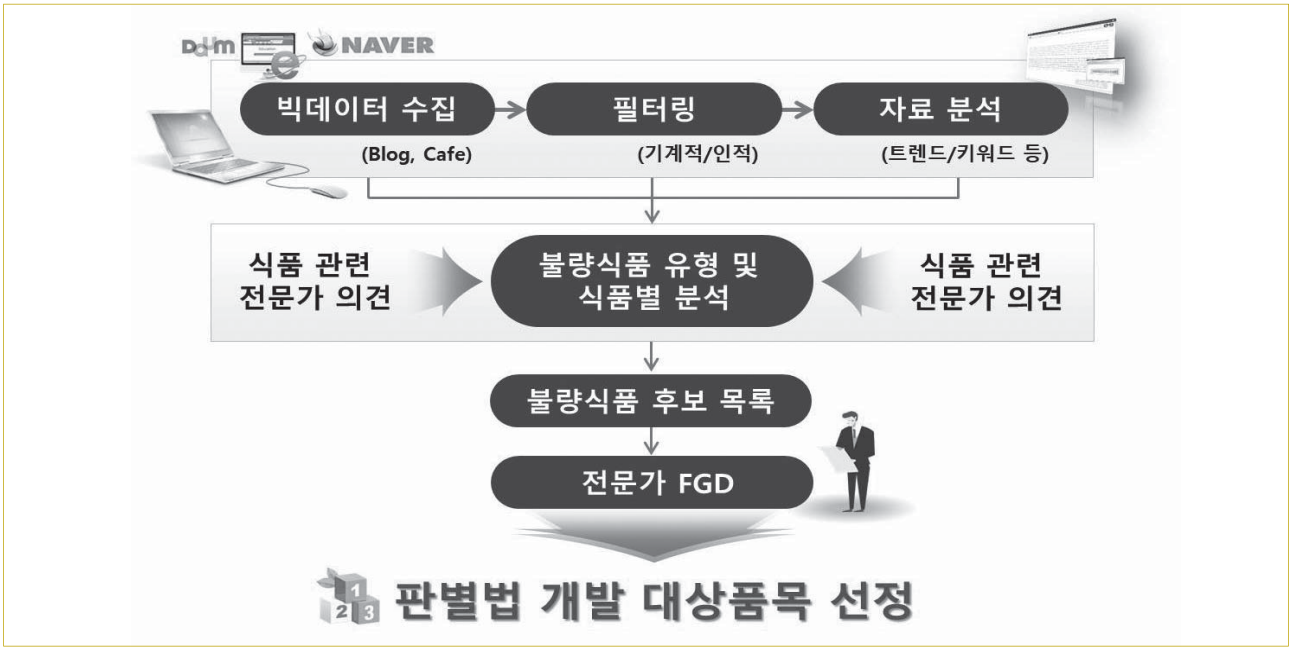


Fig. 1. 빅데이터 분석과 FGD를 통한 불량식품 판별법 개발 대상품목 선정 체계

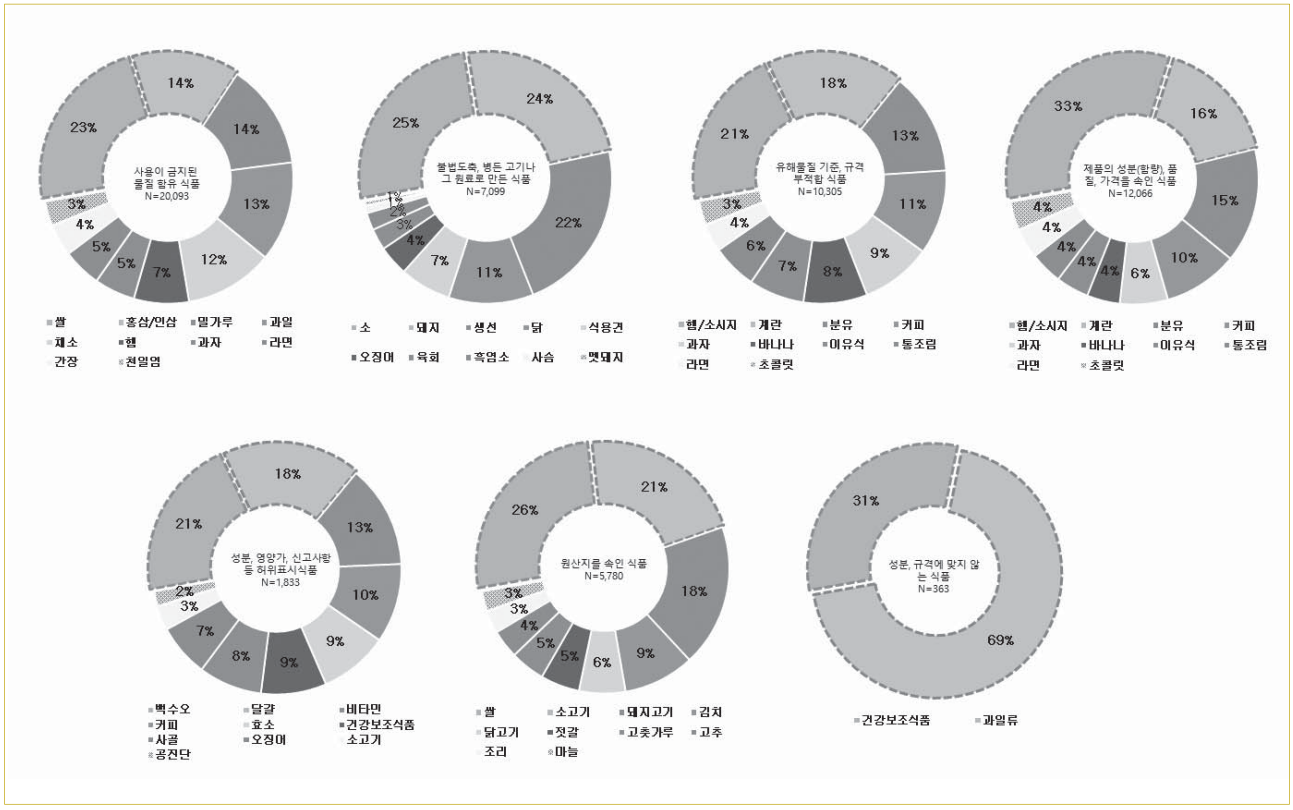


Fig. 2. 소셜 빅데이터 분석을 통한 불량식품 유형별 다빈도 10위 식품

품과 분석연구가 필요한 식품의 우선순위를 결정하고, 개발이 요구되는 판별법을 협의하였다.

3. 연구결과

3.1. 소셜 빅데이터 상세분석을 통한 사전탐색적 불량식품 판별법 개발 후보 도출

1) 불량식품 유형별 키워드 발생 다빈도 식품

7개 불량식품 유형에 대한 소셜 빅데이터를 상세분석하여 키워드 발현량이 높은 식품을 10위까지 도출한 결과는 Fig. 2와 같다. 사용이 금지된 물질이 함유된 식품유형에 해당하는 키워드가 쌀이 4,641건으로 발생빈도가 가장 높게 나타났으며, 홍삼/인삼 2,803건, 밀가루 2,745건 순으로 분석되었다. 불법도축, 병든 고기나 그 원료로 만든 식품유형에서는 소비자들이 주로 섭취하고 있는 어·육류에 대한 버즈 발생빈도가 높게 조사되어 소(1,788건), 돼지(1,713건), 생선(1,594건)순으로 분석되었다. 제품의 성분(함량), 품질, 가격을 속인 식품유형과 성분, 영양가, 신고사항 등 허위표시 식품유형에서 백수오가 각각 3,946건 및 382건의 발생빈도를 보여 2015년 이슈가 되었던 가짜 백수오 사건에 의해 키워드가 발생하는 다빈도 식품으로 조사된 것으로 판단되었다. 원산지를 속인 식품유형에서는 쌀이 1,504건으로 조사되었으며, 소고기, 돼지고기, 김치 순으로 키워드 발생빈도가 높은 것으로 나타났다. 불량식품 유형별 빅데이터 상세분석결과, 소비자들이 주식으로 섭취하

