논문 15-40-06-23 The Journal of Korean Institute of Communications and Information Sciences '15-06 Vol.40 No.06
http://dx.doi.org/10.7840/kics.2015.40.6.1166

빅 데이터 분석 기반 농 식품 위해인자 신속관리 방법

박 현◆°, 강 성 수*, 정 훈*, 김 세 한*

Rapid Management Mechanism Against Harmful Materials of Agri-Food Based on Big Data Analysis

Hyeon Park◆°, Sung-soo Kang*, Hoon Jeong*, Se-Han Kim*

요 약

단순 바코드 또는 포장 내용물 단위의 이력추적, 농 식품의 저장 창고나 배송차량의 일부 정보 추적, 직감에 의한 원격 환경 조정 등을 통해 농 식품의 위해인자를 차단하려는 노력들이 있었다. 그러나 이러한 시도는 선택적인 정보수집 및 불충분한 정보량, 현실과 수집 시점 간 시간차에 따른 정보 왜곡의 문제점 및 각 유통 기업의 자체 독립적인 정보망으로 인하여 생산지로부터 소비자까지의 총체적인 위해인자 차단이 어렵다. 본 논문에서는 농 식품의 생산지뿐만 아니라 전주기상의 주요 유통 거점, 소비지까지 정형, 반 정형, 비정형의 다양하고 대규모의 농 식품 유통 정보를 이용하여, 위해인자 발생의 실시간 상황이나 예측, 추적을 통하여, 위해인자 파급 차단과 예방을 위한 농 식품의 위해인자 신속 관리 방법을 제안한다. 제안방법은 빅 데이터 클러스터 기반, 실시간으로 정보를 수집하고, 위해인자 상황인지, 위해인자 발생 예측, 위해인자 발생지 추적 분석을 통해 위해인자를 차단하고 파급을 예측하며, 그 결과를 가시화하여 신속하게 위해인자를 관리 할 수 있도록 한다.

Key Words : Big data, Harmful material, Agri-food, Rapid management, Hadoop

ABSTRACT

There were the attempts to prevent the spread of harmful materials of the agri-food through the record tracking of the products with the bar code, the partial information tracking of the agri-food storage and the delivery vehicle, or the control of the temperature by intuition. However, there were many problems in the attempts because of the insufficient information, the information distortion and the independent information network of each distribution company. As a result, it is difficult to prevent the spread over the life-cycle of the agri-food using the attempts. To solve the problems, we propose the mechanism mainly to do context awareness, predict, and track the harmful materials of agri-food using big data processing.

Ⅰ. 서 론

식품 위해인자 검출은 식품 공정에 의해 시료에서 검출까지 1~3일 이상, 수 일정도 소요되며 검출과정에서 배양에 의한 증균 과정이 약 12~48시간 정도 대부분의 시간을 차지하여 신속한 검출이 이루어지지 못하고 있다. 이로 인해 식품 위해인자 발생의 실시간 상황이나 예측 등이 어려워 위해인자가 발생 하였음

※ 본 연구는 MSIP/IITP의 IT·SW융합산업원천기술개발산업의 일환으로 수행하였음. [R0101-15-0063(10044580),농식품의 안전한 유통을 위한 위해인자 신속관리시스템 개발].
◆° First and Corresponding Author : Electronics and Telecommunications Research Institute, hpark@etri.re.kr, 정회원
* Electronics and Telecommunications Research Institute, {(sskang, hjeong, shkim72)@etri.re.kr}, 정회원
논문번호 : KICS2015-01-017, Received January 31, 2015; Revised June 19, 2015; Accepted June 19, 2015

에도 불구하고, 이후 파급을 신속히 차단하지 못하여 소비자에게 많은 피해를 주고 있다. 이를 극복하기 위해 단순 바코드 또는 포장 내용물 단위의 이력추적, 농 식품의 저장 창고나 배송차량의 일부 정보 추적, 직감에 의한 환경 원격 조정 등을 통해 위해인자를 차단하여 소비자를 보호 하려는 시도들이 있었다. 예로써, 농 식품 배송차량 (DTG (Digital Tacho Graph)가 장착됨) 에 부착되어 있는 GPS 및 온도센서를 이용하여 온도변화에 따른 미생물의 증식을 단편적으로 예측함으로써 단지 위험 구간에 대한 정보를 관리자와 운전자에게 경고 및 개선 조치사항을 제공하고 있다.

그러나 이러한 시도는 선택적 정보수집 및 불충분한 정보량, 현실과 수집 시점 간 시간차에 따른 정보 왜곡의 문제점 및 각 유통 기업의 자체 독립적인 정보망으로 인하여 생산지로부터 소비자까지의 총체적인 위해인자 차단이 어렵다.

