# 텍스트 마이닝을 이용한 매체별 에볼라 주제 분석*

## - 바이오 분야 연구논문과 뉴스 텍스트 데이터를 이용하여 -

## Text Mining Driven Content Analysis of Ebola on News Media and Scientific Publications

안 주 영 (Juyoung An)**
안 규 빈 (Kyubin Ahn)***
송 민 (Min Song)****

목 차



## 초 록

에볼라 바이러스(Ebola virus disease)와 같은 전염병들은 사회적으로 큰 이슈가 되어 언론의 관심을 받으며 동시에 많은 연구의 대상이 되기도 한다. 이에 따라 국내외로 전염병과 관련된 텍스트 마이닝 연구가 활발하게 진행되고 있으나, 텍스트 마이닝 기법을 사용하여 상이한 특성을 가진 매체 간 주제를 분석한 연구는 아직까지 진행되지 않고 있다. 따라서 본 연구에서는 전염병 중 하나인 에볼라를 키워드로 하여 사회적 특성을 지닌 뉴스 기사와 바이오 분야의 전문적 특성을 지닌 연구 논문 간의 주제 분석을 진행하였다. 텍스트 분석에는 매체별 문헌 데이터로부터 다양한 토픽들을 추출하기 위해 토픽모델링 기법을 적용하였고, 매체 간의 구체적인 내용 분석을 위해 중요 개체를 선정하고 이를 중심으로 동시출현 단어 네트워크 분석을 수행하였다. 또한 각 매체별로 등장하는 주제를 시각적으로 표현하기 위해 토픽맵을 구축하였다. 분석 결과, 두 매체에서 다루는 주제의 차이점과 공통점을 발견할 수 있었으며 동시 출현 주제의 시계열 분석을 통해 매체 간 특성의 차이를 찾을 수 있었다. 본 연구를 통해 상이한 특성을 지닌 매체들의 주제와 개체들을 함께 제시하고, 매체 간의 공통점과 차이점을 보여줌으로써 매체별 정보 생산자들이 연구 및 현상 분석을 진행하는 데 있어 관점의 다양성을 제공할 수 있을 것이다.

## ABSTRACT

Infectious diseases such as Ebola virus disease become a social issue and draw public attention to be a major topic on news or research. As a result, there have been a lot of studies on infectious diseases using text-mining techniques. However, there is no research on content analysis of two media channels that have distinct characteristics. Accordingly, in this study, we conduct topic analysis between news (representing a social perspective) and academic research paper (representing perspectives of bio-professionals). As text-mining techniques, topic modeling is applied to extract various topics according to the materials, and the word co-occurrence map based on selected bio entities is used to compare the perspectives of the materials specifically. For network analysis, topic map is built by using Gephi. Aforementioned approaches uncovered the difference of topics between two materials and the characteristics of the two materials. In terms of the word co-occurrence map, however, most of entities are shared in both materials. These results indicate that there are differences and commonalties between social and academic materials.

키워드: 에볼라 바이러스, 텍스트 마이닝, 전염병, 매체별 분석, 토픽 모델링, 동시출현 네트워크, 토픽맵
Ebola virus, Text mining, Epidemics, Media analysis, Topic modeling, Co-occurrence network, Topic map

---

* 이 논문은 2015학년도 연세대학교 미래선도연구사업(부분적인) 지원에 의하여 작성된 것임(2015-22-0119).
** 연세대학교 문헌정보학과 석사과정(anjy@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 석사과정(kyubin308@hanmail.net) (공동저자)
**** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
논문접수일자: 2016년 5월 3일 최초심사일자: 2016년 5월 3일 게재확정일자: 2016년 5월 12일
한국문헌정보학회지, 50(2): 289-307, 2016. [http://dx.doi.org/10.4275/KSLIS.2016.50.2.289]

## 1. 서 론

에볼라 바이러스(Ebola virus disease)는 1976년 발견되어 2014년 미국에 전파되었으며 지금까지도 종식되지 않은 전염병으로 전 세계에서 에볼라 바이러스에 관한 연구와 뉴스보도가 다수 이루어져 왔다. 에볼라 바이러스를 비롯한 전염병들은 다른 질병들과는 달리 특정 기간에 사회적으로 큰 이슈가 되며 연구 역시 사회적인 영향을 받아 이루어지게 된다(Salathe et al. 2012). Pesquita 등(2014)은 전염병의 경우 다른 병들보다 더 큰 사회적 의미를 지니며 이에 따라 부수적으로 생겨나는 데이터가 많다는 것을 인식하고, 그러한 데이터를 효과적으로 이용할 수 있도록 전염병 온톨로지를 구축하기도 하였다. 따라서 전염병의 경우, 비(非)전염성 질병보다 더 많은 사회적 요소를 포함한 연구가 진행되며, 언론의 보도 역시 사회적 주제들 외의 다양한 전문적 지식들을 포함하여 이루어진다고 추론할 수 있다. 지금까지 비슷한 특성을 지닌 매체 간 전염병에 대한 관점을 비교한 연구는 수행되어 왔으나 그 특성이 매우 상이한 연구논문과 뉴스 간의 내용을 분석하고 그 특성을 비교한 연구는 진행되지 않았다. 또한, 국내에서 전염병에 대한 텍스트 마이닝 분야의 연구는 예측시스템 구축을 중심으로 진행되어 왔다. 따라서 본 연구는 위의 가정을 검증하기 위하여 전염병 중 하나인 에볼라를 키워드로 하여 최근 6년 동안 발행된(2010년~2015년) 바이오 분야의 연구논문과 뉴스 데이터를 수집하였다. 이를 이용하여 에볼라가 각 매체에서 어떠한 관점에서 다루어지는지 그 공통점과 차이점을 거시적 측면과 미시적 측면에서 분석하였다. 이를 위해 먼저 토픽모델링 기법을 사용하여 문헌 내에서 언급되는 토픽, 즉 주제를 찾고 주제별 가중치를 이용하여 토픽맵을 구축한 후 어떠한 주제들이 언급되는 지 전체적인 주제의 지형도를 파악하였다. 그 후 세부적으로 어떠한 개체들이 어떻게 다른 관점에서 다루어지는지 살펴보기 위하여 연구논문과 뉴스에 공통적으로 등장하는 개체들을 선정하고 개체와 한 문장에 출현한 단어들의 동시출현 관계를 이용한 동시출현 네트워크를 구축하였다. 이를 통해 거시적, 미시적 차원에서의 두 매체 간 차이와 공통점을 발견할 수 있었다. 이렇게 밝혀진 매체별 관점의 차이와 공통점은 각 매체별 정보 생산자들이 좀 더 다양한 양질의 정보를 생산하는데 도움을 줄 것이다.

## 2. 관련 연구

전염병에 대한 매체의 관점을 다룬 연구는 주로 국외에서 활발히 이루어지고 있다. Towers 등(2015)은 기존의 전염병 예측 시스템이 검색엔진 검색 추이만을 반영한 기존의 방법에 한계가 있다는 것을 밝혀내기 위하여, 에볼라와 관련된 미국 내의 TV 뉴스 방송과 트윗, 인터넷 검색 빈도 간의 관계를 분석하였다. 분석에는 여러 통계적 모형이 활용되었으며, 분석 결과 에볼라를 주제로 한 TV 뉴스가 방송될 경우, 에볼라를 키워드로 한 인터넷 검색양과 트윗의 양이 유의미하게 증가하는 것으로 밝혀졌다. Seltzer 등(2015) 역시 미국 내의 에볼라 창궐 기간 동안의 데이터를 분석하였는데, 사진 공유를 기반으로 하는 사회관계망 서비스(SNS)인