는 '쌀', '밀'과 단백질 급원 식품으로 주로 이용되는 '소고기', '돼지고기' 및 반찬류인 '김치'에 대한 우려가 높은 것으로 나타났으며, 건강관리를 위해 소비하는 건강기능식품의 한 종류인 '홍삼/인삼', '백수오'등에 대한 우려가 높은 것으로 판단되었다

2) 다빈도 발생 불량식품의 워드 클라우드 분석

불량식품 수집 키워드 발현량이 높은 식품과 7개의 불량식품유형에 중복되어 발현되는 주요식품에 대해 워드 클라우드 분석을 실시하고, 해당 식품에 대해 소비자가 우려하는 내용을 정리한 결과는 Table 2와 같다. '쌀'은 3개의 불량식품 유형에서 발생빈도가 높은 식품으로 조사되었으며, 그 중 사용금지 물질 함유 식품유형에서 4,641건의 키워드가 발생되었다. GMO 쌀의 제초제 성분과 농약, 중금속 등에 오염된 쌀이 사용되는 점을 우려하는 것으로 나타났다. 또한, 원산지를 속인 식품유형에서 발생된 쌀에 대한 우려로 식당에서 사용하는 경우 수입산과 국내산을 혼합하여 소비자를 속이는 행위에 대한 의견이 조사되었다. '홍삼'은 사용금지 물질 함유 식품유형에서 2,803건의 키워드가 발현되어 가공품의 원료로 사용되는 홍삼의 화학첨가제와 잔류농약에 대한 우려가 많은 것으로 나타났다.

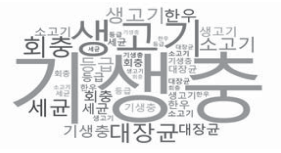
3.2. FGD를 통한 불량식품 판별법 개발 후보 식품의 우선순위

소셜 빅데이터 분석을 통해 버즈 발생빈도가 많았던 식품과 불량식품 유형에 중복되어 나타나는 식품을 포함하여 전문가들이 인식하는 불량식품 품목과 판별법 개발이 필요한 우선순위 불량식품 후보 목록을 Table 3과 같이 정리하였다. 총 12개의 불량식품 목록이 선정되었으며, 홍삼의 우선순위가 가장 높은 것으로 나타났다. 홍삼의 경우 연근 수(예. 6년근 확인 등)에 대한 지속적인 연구가 필요하다는 의견이 제시되었으며, 색소나 물엿 등의 첨가제 혼입 시 자연 발생되는 갈색 성분과 인위적인 색소 혼입을 구별하기 어려워 이에 대한 판별법이 필요한 것으로 평가되었다. 고추와 고춧가루는 두 가지 형태 모두 우선순위가 높은 식품으로 선정되었으며, 색소 혼입에 대한 판별법이 필요한 것으로 평가되었다. 특히 천연색소의 경우 고추 색소와 구별하기 어려워 이에 대한 판별법이 요구되었다. 다양한 식품이 속해 있는 건강기능식품도 우선순위가 높은 불량식품으로 도출되었는데 주로 이종원료의 혼입, 색소 첨가 등에 대한 판별법이 필요한 것으로 평가되었다. 전복은 수입산이라도 국내에서 3-6개월간 양식할 경우 국내산 표기가 가능하여 원산지 판별이 어려우므로 유통 이력 구조를 토대로 원산지를 판별하는 방법이 필요하다는 의견이 제시되었다.