최근 농 식품의 품질제고를 위한 신선유지를 위해 유통과정과 기간을 단축시키는 추세에 있어, 농 식품의 위해인자 (위해 미생물 (예: 살모렐라), 위해 독소 (예: 아플라톡신) 등)의 신속한 검출은 그 중요성이 더 강조 되고 있다. 이를 위해서는 위해인자 검출 시간을 현저히 축소시킴으로서 실시간으로 위해인자에 대한 감시가 가능해야 한다. 농 식품 위해인자 검출 장비의 발전으로 근래에는 검출 시간을 실시간 수준으로 단축시킬 수 있는 (현장에서 2시간 내 위해인자를 검출) 장비들이 개발 되고 있다. 이러한 위해인자 검출 장비들의 정보를 기반으로 대규모 농 식품 유통정보를 이용하여, 실시간으로 위해인자를 차단하여 소비자를 보호하기 위한 (피해를 최소화하기 위한) 위해인자 신속관리 기술이 요구된다. 따라서 신속한 농 식품 위해인자 관리를 위하여, 농 식품의 생산지로부터 소비자까지의 총체적인 감시와 이를 위한 농 식품의 유통 전 과정에서 다양한 정보가 요구 된다.

즉, 정확한 실시간 위해인자 발생 예측을 위해서는 이러한 검출 정보나 배송 량뿐 만 아니라 과거에서 현재 시점까지의 이력 데이터 기반, 위해인자 패턴이나 현재의 배송 환경 (온도, 습도 등), 그리고 소셜 네트워크에서 회자되고 있는 실시간 정보 등, 다양한 정보를 바탕으로 한 분석이 요구된다.

본 논문에서는 농 식품의 생산지뿐만 아니라 전주기상의 주요 유통 거점, 소비지까지 정형, 반 정형, 비정형의 다양하고, 대규모의 농 식품 유통 정보를 이용하여, 농 식품의 위해인자의 신속한 관리를 위한 방법을 제안한다. 제안방법은 실시간으로 농 식품 위해인자 상황인지, 위해인자 발생 예측, 위해인자 발생지 추적을 통해 위해인자를 차단하고 파급을 예측하며, 그 결과를 가시화하여 신속하게 위해인자를 관리 할 수 있도록 하는 방법이다. 이를 위해, 농 식품의 전 주기 상에 도출 될 수 있는 다양한 정보 수집의 연구가 선행 되었으며, 이러한 다양한 정보를 바탕으로 상황인지, 예측, 추적 분석이 수행 되었다. 이러한 대규모 데이터는 큰 데이터를 효율적으로 접근하는 방법으로 클라우드 기반의 분산병렬 컴퓨팅 방법이 저렴한 비용과 효율적인 처리방법으로 많이 연구되어 활용되고 있으며[1], 이에 본 논문에서는 빅 데이터 (Big Data) 하둡 클러스터를 통한 정보수집, 저장과 분석을 수행하였다.

본 논문은 다음과 같이 구성되어 있다. II장에서는 농 식품의 위해인자의 신속한 관리를 위한 전체적인 프레임워크 (framework)를 기술하고, III에서는 농 식품 위해인자 신속관리를 위한 분석 시 요구되는 다양한 위해인자 정보에 대해서 언급하고, IV에서는 농 식품 위해인자 상황인지, 예측, 추적 분석 기법 및 실험에 대해서 설명한다. V에서는 분석된 결과들의 가시화에 대하여 기술하며, VI장에서는 위해인자 신속관리 수행 결과를 바탕으로 본 논문을 요약 한다.

Ⅱ. 위해인자 신속관리 프레임워크

최근, 인터넷 연결과 그로 인해 도출되는 데이터를 통해 커다란 가치 창출을 얻기 위하여 빅 데이터 기술에 관심이 집중되고 있으며, 빅 데이터 원리들을 구현하려고 한다[2]. 빅 데이터는 일반적으로 데이터 규모가 크고, 다양한 데이터를 실시간 처리를 통해 훌륭한 가치 창조 및 중요한 예측 분석을 제공하는 의미에서 4V (Volume, Variety, Velocity, Value) 로 정의 되고 있다. 하둡 (Hadoop) 이라는 생태계 (Eco-system)를 기반으로 수행되며, 정형, 반 정형, 비정형의 대량의 데이터를 수집하여 데이터의 중복 분산, 분산된 네트워크 클러스터에서 병렬로 처리하여 단시간 및 가치 있는 정보의 추출이라는 기술적 의의를 부여하고 있다.

또한 메모리 기반으로 데이터 스테이지를 동적으로 적절한 메모리로 이동시키는 방법[3]들이 제안되고 있어 대규모 데이터라 하더라도 스토리지의 빠른 접근 (access) 시간과 적은 에너지 소비로 신속한 위해인자 관리가 용이하다.

