텍스트 마이닝 기법을 적용한 뉴스 데이터에서의 사건 네트워크 구축

Construction of Event Networks from Large News Data Using Text Mining Techniques

저자 (Authors)
이민철, 김혜진
Minchul Lee, Hea-Jin Kim

출처 (Source)
지능정보연구 24(1), 2018.3, 183-203(21 pages)
Journal of Intelligence and Information Systems 24(1), 2018.3, 183-203(21 pages)

발행처 (Publisher)
한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07408510

APA Style
이민철, 김혜진 (2018). 텍스트 마이닝 기법을 적용한 뉴스 데이터에서의 사건 네트워크 구축. 지능정보연구, 24(1), 183-203

ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org

J Intell Inform Syst 2018 March: 24(1): 183~203 http://dx.doi.org/10.13088/jiis.2018.24.1.183

텍스트 마이닝 기법을 적용한 뉴스 데이터에서의 사건 네트워크 구축*

이민철
연세대학교 문헌정보학과
(bab2min@yonsei.ac.kr)

김혜진
연세대학교 근대한국학연구소
(erin.hj.kim@yonsei.ac.kr)

전통적으로 신문 매체는 국내외에서 발생하는 사건들을 살피는 데에 가장 적합한 매체이다. 최근에는 정보통신 기술의 발달로 온라인 뉴스 매체가 다양하게 등장하면서 주변에서 일어나는 사건들에 대한 보도가 크게 증가하였고, 이것은 독자들에게 많은 양의 정보를 보다 빠르고 편리하게 접할 기회를 제공함과 동시에 감당할 수 없는 많은 양의 정보소비라는 문제점도 제공하고 있다. 본 연구에서는 방대한 양의 뉴스기사로부터 데이터를 추출하여 주요 사건을 감지하고, 사건들 간의 관련성을 판단하여 사건 네트워크를 구축함으로써 독자들에게 현시적이고 요약적인 사건정보를 제공하는 기법을 제안하는 것을 목적으로 한다. 이를 위해 2016년 3월에서 2017년 3월까지의 한국 정치 및 사회 기사를 수집하였고, 전처리과정에서 NPMI와 Word2Vec 기법을 활용하여 고유명사 및 합성명사와 이형동의어 추출의 정확성을 높였다. 그리고 LDA 토픽 모델링을 실시하여 날짜별로 주제 분포를 계산하고 주제 분포의 최고점을 찾아 사건을 탐지하는 데 사용하였다. 또한 사건 네트워크를 구축하기 위해 탐지된 사건들 간의 관련성을 측정을 위하여 두 사건이 같은 뉴스 기사에 동시에 등장할수록 서로 더 연관이 있을 것이라는 가정을 바탕으로 코사인 유사도를 확장하여 관련성 점수를 계산하는데 사용하였다. 최종적으로 각 사건은 각각의 정점으로, 그리고 사건 간의 관련성 점수는 정점들을 잇는 간선으로 설정하여 사건 네트워크를 구축하였다. 본 연구에서 제시한 사건 네트워크는 1년간 한국에서 발생했던 정치 및 사회 분야의 주요 사건들이 시간 순으로 정렬되었고, 이와 동시에 특정 사건이 어떤 사건과 관련이 있는지 파악하는데 도움을 주었다. 또한 일련의 사건들의 시발점이 되는 사건이 무엇이었는가도 확인이 가능하였다. 본 연구는 텍스트 전처리 과정에서 다양한 텍스트 마이닝 기법과 새로이 주목받고 있는 Word2vec 기법을 적용하여 봄으로써 기존의 한글 텍스트 분석에서 어려움을 겪고 있었던 고유명사 및 합성명사 추출과 이형동의어의 정확도를 높였다는 것에서 학문적 의의를 찾을 수 있다. 그리고, LDA 토픽 모델링을 활용하기에 방대한 양의 데이터를 쉽게 분석 가능하다는 것과 기존의 사건 탐지에서는 파악하기 어려웠던 사건 간 관련성을 주제 동시출현을 통해 파악할 수 있다는 점에서 기존의 사건 탐지 방법과 차별화된다.

주제어 : 사건 네트워크, 사건탐지, 자연어처리(NLP), 텍스트 마이닝, 토픽모델링

논문접수일 : 2017년 11월 19일 논문수정일 : 2018년 3월 9일 게재확정일 : 2018년 3월 16일
원고유형 : 일반논문 교신저자 : 김혜진

1. 서론

신문 매체는 국내외에서 발생하는 사건들을 살피는 데에 가장 적합한 매체이다. 과거에는 인쇄 뉴스 매체가 사건들을 전달하는 유일한 수단이었으나, 정보통신 기술의 발달로 온라인 뉴스

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A6A3A01079581).

매체가 다양하게 등장하면서 주변에서 일어나는 사건들에 대한 보도가 크게 증가하였고, 이것은 독자들에게 많은 양의 정보를 보다 빠르고 편리하게 접할 기회를 제공함과 동시에 감당할 수 없는 많은 양의 정보소비라는 문제점도 제공하고 있다. 따라서 방대한 양의 뉴스 데이터로부터 주요 사건들을 자동으로 요약하여 제공하는 것은 이용자들이 많은 사건들을 일목요연하게 살피는데 도움이 될 것이다. 또한 사건들 간의 관련성을 기반으로 사건 네트워크를 구축하여 제공한다면 독자가 현재의 사건을 이해하는데 있어 큰 도움을 받을 수 있을 것이다.

본 연구에서는 이것을 해결하기 위해 방대한 양의 뉴스 기사로부터 데이터를 추출하여 주요 사건을 감지하고, 사건들 간의 관련성을 판단하여 사건 네트워크를 구축함으로써 독자에게 현시적이고 요약적인 사건정보를 제공하는 기법을 제안하는 것을 목적으로 한다. 이를 위하여 2016년 3월부터 2017년 3월까지의 신문기사를 수집하고 한국 사회에서 일어났던 정치·사회 분야 사건들을 추적·요약하여 실제로 사건 네트워크를 구축하여 제시하였다.

특히 뉴스 기사의 경우 다양한 고유명사 및 합성명사가 등장하고, 이들 중 많은 경우가 형태소 분석기 사전에 등록되어 있지 않아 명사의 추출이 제대로 되지 않는 경우가 잦다. 같은 인물이나 기관, 지명이라도 다른 형태로 적는 경우도 많아 이 역시 높은 품질의 텍스트 정제 결과를 얻기 위해서는 해결해야 할 문제이다. 본 연구에서는 Word2Vec과 글자 단위 자카드 (Jaccard) 유사도를 이용하여 이형동의어 목록을 자동으로 추출·통합하는 방법을 사용하여 고유명사 및 합성명사 추출의 정확성을 높였다.

본 논문의 구성은 다음과 같다. 제2장 선행연구에서는 사건탐지 관련 선행연구를 살펴보았고, 제3장에서는 사건 네트워크를 구성하기 위한 뉴스 데이터의 수집 및 명사(구) 토큰 추출을 위한 전처리 기법에 대해서 다루었다. 제4장은 토픽모델링을 적용하여 사건을 탐지하고, 사건 간 관련성을 기반으로 사건 네트워크를 구성하는 기법을 제시하였다. 제5장은 사건 네크워크의 결과와 제6장은 결론을 제시하였다.

2. 선행연구

사건탐지(event detection)는 텍스트를 연관된 이야기로 분할하고, 새로운 사건을 탐지하여 기존의 사건이 어떻게 변화하는지를 추적하는 주제탐지 및 추적하는 기술(topic detection and tracking, TDT)의 하위 분야로, 텍스트 데이터에서 기존에 밝혀져 있지 않던 사건을 밝히는 것을 그 목적으로 한다(Atefeh and Khreich, 2015).

사건 탐지를 위한 기술은 크게 문헌 기반(document-pivot)과 자질 기반(feature-pivot) 기술로 구분할 수 있다. 문헌 기반 기법은 유사한 문헌을 클러스터링하여 사건을 탐지하는 기법으로, 여기에는 용어 벡터와 bag-of words 모형을 활용하는 전통적인 접근법(Salton, 1989)과 개체명 벡터를 이용하는 접근법(Kumaran and Allan, 2004)이 있다.

자질 기반 기법은 키워드 등 텍스트에서 추출된 자질을 이용해 사건을 탐지하는 기법으로, 트위터 등과 같은 새로운 형태의 미디어를 분석하는데 자주 사용되고 있다(He et al., 2007; Kleinberg, 2002).