4. 요약 및 결론

불량식품 유형에 따른 키워드를 기반으로 빅데이터를 수집하고 발생빈도가 높은 식품에 대한 소비자의 다양한 의견을 워드 클라우드로 분석하였다. 이를 통해 도출된 주요식품은 전문가 FGD(Focus group discussion)를 실시하여 사전탐색적 불량식품 판별법 개발 후보 품목의 우선순위를 결정하고 개발이 요구되는 판별법에 대해 협의하였다. 빅데이터를 기반으로 분석한 7개 불량식품 유형에 대하여 키워드 발현량이 높은 식품과 전문가 FGD 결과를 통해 도출된 불량식품 대상품목을 포함하여 총 12개의 판별법 개발용 불량식품 후보 목록이 선정되었으며, 홍삼의 우선순위가 가장 높은 것으로 나타났다. 홍삼의 경우 연근 수에 대한 지속적인 연구가 필요하다는 의견이 제시되었으며, 색소나 물엿 등의 첨가제 혼입 시 자연 발생되는 갈색 성분과 인위적인 색소 혼입을 구별하기 어려워 이에 대한 판별법이 필요한 것으로 평가되었다. 고추와 고춧가루는 색소 혼입에 대한 판별법이 요구되며, 건강기능식품의 경우 주로 원료의 이종혼입, 색소 첨가 등에 대한 판별이 필요한 것으로 모두 우선순위가 높은 식품으로 선정되었다. 또한 소비자들이 주식으로 섭취하는 '쌀', '밀'과 단백질 급원 식품으로 주로 이용되는 '소고기', '돼지고기' 및 다소비 반찬류인 '김치'와 원료로 사용되는 '고추/고춧가

Table 2. 불량식품 유형별 워드 클라우드 빈도 분석

식품명	불량식품 유형	발생빈도	워드 클라우드	
쌀	사용금지 물질 함유 식품	4,641		· 농약, 중금속(카드뮴), 비소 등에 오염된 쌀에 대한 우려 · GMO쌀에 사용되는 제초제 성분 글리포세이트 우려
	제품의 성분(함량), 품질, 가격을 속인 식품유형	1,166		· 햅쌀 진위 여부에 대한 우려 · 음식점에서 사용되는 쌀의 품질에 대한 우려 · 쌀등급 표시제의 실효성
	원산지를 속인 식품	1,504		· 식당에서 국산 및 수입쌀 혼합판매에 대한 우려 · 쌀 혼합 후 국산으로 표기
홍삼/인삼	사용금지 물질 함유 식품	2,803		· 어린이 홍삼제품에 사용되는 화학첨가제에 대한 우려 (산탄검, 젤란검, 유화제, 시클로덱스트린 등) · 홍삼의 잔류 농약 성분에 대한 우려
	제품의 성분(함량), 품질, 가격 등을 속인 식품	1,788		· 성분을 속인 가짜 홍삼 및 짝퉁 홍삼 브랜드에 대한 우려 · 홍삼 품질에 대한 우려
소고기	제품의 성분(함량), 품질, 가격 등을 속인 식품	1,951		· 등급이나 수입산을 섞어 파는 행태에 대한 우려 · 품종이나 부위명을 허위로 표기하여 판매되는 형태 · 2014년 방송으로 나온 주사기 마블링 제조에 대한 언급
	불법 도축, 병든 고기나 그 원료로 만든 식품	1,788		· 생고기 섭취시 기생충 및 세균 감염 우려
	원산지를 속인 식품	1,237		· 음식점이나 판매점에서 수입산을 한우로 속여 파는 행태에 대한 우려 · 명절에 소고기 원산지 둔갑에 대한 의구심(일반 판매점, 선물세트 등)

식품명	불량식품 유형	발생빈도	워드 클라우드	
과자	사용금지 물질 함유 식품	1,059		· 아이들이 주로 먹는 과자에 수많은 첨가제에 대한 부정인식 · 과자 원료에 대한 부정적 인식 (GM밀, GM옥수수, 타르/카라멜 색소 등)
	유해물질 기준 · 규격 부적합 식품	941		· 세균 기준치 초과 제품 및 비허용 첨가물 사용 제품에 대한 우려
건강보조식품	성분, 영양가, 신고사항 등 허위 표시 식품	155		· 건강보조식품의 효능을 과대, 허위 과장하는 경우가 주로 언급(다이어트 보조제, 비타민, 홍삼 등 언급) · 중국산 밀반입 건강보조 식품에 대한 우려(성기능 식품, 한약재 관련 등)
	성분 · 규격에 맞지 않는 식품	112		· 함량미달 및 기준치 이하의 건강보조식품에 대한 우려