그림 1 은 농 식품의 신속한 위해인자 관리를 위한 전체 프레임워크로써, 생산지로부터 소비지까지, 농 식품 전주기 (Life Cycle) 상의 위해인자 실시간 신속관리 시스템으로, 생산지, 저장고, 가공업체, 물류센터,

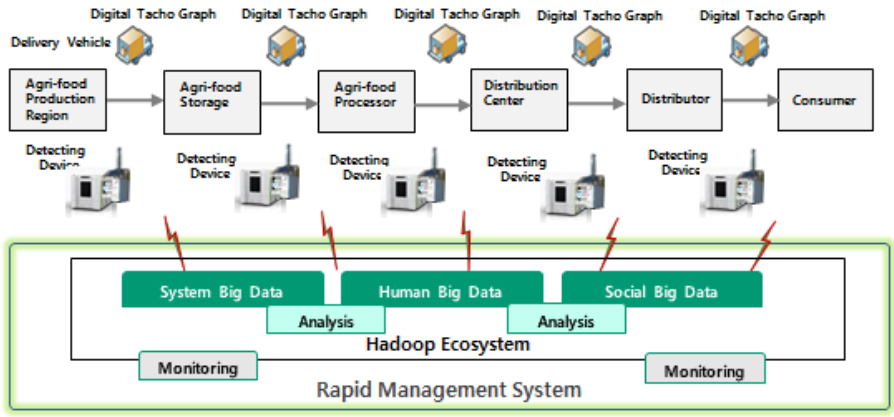

그림 1. 위해인자 신속관리 프레임워크
Fig. 1. Framework for the rapid management against harmful materials of agri-products

판매처, 소비지에 이르는 유통의 전 구간을 대상으로 한다.

신속한 위해인자 관리를 위해서는 우선, 생산지부터 소비지까지 위해인자 정보를 실시간으로 수집한 뒤 이를 하둡 기반의 분산 파일에 저장한다. 이러한 정보를 모두 활용, 분석을 위해 이용할 수 있는 정보의 원천을 다양하게 확보하여 예측 모형을 개발한다. 분산 저장된 정보를 기반으로 거점과 물품 배송 량에 관한 표준화를 수행하거나 지표들 (features) 간의 상관관계 분석, 이력 데이터를 바탕으로 위해인자 패턴 분석을 통한 상황, 예측, 추적 결과를 도출 한다. 이를 위해 빅 데이터 기반의 정보 저장, 분석이 하둡 생태시스템 (Hadoop Ecosystem)에 의해 수행된다. 하둡은 분산 파일 시스템 (distribution file system) 과 분산 컴퓨팅을 위한 맵리듀스 (MapReduce)를 포함하여 개발된 분산병렬처리 시스템이다[4]. 즉, 다양한 정보는 하둡의 생태시스템 (수집을 위한 Flume[5], Sqoop 등)에 의해 수집되며, 저장 시스템 (HDFS, HBase 등)에 의해 분산 저장되며, 수집된 정보들은 R이나 Map/Reduce의 분석 엔진을 통해 분석 된다. 분석된 결과는 위해인자 신속관리 시스템의 모니터링에 의해 가시화 된다.

Ⅲ. 위해인자 정보 수집

농 식품의 위해인자 신속관리를 위해 분석에 사용되는 수집정보는 유통 과정에서의 주요 거점 (유통단계) 들에서 실시간으로 검출 장비에 의한 직접지표 (위해인자 검출 장비에 의해 직접 검출된 검출정보)와 농 식품의 전체 유통 과정에서 발생되는 간접지표 (검출 장비에 의한 정보가 아닌 온도, 습도, SNS 등의 위해인자 분석 시 도움이 되는 정보) 정보로 나눈다. 표 1 에서는 농 식품 위해인자 상황인지, 예측, 추적 분석에 이용되는 다양한 수집정보를 정리하였다. 이러한 데이터는 위해인자 검출장비, 배송차량, 온도 유지 설비 등의 센서 노드로부터 수집 되는 시스템 빅 데이터와 SNS, 웹 로그 (VOC, 사용자 불만 등), 인터넷

표 1. 분석을 위한 위해인자 정보
Table 1. Information of the harmful materials of agri-products for the analysis

Information Category		Information Classification	Aggregation Method
Real-time Information	Direct Indication	On-site detection info.	API
	Indirect Indication	SNS info.	Web Crawling
		DTG(temp., humidity, GPS)	API
		Weather info.	API/Web Crawling
		Storehouse	API
		Safety news	Web Crawling
Non-Real Time Information		Harmful material characteristics	Off-line
		Distribution base info.	Off-line
		Harmful material profiles	Off-line
	Direct Indication	Accumulated detection info.	API

뉴스와 같은 소셜 빅 데이터, 그리고 기상청 날씨, 농림수산식품교육문화정보원과 같은 공공 기관에서 제공해주는 휴먼 빅 데이터를 포함한다.

수집정보는 실시간으로 수집되는 정보와 비 실시간으로 수집되는 정보로 분류되며, 이는 다시 직접지표와 간접지표들로 나눌 수 있다. 검출 장비와 DTG 정보는 API (Application Protocol Interface) 를 통하여 연계되며, 기상청 데이터는 API 또는 Web Crawling을 통하여 수집한다. 또한 위해인자의 리스크 프로파일과 SNS 단문은 실시간 검색을 할 수 있는 검색 사이트를 통하여 수집하도록 하며, 가공된 데이터는 핵심 단어를 사용한 Web Crawling 방식으로 수집하는 안전뉴스와 같은 형태이다.