LDA (Latent Dirichlet Allocation) 토픽 모델링 기법은 문서집단의 단어 출현빈도를 분석하여

문서가 가질 수 있는 주제와 그 주제에 포함될 단어들의 생성확률을 추정하는 기법이다(Blei et al., 2003). 토픽 모델링은 수많은 단어의 집합으로 표현되는 문헌을 비교적 적은 수의 잠재 토픽으로 압축하여 그 내용을 간결하게 보여줄 수 있다는 특징 덕분에 대량의 사용자 리뷰를 분석하여 분류하거나(Chae, 2015), 온톨로지와 함께 쓰여 지식 베이스를 구축하는데 사용되는(Jeong, 2015) 등 다양한 자연언어처리 분야에 응용되고 있다.

이 기법을 적용한 사건탐지 연구로는 문헌에 토픽 모델링 기법을 적용하여 확장성을 확보하고 주제와 관련 없는 부분을 제거하고자 한 연구(Ha-Thuc, 2009), 텍스트뿐만 아니라 이미지나 다른 형태의 데이터로부터 사건을 추출할 수 있도록 멀티 모달 토픽 모델링을 사용한 연구(Qian et al., 2016) 등이 있다.

국내의 사건 탐지와 관련된 연구에는 소셜 미디어를 수집하고 개체명 인식을 통해 사건 정보를 추출하고 시공간적으로 분석한 연구(Oh et al., 2014)가 있다. 특히 토픽 모델링을 사용하여 사건 탐지를 시도하고자 했던 연구에는 사용자 행동 및 시간 정보를 반영하도록 LDA를 개량하여 사용한 연구(Tsolmon, 2013)와 트위터로부터 이슈 트래킹 시스템을 구축한 연구(Bae, 2014)가 있었다.

본 연구는 아직까지 국내에서 시도한 사례가 없는 텍스트 마이닝 기법을 적용하여 대량의 뉴스 데이터를 대상으로 사건을 탐지하고 사건기반 연관 네트워크를 구축하는 방법을 제시하고자 한다.

3. 데이터 수집 및 전처리 과정

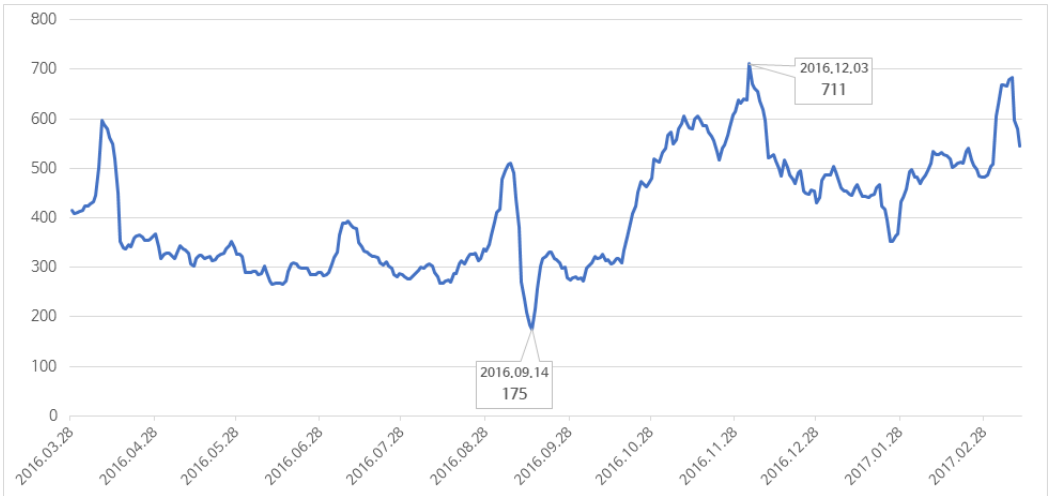
3.1 데이터 수집

본 연구에서는 지난 1년간 한국사회에서 발생한 정치·사회적 사건을 탐지하기 위해서 온라인으로 제공되는 뉴스 기사를 활용하였다. 뉴스 기사는 한국언론진흥재단에서 제공하는 뉴스 기사 데이터베이스인 BigKinds (www.bigkinds.or.kr)에서 방송사 4개사와 중앙지 8개를 대상으로 하였고 파이썬(www.python.org)을 활용하여 자동수집 하였다. 본 연구에서 수집한 뉴스 매체와 수집 건 수는 <Table 1>과 같다.

〈Table 1〉 The overview of data collection

Date	2016.03.28～2017.03.20
Target	- 4 Broadcasters: MBC, OBS, SBS, YTN - 8 Newspapers: Kyunghyang Shinmun, Kukmin Ilbo, Naeil Shinmoon, Munhwa Ilbo, Seoul Shinmun, Segye Times, Hankyoreh, Hankook Ilbo
Section	Politics, Society
Field	Title, Content, Organization, Publication Date
Total	146,703 News articles

<Figure 1>은 수집한 데이터의 일별 분포도를 나타낸 것이다. 일별 최소 175건에서 최대 711건까지 다소 변동이 있으나, 특정 구간에 집중되어 있지 않고 고른 분포도를 보이고 있다.



〈Figure 1〉 The number of articles by date

3.2 PMI를 활용한 토큰 결합

수집된 기사 본문에서 사건 네트워크를 구성하기 위해서는 제일 먼저 사건 구성의 단위인 명사(구)를 추출해야 한다. 본 연구에서는 코모란(Komoran) 형태소 분석기를 활용하여 총 1,135,190개의 명사 토큰을 추출하였다. 그러나 형태소 분석기를 통하여 자동 추출한 명사 토큰에서는 뉴스 기사가 가지고 있는 다양한 고유명사 및 합성명사가 제대로 반영되지 않았는데, 이는 형태소 분석기에서 제공하는 사전에 이러한 단어들이 포함되어 있지 않기 때문이다. <Table 2>는 잘못 추출된 토큰의 예시이다. 여기에서 NNP는 고유명사, NNG는 일반명사, MM은 관형사를 나타내는 형태소 태그이다.

〈Table 2〉 Examples of tokenization error

Form	Tokenization Result
최순실	최순/NNP + 실/NNG
조응천	조/NNB + 응천/NNP
우병우	우/NNP + 병우/NNP
두테르테	두/MM + 테/NNG + 르/NNG + 테/NNG

이 문제를 해결하기 위하여 연속된 n개의 토큰을 하나로 묶어 처리하는 n-gram 기법을 사용할 수 있으나, 이 경우 n의 값이 커짐에 따라 전체 어휘의 수가 기하급수적으로 증가하고, 그에 따라 문헌용어행렬이 희소해진다(sparse)는 한계가 있다. 또한 <Table 2>에서 볼 수 있듯 '최순실'은 2개의 토큰으로 분할되지만, '두테르테'는 4개의 토큰으로 분할되기에 n을 2로 설정할 경우 '최순실'은 바르게 묶을 수 있지만, '두테르테'를 바르게 묶어줄 수 없으며, 4로 설정할 경우, '두테르테'는 바르게 묶어주지만, '최순실'은 뒤따라오는 다른 토큰과 함께 묶이게 된다. 따라서 단순히 n-gram만으로는 잘못 분할된 토큰을 병합하는데 한계가 있다.

본 연구에서는 잘못 분할된 토큰의 병합을 해결하기 위해서 점별 상호정보량(Pointwise Mutual Information, 이하 PMI)를 이용하였다.

