키워드 네트워크를 이용한 구제역 파급효과의 트렌드 분석

노병준*, 서정순*, 이종욱*, 박대희*†, 정용화*
*고려대학교 컴퓨터정보학과
e-mail:{powernoh, jssoon77, eastwest9, dhpark, ychungy}@korea.ac.kr

Trend Analysis of Repercussion Effect of Foot-and-Mouth Disease Using Keyword Network

Byeongjoon Noh*, Zhenshun Xu*, Jonguk Lee*, Daihee Park*, Yonghwa Chung*
*Dept of Computer Information Science, Korea University

요 약

최근 구제역의 발생으로 인해 농·축산업계 및 관련 산업분야에 막대한 피해를 야기함에 따라, 구제역의 발병에 따른 다양한 사회적 파급효과의 분석이 필요하다. 본 논문에서는 온라인 뉴스를 대상으로 텍스트 마이닝 방법들을 사용하여 구제역으로 인한 경제적, 환경적, 그리고 정책적 파급효과를 분석하는 공학적 방법론을 제안한다. 제안하는 시스템은 먼저, 구제역 관련 온라인 뉴스를 수집한 후, 토픽 모델링의 대표적인 방법 중 하나인 LDA(Latent Dirichlet Allocation)를 활용하여 뉴스 기사로부터 키워드들을 추출한다. 둘째, 추출된 키워드들로부터 구제역으로 인한 파급효과의 분석을 위해 동시출현 키워드 네트워크를 구성한다. 셋째, 키워드 네트워크 타임라인을 통해 각 파급효과들의 변화를 분석한다. 마지막으로, 사례분석을 통해 2010년 7월부터 2011년 12월까지 한국에서 발생한 구제역으로 인한 사회적 파급효과의 분석을 수행하였다.

1. 서론

지난 2010년 발생한 구제역은 전국적으로 확산되어 농·축산업계 및 소비자들에게 막대한 피해를 야기하였다[1][2]. 이에 따라 정부의 주도하에 구제역 등과 같은 가축질병의 예방 및 확산방지를 위한 공공 데이터의 수집이 이루어졌으며, 다양한 학술적 연구들이 선진외국을 중심으로 활발히 진행되고 있다[2][3]. 그러나 공공 데이터와 같은 정형적 데이터만으로는 구제역으로 인한 다양한 사회적 이슈들을 종합적으로 분석하기에는 한계가 존재한다.

반면, 온라인 뉴스는 정형화된 공공 데이터베이스에서 다루지 못하는 다양한 정보들이 빠르게 전파되며, 사회적 이슈에 신속하게 반응하는 특성을 갖는다[4]. 또한, 이와 같은 텍스트 데이터를 공학적으로 처리하여 새로운 정보를 획득하고 효과적으로 요약 및 시각화하기 위한 토픽 모델링, 네트워크 분석 등과 같은 다양한 방법론들이 사회문제에 적용 가능할 만큼 충분히 성숙되었기에, 온라인 뉴스를 대상으로 구제역 등과 같은 가축질병으로 인한 다양한 사회적 이슈를 분석하는 시도는 실천가능하다고 평가된다.