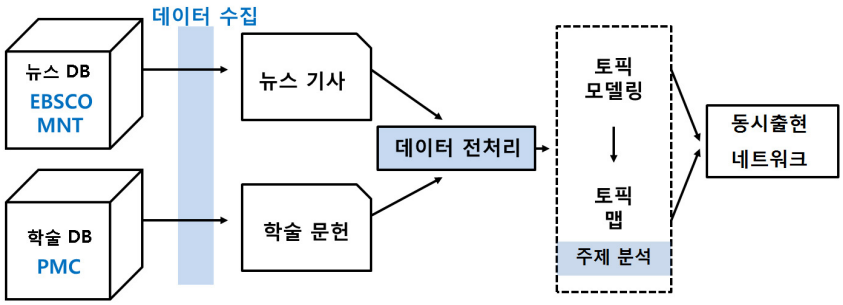
Instagram과 Flicker에 게시된 사진 데이터를 대상으로 코딩분석을 수행하였다. 분석 결과 같은 사진 기반 SNS라 하더라도 에볼라에 관련된 서로 다른 측면을 담은 사진들을 게시하는 것으로 나타났다. Kim 등(2015)의 경우, 에볼라가 가장 창궐한 3개월(2014년 6월~2014년 8월)에 대한 뉴스와 트위터 데이터를 수집하여, 두 매체 간 주제 분석을 통해 에볼라라는 질병이 어떻게 다르게 이야기되고 있는 지 분석하고, 감성분석을 통해 각 매체의 감성 추이를 추적하였다. 이와 비슷한 연구로는 Househ(2015)의 연구를 들 수 있는데, 해당 연구는 에볼라에 대한 정보를 교환하는 과정에서 뉴스 미디어와 트위터가 사용되는 양상을 분석하고 그 관계를 밝혀냈다. 기술통계를 적용하여 분석한 결과, 트위터는 뉴스매체의 연장선상에서 사용되는 경향이 있으나, 그 여파가 24시간 이내인 것으로 드러났다. 전염병 외의 질병에 대한 다른 관점을 분석한 연구들 또한 진행되었다. Lee 등(2015)은 전문가와 대중이 당뇨병에 대해 갖는 다른 관점을 파악하기 위하여 Pubmed가 제공하는 당뇨병 연구 문헌의 초록과 당뇨병 관련 커뮤니티의 게시글을 수집하였다. 이후, 개체들 간의 관계를 추출하고, 그래프 분석을 통해 전문가와 대중들이 당뇨병에 대해 어떠한 다른 관점을 가지고 있는지 밝혀냈다.

한편, 국내에서 이루어진 전염병에 대한 연구는 주로 그 예측과 전파경로 모델을 구축하기 위해 수행되었다. 최정실(2008)은 실제 의료현장에서 발생되는 임상정보를 활용하여 법정전염병의 전파경로를 분류하고 전염병 발병에 관련된 정보를 정부기관에 자동으로 전송하는 프로그램을 구축하고 평가하였다. 김은경 등(2013)은 현대사회의 전염병 전파가 갖고 있는 복잡성에 주목하여, 일일 교통량, 인구 통계, 항공 통계 자료, 질병자료를 아우르는 데이터들을 사용하여 전염병 확산이 되어가는 과정을 시뮬레이션 했다. 이를 통해 전염병이 감염되는 경로를 추적하고 나아가 확산 경로와 그 정도를 예측하고자 하였다. 황교상 등(2014)은 특정 지역의 인구 주택 총 조사 자료를 사용하여 개인별 이동 패턴을 고려한 전염병 확산 시뮬레이션 모델을 구현하고 그 효과를 예측하였다. 지금까지 살펴본 국내외의 연구들은 전염병이 가진 사회적 측면을 인식하고 서로 다른 매체 간 관점을 비교하거나 사회적 특성을 고려한 전염병의 전파를 추정하였으나, 앞서 말한 바와 같이 그 특성이 확연히 다른 뉴스와 학술문헌에서 전염병을 각각 어떻게 다루는지 분석하고 비교한 연구는 국내외적으로 이루어지지 않았다.

## 3. 연구 설계

### 3.1 연구 모형

본 연구는 〈그림 1〉과 같이 데이터 수집, 전처리, 분석 순서로 진행하였다. 가장 먼저 사회적 관점과 전문 분야의 관점을 가장 잘 대변할 수 있는 데이터로 각각 뉴스와 학술 논문을 선정하였다. 이 중 뉴스 데이터의 경우, Ebsco(https://www.ebscohost.com/)와 의학전문기사 사이트인 MNT(http://www.medicalnewstoday.com/)에서 총 42,330개의 뉴스를 수집하였다. 바이오 전문 분야 데이터의 경우, PMC(http://www.ncbi.nlm.nih.gov/pmc/)에서 제공하는 학




〈그림 1〉 연구 모형

술 논문 전문 4,222개를 수집하였다. 두 데이터 모두 검색 키워드로 'ebola'를 사용하였다. 데이터 수집 이후 분석을 위한 전처리를 수행하였으며 StanfodNLP(Manning et al. 2014)의 원형복원, 불용어 처리, 품사태깅 기능이 사용되었다. 불용어 처리에는 일반적인 영어 불용어 사전을 이용하였으며 품사태깅을 거쳐 동사, 명사, 형용사, 부사에 해당하는 단어들만 분석에 포함하였다. 데이터 분석에는 토픽 모델링 기법과 동시출현 네트워크 구축 기법을 적용하였다.

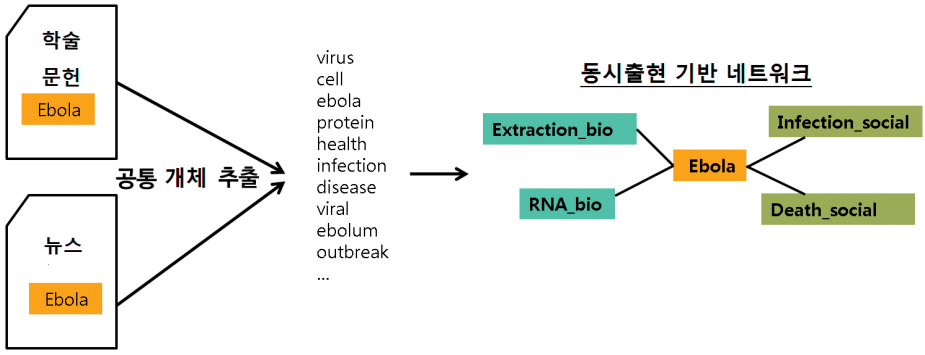
### 3.2 분석 방법

#### 3.2.1 토픽 모델링

본 연구에서 사용된 토픽 모델링 기법은 DMR(Dirichlet-multinomial Regression)이다. 이는 Blei 등(2013)이 제시한 토픽 모델링 기법인 LDA(Latent Dirichlet Allocation)를 기반으로 문헌-주제 분포에 근거한 log-linear prior를 사용하여 저자, 발행처, 참고문헌, 날짜와 같은 문헌의 자질들을 토픽모델링 분석에 추가한 것이다(Mimno and McCallum 2012). 즉 DMR은 주어진 문헌에 대하여 각 문헌에 어떤 주제들이 존재하는지에 대한 확률 모형을 만든 LDA에서 분석에 사용될 변수를 추가하여 분석을 수행한다. 본 연구에서는 추가된 문헌의 자질로 연도와 달 데이터를 선정하여 토픽모델링을 수행하였다. LDA와 DMR 토픽모델링은 기본적으로 문헌집단의 주제 분포와 주제별 단어의 생성 확률에 기초한 분석이기 때문에 문헌집단이 잠재적으로 가지고 있는 주제의 개수와 그 주제를 나타내는 단어의 개수를 연구자가 임의로 설정하여 분석을 진행하게 된다. 본 연구의 경우 분석 대상이 되는 학술문헌과 뉴스 데이터가 각각 500만, 100만개 이상의 어절을 지닌 대량의 데이터이므로 그에 적합한 잠재 토픽 개수로 30개를 설정하였다. 이때, 토픽 모델 생성 반복횟수와 분석 시간은 반비례 관계에 있기 때문에 효율성과 정확성을 모두 얻기 위해서는 적당한 수의 반복횟수를 설정해 주어야 한다. 이를 위해 보통 1,000에서 2,000 사이의 값으로 설정하는 것을 권고하고 있고, 본 연구에서는 반복횟수를 1,000으로 설정하여 효율적으로 토픽 모델을 생성함과 동시에 모델의 정확성도 얻고자 하였다. 토픽 모델링 수행 이후 토픽별 주제 선정 과정에는 바이오 분야 주제 전문가의 조언을 참고하였다.

3.2.2 동시출현 네트워크

토픽 모델링 라벨링 결과와 토픽맵은 각 분야의 매체에서 어떤 주제가 주로 이야기되고 있는 지 거시적인 측면을 보여주지만, 구체적으로 어떤 키워드들이 어떻게 다른 관점에서 다루어지는지 알 수 없다는 한계가 있다. 따라서, 키워드, 주제, 데이터, 주제분야의 항목 등을 통해 해당 분야의 지식구조를 파악할 수 있다는 '개체(entity)' 개념(Ding et al. 2013)을 차용하여 상위 출현빈도를 가진 개체들을 추출하고, 상위 개체 3개(ebolum, virus, cell)를 중심으로 명사, 동사, 형용사만을 추출하여 동시출현 기반 그래프를 구축하였다. 〈그림 2〉는 그 과정을 도식화 한 것이다. 이를 통해 각 개체들이 학술문헌과 뉴스에서 어떤 단어들과 함께 출현하는지 살펴, 각 매체에서 비롯된 구체적인 관점의 차이를 파악할 수 있었다. 총 24,820개의 단어가 두 문헌에서 동시에 출현하였으며 〈표 1〉은 상위 출현 빈도 단어 10개를 나열한 것이다.