PMI는 두 토큰이 함께 등장하는 정도를 정보량을 바탕으로 계산한 지수이다. 두 토큰 A, B에 대해 PMI는 다음과 같이 정의된다(Bouma, 2009).

$$PMI(A,B) = \log\frac{p(A\cap B)}{p(A)p(B)} \quad (1)$$

이 때 p(A)와 p(B)는 각각 A와 B가 등장할 확률이고, p(A∩B)는 A와 B가 동시에 등장할 확률이다. 전체 문헌의 토큰 개수를 n이라고 하고, A가 등장하는 횟수를 a, B가 등장하는 횟수를 b, A와 B가 동시에 등장하는 횟수를 z라고 할 경우 PMI를 다음과 같이 계산할 수도 있다.

$$PMI(A,B) = \log\frac{nz}{ab} \quad (2)$$

이 값이 클수록 두 토큰은 함께 등장하는 경향이 강하고, 작을수록 함께 등장하지 않으려는 경향이 강하다. 하지만 이 PMI값의 범위는 해당 문헌의 크기에 따라 달라지므로 서로 다른 쌍들 간의 값을 비교하는 데에는 적합하지 않다. 서로 다른 쌍들 간의 비교를 위해 PMI값을 [-1, 1]의 범위로 정규화할 수 있는데(Lee, 2003), 본 연구에서는 다음과 같은 정규화 방법을 사용하였다.

$$NPMI(A,B) = \frac{PMI(A,B)}{-\log\ p(A\cap B)} = \frac{\log\frac{nz}{ab}}{\log\frac{n}{z}} \quad (3)$$

위에서 제시한 PMI 및 NPMI는 두 개의 토큰을 대상으로만 정의되기 때문에, 본 연구에서는 추가적으로 Van de Cruys가 제시한 일반화된 다변수(N) PMI 정의를 참고하여 3개 토큰에 대하여 PMI를 다음과 같이 계산하였다(Van, 2011)

$$PMI(A,B,C) = \log\frac{p(A\cap B\cap C)}{p(A)p(B)p(C)} = \log\frac{n^2z}{abc} \quad (4)$$

마찬가지로 여기서 a, b, c는 각각 토큰 A, B, C가 등장하는 횟수이며, z는 A, B, C가 동시에 등장하는 횟수를 의미한다. 유사하게 3개 토큰에 대한 NPMI는 다음과 같이 계산할 수 있다.

$$NPMI(A,B,C) = \frac{PMI(A,B,C)}{-2\log p(A\cap B\cap C)} = \frac{\log\frac{n^2z}{abc}}{2\log\frac{n}{z}} \quad (5)$$

4개 이상의 토큰에 대한 PMI 및 NPMI도 같은 방법으로 정의할 수 있으나, 길이 4 이상의 토큰 쌍마다 NPMI를 계산하는 것은 연산 비용이 크므로 길이 3 까지에 대해서만 NPMI를 계산하고 이를 씨앗으로 사용하여 확장해 나가는 방법을 사용하였다. 이를 위해 먼저 형태소분석의 결과로 얻은 모든 토큰을 대상으로 토큰 길이 2와 길이 3으로 결합하는 NPMI를 계산하였다.

<Table 3>에서 확인할 수 있듯 하나의 토큰으로 분석되었어야 할 단어들이 큰 쌍에서 높은 NPMI 값을 갖는 것을 알 수 있다. NPMI로 얻은 길이 2 와 길이 3인 토큰쌍을 기반으로 길이 4 이상의 연속된 토큰 쌍을 찾기 위해 본 연구에서는 패턴기반 알고리즘을 고안하여 활용하였다. <Table 3>의 **엔리케/NNP, 페냐/NA, 니/NNG**와 **페냐/NA, 니/NNG, 에토/NNP**를 보면 **페냐/NA, 니/NNG**가 겹친다는 것을 확인할 수 있다. 이는 문헌 내에서는 저 토큰들이 **엔리케/NNP, 페냐/NA, 니/NNG, 에토/NNP**와 같은 형태로 나타났을 거라는 사실을 간접적으로 보여준다.

이러한 패턴에 근거하여 길이 3의 토큰 쌍을 바탕으로 더 긴 토큰 쌍을 추론하는 알고리즘을 고안하였는데, 첫 번째 단계는 길이가 3인 토큰 쌍 목록을 바탕으로 길이 4 이상인 토큰 쌍 후보를 생성하는 알고리즘이 사용되었고, 두 번째 단계에서는 생성된 토큰 쌍 후보에서 부분적으로 겹치는 것을 제거하고 유의미한 것만 남겨 후보를 추려내는 알고리즘을 적용하였다.

〈Table 3〉 Token groups with high NPMI

Token Groups (length:2)	NPMI
우/NNP, 병우/NNP	0.9935
멜/NNP, 라니아/NNP	0.9889
최순/NNP, 실/NNG	0.9844
요시/NNP, 히데/NNP	0.9841
KD/SL, 코퍼레이션/NNP	0.9761
Token Groups (length:3)	**NPMI**
쿵/MAG, 쉬/MAG, 안유/NNG	0.9677
스가/NA, 요시/NNP, 히데/NNP	0.9571
패트/NNP, 리/XSN, 엇/EP	0.9534
엔리케/NNP, 페냐/NA, 니/NNG	0.9488
페냐/NA, 니/NNG, 에토/NNP	0.9375

● 후보생성 알고리즘: 길이 3 토큰 쌍 목록을 입력 받아 끝부분 2개와 앞부분 2개가 일치하는 같은 토큰 쌍을 묶어 더 긴 토큰 쌍을 생성한다.

초기화

1. 길이 3 토큰쌍 중 NPMI가 T 이상인 것을 모두 리스트 seed와 리스트 result에 넣는다.

확장

2. result에서 토큰 쌍 하나를 꺼내어 k라 하고, 그 토큰쌍의 NPMI를 v라 하자.
3. 토큰 쌍 k의 뒷부분 토큰 2개를 suffix라고 하자. seed에서 suffix로 시작하는 모든 토큰쌍을 찾는다.
   3.1. 찾은 토큰 쌍 중 하나를 꺼내어 c라 하고, 그 토큰 쌍의 NPMI를 u라 하자.
   3.2. k와 c를 연결하여(겹치는 suffix는 제거) 새로운 토큰 쌍을 만들고, 그 토큰쌍의 NPMI는 v*u로 계산하여 result에 추가한다.
   3.3. **3.1번**으로 돌아가 찾은 토큰 쌍 모두에 대해 반복한다.
4. **2번**으로 돌아가 모든 result내의 토큰 쌍에 대해 반복한다.

결과

5. result에는 길이 3 이상의 모든 토큰 쌍 후보가 포함되어 있다.

● 후보선별 알고리즘: 길이 3 이상의 토큰 쌍 목록을 입력 받아 그 중 의미 있는 토큰 쌍만 남긴다.

초기화

1. 길이 3 이상의 토큰쌍 목록의 모든 토큰쌍에 대해 문헌내 등장 빈도를 계산하여 토큰쌍 및 그 NPMI와 함께 리스트 eval에 넣는다.
2. eval은 토큰쌍의 길이에 대해 오름차순으로 정렬한다.

부분중복 제거

3. eval에서 토큰쌍 하나를 꺼내어 k라 하고, 그 토큰쌍의 NPMI는 v, 문헌내 출현빈도는 c라고 하자.
4. k에서 앞부분의 토큰 1개를 제거한 것을 f라 할 때, f가 eval 안에 존재하고, f의 문헌 내 출현빈도 d에 대해 d > c * R이면, eval에서 f를 제거한다.
5. k에서 뒷부분의 토큰 1개를 제거한 것을 r라 하고, 4번과 마찬가지 조건을 만족할 경우 eval에서 r을 제거한다.
6. 4번, 5번에서 f와 r을 제거하지 않고 모두 남아있는 경우, v < T이면 eval에서 k를 제거한다.
7. 3번으로 돌아가 모든 eval내의 토큰쌍에 대해 반복한다.

결과

8. eval에는 부분 중복이거나 NPMI가 낮은 토큰쌍이 제거되어 있다.

〈Table 4〉 Examples of token groups to be combined

Token Groups	Product of NPMI	Frequency
나/NP, 가미/NNG, 네/XSN, 야스/NNP, 마사/NNP, 주한/NNG, 일본/NNP, 대사/NNG (length: 8)	0.2492	188
엔리케/NNP, 페냐/NA, 니/NNG, 에토/NNP, 멕시코/NNP, 대통령/NNG (length: 6)	0.3910	100
두/MM, 테/NNG, 르/NNG, 테/NNG, 필리핀/NNP, 대통령/NNG (length: 6)	0.2302	394
동부전선/NNP, GP/SL, 기관총/NNG, 오발/NNG, 사고/NNG (length: 5)	0.7046	92
고/XPN, 고도/NNG, 미사일방어/NNP, 체계/NNG (length: 4)	0.8144	4736
서비스/NNG, 산업/NNG, 발전/NNG, 기본법/NNG (length: 4)	0.5049	443

상수 T값과 R값에 따라 두 알고리즘이 산출하는 결과가 달라진다. T값이 낮을수록 더 많은 후보가 뽑히고, R값이 작을수록 뽑힌 후보 중 부분적으로 겹치는 후보가 많이 제거된다. <Table 4>는 T = 0.5, R = 0.9로 설정하였을 때 결합 대상 토큰 쌍의 결과이다.

형태소 분석 이후, 텍스트 내에서 결합대상으로 선정된 토큰을 찾아서 언더 바(_)로 연결하여 결합하고 NNE라는 태그로 식별하였다.

3.3 Word2vec을 활용한 이형동의어 통합

많은 경우의 뉴스 기사에서 동일한 인물이나 기관, 대상이 여러 가지 형태의 다른 표현으로 작성되는 것을 흔히 발견할 수 있다.