본 논문에서는 구[1]제역과 관련된 다양한 사회적 이슈들을 심층 보도하고 있는 뉴스기사들을 대상으로 토픽 모델링 및 네트워크 분석을 활용하여 구제역으로 인한 다양한 파급효과의 트렌드를 분석하였다. 본 연구에서는 1) 가축의 폐사로 인해 농·축산업계에 악영향을 미치는 경제적 파급효과; 2) 감염 가축의 매몰로 인한 지하수의 오염과 같은 환경적 파급효과; 3) 구제역의 발생으로 인한 정부의 대응 정책 및 대책과 같은 정책적 파급효과를 그 분석 대상으로 한다. 분석과정은 먼저, 구제역과 관련된 온라인 뉴스를 수집하고, 전처리 과정을 거친다. 다음으로 토픽 모델링의 대표적인 방법 중 하나인 LDA(Latent Dirichlet Allocation)를 활용하여 키워드들을 추출하고, 동시출현 키워드 네트워크를 구성한다. 마지막으로 구성된 네트워크 및 네트워크 타임라인을 통해 파급효과의 트렌드 및 키워드들의 변화를 종합적으로 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 동시출현 키워드 네트워크를 통한 구제역의 파급효과 분석 시스템에 대해 설명한다. 3장에서는 제안한 방법을 활용한 사례 분석을 통해 구제역의 세 가지 파급효과의 분석 결과를 살펴보고, 마지막으로 4장에서 본 연구의 결론을 맺는다.

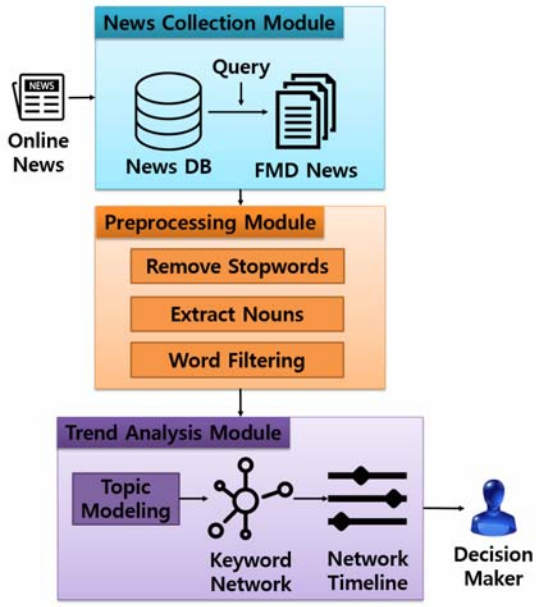
2. 키워드 네트워크 기반의 구제역 파급효과 트렌드 분석 시스템

본 논문에서 제안하는 키워드 네트워크 기반의 구제역 파급효과 분석 시스템은 크게 온라인 뉴스 수집 모듈, 전처리 모듈, 트렌드 분석 모듈로 구성되며, 시스템 구조는 그림 1과 같다.

2.1 데이터 수집 모듈

데이터 수집 모듈에서는 크롤러를 활용하여 웹 포털 사이트에 게재된 온라인 뉴스 기사를 수집한다. 데이터 수집 시, 시간에 따른 분석 등 다양한 분석에 활용하기 위하여 뉴스의 게재시간, 기사의 제목, 내용 등을 함께 수집한다.

† 교신저자:dhpark@korea.ac.kr



(그림 1) 키워드 네트워크 기반의
구제역 파급효과 분석 시스템

2.2 전처리 모듈

전처리 모듈에서는 뉴스 기사의 불용어 제거, 명사 추출, 단어의 필터링 및 변환과정을 수행한다. 먼저, 불용어 제거 과정에서는 선택된 뉴스에 포함된 불용어 및 광고 등을 제거한다. 다음으로 명사추출기를 활용하여 뉴스에 사용된 명사를 추출하고, 정확한 키워드의 추출을 위해 해당 단어들의 필터링 및 변환을 수행한다. 단어의 필터링 및 변환 규칙은 표 1과 같다.