〈그림 2〉 개체 중심 동시출현 네트워크 구축 과정

〈표 1〉 상위 10개 동시 출현 단어

| 단어 | 빈도 (학술 문헌 + 뉴스) | 빈도 (학술 문헌) | 빈도 (뉴스) |
|---|---|---|---|
| ebolum | 149,885 | 10,516( 7.02%) | 139,369(92.98%) |
| virus | 114,050 | 55,979(49.08%) | 58,071(50.92%) |
| cell | 80,067 | 71,382(89.15%) | 8,685(10.85%) |
| health | 55,931 | 17,260(30.86%) | 38,671(69.14%) |
| disease | 51,840 | 21,038(40.58%) | 30,802(59.42%) |
| infection | 43,266 | 30,720(71.00%) | 12,546(29.00%) |
| protein | 42,341 | 35,513(83.87%) | 6,828(16.13%) |
| outbreak | 37,425 | 10,509(28.08%) | 26,916(71.92%) |
| viral | 32,720 | 24,302(74.27%) | 8,418(25.73%) |
| ebola | 6,623 | 2,469(37.28%) | 4,154(62.72%) |

# 4. 연구 결과

## 4.1 기술통계

〈표 2〉는 수집된 데이터의 기술통계이다. 뉴스와 연구 논문의 형태적 특성상, 연구 논문의 수가 뉴스의 수보다 매우 적으나, 어절의 개수는 2배 이상 많은 것을 확인할 수 있다. 〈그림 3〉은 수집 데이터의 개수가 시계열 추이에 따라 변화하는 것을 나타낸 것이다. 에볼라가 창궐하기 이전에는 연구논문의 개수가 뉴스의 개수보다 많은 경향을 보이나, 에볼라가 창궐한 2014년에는 뉴스의 개수가 급증하는 것을 발견할 수 있다. 한편, 뉴스만큼 큰 폭은 아니지만 연구논문 역시 에볼라가 창궐함에 따라 그 수가 급격히 늘어났으며 특히 2015년에 매우 급증하는 경향을 보이는데, 이는 연구 논문의 경우 출판되기까지 뉴스보다 훨씬 더 오랜 시간이 걸린다는 매체의 특성이 반영된 것으로 해석할 수 있다.

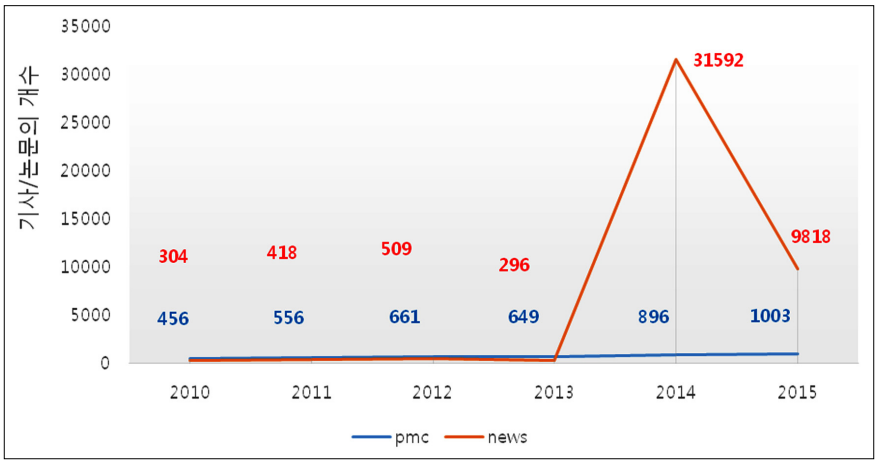
## 4.2 토픽모델링 결과

### 4.2.1 토픽 모델링 라벨링

〈표 3〉은 토픽 모델링 수행 후 토픽별 단어 분포와 토픽별 구(phrase) 출현 빈도, 바이오 전문가의 의견을 참고하여 각각의 토픽에 주제명을 부여한 결과이다. 음영 표시를 한 주제명들은 학술 문헌과 뉴스에서 동시에 등장한 주제들이다(Immune Response, Public Health, Medicine, Bat as a Host, Outbreak, Ebola Virus,

〈표 2〉 데이터 기술통계

| | 언어 | 개수 | 수집기간 | 어절의 개수 |
|---|---|---|---|---|
| EBSCO | 영어 | 42,040 | 2010~2015 | 1,561,811 |
| MNT | | 290 | | |
| Pubmed Central | | 4,222 | | 5,546,455 |



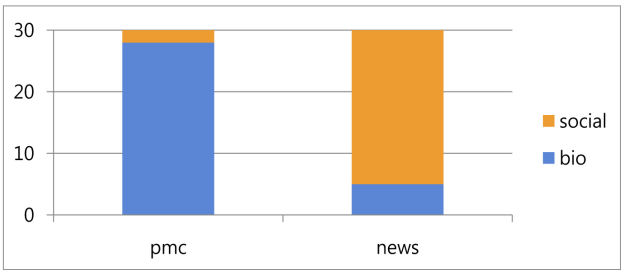

〈그림 3〉 데이터 양의 시계열 추이

〈표 3〉 토픽 모델링 라벨링 결과

| | 학술 문헌 | 뉴스 | | 학술 문헌 | 뉴스 |
|---|---|---|---|---|---|
| Topic 0 | immune response | duncan | Topic 15 | phylogenetic tree | infected nurse |
| Topic 1 | public health | liberia | Topic 16 | animal model | lesson learned |
| Topic 2 | rt pcr | medical correspondence | Topic 17 | medicine | report |
| Topic 3 | immunology | military | Topic 18 | membrane fusion | airport screening |
| Topic 4 | HPV | world cup | Topic 19 | global health | the international situation |
| Topic 5 | influenza virus | bat as a host | Topic 20 | dengue virus | sierra leone |
| Topic 6 | Virus cell fusion | outbreak | Topic 21 | escrt | ebola virus |
| Topic 7 | clinical trial | concept | Topic 22 | Cell culture | immune response |
| Topic 8 | HIV | medicine | Topic 23 | metadata of article | American administration |
| Topic 9 | monoclonal antibody | bridal shop closed | Topic 24 | genetic diversity | health organization |
| Topic 10 | Experimental datum set | ebola virus | Topic 25 | outbreak | prevention |
| Topic 11 | genome sequence | American politics | Topic 26 | ebola virus | discussion |
| Topic 12 | RNA-virus | fight | Topic 27 | gene expression | expert |
| Topic 13 | bat as a host | public health | Topic 28 | Iterferon | infected nurse |
| Topic 14 | infectious disease | health organization | Topic 29 | Antiviral activity of plant lectins | lesson learned |

Immune Response).

토픽 모델링 결과를 통해 각 매체에서 에볼라에 대해 이야기하고 있는 주제들을 확인할 수 있으며, 각 매체에서 논의되고 있는 에볼라 관련 주제들이 서로 관련을 맺고 있다는 점도 발견할 수 있다. 〈그림 4〉는 각 매체에서 추출된 주제들 중 바이오와 사회적 분야 중 어느 분야에 치우친 주제인지 주제 전문가와 함께 판정하여 그 분포를 도식화한 것이다. 학술 문헌에서 2개의 주제는 사회적인 요소가 포함된 'Public Health'와 'Outbreak'이며 〈그림 5〉는 실제 두 요소를 포함하고 있는 논문의 예이다. 뉴스의 경우 4개의 주제(Ebola Virus, Bat as a Host, Immune Response, Medicine)가 바이오 분야에서 주로 다루는 주제들로 판정되었으며 〈그림 6〉에서 실제 기사를 확인할 수 있다.