예를 들어 헌법재판소는 종종 '헌법 재판소'라고 띄어쓰기를 포함해서 쓰이기도 하고, '헌재'라고 줄여서 쓰이기도 한다. 이렇게 같은 대상을 가리키지만 줄여 쓰거나 띄어쓰기를 달리하여 표현된 단어들을 '이형동의어'라고 한다. 이형동의어들은 전체 어휘 집합을 키우고, 문헌 용어행렬을 희소하게 하는 결과를 가져온다. 따라서 이형동의어를 통합하는 작업인 어형통제(vocabulary control)를 수행하는 것이 전반적으로 향상된 단어 추출의 결과를 가져온다. 어형통제는 수작업으로 작성된 이형동의어 목록을 통해 실시할 수도 있지만, 큰 문헌 집합을 대상으로 이 목록을 수작업으로 작성하는 것은 쉽지 않은 일이다. 따라서 본 연구에서는 Word2Vec과 글자 단위 자카드(Jaccard) 유사도를 이용하여 이형동의어 목록을 자동으로 추출하여 통합하는 방법을 사용하였다.

Word2Vec은 단어 임베딩 과정을 통해 개개의 단어를 특정한 차원의 벡터로 대응시키는 알고리즘(Goldberg and Levy, 2014)으로 이렇게 벡터로 대응된 단어들은 의미적으로 유사한 단어들끼리 유사한 값을 갖게 되는데 이것을 활용하여 해당 단어와 의미적으로 유사한 단어를 찾아낼 수 있다. Word2Vec의 입력 데이터로는 3.2에서 제시한 방법을 따라 형태소 분석 후 분할된 토큰을 결합한 텍스트 데이터를 사용하였고, 차원 수는 100, Skip-gram 알고리즘을 사용하여 학습하였다. 학습 후 일부 단어에 대해 유사한 단어를 추출한 결과 중 "헌법재판소"와 "문 전 대표"를 보면 다음 <Table 5>와 같다.

〈Table 5〉 Top 5 similar words of '헌법재판소' and '문_전_대표' in semantic

헌법재판소/NNP	
헌재/NNG	0.936
탄핵_심판/NNE	0.868
대통령_탄핵_심판/NNE	0.845
탄핵_심판_사건/NNE	0.782
헌법_재판소/NNP	0.754
문_전_대표/NNE	
안_지사/NNE	0.867
문재인_전_대표/NNE	0.854
문재인/NNP	0.790
안희정_충남지사/NNE	0.754
안희정_지사/NNE	0.712

Word2vec 기법을 활용하여 "헌법재판소"와 유사한 토큰으로 추출된 상위 5개의 단어 중에는 "헌재" 및 "헌법_재판소"가, "문 전 대표"와 의미적으로 유사한 토큰은 "문재인_전_대표"와 "문재인"이 포함된 것을 알 수 있다. 그러나 의미상으로는 유사하지만 이형동의어라고 할 수 없는 단어가 결과에 포함되어 있어서 이 자체로는 이형동의어를 선별해내는데 어려움이 있다. 이를 보완하기 위해 두 단어의 형태적 유사도를 측정하여 이형동의어를 가려내는 추가 자질로 활용하였다. 두 단어 A와 B의 형태적 유사도는 자카드 유사도의 변형을 이용하여 다음과 같이 정의한다.

$$FS(A,B) = \frac{|A \cap B|}{|A|+|B|-|A \cap B|+z} \quad (6)$$

이때 |A|, |B|는 각각 A, B단어에 포함된 서로 다른 글자의 수, |A∩B|는 A, B단어에 공통적으로 포함된 글자의 수이다. z는 전체 글자 수가 적은 단어 간의 유사도가 높게 나오는 것을 막기 위한 편향값으로 본 연구에서는 2로 설정했다. 글자의 수를 계산할 때 공백 및 특수기호는 모두 제거하였다. <Table 6>은 <Table 5>를 대상으로 형태적 유사도(spelling similarity)를 계산한 결과이다. 두 단어 A, B의 Word2Vec을 통한 의미적 유사도를 *SS(A, B)*라고 할 경우 최종적으로 A, B의 이형동의어 점수 *S(A, B)*는 *SS(A, B)*와 *FS(A, B)*의 조합으로 표현할 수 있다.

$$S(A,B) = SS(A,B)^{p} \cdot FS(A,B)^{1-p} \quad (7)$$

이 때 지수 p는 의미적 유사도와 형태적 유사도를 혼합하는 비를 계산하는 계수를 의미한다. 즉, p=1이면 의미적 유사도만 고려하고, p=0이면 형태적 유사도만 고려하게 된다.

〈Table 6〉 Spelling similarities between the similar words in the cases of '헌법재판소' and '문_전_대표'

헌법재판소/NNP	
헌재/NNG	0.286
탄핵_심판/NNE	0.100
대통령_탄핵_심판/NNE	0.077
탄핵_심판_사건/NNE	0.083
헌법_재판소/NNP	0.714
문_전_대표/NNE	
안_지사/NNE	0.000
문재인_전_대표/NNE	0.500
문재인/NNP	0.125
안희정_충남지사/NNE	0.000
안희정_지사/NNE	0.000

마지막으로 이형동의어 점수가 높은 단어쌍들을 묶어주기 위하여 클러스터링을 실시하였다. 이 과정을 통해 p=0.66으로 두고 이형동의어 목록을 생성하고, 이에 해당하는 단어들은 하나로 합침으로써 단어 집합 중 총 9278개의 이형동의어를 3583개로 통합하였다.

4. 사건탐지 과정

이 장에서는 전처리를 통하여 (합성)명사 및 이형동의어 처리과정을 거친 뉴스 데이터의 주요 사건을 탐지하고 사건들 간의 관련성을 산출하는 방법에 대해 논의하도록 하겠다.

본 연구에서 “사건”이라 함은 “특정한 시기에 발생하여 언론 보도의 주제를 변화시킨 일”이라고 정의한다. 이렇게 정의된 사건을 탐지하기 위해 본 연구에서는 다음과 같은 가정하였다.

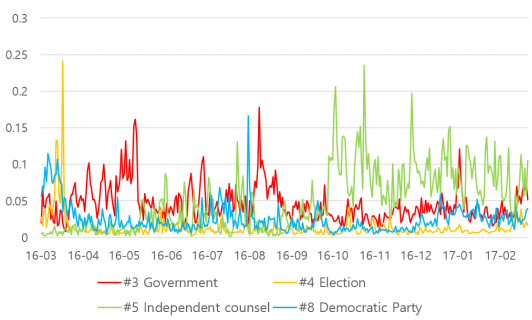
- 하나의 사건은 특정 주제에 속하며, 그 주제 분포에만 영향을 미칠 것이다.
- 언론의 반응은 사건이 발생한 직후에 가장 크게 나타나며, 시간이 지날수록 해당 사건에 의한 주제 분포 변동은 점차 줄어들 것이다.
- 둘 이상의 사건이 한 기사에 등장할 경우 서로 관련이 있을 것이다.

사건 탐지는 LDA 토픽 모델링 기법을 이용하였다. LDA는 문헌 집합 내에 등장하는 주제를 파악하는 확률기반 알고리즘으로 토픽 모델링 중 한 기법이다. LDA에서는 문헌별 주제 분포와 주제별 단어 분포가 디리클레 분포를 따른다고 가정하고, 생성 모델을 통해 관측되는 단어를 가지고 문헌의 주제를 추론해 나간다. 그 결과 하나의 문헌이 가지는 여러 가지의 주제의 분포와 문헌집단 내 포함된 주제의 비율을 산출할 수 있게 된다(Blei et al., 2003).

4.1 토픽모델링

LDA 기법을 적용하여 뉴스 기사의 주제를 추적하기 위해서 LDA 하이퍼 파라미터로 $\alpha = 0.1$, $\beta = 0.001$를 설정하고, 주제 개수 K를 16, 24, 32, 40로 바꾸어 가며 총 4회 LDA를 실시하였다. 결과 중 K = 32일 때가 유사한 주제가 나타나는 횟수가 적으면서도, 한 주제에 너무 많은 대상이 포함되지 않아 분석에 용이하다고 판단하여 이 값으로 주제분석을 진행하였다. <Table 7>는 K=32일 때 주제별 단어와 그 단어의 중요도를 나타낸다. 수집한 뉴스 기사의 날짜 정보를 이용하여 LDA를 통해 얻어진 토픽별 분포도를 시계열로 산출하여 중요한 토픽들이 날짜별로 어떤 분포를 나타내는지 살펴보았다.