Ⅳ. 위해인자 분석

앞 장에 기술된 빅 데이터 정보를 기반으로 위해인자의 신속한 관리를 위해서는 생산지, 유통, 소비지까지 전주기 추적을 통해 위해인자 차단 및 위해인자 발생 위험도 예측을 위한 실시간 위해인자 발생 예측, 위해인자 확산 예측, 위해인자 발생지 추적, 위해인자 발생 상황인지가 이루어져야 한다. 다음의 각 절에서는 각각의 분석 기법에 대하여 기술한다.

4.1 위해인자 예측

위해인자 예측에는 실시간으로 검출장비에 의해 검출되어 이미 발생된 위해인자의 확산 지역을 예측하는 확산 예측과 아직 위해인자가 발생하지는 않았지만 향후 위해인자가 발생될 것으로 예상되는 발생 예측이 있다.

4.1.1 위해인자 발생 예측

위해인자 발생과 관련되는 다양한 변수를 고려하여 예측하되 상관성이 있는 거점을 통한 위해인자 발생을 추정하였다. 이산 형 반응변수 (discrete response variable) 인 위해인자 발생에 대한 모델링으로써 다중 로지스틱 회귀분석을 사용하였다. 위해인자 발생 여부가 0,1 로 binary이며 발생할 확률은 p이며, 설명변수는 거점, 시각, 위해인자, 온도 등이 설명 변수로 사용되었다 (식품안전 사건사고와 기후 요소와의 관련성 분석 결과, 강수량 (0.48) 과 음의 상관관계를, 그리고 최저기온 (0.45) 과는 양의 상관관계[6]가 있어, 온도는 설명변수로써의 의미를 갖는다.). 이때 설명변수의 계수는 위해인자 발생에 각 설명변수가 미치는 영향의 크기를 나타낸다. 다중 로지스틱 회귀분석의 기본 모형은 (1) 과 같이 주어진다.

$$logit(p) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_n \times x_n \quad (1)$$

이 경우 p 의 로짓(logit)은 $logit(p) = log\ \frac{p}{1-p}$ 이다 (p 의 추정량이 언제나 0과 1사이가 될 수 있는 함수로의 변형). 그리고 사건 발생 확률인 p 의 추정은 $p = \frac{e^{logit(p)}}{1+e^{logit(p)}}$ 와 같다. 여기에서 $p = \frac{e^{\beta_0+\beta_1\times x_1+\beta_2\times x_2+\cdots+\beta_n\times x_n}}{1+e^{\beta_0+\beta_1\times x_1+\beta_2\times x_2+\cdots+\beta_n\times x_n}}$ 이다.

거점단위의 실험 검출 정보를 기반으로 빅 데이터의 하둡 생태 체계의 R[7,8] 기반 위해인자 발생을 추정하도록 그림 2 와 같이 구현하였다.

위에서 도출된 위해인자 발생 확률 p를 기반으로 분석 모형의 타당도 (validity)의 정확도 (accuracy)를 산출하기 위해 Receiver Operating Characteristics Curve (ROC Curve)를 사용하였다. 그림 3의 x축은 비 위해인자 오 예측 율 (False Positive Rate), y 축은 위해인자 정 검출 율 (True Positive Rate)을 나타내며, 이는 실제 위해인자가 아닌 경우를 위해인자로 예측하는 비율 대비 실제 위해인자를 위해인자로 검출

```
>predict.logit = predict(result.logit, newdata=aulc_new)
>aulc_new$Div = "Prediction"
>if(nrow(aulc_new)==0)
error2=TRUE
>write.csv(aulc_new[, c(31, 32, 24, 7, 1, 2, 8, 30)], file=result_file,
quote=FALSE, row.names=FALSE
```

그림 2. 위해인자 발생예측 분석 모형 실행 code
Fig. 2. Source code to implement the model of the prediction of the harmful material

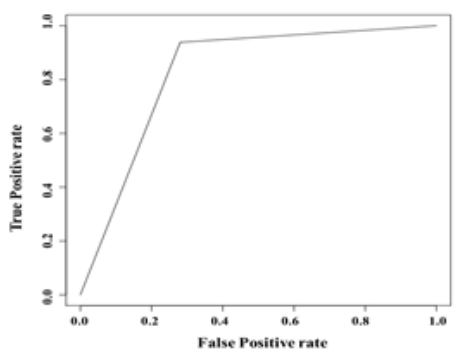

그림 3. 위해인자 발생 예측 모형의 ROC
Fig. 3. Receiver Operating Characteristics chart for the prediction of the harmful material outbreak

하는 비율의 변화를 관찰할 수 있는데, 모형이 데이터를 분석하여 위해인자로 잘 예측할수록 위해인자 정 검출 율이 높아지지만 반대로 비 위해인자를 위해인자로 예측하는 오류도 높아질 수 있다. 그러므로 좋은 모형이 되기 위해서는 높은 위해인자 정 검출 율과 낮은 비 위해인자 오 예측 율을 가져야한다. 실험 결과의 ROC 차트는 그림 3과 같으며, 본 실험은 비교적 ROC 커브가 그래프가 좌측 상단으로 치우침으로써 비교적 좋은 위해인자 발생 예측 모형임을 알 수 있다.