〈그림 4〉 매체별 주제 분포

**The Emergence of Ebola as a Global Health Security Threat: From 'Lessons Learned' to Coordinated Multilateral Containment Efforts**

Sarathi Kalra, Dhanashree Kelkar,[1] Sagar C. Galwankar,[1] Thomas J. Papadimos,[2] Stanislaw P. Stawicki, Bonnie Arquilla,[3] Brian A. Hoey, Richard P. Sharpe, Donna Sabol, and Jeffrey A. Jahre

Author information ▶ Copyright and License information ▼

Copyright : © Journal of Global Infectious Diseases

This is an open-access article distributed under the terms of the Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

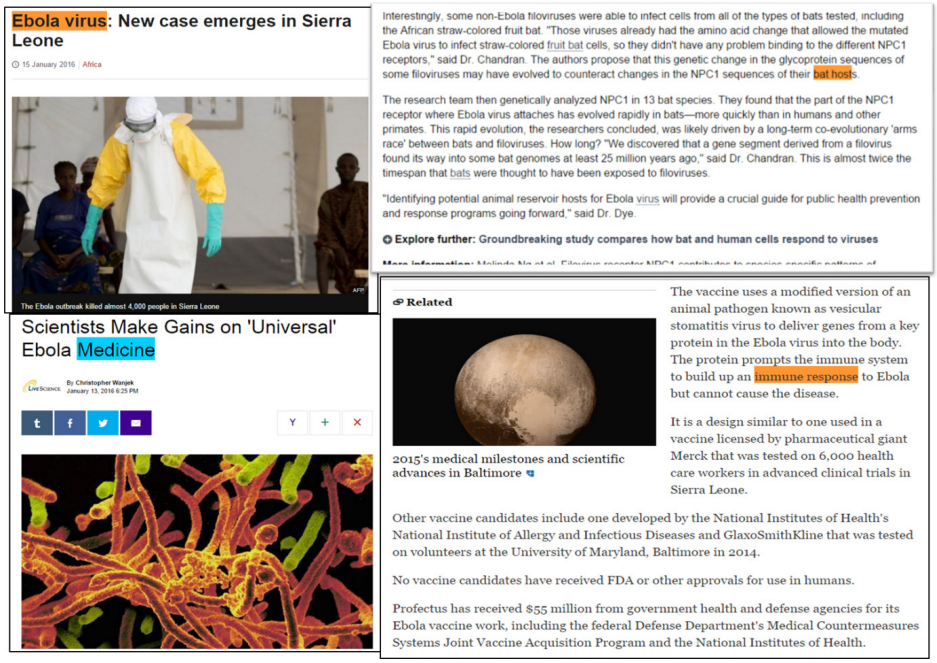
**Mapping the zoonotic niche of Ebola virus disease in Africa**

David M Pigott,[1,†] Nick Golding,[1,†] Adrian Mylne,[1] Zhi Huang,[1] Andrew J Henry,[1] Daniel J Weiss,[1] Oliver J Brady,[1] Moritz UG Kraemer,[1] David L Smith,[1,2] Catherine L Moyes,[1] Samir Bhatt,[1] Peter W Gething,[1] Peter W Horby,[3] Isaac I Bogoch,[4,5] John S Brownstein,[6,7] Sumiko R Mekaru,[8] Andrew J Tatem,[9,10,13] Kamran Khan,[4,11] and Simon I Hay[1,12,*]

Prabhat Jha, Reviewing editor
Prabhat Jha, University of Toronto, Canada;

Author information ▶ Article notes ▶ Copyright and License information ▶

〈그림 5〉 사회적 주제를 다루고 있는 논문



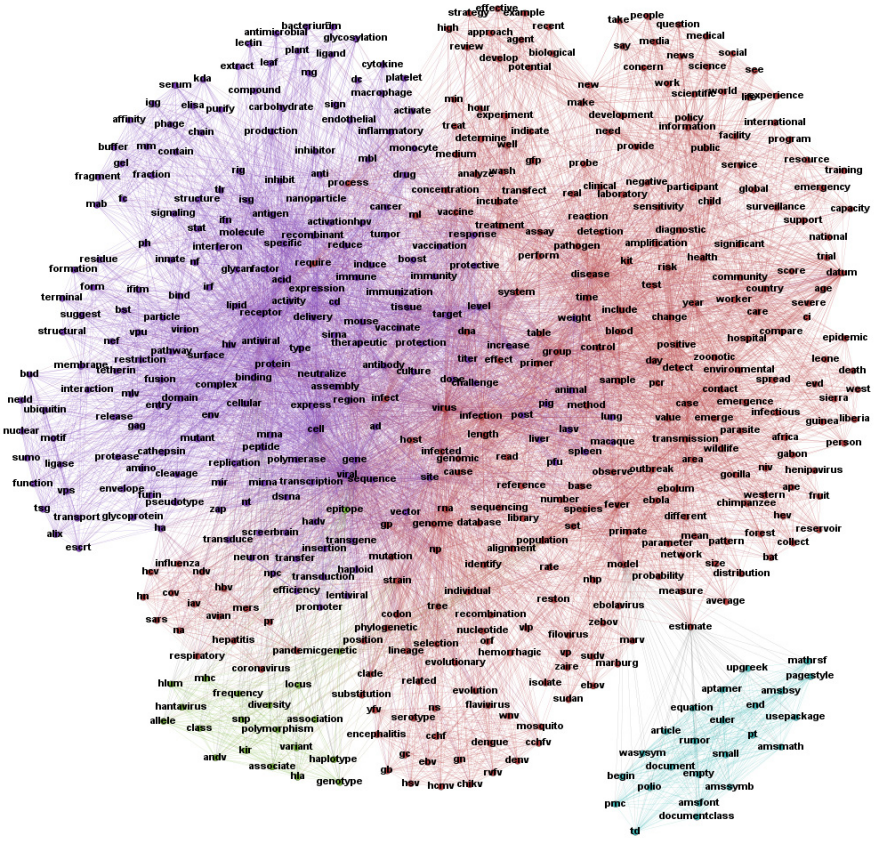

〈그림 6〉 바이오 관련 주제를 다루고 있는 뉴스 기사

이처럼 각 매체들은 서로 특정 주제들을 공유하기도 하며, 서로의 분야에 해당하는 주제들이 자신의 분야에 혼재되는 양상을 보이고 있다.

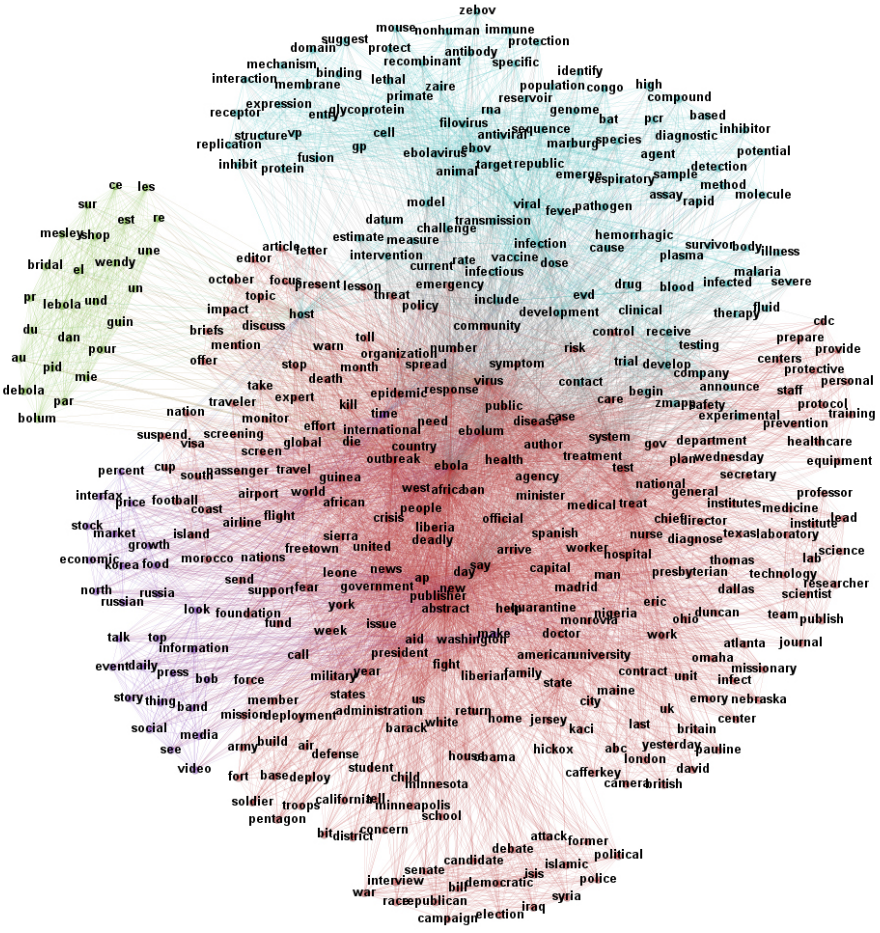
4.2.2 토픽맵

〈그림 7〉과 〈그림 8〉은 각각 PMC와 뉴스 데이터에서 도출된 토픽 구성 단어들을 노드로, 토픽별 단어의 가중치를 엣지로 설정하여 토픽맵을 그린 것이다. 두 단어가 같은 토픽에 등장할 경우 두 단어는 연결되어 있다는 가정하에 노드들이 연결되었으며 시각화에는 Gephi(Bastian et al. 2009)를 사용하였다. 아울러, 각 토픽맵 간의 상세한 비교를 위하여 네트워크 별로 기초통계분석을 수행하였으며 연결중심성 값을 기준으로 각 네트워크에서 중요한 단어들을 추출하였다. 연결중심성이란, 네트워크를 구성하고 있는 하나의 노드가 네트워크 상의 다른 노드들과 연결되는 정도를 나타내는 값으로, 그 값이 높을수록 다른 노드들과 많이 연결되어 네트워크에서 높은 영향력을 가지는 것으로 해석할 수 있다. 〈표 4〉는 그래프별 기초통계분석과 연결중심성 값으로 추출한 상위 30개의 단어들을 정리한 것이다.