<Figure 2>는 정부(#3 Government), 선거(#4 Election), 특검 수사(#5 Independent counsel), 더불어민주당(#8 Democratic Party) 토픽의 날짜별 비중 변화 그래프이다. <Figure 2>를 보면 각 토픽의 분포도가 날짜별로 다르게 나타나고 각 토



〈Figure 2〉 Distributions of four major topics by date

〈Table 7〉 Top 5 words and word weights for each topic (K=32)

#1 Assassination of Kim Jong-nam		#2 Constitutional amendment		#3 Government		#4 Election		#5 Independent counsel	
김정남 암살	.039	개헌	.033	방문	.027	선거	.041	특검	.052
경찰	.022	부분	.026	황교안 권한대행	.024	투표	.038	수사	.050
사건	.021	대통령	.024	논의	.017	선거일	.015	검찰	.046
남성 용의자	.017	말씀	.021	예정	.017	선관위	.014	조사	.031
확인	.011	상황	.018	만나다	.017	참여	.012	최순실 씨	.020
#6 Opposition parties		#7 Press release		#8 Democratic Party		#9 Economic policy		#10 DPRK	
안희정 충남도지사	.031	보도	.021	문재인 전 대표	.087	경제	.052	주재 북한 대사관	.014
대선	.028	밝히다	.020	대표	.067	정책	.040	리정철	.013
바른 정당	.025	언론	.016	더불어 민주당	.053	정부	.038	암살	.013
전 대표	.022	내용	.015	야권 대선 주자	.018	공약	.023	전하다	.010
반기문 전 총장	.022	확인	.014	강조	.018	일자리	.021	당국	.010
#11 Samsung		#12 Impeachment		#13 Anti-Park congressman		#14 THAAD Placement		#15 Economic support	
발언	.056	대통령 탄핵 심판	.035	비박계	.066	정부	.020	지원	.014
삼성전자 이재용	.047	대통령	.014	당	.061	한반도 사드 배치	.015	경제	.009
비판	.031	촛불	.012	의원	.060	외교	.013	사업	.008
논란	.029	쓰다	.007	탈당	.030	안보	.012	기업	.007
주장	.023	태극기	.007	지도부	.014	대북 제재 결의안	.012	회사	.006
#16 Conflicts between parties		#17 Speech of Park		#18 Election		#19 Congress		#20 Presidential candidates	
국민	.043	대통령	.189	지역	.081	의원	.239	후보	.124
황교안 권한대행	.020	박근혜 대통령	.160	의원	.023	국회	.031	경선	.063
책임	.013	밝히다	.015	표	.022	대통령 대리인단	.029	대선	.052
상황	.012	국정	.010	13 총선	.021	더불어 민주당	.018	최 씨	.021
정치	.012	입장	.009	선거	.020	국회의원	.015	대선 후보	.020
#21 Political reform		#22 Foreign politics		#23 Provocation of DPRK		#24 Protest		#25 Trial and hearings	
국민	.027	대통령	.081	발사	.026	집회	.059	출석 사유서	.103
정치	.022	국무장관	.016	노동미사일	.024	박 대통령	.026	증인	.055
개혁	.017	현지 시각	.010	훈련	.016	시민들	.022	청문회	.043
국가	.012	렉스티	.010	군	.010	열리다	.020	질문	.032
정책	.011	장관	.010	탄도미사일 시험 발사	.010	경찰	.020	국회	.029
#26 Election campaign		#27 Law		#28 History issues		#29 Constitutional court		#30 Congress	
회장	.039	법	.015	블랙리스트	.033	탄핵	.085	국회	.046
캠프	.030	예산	.009	재단	.025	대통령	.078	합의	.022
인사	.024	공무원	.008	정부	.025	헌법 재판소	.053	더불어 민주당	.019
출신	.023	지정	.008	위안부 소녀상 설치	.023	결정	.031	논의	.019
맡다	.023	규정	.007	서울중앙지검 특별수사본부	.017	박근혜 대통령	.021	자유 한국당	.017
#31 Scandal of Park & Choi		#32 Approval rating							
최순실 국정농단	.040	지지율	.042						
교수	.019	조사	.033						
인사	.012	포인트	.016						
이사장	.009	지지층	.015						
장관	.008	높다	.015						

픽이 발생한 시기에 분포값이 크게 나타남을 알 수 있다. 이는 해당 시기에 그 주제와 관련된 뉴스 기사의 양이 증가하였기 때문이다.

4.2 사건탐지

하나의 사건이 하나의 주제 분포에 영향을 미친다는 가정에 의해 각각의 주제 분포가 급등하는 시점에 특정 사건이 발생했다고 추론할 수 있다. 즉, 날짜별 주제 비중 변화 그래프에서 극대점을 찾으면 잠재적으로 사건이 있었을 시점을 탐지할 수 있다.

이를 위해 각 주제에 대해 날짜별 비중 변화 그래프를 하나의 수열로 보고 차분을 계산하였다. 차분의 부호가 양(+)에서 음(-)으로 바뀌는 지점이 극대점이므로 이를 만족하는 모든 지점을 찾는 것으로 사건발생을 추적할 수 있다. 또한 극소점으로부터 극대를 지나 극소점으로 돌아오기까지 그래프의 밑넓이를 구함으로써 사건이 해당 시점에 얼마나 큰 비중을 가졌는지를 계산하였고, 가우시안 스무딩(Gaussian Smoothing) 기법을 적용하여 사건을 필터링하였다.

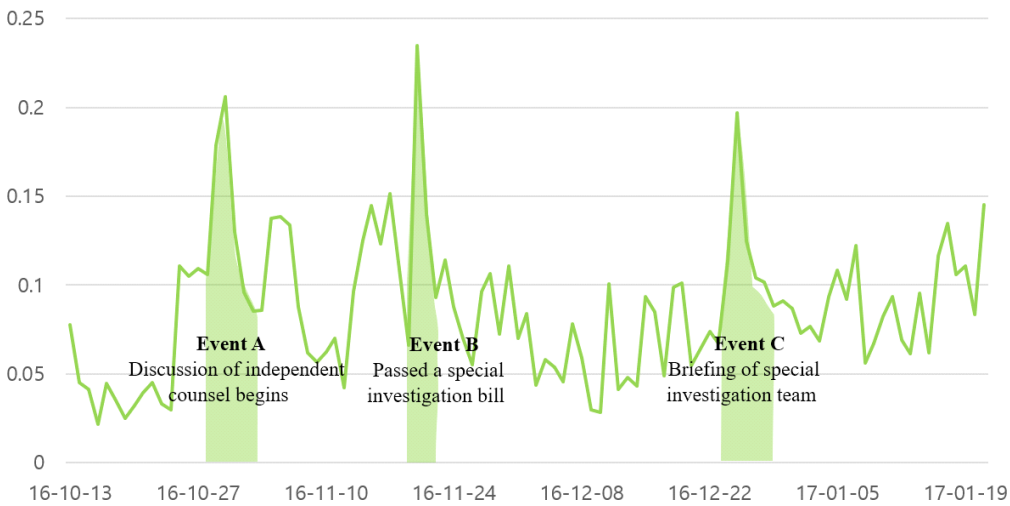
<Table 8>은 필터 크기에 따른 탐지된 사건 수 및 평균 사건 비중을 나타내고 있다. 필터 크기가 0일 때, 즉 필터링을 전혀 실시하지 않았을 때는 총 129개의 사건이 탐지되고, 평균 사건의 비중이 0.428로 작았던 반면, 필터 크기를 차츰 키워감에 따라 탐지되는 사건의 수가 줄어들고, 그 사건들의 평균 비중은 높아져가는 것을 확인할 수 있다.

<Figure 3>에는 “특검 수사(#5 Independent counsel)” 주제를 가지고 사건들을 탐지한 결과이다. 필터 크기를 2로 잡고 비중이 0.25 이상인 사건들만 추출하면 2016년 10월 29일에 극댓값을 보여주는 비중 1.245의 사건 A, 2016년 11월 16일에 극댓값을 보여주는 비중 0.663의 사건 B, 마지막으로 2016년 12월 25일에 극댓값을 보여주는 비중 1.166의 사건 C가 탐지된다.