4.1.2 위해인자 확산 예측

위해인자 확산 지역을 예측하는 것은 검출된 위해인자를 기반으로 해당 위해인자의 파급이 예상되는 지역을 예측하여, 예측된 지역에서 위해인자가 발생하지 않도록 차단하는 것을 목적으로 한다.

사용한 데이터는 4.2 절의 위해인자 발생지 추적에서의 거점 단위 검출데이터 기반, 거점 간 유사도를 이용하며, Link Prediction을 수행하였다. 관측된 네트워크에서 노드 간 연결 정보를 바탕으로 향후 발생할 수 있는 연결을 예측하는 것으로, 네트워크에서 노드 간 유사도는 이전에 구한 품목으로 부터 얻어지는 거점 간 관계 또는 Jaccard Similarity를 사용한다. Jaccard Similarity 는 두 대상이 서로 공통된 항목을 얼마나 갖고 있는지를 측정 하는데, 식 (2) 에서 거점 A, B 에 관해 거점-품목으로 구성되어 분석의 자료로 적절 할 수 있다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Link Prediction 수행은 그래프 기반 분석을 통해 이루어지며, 기존의 그래프와 유사도 그래프의 차이를 바탕으로 Link를 예측한다. 즉, 기존 거점 간 배송관계 또는 품목 유사도를 바탕으로, 한 그래프를 구축하고, 이 그래프로 부터 각 거점간의 유사도를 산출하여 그래프 구축한다. 두 그래프의 차이로 연관도가 높은 거점들의 관계를 찾으며 이 관계를 다시 그래프로 구축하며 그 결과가 Link Prediction의 결과로 도출 된다.

그림 4는 농 식품 유통 과정에서의 8개 거점에 대한 Link Prediction을 통한 위해인자 확산 예측의 결과를 나타내며, 이러한 결과는 실제 신속관리 시스템의 모니터링을 통하여 지역 지도로 맵핑되어 확산지역을 용이하게 파악 할 수 있다.

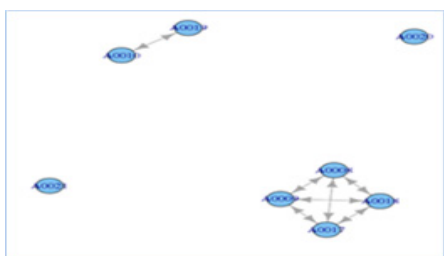

그림 4. 위해인자 확산 예측 그래프
Fig. 4. Graph for the prediction of the harmful material diffusion

4.2 위해인자 발생지 추적

위해인자 발생지 추적은 배송되는 식품 품목에서 검출된 위해인자에 대한 원래의 발생지를 알아내기 위한 추적 기능으로서 거점 간 관계 기반 Clustering 과 Graph mining 기법을 사용하여 각 거점들 사이의 관계를 도출하는 모델을 사용하여 현재 검출된 거점에 대한 이전의 배송지 거점 정보를 도출하여 결과를 제공하도록 하였다.

위해인자 발생 추적은 위해인자 발생 거점 A와 직접 관련이 없는 거점 B의 정보를 분석하여 검출 데이터를 수집함과 동시에 거점 B에 경보를 주는 기능으로서 단일 품목을 기준으로 하는 Food chain의 경우에도 푸드 체인의 복잡성과 DTG에서 수집되는 데이터 시점의 불일치 해결이 필요하며, 어떤 유통회사에 상관없이 통합 적용 가능하도록 해야 한다.

따라서 식품을 배송하는 각 거점 사이의 관계를 나타내는 거리 산출과 거점 사이의 관계를 분석하고 그래프로 표현하기 위하여 거점과 품목 사이의 테이블 형식을 사용하였다. 이러한 테이블에 표시되는 품목에 대한 유사도 계산 방식은 품목을 기준으로 하는 경우 (그림 5), 거점별로 거의 모든 품목을 취급하므로 거점사이의 차이가 거의 없다. 그렇지만 각 품목에 대한 사용 빈도, 즉 배송회수를 기준으로 하게 되면 (그림 6) 유사도는 1보다 작은 값이 되지만 배송회수×유사

Distribution Base	pork1	pork2	pork3
A0020	1	1	1
A0021	1	1	1
...			
A00XX	1	1	0

$$V = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \qquad similarity = \begin{bmatrix} 1 & 1 & 0.81 \\ 1 & 1 & 0.81 \\ 0.81 & 0.81 & 1 \end{bmatrix}$$

그림 5. 품목 기준인 경우, V: 품목, similarity: 각 거점별 유사도 수치
Fig. 5. Food item based case, V: Food item, similarity: similarity value for each distribution base

Distribution Base	pork1	pork2	pork3
A0020	2	1	3
A0021	4	2	1

A00XX	8	1	0

$$V=\begin{bmatrix}2&1&3\\4&2&1\\8&1&0\end{bmatrix} \qquad similarity=\begin{bmatrix}1&0.76&0.56\\0.76&1&0.92\\0.56&0.92&1\end{bmatrix}$$

그림 6. 배송횟수 기준인 경우, V: 배송횟수, similarity: 각 거점별 유사도 수치
Fig. 6. Number of shipping based case, V: Number of shipping, similarity: similarity value for each distribution base

도 (similarity)를 하면 유사도 값에 대한 수치는 얻을 수 있지만 배송 건수에 대한 민감도가 너무 크다.