<표 1> 단어 필터링 및 변환 규칙

	변환규칙	필터링/변환 예	변환 후
1	불필요 단어 제거	***기자	(제거)
		**슈퍼 파격할인	
2	명사형 변환	관찰하고	관찰
		관찰하면서	
3	지역명 통일	충청남도	충남
		충남일대	

2.3 트렌드 분석 모듈

트렌드 분석 모듈에서는 먼저, 전처리된 뉴스 기사들을 대상으로 토픽 모델링 방법 중 하나인 LDA를 사용하여 키워드들을 추출한다. 다음으로 추출된 키워드를 활용하여 동시출현 키워드 네트워크를 구성하고, 구제역으로 인한 세 가지 파급효과들을 분석한다. 동시출현 키워드 네트워크는 키워드가 해당 뉴스 기사에 출현한 횟수를 나타내는 문서-키워드 행렬(document-keyword matrix)을 키워드간의 인접 행렬(adjacency matrix)로 변환하고, 이를 네트워크로 구성한다. 그 과정은 그림 2와 같으며, 생성된 인접행렬을 통해 동시출현 키워드 네트워크를 시각화한다. 동시출현 키워드 네트워크의 노드(node)의 크기는 키워드의 출현 빈도를 나타내고, 엣지(edge)의 두께는 키워드간의 동시출현 빈도를 의미하며, 두 키워드의 동시출현 빈도가 높을수록 두 키워드의 연관성이 높음을 알 수 있다.

마지막으로 각 네트워크에서 나타난 파급효과와 관련된 키워드들을 통해 타임라인을 구성하고, 구제역으로 인한 파급효과의 트렌드 및 키워드의 변화를 분석한다.

Document-Keyword Matrix

	$keyword_1$	$keyword_2$	$keyword_3$	$keyword_4$
doc_1	1	0	1	2
doc_2	0	17	0	2
doc_3	3	1	0	0
doc_4	4	1	3	0
doc_5	0	4	8	1

Transform

	$keyword_1$	$keyword_2$	$keyword_3$	$keyword_4$
$keyword_1$	.	.	.	.
$keyword_2$	2	.	.	.
$keyword_3$	2	2	.	.
$keyword_4$	1	2	2	.

Keyword-Keyword Adjacency Matrix

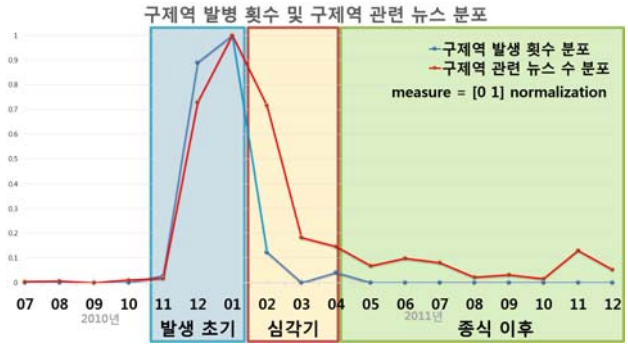
(그림 2) 인접행렬 변환의 예

3. 실험 및 분석 결과

3.1 실험 설계

본 절에서는 실험에서 사용한 온라인 뉴스 데이터의 소개 및 실험 설계방법을 서술한다. 먼저, 온라인 뉴스 데이터의 수집은 웹 포털사이트 'N'사 뉴스 페이지에 게재된 정치, 경제, 사회, 문화 카테고리의 뉴스기사를 수집하였다. 수집 기간은 구제역이 심각하게 발생한 기간인 2010년 7월부터 2011년 12월까지로 설정하였으며, '구제역' 키워드를 포함하는 뉴스기사만을 선택하였다. 명사 추출 및 키워드 추출은 통계프로그램 R의 KoNLP 패키지와 topicmodels 패키지를 사용하였으며, 네트워크 시각화 및 분석 패키지인 igraph 패키지를 사용하여 '구제역' 키워드를 중심으로 성형(star) 구조의 동시출현 키워드 네트워크를 구성하였다. 또한, 그림 3과 같이 구제역 발생 시기를 세 구간('발생 초기', '심각기', '종식 이후')로 구분하고, 각 구간에서의 세 가지 파급효과들(경제적, 환경적, 정책적 파급효과)을 동시출현 키워드 네트워크 분석을 통해 계층적으로 분석하도록 설계하였다. 마지막으로 주요 키워드들을 네트워크 타임라인으로 표현함으로써 키워드들의 변화 및 파급효과의 트렌드를 종합적으로 분석하였다.