마지막으로 각각의 사건 내용을 파악하기 위해서 토픽 모델링 결과를 활용하였다. 각 사건의 영향범위에 포함된 모든 문헌들에 대해 사건이 속한 주제의 단어들을 활용하여 실제로 뉴스 기사에서 해당 사건이 어떠한 단어들로 이루어졌는지 제시하였다. 예를 들어 <Figure 3>에서 나타난 사건 A의 영향범위에 속하는 뉴스 기사들에서 “특검 수사” 주제에 포함되는 단어를 추출하면, 최순실씨(14765회), 검찰(10413회), 수사(8044회), 의혹(7952회), 특검(4913회), 혐의(1470회) 등이 발견된다. 이것으로 A 사건은 “특검 수사 논의 시작”과 관련되었다는 것을 알 수 있다.

〈Table 8〉 The number of detected events by filter size

Filter size	Number of detected events	Average weight of events
0	129	0.428
1	118	0.542
2	85	0.757
3	63	0.968



〈Figure 3〉 Detected events in the topic "independent counsel" (topic #5)

본 연구에서는 32개 모든 주제에 대해 필터 크기를 2로 설정하여 위와 같은 방법으로 사건을 탐지하였고, 그 결과 총 85개의 사건을 탐지해낼 수 있었다. 그 중 비중이 1.0 이상인 사건들은 총 16개로 <Table 9>에 제시된 바와 같다. 사건 피크는 그 사건과 관련된 뉴스 기사가 가장 많았던 시점을 의미하므로, 반드시 실제 사건 발생일과 일치하지는 않는다. 예를 들어 "2차 대국민 담화(2nd speech of Park)" 사건 같은 경우 실제 담화는 11월 4일에 있었지만, 사건 피크는 11월 3일에 위치한다. 이는 언론들이 앞서서 해당 사건을 예의주시하며 예보를 하였기 때문으로 보인다. 마찬가지로 "김정남 암살 사건(Assassiantion of Kim Jong-nam)"의 경우 실제 사건이 일어난 날은 2월 13일이지만, 언론의 관심은 그보다 늦은 19일에 최고점을 찍었다.

4.3 사건 간 관련성 계산

본 연구에서는 LDA 토픽 모델링으로 사건을 탐지한 뒤, 사건이 속한 주제가 특정 뉴스 기사에 함께 등장하는 정도를 기반으로 하여 두 사건 간의 관련성을 측정하였고 이 결과를 이용하여 관련 사건들 간의 사건 네트워크를 구성하였다.

수식(8)은 코사인 유사계수를 활용하여 주제 p에 속하는 사건 A와 주제 q에 속하는 사건 B의 관련성을 계산하는 것으로, 두 사건 A, B의 관련성 점수 R(A, B)을 다음과 같이 정의될 수 있다.

$$R(A,B) = \frac{\sum_{e \in D(A) \cap D(B)} T_p(e) \cdot T_q(e)}{\sqrt{\sum_{e \in D(A)} T_p(e)^2 \cdot \sum_{e \in D(B)} T_q(e)^2}} \quad (8)$$

- D(A) : 사건 A의 영향 구간에 속하는 모든 뉴스 기사들의 집합
- D(B) : 사건 B의 영향 구간에 속하는 모든 뉴스 기사들의 집합
- $T_p(e)$: 뉴스 기사 e의 주제 p 비중

〈Table 9〉 16 detected events over event weight 1.0

Event Name	Topic	Peak Date	Effect Date	Weight	Top 5 Words
Legislative election, 2016	#4 Election	16.04.13	16.04.08~04.23	1.26	13총선, 지역, 당선, 선거, 무소속
National meetings of the Workers' Party of N. Korea	#10 DPRK	16.05.07	16.04.24~05.18	1.49	노동당대회, 제1비서, 대회, 위원장, 인민
Abolishment of emergency committee of Saenuri Party	#13 Anti-Park congressman	16.05.17	16.05.06~0.522	1.14	비박계, 당, 혁신, 혁신위, 새누리당 정진석 원내대표
Returning congressman issue of Saenuri Party	#13 Anti-Park congressman	16.06.17	16.06.07~0.623	1.31	복당, 비박계, 당, 비대위구성, 의원
N. Korea discharges water from dam without notice	#23 Provocation of DPRK	16.07.09	16.07.04~07.16	1.01	황강댐무단, 요격, 배치, 노동미사일, 국방부
Pukkuksong missile test of N. Korea	#23 Provocation of DPRK	16.08.24	16.08.17~09.01	1.01	신포급잠수함, 발사, SLBM시험발사, 제3후보지, 잠수함발사탄도미사일
Discussion of THAAD at G20 Summit	#3 Government	16.09.04	16.08.27~09.11	1.12	정상, 시진핑 중국국가주석, 회담, 방문, 참석
Discussion of sanctions against North Korea	#14 THADD placement	16.09.10	16.08.30~09.17	3.14	대북제재결의안, 한반도사드배치, 핵, 핵실험이후, 국제사회
N. Korea's 5th nuclear test	#23 Provocation of DPRK	16.09.10	16.09.01~09.17	1.40	핵실험이후, 노동미사일, 지진, 발사, 핵탄두
Discussion of independent counsel began	#5 Independent counsel	16.10.29	16.10.19~11.02	1.25	최순실씨, 검찰, 수사, 의혹, 특검
People call for action in opposition	#16 Conflicts between parties	16.11.03	16.10.21~11.10	1.74	국민, 국정, 야당, 책임, 사태
2nd speech of President Park	#17 Speech of Park	16.11.03	16.10.28~11.11	1.44	박근혜, 대통령, 거국중립내각, 총리, 김병준 국무총리내정자
Briefing of special investigation team	#5 Independent counsel	16.12.25	16.12.20~17.01.01	1.17	특검, 수사, 박영수 특별검사팀,최순실씨, 검찰
Ban Ki-moon pulled out of presidential race	#6 Opposition parties	17.01.31	17.01.28~02.18	1.51	안희정 충남도지사, 반기문, 바른정당, 불출마선언, 보수
Assassination of Kim Jong-nam	#1 Assassination of Kim Jong-nam	17.02.19	17.02.07~03.08	1.19	김정남 암살사건, 남성용의자, 말레이시아 경찰청장, 사건, 시신
Predent Park's final argument	#29 Opposition parties	17.02.26	17.02.15~03.02	1.02	헌법재판소, 탄핵, 심판최종변론, 이정미재판관 퇴임, 변론
Impeachment of President Park	#29 Opposition parties	17.03.10	17.03.02~03.21	1.33	탄핵, 헌법재판소, 선고, 사저, 이정미재판관 퇴임

이 정의에 따르면 두 사건 *A, B*의 영향 구간이 완전히 일치하고 주제 *p, q*의 비중이 해당 구간에서 서로 비례할 경우 관련성 *R(A, B)* = 1로 최대가 되고, *A, B*의 영향 구간이 전혀 일치하지 않거나, *p, q*의 비중이 전혀 상관이 없으면 *R(A, B)* = 0으로 최소가 된다.

<Table 10>은 2016년 10월 29일의 "특검 논의 본격 시작(Discussion of independent counsel began)" 사건과 다른 사건들 간의 관련성 점수를 계산하여 제시한 것이다. 관련성 점수 상위 4개의 사건 중 "국민들 야당에 행동 촉구(People call for action in opposition)"와 "2차 대국민 담화(2nd speech of President Park)"는 비교적 높은 관련성 점수를 보이는 반면, "2차 촛불집회(2nd Candlelight rally)"와 "트럼프 당선(Election of Trump as U.S. President)"은 관련성 스코어가 매우 낮음을 알 수 있다.

다만 위의 관련성 점수 공식은 몇 가지 한계를 가지고 있다. 먼저 서로 다른 *p*와 *q*에 대해서만 정의된다. 즉 같은 주제에 속하는 사건들에 대해서는 관련성을 계산할 수 없다는 한계점이 있다. 또한 두 사건의 영향 구간이 겹치지 않는 경우에도 분자가 0이 되므로 관련성을 계산하는 것이 불가능하다.