학술 문헌 데이터로 구축한 토픽맵(그림 7)의 경우 해상도 값을 1.7로 설정한 모듈리티 알고리듬(modularity algorithm; Blondel et al. 2005) 수행 결과 4개의 커뮤니티가 감지되었으며 514개의 노드와 11,058개의 엣지로 구성되었




〈그림 7〉 학술 문헌의 토픽맵




〈그림 8〉 뉴스의 토픽맵

다. 좌측 상단의 커뮤니티의 경우, 주로 유전자, 단백질과 관련된 바이오 전문 분야에서 출현하는 단어들이 다수 출현하여 바이오 전문 분야의 주제들로 구성되어 있음을 알 수 있다. 오른쪽 상단의 커뮤니티의 경우 바이오 전문 분야에서 추출된 주제들임에도 불구하고 'People, International, Social, Emergence'와 같은 사회적 주제에 대한 단어들이 다수 출현하는 것을 볼 수 있다. 이 두 주요 커뮤니티 외에 좌측 하단의 커뮤니티의 경우, 에볼라와 연관이 있는 다른 바이러스의 상세한 정보들을 담고 있는 단어들로 구성되어 있는 것을 확인할 수 있다. 우측 하단의 커뮤니티에는 다소 에볼라와 연관성이 떨어지는 단어들이 커뮤니티를 구성하고 있는데, 이는 문헌상에서 자주 등장함에도 불구하고 토픽 모델링 결과를 구성하고 있는 주제들에서 중요한 의미를 갖고 있지 않아 낮은 가중치를 가진 주제들이 토픽맵 시각화 과정에서 커뮤니티를 이룬 것으로 보인다. 이를 통해 토픽맵 구축이 기존 토픽 라벨링 과정에서는 잡

〈표 4〉 네트워크 기초통계와 연결중심성 상위단어

| 학술 문헌의 토픽맵 | | | 뉴스의 토픽맵 | | |
|---|---|---|---|---|---|
| 노드(개) | 상위단어 | 연결중심성 | 노드(개) | 상위단어 | 연결중심성 |
| 514 | virus | 316 | 417 | ebola | 404 |
| 엣지(개) | cell | 264 | 엣지(개) | ebolum | 342 |
| 11085 | viral | 253 | 9358 | virus | 316 |
| 해상도 | infection | 238 | 해상도 | abstract | 254 |
| 1.7 | protein | 224 | 1.5 | publisher | 254 |
| 커뮤니티 개수 | expression | 180 | 커뮤니티 개수 | say | 235 |
| | disease | 163 | | disease | 226 |
| 4 | level | 153 | 4 | health | 187 |
| 평균 연결 중심성 | gene | 152 | 평균 연결 중심성 | ap | 192 |
| | rna | 140 | | africa | 221 |
| 43.027 | group | 137 | 44.882 | west | 186 |
| | table | 137 | | outbreak | 178 |
| | binding | 134 | | new | 172 |
| | include | 134 | | country | 122 |
| | mouse | 132 | | case | 123 |
| | time | 123 | | liberia | 129 |
| | strain | 123 | | treat | 133 |
| | antibody | 122 | | hospital | 116 |
| | datum | 122 | | official | 103 |
| | site | 118 | | world | 116 |

아낼 수 없었던 불필요한 단어들을 탐지할 수 있는 한 방법으로 적용될 수 있다는 것을 발견할 수 있었다. 〈그림 8〉의 뉴스 토픽맵은 417개의 노드와 9,358개의 엣지로 이루어져 있으며 1.5의 해상도 값을 지정한 모듈리티 알고리듬 수행 결과 4개의 커뮤니티가 감지되었다. 좌측 상단의 소규모 커뮤니티 역시 앞서 말한 불필요한 단어들로 구성되어 있으며, 뉴스에서 추출된 토픽들로 구성된 토픽맵임에도 불구하고 우측 상단의 커뮤니티의 경우 Cell, Immune, Genome, RNA 등 바이오 전문분야의 용어들이 밀집해 있는 것을 발견할 수 있다. 중앙의 커뮤니티는 특정 지명, 발발, 전염 등을 나타내는 단어들로 구성되어, 뉴스의 대부분을 차지하는 주제들이 해당 커뮤니티에 속해 있는 것을 알 수 있다. 좌측 측면의 커뮤니티에서는 에볼라 발병 국가 외의 국가명, 정보와 경제에 관련된 단어들이 다수 발견되어, 에볼라와 직접적인 관련은 없으나 뉴스에 등장할 법한 간접 파급력에 대한 주제들이 나타나고 있음을 알 수 있다.

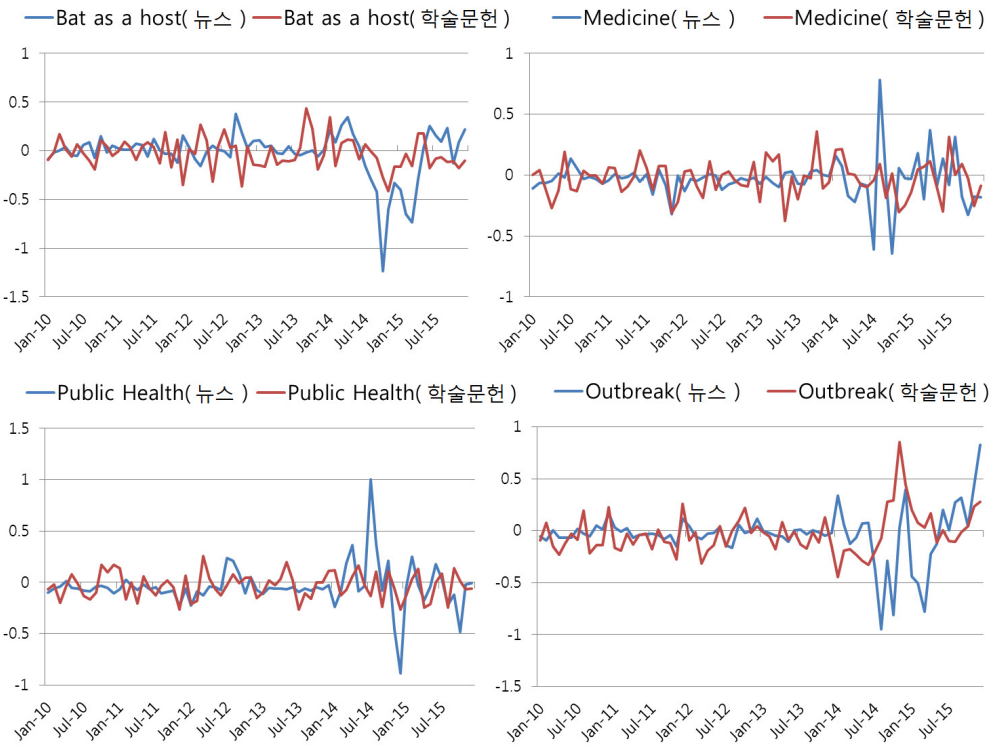
한편, 〈표 4〉에 기술된 기초통계를 따르면, 두 네트워크 간의 평균 연결 중심성 정도는 크게 다르지 않으며, 그 외의 특성들도 큰 차이를 보이고 있지 않다. 그러나 연결 중심성을 기준으로 추출한 20개의 단어들을 살펴보면, 두 데이터에서 주요하게 다루고 있는 토픽 간의 차이를 분명히 확인할 수 있다. 가장 큰 차이는 바이오 분야의 학술문헌에서는 'Ebola'라는 단