뉴스 데이터 수집 결과, 2010년에 약 3,520건, 2011년에 약 12,058건의 구제역 관련 뉴스가 게제 되었으며, 전처리 과정을 통해 약 17,000개의 단어를 획득하였다. 트렌드 분석 단계에서는, 각 분석구간별 30개의 키워드들을 추출하였다.



(그림 3) 구제역 관련 뉴스 분포

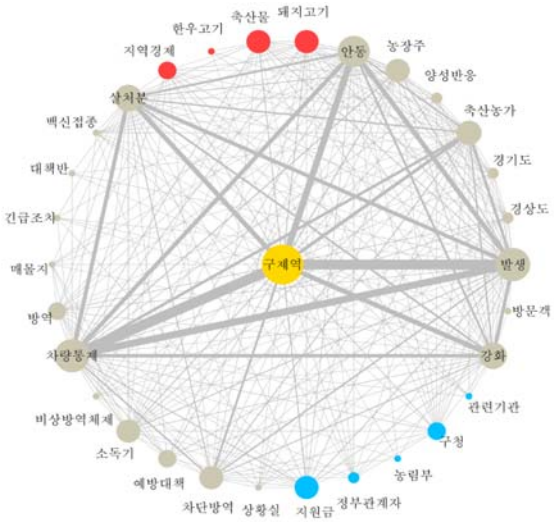
3.2 실험 결과 분석

3.2.1 구제역 발생 구간별 파급효과 분석

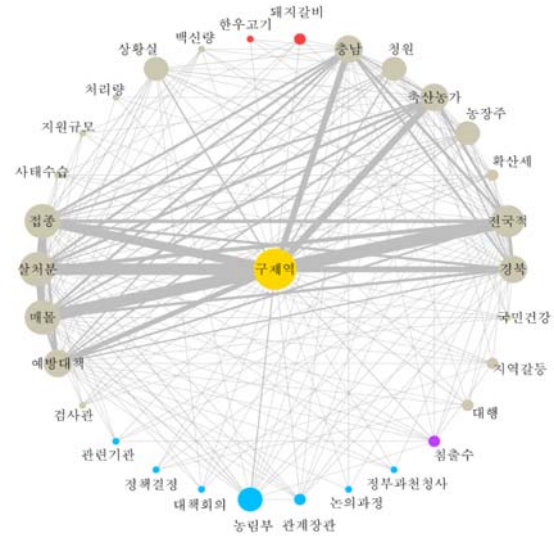
본 절에서는 동시출현 키워드 네트워크를 활용하여 구제역 발생 구간별 파급효과의 트렌드 분석 결과를 서술한다. 먼저, '구제역 발생 초기 구간'의 키워드 네트워크 분석 결과는 그림 4와 같으며, '발생', '차량통제', '살처분' 등과 같은 국가 위기상황 매뉴얼에 기초한 구제역의 초기대응과 관련된 키워드들이 주로 나타나고 있다. 반면, 경제적 파급효과와 관련된 '돼지고기', '지역경제' 등의 키워드들과 '지원금', '구청' 등과 같은 정책적 파급효과와 관련된 키워드들이 확인된다. 그러나 환경적 파급효과와 관련된 키워드들은 사회적 이슈로 대두되기에는 시기적으로 아직 이른 감이 있음을 보여준다.

'구제역 심각기 구간'의 키워드 네트워크는 그림 5와 같이 나타나며, '확산세', '전국적' 등과 같은 키워드를 통해 구제역이 전국적으로 확산되고 있음을 알 수 있다. 뿐만 아니라, 국가 위기상황 매뉴얼에 기초한 구제역의 적극적 대응과 관련된 '접종', '살처분', '매몰' 등의 키워드들이 나타난다. 또한, 구제역이 장기간 지속됨에 따라 '돼지갈비', '대책회의', '침출수' 등과 같은 경제적, 정책적, 환경적 파급효과와 관련된 다양한 키워드들이 나타남을 알 수 있다.