5. 사건 네트워크 구성 결과

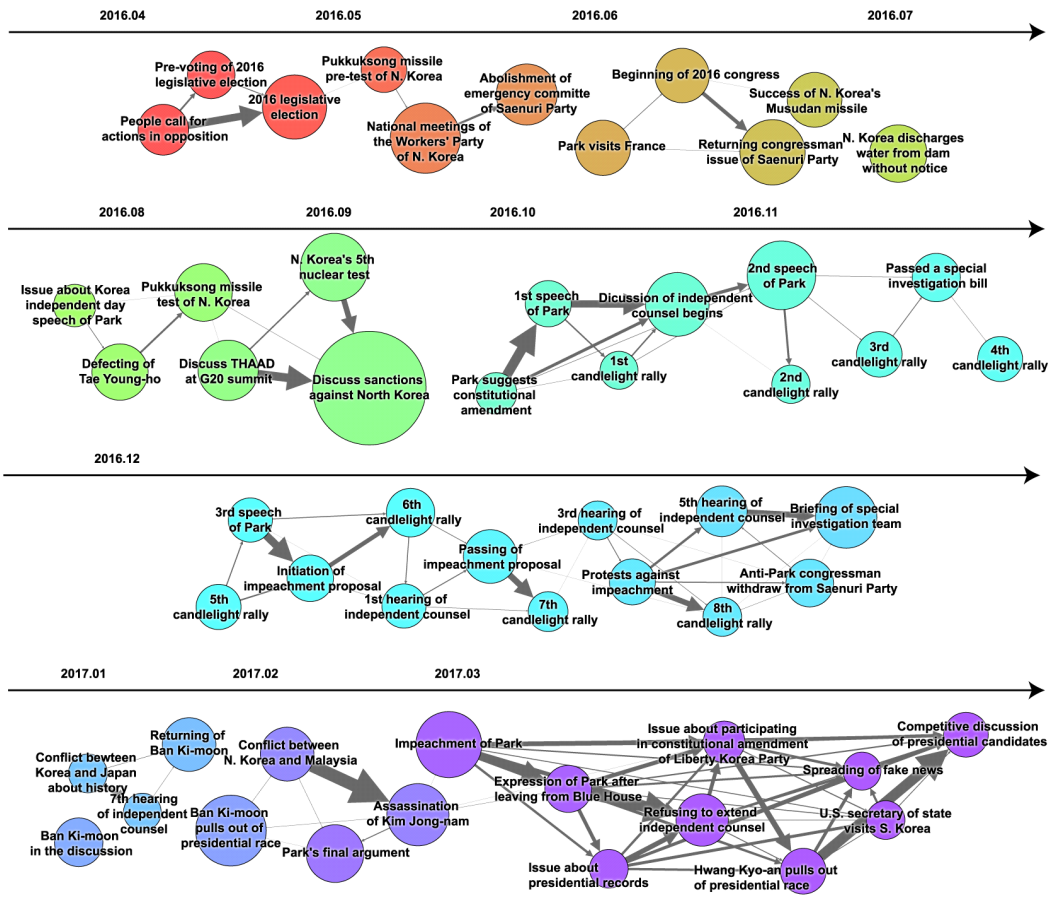
4장에서 제시된 방법을 통해 2016년 3월부터 2017년 3월 1년 간의 사건을 탐지하고, 사건 네트워크를 구성하였다(Figure 4). 사건 네트워크에서 각 노드는 사건을 의미하고, 노드의 크기는 해당 사건의 비중을, 각 노드간의 연결 강도는 관련성 점수의 크기를 의미한다.

<Figue 4>의 사건 네트워크를 살펴보면 각 사건들이 발생한 시간 순으로 배열되고, 연관된 사건들은 서로 연결된 것을 확인할 수 있다. 예를 들어, 2016년 8~9월을 살펴보면 G20 정상회의에서 사드 배치에 관한 논의(Discussion of THAAD at G20 summit)가 있었고, 이는 북한의 5차 핵실험(N. Korea's 5th nuclear test)과 연관이 깊다는 걸 알 수 있다. 그리고 이 두 사건은 최종적으로 9월부터 본격적으로 확정된 대북 제재 논의(Discussion of sanctions against North Korea)의 결정적인 원인이라고 볼 수 있는데, 이 관계가 네트워크를 통해 잘 드러난다.

2016년 12월에는 3차 대국민 담화(3rd speech of President Park)와 탄핵 소추안 발의(Initiation of impeachment proposal), 6차 촛불 집회(6th Candlelight rally) 등의 사건이 등장하는데, 연결 강도를 통해 3차 대국민 담화가 탄핵 소추안 발의에 주요한 영향을 미쳤고, 탄핵 소추안 발의는

〈Table 10〉 Top 4 related events of "Discussion of independent counsel began"

Event name	Relatedness	Topic	Peak date	Weight
People call for action in opposition	0.204	#16 Conflicts between parties	2016.11.03	1.741
2nd speech of President Park	0.124	#17 Speech of Park	2016.11.03	1.435
2nd Candlelight rally	0.007	#24 Protest	2016.11.05	0.281
Election of Trump as U.S. President	0.006	#22 Foreign politics	2016.11.09	0.644



〈Figure 4〉 The event networks during the year of 2016 to 2017

6차 촛불집회와 깊은 관계가 있다는 것을 확인할 수 있다.

2017년 3월에는 박 전 대통령이 탄핵되는 사건(Impeachment of President Park)이 있었고, 이 사건 이후 발생하는 사건들은 모두 해당 사건과 긴밀하게 연결되어 있다는 사실이 네트워크를 통해 잘 드러난다. 이를 바탕으로 박 전 대통령 탄핵 사건이 이후 발생하는 일련 사건들의 발단이라는 것을 쉽게 알 수 있다. 또한 이 시기는 앞선 다른 시기와 다르게 발생 사건들이 굉장히 촘촘하게 연결되어 있는데, 이는 해당 시기의 사건들이 기사에 동시 출현한 횟수가 많기 때문이고, 이를 통해 언론에서 이 사건들 간의 연관성을 비중 있게 다뤘음을 짐작할 수 있다.

몇 가지 주의해야 할 점은 구축된 사건 네트워크의 해석에 있어서 두 사건이 관련성을 가지고 있다고 해서 그것이 반드시 인과관계를 의미하는 것은 아니라는 것이다. 또한 사건 피크 시점

이 항상 사건 발생 일시와 일치하지 나타나는 것은 아니기 때문에 실제 사건의 선후관계와 네트워크에 나타나는 관계가 다를 수 있다. 예를 들어 2017년 2월의 김정남 암살 사건(Assassination of Kim Jong-nam)과 북한-말레이 갈등(Conflict between N. Korea and Malaysia)의 경우, 실제로 전자로 인해 후자가 발생하였지만, 전자에 대한 언론의 반응이 길게 지속되어 피크가 후자보다 밀렸기에 네트워크 상에서는 노드 간의 연결이 후자에서 전자로 연결되어 있는 것을 확인할 수 있다.

6. 결론

본 연구에서는 텍스트 마이닝 분야에서 자주 활용되는 LDA 토픽 모델링 기법을 활용하여 뉴스 기사에서 추출한 데이터로부터 주요 사건을 감지하고, 사건들 간의 관련성을 계산하여 사건 네트워크를 구축하는 방법을 제안하였다.

본 연구에서 “사건”을 “특정한 시기에 발생하여 언론 보도의 주제를 변화시킨 일”이라고 정의하고, (1)하나의 사건은 특정 주제에 속하며, 그 주제 분포에만 영향을 미칠 것이며, (2) 언론의 반응은 사건이 발생한 직후에 가장 크게 나타나며, 시간이 지날수록 해당 사건에 의한 주제 분포 변동은 점차 줄어들 것이며, (3) 둘 이상의 사건이 한 기사에 등장할 경우 서로 관련이 있을 것이라는 가정을 세우고, LDA 토픽 모델링 기법을 적용하여 뉴스 기사의 주제를 추적하고 사건을 탐지하여 제시하였다. 토픽 모델링 결과 총 32개의 주제가 추출되었고, 각각의 주제 분포가 급등하는 시점을 찾아 사건 발생의 시점을 추론하였다. 결과적으로 총 85개의 사건이 탐지되었으나 추가적으로 가우시안 스무딩 기법을 적용하여 최종 16개의 사건을 필터링하여 제시하였다.

마지막으로 코사인 유사계수를 활용하여 사건 간 관련성을 계산하고 사건들을 연결하여 사건 네트워크를 구성하였다.

본 연구는 텍스트 전처리 과정에서 다양한 텍스트 마이닝 기법과 새로이 주목받고 있는 Word2vec 기법을 적용하여 봄으로써 기존의 한글 텍스트 분석에서 어려움을 겪고 있었던 고유명사 및 합성명사 추출과 이형동의어의 정확도를 높였다는 것에서 학문적 의의를 찾을 수 있다. 기존의 사전기반 고유명사 및 합성명사의 추출은 형태소 분석기 사전에 등록되어있지 않은 경우 정확성이 떨어지는 문제점을 가지고 있었다.

본 연구에서 제시한 사건의 탐지 및 네트워크 구성 기법은 실무적 적용에 있어서 다음과 같은 장점을 가진다.