위와 같이 배송횟수를 기준으로 하는 경우 건수에 민감하게 반응하므로 횟수가 많은 거점과 적은 거점과의 편차를 제거하는 표준화 작업을 하였다. 이 경우 품목별 표준화와 거점별 표준화 중에서 연구개발 수집된 거점 데이터가 불충분하므로 거점별 표준화를 추진하였다 (그림 7).

이어 임의로 설정한 서로 다른 K개 군집으로 클러스터링을 반복 수행하였으며 (본 논문에서는 K-means를 사용하였으며, K-means는 임의의 초기 값에서 정해진 수렴 값에 수렴할 때까지 추정, 최대화 과정을 반복하면서 중심을 찾는 알고리즘이다[9]. 자료 수집 및 가공 단계에서 시간적, 공간적으로 동기화 된 데이터는 몇 개의 클러스터로 나누어지는데, 일반적으로 n개의 데이터를 k(≥2)개의 클러스터로 나눈다면 O(kn)의 경우의 수가 존재하기 때문에 모든 경우를 비교하여 최적의 클러스터를 찾기는 어렵다[10].) 2개 거점이 동일한 클러스터에 들어간 횟수는 심각도로 사용하였다. 또한 위와 같이 생성된 거점-품목 행렬을 바탕으로 거점-거점 사이의 관계 행렬을 생성하여 Graph mining에 사용하였다. 이와 같이 생성된 그래프의 degree 및 centrality 근접성과 betweenness 계산을 하였다.

Distribution Base	pork1	pork2	pork3
A0020	-0.87	-0.58	1.09
A0021	-0.22	1.15	-0.22

A00XX	1.09	-0.58	-0.87

Distribution Base	Pork1	pork2	pork3
A0020	0.00	-1.00	1.00
A0021	1.09	-0.22	-0.87

A00XX	1.15	-0.46	-0.69

그림 7. 품목별 vs 거점별 표준화 기준
Fig. 7. Food item vs distribution base standardization

그림 8 과 같이 왼쪽의 행렬은 5개 거점의 발생지 추적의 관계를 나타내는 예를 나타낸 것이며, 맨 오른쪽의 그래프는 앞장의 데이터를 기반으로 분석한 결과의 거점간의 추적 관계 도를 나타낸다. 화살표 관계의 거점은 위해인자가 발생한 경우 그 관계를 검토하면 추적지를 용이하게 파악 할 수 있다.

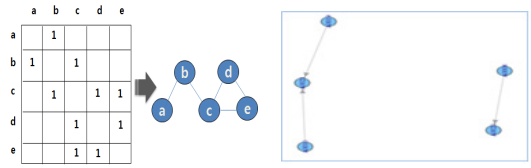

그림 8. Graph mining을 위한 Adjacency matrix 생성 및 분석 그래프
Fig. 8. Matrix for graph mining and the result graph

4.3 위해인자 발생 패턴

위해인자 발생 패턴 분석은 발생된 위해인자의 이력 정보를 기반으로 발생 시기와 지역 패턴을 분석하는 것으로서 기본적으로 시계열 패턴 분석 모델을 사용하였으며, 정확성 제고를 위하여 공공 기관이 제공하는 자료를 기반, Association Rule (AR)을 적용하였다.

위해인자 발생 패턴은 발생 정보를 시간 기준으로 정리하여 시계열 패턴을 확인하는 것으로 시계열 패턴 정보를 거점/계절/시간대역 (오전, 오후, 야간) 제품류/동일지역 배송여부/배송단계정보 등으로 범주화하여 품목화 시킨 후에 발생 패턴을 분석하였다. 다수의 거래 내역 모두에 포함된 품목의 관찰에 의한 규칙을 발견하기 위하여 모든 데이터를 범주 형이라고 가정하여 모든 규칙을 찾는 방식의 AR을 통하여 의미 있는 패턴 마이닝 (pattern mining)을 수행하였다. 이러한 AR의 기준으로 지지도, 신뢰도, 향상도를 사용하였다.