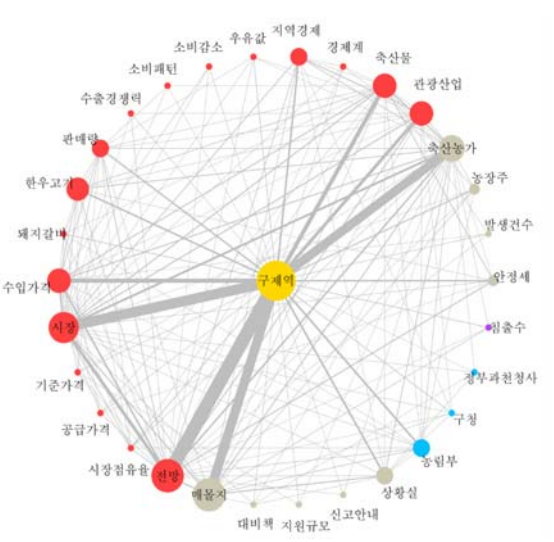
마지막으로, '구제역 종식 이후 구간'의 키워드 네트워크는 그림 6과 같이 나타난다. 그림 6에 의하면 세 가지 파급효과와 관련된 키워드들이 모두 나타나지만, '수입가격', '돼지갈비' 등과 같은 경제적 파급효과와 관련된 키워드들이 다수 등장함을 알 수 있다. 이를 통해 구제역으로 인한 돼지고기 가격의 하락과 국내 축산물 시장에 미치는 악영향, 그리고 '우유값' 키워드를 통해 젖소의 폐사로 인한 우유의 가격 형성에 미치는 경제적 영향 등을 반영함을 알 수 있다. 반면, 경제적 파급효과가 농·축산업계에만 국한된 것이 아니라 관광산업에 까지도 경제적 영향력을 미치고 있음을 '관광산업' 키워드를 통해 알 수 있다. 이는 구제역으로 인한 지역 축제의 취소가 빈번했으며, 관광객을 주요 고객으로 하는 음식점, 숙박시설, 운송업체 등의 관광 서비스 산업계에 연쇄적으로 피해를 야기했음을 시사한다.



(그림 4) 구제역 발생초기의 네트워크



(그림 5) 구제역 심각기의 네트워크



(그림 6) 구제역 종식이후의 네트워크

3.2.2 키워드 네트워크 타임라인 분석

네트워크 타임라인 분석에서는 그림 7과 같이 구제역 발생 기간 동안의 세 가지 파급효과의 변화를 분석하였다. 분석 결과 '시장', '축산물' 등과 같은 경제적 파급효과와 관련된 키워드들은 구제역 발생의 전 구간에서 꾸준히 등장하고 있으며, 특히 구제역 종식 이후의 구간에서 높은 빈도로 등장함을 알 수 있다. 정책적 파급효과와 관련된 키워드인 '구청', '대책회의' 등은 구제역 심각기 구간에서 등장하였다가 구제역 종식과 동시에 소멸하였다. 한편, '침출수'와 같은 환경적 파급효과는 구제역 발생 기간 동안 중요한 사회적인 문제로 인식되지 못했음을 알 수 있다.