Table 3. FGD를 통한 판별법 개발용 우선순위 불량식품 후보 목록

순위	식품명	도출 근거		개발 요구 판별법
		빅데이터 분석	전문가 의견	
1	홍삼	○		· 홍삼의 연근 판별법 · 카라멜 색소 및 물엿 혼입 판별법
2	고춧가루/건고추	○		· 색소 혼입 판별법
3	건강기능식품	○		· 복분자 색소 첨가 및 혼입 판별법 · 이종 오가피 혼입 판별법
4	고추가공품		○	· 다대기 이종원료 사용 판별법 · 색소 혼입 판별법
5	전복		○	· 전복의 원산지 위 · 변조 판별법
6	쌀	○		· 햅쌀 진위 여부 판별법
7	밀가루	○		· 잔류농약 분석법
8	치즈		○	· 모조 치즈와 천연치즈의 판별법
9	김치	○		· 김치의 원산지 위 · 변조 판별법
10	쇠고기	○		· 부위 판별법
11	꿀		○	· 꿀의 원산지 위 · 변조 판별법
12	흑마늘, 과자		○	· 색소(타르/카라멜) 혼입 및 성분 함량 미달 판별법

루'에 대한 우려가 높은 것으로 나타났으며, 건강기능식품 중 '홍삼/인삼', '백수오' 등 식품안전관리에 대한 소비자 및 전문가들의 우려가 높은 것으로 판단되었다.

불량식품 판별법 개발을 위해 소비자의 방대한 의견을 수집, 분석하고 개발이 요구되는 판별법에 대한 전문가의 의견을 통하여 판별법 개발 대상품목을 도출함으로써 향후 불량식품의 사전관리 및 현실적인 정책대응을 위한 기초자료로 활용될 수 있을 것으로 판단된다.

감사의 글

본 연구는 2017년도 식품의약품안전처의 연구개발비(17162미래사065)로 수행되었으며 이에 감사드립니다.

참고 문헌

1. 전향숙, 하상도, 이헌주, 조승룡, 김건희, 강건욱. 국내외 불량식품의 발생 및 대응현황. 식품과학과 산업, 2014, 6, 3-12.

2. 장현후, 장수경, 조건희, 김주란, 조천호, 권기성, 전향숙. 소셜 빅데이터를 활용한 불량식품에 대한 국내 소비자 인식 분석 -성분, 품질, 가격을 속인 식품유형을 중심으로-. 한국식품위생안전성학회지, 2017, 3, 60-67.

3. 장현후, 우소영, 허성은, 정소영, 김주란, 권기성, 전향숙. 소셜 빅데이터를 활용한 원산지 위 · 변조 불량식품에 대한 국내 소비자 인식 분석. 한국식품위생안전성학회지, 2017, 3, 68-74.

4. Roh, S. K. Big data analysis of public acceptance of nuclear power in Korea. Nuclear engineering and technology, 2017, 49, 850-854.

5. Lee, S. J., Kim, K. M., Cho, S. M., Lee, J. S., Choi, S. S. Analysis of disaster and safety management SOP manuals in social welfare facilities. Crisisonomy, 2017, 5, 1-14.

6. Park, S. J., Li, H. M., Park, H. W. Can sustainability information on corporate websites reflect CSR strategies?: A semantic network analysis of top companies in south Korea and China. J. Korean data analysis society, 2017, 1, 43-54.

7. 김예슬, 김병주. 누리과정에 대한 신문 사설의 의미 연결망 분석. 교육행정학연구, 2017, 3, 1-32.

8. 박창현. 누리과정 정책 수요에 대한 빅데이터 분석: 워드 클라우드와 의미망 분석을 중심으로. 유아교육연구, 2017, 3, 73-91.

9. Raghupathi, W., Raghupathi, V. Big data analytics in healthcare: promise and potential. Health information science and systems, 2014, 2, 3.

10. Spink, J., Moyer, DC. Defining the public health threat of food fraud. J. Food Sci, 2011, 76, 157-163.