그림 9는 위해인자 발생 패턴의 시계열 그래프로써 시간 (월 단위) 별, 위해인자 발생 건수 (건)로, 시계열

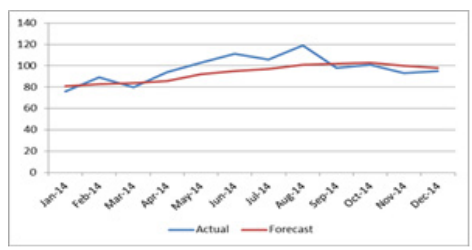

그림 9. 위해인자 발생 건수 트렌드 시각화
Fig. 9. Graph of the number of harmful material outbreak

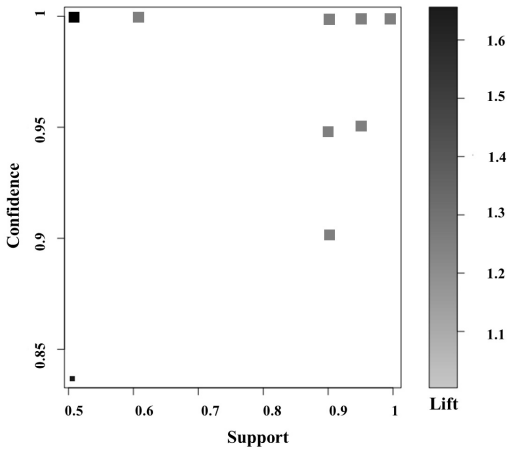

그림 10. 위해인자 발생 패턴, Association Rule 결과의 시각화
Fig. 10. Graph of the number of harmful material outbreak

모형에 의해 예측된 위해인자 건수 (빨간색)와 실제 발생 건수 (파란색)를 나타낸다.

또한 그림 10은 AR을 통한 패턴 마이닝으로 Support는 전체 데이터 중에서 해당 위해인자 관련 패턴이 발생하는 확률, Confidence는 위해인자 발생 패턴의 조건부 확률 (위해인자 발생에 대한 검출내역에서, 어떤 항목 x가 발생했을 때, y가 추가로 발생할 조건부 확률), Lift 는 위해인자 발생 패턴 내 항목 간 관련 정도 (검출내역에서 x를 고려한 y의 발생확률을 x를 고려하지 않은 y의 발생확률로 나눈 것으로)를 나타낸다.

4.4 위해인자 발생 상황

위해 인자 발생 현황은 지역별 당일의 식품 위해 인자 발생 건수와 위해인자 검출 내역에 대한 상세 정보를 주기적으로 (예: 4시간 내) 제공한다. 일일 현황에 대한 위해인자 발견과 분포를 나타내기 위한 것으로, 위해인자 검출장비의 검출 내역과 텍스트의 처리를 통하여 발생 상황에 대한 자료를 제공하는 것으로써, Text-based documents을 대상으로 텍스트 마이닝을 적용 하였다. 처리할 자료의 대상이 되는 문서(안전뉴스/리스크 프로파일 등)를 식별할 수 있는 용어를 추출 하고, 단순한 빈도 기반의 고빈도 용어의 추출보다 다른 가중치를 적용해 용어를 추출하고, 리스크 프로파일이나 안전뉴스 각각을 하나의 문서로 가정하고, 이 문서들에서 발견되는 용어와의 관계를 행렬로 정리한 Document Term Matrix를 생성했다. 단순 빈도 대신 식별력이 우수한 용어 선택을 위해 역 문헌 빈도를 적용 하였다.

V. 위해인자 분석 결과 가시화

앞 장에 서술된 빅 데이터 수집 및 위해인자 상황 인지, 예측, 추적 분석을 수행 후, 그 결과 들은 시스템 운용자, 농 식품 사용자들을 위하여 실시간으로 가시화 하여야 한다. 가시화를 위해 내부 하둡 클러스터 서버에 의한 가시화가 이루어지거나, 원거리의 통합 관리 시스템과의 RESTful API 연동을 통하여, 하둡 클러스터 독립 운용 및 타 시스템 상호 운용이 용이하도록 하고, 모니터링 서버를 통한 하둡 클러스터의 load 및 작업 내용, 장애 정보 감시, 하둡 클러스터들의 고 신뢰의 자원 관리를 수행하도록 하는 것이 필요하다.

그림 11 은 실시간으로 위해인자의 상황이나 각종 분석 결과를 모니터링 화면에 가시화 하는 화면을 보여준다. 앞장에서 분석 된 결과들은 지도상의 위치에 각각 맵핑되어 도시되며, 안전뉴스나 SNS 등을 통한 Text mining 결과 또한 가시화 된다.

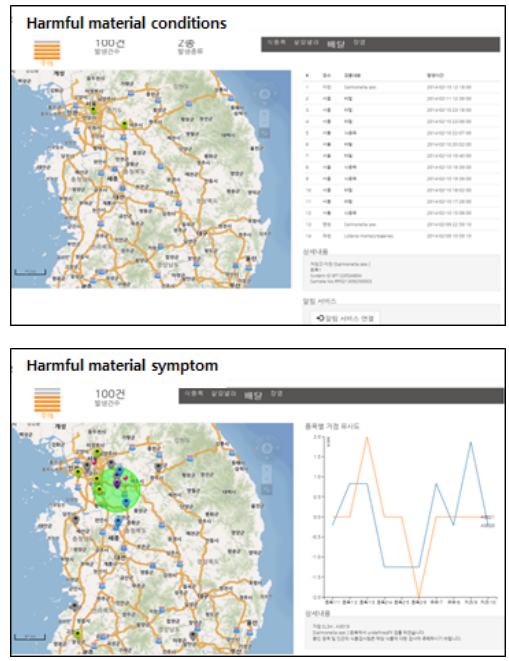

그림 11. 위해인자 예측 및 상황 결과 가시화 화면
Fig. 11. Visualization of the result of the harmful material prediction and conditions

Ⅵ. 결 론

본 논문에서는 농 식품의 생산지뿐만 아니라 전주기상의 주요 유통 거점, 소비지까지 정형, 반 정형, 비 정형의 다양하고, 대규모의 농 식품 유통 정보를 이용하여, 농 식품의 위해인자의 신속한 관리를 위한 방법을 제안하였다.