어가 중요하게 쓰이고 있지 않다는 점이다. 이는 'Ebola, Ebolum'가 뉴스 데이터에서는 가장 중요한 단어로 추출된 것에 대비되는 부분으로, 학술문헌에서는 에볼라 바이러스보다는 'Virus' 자체가 중요한 단어라고 해석할 수 있다. 또한, 뉴스의 경우 'Africa, West, Country, Liberia'와 같이 지명을 구체적으로 나타내는 단어가 사용되고 있는 반면, 학술문헌에서는 'Site'라는 상위의 단어로 이러한 지명을 포괄하고 있는데, 이를 통해 학술문헌에서는 뉴스와는 달리 발병 지역이나 현재 진행 상황이 크게 중요하지 않다는 것을 알 수 있다. 한편, 뉴스의 토픽맵에서 상위에 등장한 일반적인 단어(abstract, publisher, say)들은 뉴스에서 관용적으로 사용하는 표현으로 뉴스가 가지고 있는 매체적 특성이라 할 수 있다. 마지막으로, 학술문헌에서는 'Cell, Protein, Gene, RNA'와 같이 바이러스가 직접적으로 작용하는 생명체 내의 개체단위가 중요한 단어들로 추출된 반면, 뉴스에서는 'Country, Case, Treat, Hospital'과 같이 사람을 연상시키는 단어들이 중요 단어로 추출되었다. 이는 각 매체가 에볼라가 적용되는 단위에 대한 각 매체별 관점의 차이를 반영한 것으로 볼 수 있다.

#### 4.2.3 토픽 시계열 분석

본 연구에서는 DMR을 사용한 토픽모델링의 변수로 날짜데이터를 선정하였다. 학술 문헌과 뉴스 데이터 모두 에볼라가 가장 유행했던 기간인 2010년 1월부터 2015년 12월까지 데이터를 대상으로 분석을 실시하였다. 〈그림 9〉는 두 매체에서 동시에 등장한 주제들(Bat as a host, Medicine, Public Health, Ebola Outbreak)을 대상으로 시간에 따라 문헌집단에서 해당 토픽이 가진 가중치가 어떻게 변화하는 지 그 추이를 시계열 그래프로 나타낸 것이다.

〈그림 9〉에 따르면, 학술 문헌과 뉴스 모두 Bat as a host라는 주제의 가중치가 기준점에서 크게 변동하지 않다가 학술 문헌의 경우 13년 하반기 그 비중이 높아지는 것을 볼 수 있다. 박쥐가 에볼라 바이러스의 숙주로서 인식되고 관련 연구가 진행되어 왔으며 2013년 하반기에 이와 관련한 논문이 바이오 분야 저널에 등재되었음을 추론해 볼 수 있다. 또한 뉴스의 경우 2014년 중반부터 Bat as a host 주제 가중치의 비중이 크게 떨어진다. 이는 해당 시기에 뉴스에서 Bat as a host 주제보다 다른 주제를 중요하게 다루고 있음을 시사한다. 그 예로 우측 상단의 뉴스 내의 Medicine 시계열 그래프를 살펴보면 2014년 7월부터 그 비중이 급격하게 높아졌는데, 이러한 주제들에 비해 Bat as a host 주제의 비중이 급격하게 하락한 것을 의미한다. Medicine 주제는 학술문헌에서는 그 비중의 변화가 작지만 뉴스에서는 2014년 7월을 기점으로 크게 증가하는 모습을 보인다. 이는 에볼라의 첫 발견 이후 역사상 가장 큰 규모의 에볼라 유행이 2014년에 발생했기 때문으로 보인다. 2014년 10월 1일 WHO 발표 기준으로 23,406명이 감염되었고 이 중 9,467명이 사망한 서아프리카 에볼라 유행으로 인해 뉴스에서 에볼라를 치료할 수 있는 약과 관련된 기사를 높은 비중으로 다루었을 것이라 추정할 수 있다. 한편, 〈그림 9〉의 좌측 하단 Public Health 주제의 시계열 그래프는 Medicine 주제와 비슷한 양상을 보이고 있다. 학술 문헌의 경우 관련 주제의 비중이 크게 변하는 지점이 없으나 뉴스의 경우 2014년 서아프리카 에볼라 유행 기간 동안 공중 보건과 관련된




〈그림 9〉 문헌집단 내 주제 가중치의 시계열 그래프

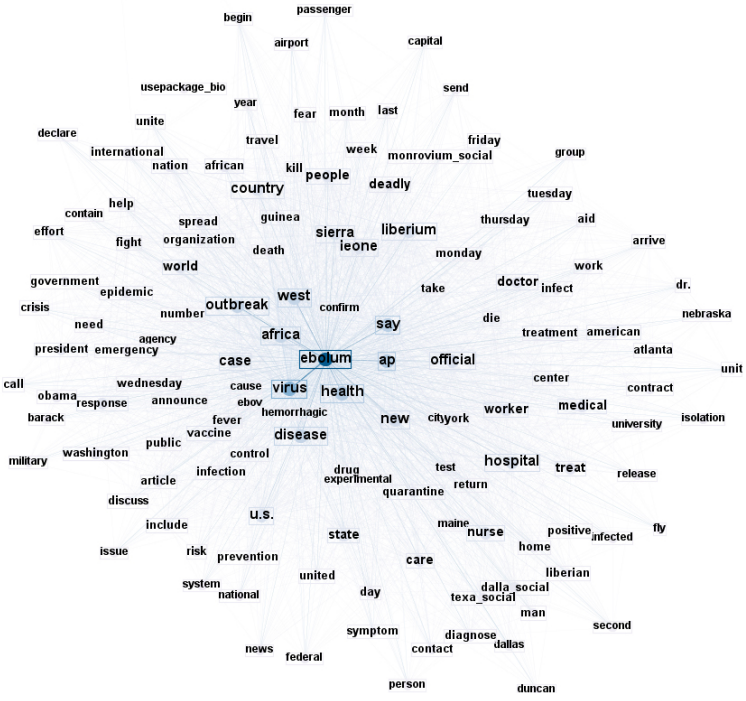
주제의 비중이 크게 올라갔음을 볼 수 있다. 이러한 그래프 양상을 보이는 이유는 2014년 에볼라가 유행한 지역인 기니, 라이베리아, 시에라리온 3국은 인구 이동이 활발한 반면 공중 보건 의료 체계가 부실하여 에볼라 바이러스가 유행하게 되었다는 주장이 언론을 통해 많이 보도되었기 때문이다. 우측 하단의 Ebola Outbreak 주제 가중치 시계열 추이를 살펴보면 학술 문헌의 경우 2014년 후반에 그 비중이 높아졌음을 볼 수 있다. 이는 2014년 초에 에볼라가 서아프리카 지역에서 대규모로 유행하고 이와 관련된 논문이 2014년 후반부에 나오면서 에볼라 발병에 관한 내용을 논문에서 다뤘기 때문으로 보인다. 반면 뉴스의 경우 2014년 초, 에볼라 바이러스가 서아프리카 지역에서 유행하기 시작하자 바로 에볼라 발병과 관련된 주제가 증가했음을 시계열 그래프를 통해 확인할 수 있다. 또한 에볼라 발병에 관한 주제의 비중은 2014년 중반 다시 급격히 줄어드는데 이는 발병에 대한 기사가 줄어들고 다른 주제와 관련된 기사, 즉 앞서 살펴본 공중 보건, 의약품과 같은 토픽에 대한 기사 비중이 높아졌음을 알 수 있다. 이를 통해 뉴스 기사는 논문에 비해 어떠한 사건에 대해 즉각적인 반응을 보이며 주제의 전환이 빠른 속도로 이뤄지는 특성을 가지고 있다는 것을 재확인 할 수 있다. 한편, 학술문헌의 경우 뉴스 정도의 변화는 아니지만, 에볼라의 발병 시기에 따라 각 주제의 가중치가 변화하는 것을 발견할

수 있었다. 특히, 창궐을 의미하는 'Outbreak' 주제에 민감한 양상을 보이는데, 이는 전문적인 연구 분야라 할지라도 사회적 상황에 영향을 받을 수밖에 없다는 전염병의 특성이 반영된 것으로 판단할 수 있다.

### 4.3 동시출현 네트워크

〈그림 10〉 ~ 〈그림 12〉는 동시출현 빈도 상위 개체 3개(Ebolum, Virus, Cell)를 이용하여 동시출현 네트워크를 구축한 결과이다. 각 네트워크별로 기초통계분석을 수행하였으며 해당 개체와 함께 한 문장 내에 출현하는 단어가 학술문헌에만 출현하는 지, 뉴스에만 출현하는지 확인하기 위하여 단어들에 각각 bio, social 라벨을 붙여 구분하였다. 라벨이 없는 단어는 학술문헌과 뉴스에서 모두 출현한 개체로 한 매체에만 국한된 것이 아니라 두 매체에서 모두 사용되는 단어이다.