4. 결론 및 향후 연구

본 논문에서는 다양한 사회적 이슈들에 즉각적으로 반응하는 온라인 뉴스를 대상으로 텍스트 마이닝 방법론을 활용하여 구제역 등과 같은 가축질병의 파급효과들을 공학적으로 분석하는 새로운 학술적 분석을 수행하였다. 본 논문의 분석실험에서는 구제역 발생 시기를 '발생 초기', '심각기', '종식 이후'로 구분하고, 각 구간에서의 경제적, 정책적, 환경적 파급효과를 동시출현 키워드 네트워크를 사용하여 계층적으로 분석하였으며, 네트워크 타임라인을 구성하여 구제역 발생 전 구간에서의 파급효과의 변화를 분석하였다. 분석 결과를 요약하면, 경제적 파급효과는 구제역 발생의 전 구간에 걸쳐 꾸준히 등장하였으며, 정책적 파급효과는 구제역 심각기 구간에서 등장하고 구제역 종식과 동시에 소멸함을 확인하였다. 반면, 환경적 파급효과는 구제역 발생의 전 구간에 걸쳐 온라인 뉴스 상에서 중요한 사회적인 문제로 인식되지 못했음을 확인하였다. 향후 연구로는 실용화가 가능한 분석 시스템으로의 고도화가 요구된다.

감사의 글

본 연구는 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2015R1D1A3A01018731)

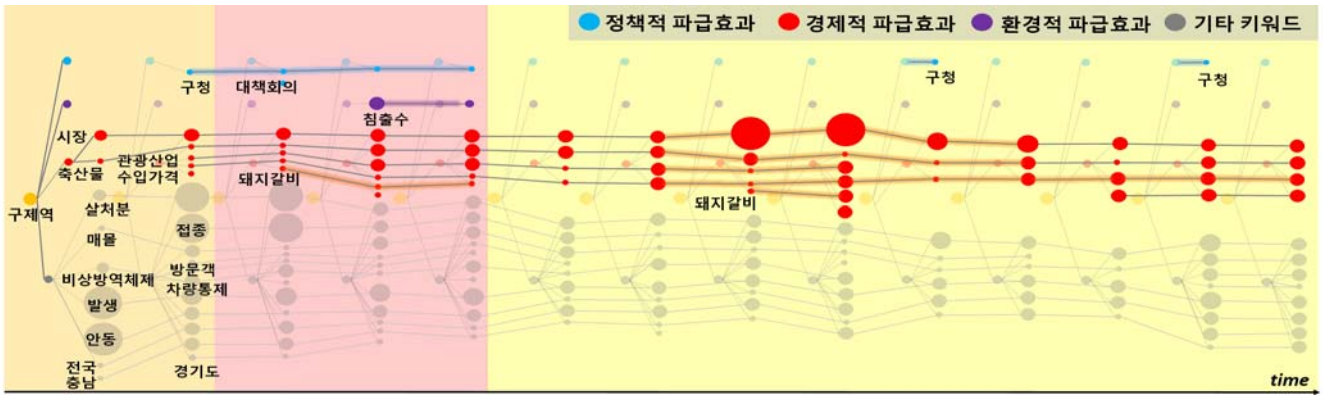
References

[1] D. Mdetel, C. Kasanga, M. Seth, and K. Kayunze, "Socio-economic impact of foot and mouth disease in wildlife-livestock interface," *World*, Vol. 5, No. 3, pp. 31-35, 2015.

[2] M. Kyung and J. Yom, "Implementation of open source SOLAP decision-making system for livestock epidemic surveillance and prevention," *Journal of the Korean Society of Surveying, Geodesy*, Photogrammetry and Cartography, Vol. 30, No. 3, pp. 287-294, 2012.

[3] H. Lee, K. Suh, N. Jung, I. Lee, I. Seo, O. Moon, and J. Lee, "Prediction of the spread of highly pathogenic avian influenza using a multifactor network: Part 2-comprehensive network analysis with direct/indirect infection route," *Biosystems Engineering*, Vol. 118, pp. 115-127, 2014.

[4] L. Hou, J. Li, Z. Wang, J. Tang, P. Zhang, R. Yang, and Q. Zheng, "NewsMiner: multifaceted news analysis for event search," *Knowledge-Based Systems,* Vol. 76, pp. 17-29, 2015.



(그림 7) 네트워크 타임라인