정확한 실시간 위해인자 발생 예측, 추적을 위해 검출 정보나 배송 량뿐 만 아니라 과거에서 현재 시점까지의 이력 데이터 수집 및 그 데이터 기반, 위해인자 패턴이나 현재의 배송 환경 (온도, 습도), 그리고 소셜 네트워크에서 회자되고 있는 실시간 정보 등, 다양한 정보를 바탕으로 빅 데이터 (Big Data) 하둡 클러스터 상에서 분석을 수행 하였으며, 클러스터내에는 실시간으로 정보를 전송하는 메카니즘이 추가되어, 결과를 가시화 하였다.

이는 기존의 선택적인 정보수집 및 불충분한 정보량, 현실과 수집 시점 간 시간차에 따른 정보 왜곡의 문제점 및 각 유통 기업의 자체 독립적인 정보망으로 인하여 생산지로부터 소비자까지의 총체적인 위해인자 차단의 어려움을 해결하여, 위해인자를 신속관리 할 수 있으며, 어떤 유통회사에 상관없이 통합 적용이 가능하도록 하였다.

References

[1] S. Park, Y. I. Kim, and S. R. Lee, “Hierarchical visualization of cloud-based social network service using fuzzy,” *J. KICS,* vol. 38B, no. 07, pp. 501-511, Jul. 2013.

[2] A. Thurai, *Big Data, IoT, API: Newer technologies protected by older security(2013)*, Retrieved March, 19, 2015, from http://blog.programmableweb.com/2013/05/17/big-data-iot-api-newer-technologies-protected-by-older-security/.

[3] H. T. Mai, K. H. Park, H. S. Lee, C. S. Kim, M. Lee, and S. J. Hur, “Dynamic data migration in hybrid main memories for in-memory big data storage,” *ETRI J.,* vol. 36, no. 6, pp. 988-998, Dec. 2014.

[4] J. W. Jeong, *Beginning Hadoop Programming: Development and Operations,* Wikibooks, 2012.

[5] S. Hoffman, *Apache Flume: Distributed Log Collection for Hadoop,* PACKT publishing, Jun. 2013.

[6] J. H. Lee, Y. S. Kim, H. J. Baek, and M. S. Chung, “The relationship between climate and food incidents in Korea,” *Climate Change Research,* vol. 2, no. 4, pp. 297-307, 2011.

[7] D. H. Rim, *Statistics using R*, Free Academy Press, Feb. 2013.

[8] M. Norman, *The Art of R Programming,* No Starch Press, 2011.

[9] H. J. Lee, D. I. Shin, and D. K. Shin, “The classification algorithm of users’ emotion using brain-wave,” *J. KICS,* vol. 39C, no. 02, pp. 122-129, Feb. 2014.

[10] Y. H. Choi, S. H. Yoo, and S. W. Seo, “Heuristic algorithm for high-speed clustering of neighbor vehicular position coordinate,” *J. KICS,* vol. 39C, no. 04, pp. 343-350, Apr. 2014.

박 현 (Hyeon Park)

1985년 2월 : 전남대학교 전산통계학과 학사

1987년 2월 : 서울대학교 계산통계학과 이학석사

2005년 8월 : 충남대학교 컴퓨터과학과 이학박사

1988년 1월~현재 : 한국전자통신연구원 농업·환경IoT연구실 책임연구원, 과제책임자

<관심분야> M2M/IoT/WoT 상황인지, 빅 데이터 마이닝, 클라우드 컴퓨팅

강 성 수 (Sung-soo Kang)

1977년 2월 : 한국항공대학교 항공통신공학과 졸업

1980년 2월 : KAIST 전자및전자 공학과 석사

1999년 8월 : 전북대학교 전자공학과 박사

1980년 3월~현재 : 한국전자통신연구원 에너지하베스팅IoT연구실 책임연구원

<관심분야> IT융합, IoT, 광통신 공학

정 훈 (Hoon Jeong)

1997년 2월 : 전남대학교 전자공학과 졸업
1999년 8월 : 전남대학교 컴퓨터공학과 석사
1999년 8월~현재 : 한국전자통신연구원 사물통신연구실 선임연구원 근무

<관심분야> 전자공학, 통신공학, IoT

김 세 한 (Se-Han Kim)

1998년 2월 : 한국항공대학교 컴퓨터공학과 학사
2000년 2월 : 한국항공대학교 정보통신공학과 공학석사
2000년 7월~현재 : 한국전자통신연구원 농업·환경IoT연구실 실장

<관심분야> M2M/IoT, 클라우드 컴퓨팅, 농업/환경 IT 융합기술