동시출현 네트워크를 통해 분석한 결과는 다음과 같다. Ebolum은 전체 출현 빈도 중 90퍼센트 이상이 뉴스에 해당되며, cell은 이와 반대



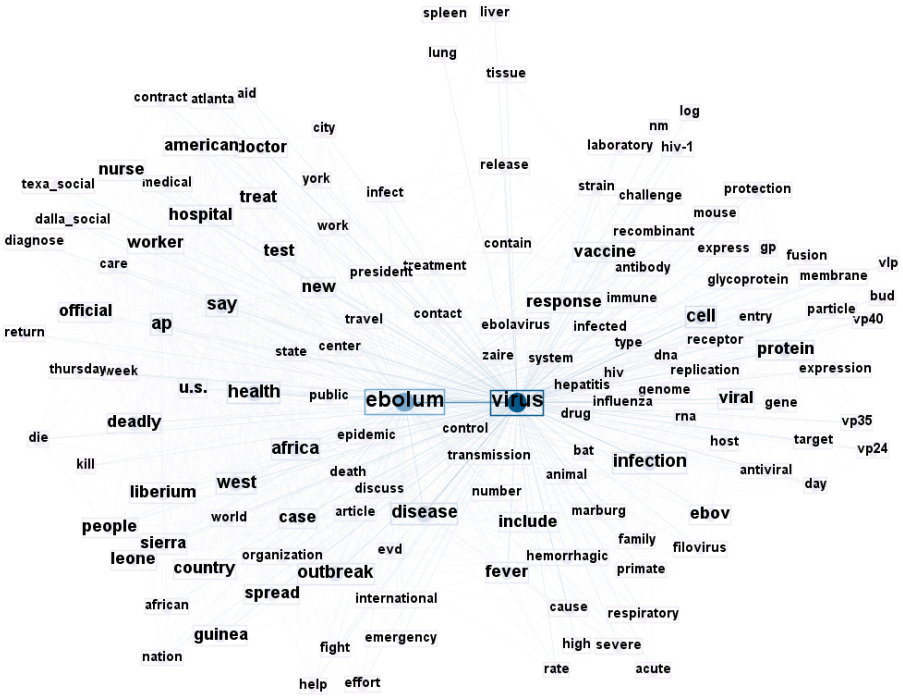

| 노드 개수 | 엣지 개수 | 학술 문헌에만 출현한 단어의 개수(_bio) | 공통출현 단어 | 뉴스에만 출현한 단어의 개수(_socio) | 평균 연결중심성 |
|---|---|---|---|---|---|
| 1,350 | 6,541 | 21 | 1,296 | 33 | 9.69 |
| 연결중심성 상위 20개 단어 | Ebolum, Virus, Ap, Health, Say, West, Africa, Outbreak, Disease, New, Hospital, Liberium, U.S., Official, Case, Sierra, Leone, Country, nurse, State | | | | |

〈그림 10〉 Ebolum 동시출현 네트워크

로 90퍼센트의 빈도가 학술 문헌에 치우쳐 있다(〈표 1〉 참조). 그럼에도 불구하고 이러한 개체들과 함께 출현한 단어들은 두 매체 모두에서 사용되고 있는 것이다. 이러한 매체 간 유사성은 각각의 동시출현 네트워크에 더 구체적으로 나타나 있다.

〈그림 10〉에서 가장 상위에 나타난 단어들은 주로 에볼라의 발병 지역, 상황 등을 묘사하는 단어들이다. 이는 개체 Ebolum이 주로 등장한 매체가 뉴스라는 사실과 관련 있는 것으로, 단어 뒤의 social 라벨을 통해 Monrovium와 Texa와 같이 지명을 나타내는 구체적인 단어들은 뉴스에서만 등장한 것을 확인할 수 있다. 한편, Usepackage의 경우 bio 라벨이 붙어있는 것으로 보아, 해당 개체는 Ebolum과 높은 동시출현 빈도를 가지면서도 학술 문헌에만 등장하였다. 이는 학술 문헌에서도 Eblum이라는 개체가 Usepackage와 함께 한 문장에 여러 번 사용되었다는 것을 의미한다.

Virus 동시출현 네트워크인 〈그림 11〉에 따르면, 학술 문헌과 뉴스에서 자주 사용될 법한 단어들이 혼재되어 높은 동시출현 빈도를 보였

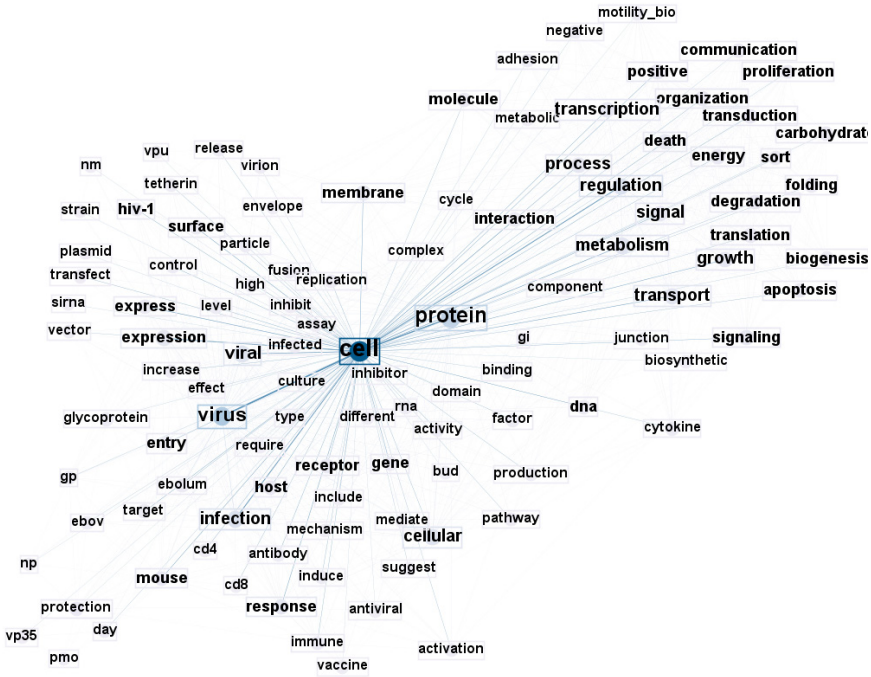



| 노드 개수 | 엣지 개수 | 학술 문헌에만 출현한 단어의 개수(_bio) | 공통출현 단어 | 뉴스에만 출현한 단어의 개수(_socio) | 평균 연결중심성 |
|---|---|---|---|---|---|
| 1,259 | 3,552 | 43 | 1,210 | 6 | 5.643 |
| 연결중심성 상위 20개 단어 | Virus, Ebolum, Cell, Disease, Health, Infection, Ap, Say, Protein, Outbreak, Africa, West, Viral, Vaccine, Hospital, Fever, Deadly, Case, New, u.s. | | | | |

〈그림 11〉 Virus 동시출현 네트워크

다. 이는 다른 개체들과 달리, Virus 개체 자체가 두 매체 중 하나에 치우치지 않고 유사한 비율로 출현했기 때문이라고 해석할 수 있다. 한편, 〈그림 10〉의 Ebolum 동시출현 네트워크의 평균 연결중심성(9.69)에 비해 낮은 연결중심성을 보이고 있는데, 이는 Ebolum 동시출현 네트워크의 대부분을 구성하고 있는 뉴스 텍스트의 문장 길이가 학술 문헌을 이루고 있는 문장 길이보다 현저히 짧은 데에서 기인한 것으로 볼 수 있다. 또한, Ebolum 동시출현 네트워크와 같이 Texa, Dalla와 같은 지명을 나타내는 단어에 social 라벨이 붙어있어, 이 단어들은 뉴스에서만 사용되었다는 것을 확인할 수 있다.