첫째, 비지도학습인 LDA 토픽 모델링을 활용하기에 방대한 양의 데이터로부터 주제 및 주제별 단어와 분포도를 쉽게 분석할 수 있다. 또한 수집한 뉴스 기사의 날짜 정보를 활용하여 토픽별 분포도를 시계열로 표현할 수 있다.

둘째, 기존의 사건 탐지에서는 파악하기 어려웠던 사건 간 관련성을 문헌 내 주제 동시출현을 이용하여 코사인 유사계수 기반 관련성을 산출하고 사건 네트워크를 구성하여 제시함으로써 현시적이고 요약적인 형태로 사건의 연결을 파악할 수 있다. 이것은 본 연구에서 제시한 사건 간 관련성 기반 사건 네트워크가 실제로 사건이 발생한 시간 순으로 구축되었음을 통해 알 수 있다. 또한 일련의 사건들의 시발점이 되는 사건이 무엇이었는가도 사건 네트워크를 통해 확인이

가능하다.

본 연구의 한계점은 LDA 토픽 모델링의 특성상 초기 파라미터와 주제 개수에 따라 결과가 달라지며, 분석 결과 나온 주제와 사건명을 연구자의 주관적 판단으로 부여해야하는 단점을 들을 수 있다. 또한 각각의 주제가 배타적이고 독립적이라고 가정하였으므로 주제 간 연관성을 고려하지 못한다.

후속 연구로 본 연구에서 다루어지지 않은 시간적으로 멀리 떨어진 사건이나 동일한 주제에 속하는 사건 간의 관련성 산출 등이 필요하다.

참고문헌(References)

Atefeh, F., and W. Khreich., “A survey of techniques for event detection in twitter,” Computational Intelligence, Vol. 31, No. 1 (2015), 132-164.

Bae, J, H., N. G. Han and M Song, “Twitter Issue Tracking System by Topic Modeling Techniques," Journal of Intelligence and Information Systems, Vol. 20, No. 2 (2014), 109~122.

Blei, D. M., A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” Journal of machine Learning research, Vol. 3, No. 1 (2003), 993-1022.

Bouma, G. “Normalized (pointwise) mutual information in collocation extraction,” Proceedings of the Biennial GSCL Conference Vol. 156. (2009), 31~40.

Chae S. H., J. I. Lim and J Kang, “A Comparative Analysis of Social Commerce and Open Market Using User Reviews in Korean Mobile Commerce,” Journal of Intelligence and Information Systems, Vol. 21, No. 4 (2015), 53~77.

Goldberg, Y., and O. Levy, “Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method,” arXiv preprint arXiv:1402.3722. 2014.

Ha-Thuc, V., Y. Mejova, C. Harris, and P. Srinivasan, “A relevance-based topic model for news event tracking.” Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, (2009) 764-765.

He, Q., K. Chang, E. P. Lim, and J. Zhang, “Bursty feature representation for clustering text streams,” Proceedings of the 2007 SIAM International Conference on Data Mining, (2007), 491-496.

Jeong, H., “A Study on Ontology and Topic Modeling-based Multi-dimensional Knowledge Map Services,” Journal of Intelligence and Information Systems, Vol. 21, No. 4 (2015), 79~92.

Kleinberg, J. “Bursty and hierarchical structure in streams,” Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (2002), 91-101.

Kumaran, G., and J. Allan, “Text classification and named entities for new event detection,” Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. (2004), 297-304.

Lee, J. Y., “A Study on Relative Mutual Information Coefficients,” Journal of the Korean Society for Library and Information Science, Vol. 34., No. 4 (2003), 177~198.

Oh, H. J., B. H. Yun, C. J. Yoo, and Y. Kim, "Trend Analysis using Spatial-Temporal Visualization of Event Information based on Social Media," Journal of Internet Computing and Services, Vol. 15, No. 6 (2014), 65~75.

Qian, S., T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis." IEEE Transactions on Multimedia, Vol. 18, No. 2 (2016), 233-246.

Salton, G. "Automatic text processing. Reading." MA: Addison-Wesley. 1989.

Tsolmon, B. "Extracting Social Events based on LDA Topic Model with Timeline and User Behaviour Analysis in Twitter Corpus," MS Thesis, Chonbuk University, 2013.

Van de Cruys, T. "Two multivariate generalizations of pointwise mutual information," Proceedings of the Workshop on Distributional Semantics and Compositionality, (2011), 16-20.

Abstract

Construction of Event Networks from Large News Data Using Text Mining Techniques

Minchul Lee* · Hea-Jin Kim**

News articles are the most suitable medium for examining the events occurring at home and abroad. Especially, as the development of information and communication technology has brought various kinds of online news media, the news about the events occurring in society has increased greatly. So automatically summarizing key events from massive amounts of news data will help users to look at many of the events at a glance. In addition, if we build and provide an event network based on the relevance of events, it will be able to greatly help the reader in understanding the current events.

In this study, we propose a method for extracting event networks from large news text data. To this end, we first collected Korean political and social articles from March 2016 to March 2017, and integrated the synonyms by leaving only meaningful words through preprocessing using NPMI and Word2Vec. Latent Dirichlet allocation (LDA) topic modeling was used to calculate the subject distribution by date and to find the peak of the subject distribution and to detect the event. A total of 32 topics were extracted from the topic modeling, and the point of occurrence of the event was deduced by looking at the point at which each subject distribution surged. As a result, a total of 85 events were detected, but the final 16 events were filtered and presented using the Gaussian smoothing technique. We also calculated the relevance score between events detected to construct the event network. Using the cosine coefficient between the co-occurred events, we calculated the relevance between the events and connected the events to construct the event network. Finally, we set up the event network by setting each event to each vertex and the relevance score between events to the vertices connecting the vertices. The event network constructed in our methods helped us to sort out major events in the political and social fields in Korea that occurred in the last one year in chronological order and at the same time identify which events are related to certain events.

* Master's Course, Graduate School of Library and Information Science, Yonsei University
** Corresponding Author: Hea-Jin Kim
HK Plus Research Professor, Institute of the Study for the Korean Modernity, Yonsei University
1 Yonseidae-gil, Wonju, Gangwon-do 26493, Korea
Tel: +82-33-760-2531, Fax: +82-033-760-2532, E-mail: erin.hj.kim@yonsei.ac.kr

Our approach differs from existing event detection methods in that LDA topic modeling makes it possible to easily analyze large amounts of data and to identify the relevance of events that were difficult to detect in existing event detection. We applied various text mining techniques and Word2vec technique in the text preprocessing to improve the accuracy of the extraction of proper nouns and synthetic nouns, which have been difficult in analyzing existing Korean texts, can be found.

In this study, the detection and network configuration techniques of the event have the following advantages in practical application. First, LDA topic modeling, which is unsupervised learning, can easily analyze subject and topic words and distribution from huge amount of data. Also, by using the date information of the collected news articles, it is possible to express the distribution by topic in a time series. Second, we can find out the connection of events in the form of present and summarized form by calculating relevance score and constructing event network by using simultaneous occurrence of topics that are difficult to grasp in existing event detection. It can be seen from the fact that the inter-event relevance-based event network proposed in this study was actually constructed in order of occurrence time. It is also possible to identify what happened as a starting point for a series of events through the event network.

The limitation of this study is that the characteristics of LDA topic modeling have different results according to the initial parameters and the number of subjects, and the subject and event name of the analysis result should be given by the subjective judgment of the researcher. Also, since each topic is assumed to be exclusive and independent, it does not take into account the relevance between themes. Subsequent studies need to calculate the relevance between events that are not covered in this study or those that belong to the same subject.

Key Words : event detection, latent Dirichlet allocation (LDA), natural language processing (NLP), text mining, topic modeling

Received : November 19, 2017 Revised : March 9, 2018 Accepted : March 16, 2018
Publication Type : Regular Paper Corresponding Author : Hea-Jin Kim

저 자 소 개

이 민 철

연세대학교에서 사학과 문헌정보학을 복수전공하고, 현재는 동 대학교 대학원에서 문헌정보학 석사과정을 밟고 있다. 주요 관심분야는 텍스트 마이닝과 자연언어처리 분야이며, 빅데이터를 기반으로 한 소셜미디어 마이닝, 감성분석, 머신러닝을 전공하고 있다.

김 혜 진

숙명여자대학교에서 문헌정보학을 전공하고, 연세대학교 대학원 문헌정보학과에서 문헌정보학 석사, 문학박사를 취득하였다. 현재는 연세대학교 근대한국학연구소 HK플러스 연구교수로 재직 중이다. 주요 연구영역은 빅데이터를 기반으로 한 text mining, social media mining, sentiment and trend analysis, machine learning이다.