〈그림 12〉는 Cell 동시출현 네트워크로, 상위 단어들의 대다수가 바이오 전문 분야에서 주로 사용되는 단어들이다. 그러나, 단어들에 아무 라벨도 붙어있지 않은 것으로 보아, 비록 학술 문헌에 등장할 법한 단어들이지만 뉴스에도 사용되고 있다는 사실을 확인할 수 있다. 한편, 본 네트워크의 가장 큰 특징은 그래프를 구성하고 있는 주요 단어들 중 bio나 social 라벨이 붙은 경우가 없다는 점인데, 이는 학술용어라 하더라




| 노드 개수 | 엣지 개수 | 학술 문헌에만 출현한 단어의 개수(_bio) | 공통출현 단어 | 뉴스에만 출현한 단어의 개수(_socio) | 평균 연결중심성 |
|---|---|---|---|---|---|
| 920 | 2,198 | 55 | 865 | 0 | 4.778 |
| 연결중심성 상위 20개 단어 | Cell, Virus, Protein, Infection, Regulation, Cellular, Viral, Metabolism, Transport, Response, Expression, Process, Transcription, Growth, Signal, Membrane, Energy, Death, Positive, Mouse | | | | |

〈그림 12〉 Cell 동시출현 네트워크

도 뉴스에도 사용된다는 앞서 말한 사실을 재확인 시켜준다. 이를 통해 뉴스 매체의 보도가 "전문적인 지식보다 일반적인 정보 전달에 머무를 것이다."라는 기존의 생각과는 달리 에볼라와 관련된 전문적인 바이오 분야의 정보 또한 다수 포함하고 있는 것으로 해석할 수 있다.

## 5. 결 론

본 연구에서는 에볼라와 관련하여 두 매체(학술문헌, 뉴스) 간의 주제 분석을 위해 토픽 모델링과 동시출현 네트워크 분석 기법을 수행하고 두 매체에서 나타나는 특성들을 살펴보았다. 우선 토픽모델링을 통해서는 학술문헌과 뉴스 간의 주제적 차이를 극명하게 알 수 있었다. PMC의 경우 에볼라와 관련된 주제로 RT PCR, siRNA, DC Sign, Monoclonal Antibody와 같은 바이오 분야의 전문적인 주제가 나타나는 반면, 뉴스의 경우 Health Organization, American Politics, Sierra Leone, Military 등과 같은 사회적 주제가 많이 나타나는 것을 볼 수 있었다. 또한 토픽 모델링 결과를 이용하여 시계열 분석을 해 본 결과 기존에 알려진 것처럼 뉴스는 학술문헌보다 이슈에 민감한 양상을 갖고 있음을 알 수 있었다. 이는 에볼라와 관련된 어떠한 주제가 이슈가 되었을 때와 그러한 이슈와 관련된 학술문헌이 나오기까지의 시간이 길다는 매체의 특성에서 기인한 것으로 볼 수 있다. 추후 이를 보다 정확히 밝히기 위하여 상관성 분석 등의 추가적 검증을 수행할 예정이다.

한편 토픽맵과 개체별 동시출현 네트워크 분석을 통해서는 뉴스 매체와 학술 문헌 사이의 공통점을 발견할 수 있었다. 학술 문헌에서는 people, international, social, emergence와 같은 사회적 이슈와 관련된 단어들이 주요 주제를 구성하고 있었으며, 뉴스에서 cell, immune, genome, RNA와 같은 바이오 분야의 전문단어들이 주요 주제를 구성하고 있었기 때문이다. 또한 출현빈도가 높은 개체(Ebolum, Virus, Cell)를 중심으로 구축한 동시출현 네트워크에서는 학술 문헌과 뉴스가 거의 대부분의 단어들을 공통(Ebolum, Virus: 약 96%, Cell: 약 94%)으로 가지고 있었다. 이를 통해 상이한 매체라 생각되었던 학술 문헌과 뉴스가 미시적 차원에서는 유사한 내용을 가지고 있다는 것을 발견할 수 있었다. 다만, 이것이 전염병 분야의 매체에서만 나타나는 특성인지 명확히 밝히기 위해서는 다른 분야의 매체별 분석 결과와 비교하는 작업이 필요할 것이다.

이렇듯 텍스트 마이닝 기법을 적용하여 학술문헌과 뉴스에서 에볼라 관련 주제와 중요 개체들을 함께 제시하고, 매체 간 특성을 보여줌으로써 바이오 전문 분야와 사회적 분야의 정보 생산자들이 연구를 하거나 현상을 분석할 때 관점의 다양성을 제공해 줄 수 있다는 점이 본 연구가 가지는 의의라 할 수 있다.

# 참 고 문 헌

[1] 김은경 외. 2013. 전염병의 경로 추적 및 예측을 위한 통합 정보 시스템 구현.『인터넷정보학회논문지』, 14(5): 69-76.

[2] 최정실. 2008. 법정전염병 감염관리를 위한 정보시스템 개발 및 효과.『기본간호학회지』, 15(3): 371-379.

[3] 황교상, 이태식, 이현록. 2014. 센서스 데이터를 기반으로 만든 전염병 전파 시뮬레이션 모델.『대한산업공학회지』, 40(2): 163-171.

[4] Bastian, M., Heymann, S. and Jacomy, M. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In *Proceedings of International AAAI Conference on Weblogs and Social Media*, May 17-20, 2009, San Jose, CA: 8: 361-362.

[5] Blei, D. M., Andrew Y. N. and Michael I. J. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3: 993-1022.

[6] Blondel, V. D. et al. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment*. [online] [cited 2016. 4. 20.] 〈http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/pdf〉

[7] Ding, Y. et al. 2013. "Entitymetrics: Measuring the Impact of Entities." *PLoS ONE*, 8(8): 1-14, e71416. [online] [cited 2016. 4. 20.] 〈http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071416〉

[8] Househ, M. 2015. "Communicating Ebola through Social Media and Electronic News Media Outlets: A Cross-Sectional Study." *Health informatics journal*. Advance online publication. [online] [cited 2016. 4. 20.] 〈http://jhi.sagepub.com/content/early/2015/02/03/1460458214568037.full.pdf〉

[9] Kim, E. H. J. et al. 2015. "Topic-based Content and Sentiment Analysis of Ebola Virus on Twitter and in the News." *Journal of Information Science*. Advance online publication. [online] [cited 2016. 4. 20.] 〈http://jis.sagepub.com/content/early/2015/10/05/0165551515608733.full.pdf+html〉

[10] Lee, D., Kim, W. C. and Song, M. 2015. "Finding the Differences between the Perceptions of Experts and the Public in the Field of Diabetes." In *Proceedings of the 24th International Conference on World Wide Web Companion*, May 18-22, 2015, Florence, Italy: 57-58.

[11] Manning, C. D. et al. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics:*

*System Demonstrations*, June 22nd-27th, 2014, Baltimore, Maryland: 55-60.

[12] Mimno, D. and McCallum, A. 2012. "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression." *arXiv preprint arXiv*: 1206.3278. [online] [cited 2016. 4. 20.] 〈https://arxiv.org/ftp/arxiv/papers/1206/1206.3278.pdf〉

[13] Pesquita, C. et al. 2014. "The Epidemiology Ontology: An Ontology for the Semantic Annotation of Epidemiological Resources." *J. Biomedical Semantics*, 5(4): 1-7. [online] [cited 2016. 4. 20.] 〈https://www.researchgate.net/profile/Francisco_Couto/publication/259805277_The_epidemiology_ontology_an_ontology_for_the_semantic_annotation_of_epidemiological_resources/links/0a85e532030f847104000000.pdf〉

[14] Salathe, M. et al. 2012. "Digital Epidemiology." *PLoS Comput Biol*, 8(7): 1-5. [online] [cited 2016. 4. 20.] 〈http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002616〉

[15] Seltzer, E. K. et al. 2015. "The Content of Social Media's Shared Images about Ebola: A Retrospective Study." *Public Health*, 129(9): 1273-1277.

[16] Towers, S. et al. 2015. "Mass Media and the Contagion of Fear: The Case of Ebola in America." *PLoS ONE*, 10(6): e0129179. [online] [cited 2016. 4. 20.] 〈http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129179〉

**• 국문 참고자료의 영어 표기**
(English translation / romanization of references originally written in Korean)

[1] Kim, Eungyeong et al. 2013. "Implementation of Integrated Monitoring System for Trace and Path Prediction of Infectious Disease." *Journal of Korean Society for Internet Information*, 14(5): 69-76.

[2] Choi, Jeong Sil. 2008. "Development and Evaluation of a Legal Communicable Disease Electronic System for Infection Control." *Journal of Korean Academy of Fundamentals of Nursing*, 15(3): 371-379.

[3] Hwang, Kyosang, Lee, Taesik and Lee, Hyunrok. 2014. "Epidemic Disease Spreading Simulation Model Based on Census Data." *Journal of the Korean Institute of Industrial Engineers*, 40(2): 163-171.