언어네트워크를 통한 대통령기록물 관련 보도자료 이슈 분석*

Analysis of Presidential records issue of the Newspaper articles through sementic network

정상준, 전북대학교 대학원 기록관리 전공, fpxltm07@naver.com
오효정, 전북대학교 기록관리학과, ohj@jbnu.ac.kr

Sang-Jun, Jung, Chonbuk National University
Hyo-Jung, Oh, Chonbuk National University

본 연구는 언어네트워크 분석기법을 활용하여 언론보도자료에 나타난 대통령기록물과 관련된 사회적 이슈를 분석하였다. 분석결과를 통해 대통령기록물 관련 이슈의 발생 현황 및 이슈의 구성요소를 파악할 수 있었으며, 대통령기록물 관련 이슈에 대한 시사점 파악 및 관련 연구의 기초자료를 제공하는 것을 목적으로 한다. 이를 위하여 국내 주요 언론사 중 하나인 조선일보를 대상으로, 주제어인 "대통령기록물"을 포함하는 관련 기사를 수집하였다. 총 780건의 수집된 보도자료를 대상으로 언어네트워크 분석을 수행하였으며, 분석결과에 대한 시각화를 진행하였다.

1. 서 론

최근 기록관리학의 분야에서 사회적 이슈로 대두되고 있는 키워드는 '대통령기록물'이다. 2007년 「대통령기록관리법」의 제정 이후, 법제화를 통해 관리되고 있는 대통령기록물은 이명박 정부가 출범한 2008년, 대통령기록물 유출이라는 하나의 사회적 사건으로 이슈화되어 언론 보도를 통해 대중에게 노출되었다. 10년이 흐른 지금까지도 대통령기록물 관련 이슈들은 꾸준히 생산되고 있었다. 이와 같은 사회적 사건은 사회과학 분야의 주된 관심사였으며 사건 그 자체에 대한 내용을 파악하는 이슈 분석, 사건 속에 내재 된 다양한 쟁점 및 갈등 현상을 파악하는 쟁점 분석 및 갈등분석을 통해 연구가 진행되어왔다(이수상, 2017). 대통령기록물은 공공기록물로서 가장 중요한 기록물임에도 불구하고 10여 년간 이슈의 중심에서 끊임없이 노출되고 있다. 이는 사회적 이슈로써 대통령기록물 관련 이슈를 분석해야 할 당위성을 제공한다. 그러나 국내의 대통령기록물 관련 이슈를 대상으로 한 분석사례는 이루어지지 않은 실정이다. 이에 본 연구에서는 대통령기록물 관련 이슈 내용의 분석을 위해 조선일보를 대상으로 대통령기록물 관련 보도자료를 수집하여 분석을 실행한다. 이슈에 대한 객관적인 내용분석을 위하여 언어네트워크 분석기법을 활용한다. 텍스트의 수집, 키워드의 선정, 키워드 간의 관계 파악, 네트워크 구성 및 분석의 과정을 거치는 언어네트워크 분석은(이수상, 2014). 단어의 조합을 통해 특정한 의도를 전달하는 보도자료의 내

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2016S1A5B8913575)

용을 분석하는데 효과적이다. 이처럼 대통령 기록물 관련 보도자료를 대상으로 한 네트워크 분석은 이슈의 해석 및 이슈의 구성 관계를 살펴볼 수 있으며, 이슈의 전달 과정을 통해 대중들에게 어떠한 영향을 미치는지 추론을 가능케 한다.

2. 연구방법

2.1 데이터 수집

본 연구는 대통령기록물 관련 이슈의 분석을 위한 데이터의 수집을 위해 국내 주요 언론사 중 하나인 조선일보를 수집대상으로 선정하였다. 대상의 선정 요인은 매체의 영향력과 이데올로기 성향을 토대로 선정하였다(김경희 외, 2011). 자료수집 기간의 2008년 1월 1일부터 2017년 12월 31일까지의 대통령기록물의 관리 법률이 제정된 이후부터 10년의 기간이다. 수집에 사용한 주제어는 유의미한 결과를 보인 '대통령기록물'을 주제어로 한정하여 자료를 수집하였다. 자료수집은 신문사 웹사이트 검색 서비스를 통해 '제목+본문'으로 검색범위를 설정한 후 검색되는 신문기사 리스트를 Python 프로그램의 requests와 BeautifulSoup4 라이브러리를 활용한 웹크롤러 사용하여 기사의 전문 텍스트(full text)의 형태로 수집하였다. 최종적으로 수집된 기사들의 전처리과정을 통해 주제어인 '대통령기록물'과 연관성이 떨어지는 기사와 제목과 본문의 내용이 중복되는 기사들을 제거하였다.

2.2 데이터 전처리 및 언어네트워크 방법

본 연구에서는 최종적으로 수집된 보도자료의 제목과 본문을 대상으로 언어네트워크분석을 실시하였으며 언어네트워크 분석을 위해 사용한 방법은 다음과 같다.

먼저 수집된 보도자료의 데이터셋을 구축하여 KnowledgeMatrix Plus (KM+) 소프트웨어의 한글 및 영문 명사추출기능을 통해 명사를 추출하였다. 그 후 stemming 기능을 활용하여 추출된 키워드에 대한 정제과정을 실시하였다. 키워드의 정제 작업은 연구자의 주관이 개입될 가능성이 높기때문에 <표 1>의 키워드 통제 기준에 따라 작업이 수행되었다.

<표 1> 키워드 통제 기준과 사례

기준	사례
동의어/유사어	靑 → 청와대, 與 → 여당
복합명사	하드, 디스크 → 하드디스크
조사유무/생략어	남북의 정상회담 → 남북정상회담

두 번째로는 정제가 완료된 키워드를 빈도순으로 나열하여 핵심키워드를 도출하였으며, 이를 기반으로 Knowledge Matrix Plus (KM+) 프로그램의 1-mode matrix 기능을 활용하여 동시출현매트릭스 데이터를 생성하였다. 이는 핵심단어 간의 관계를 수량화해 나타내는 데이터로서 핵심단어 간의 관계를 파악하는 자료로 활용된다.

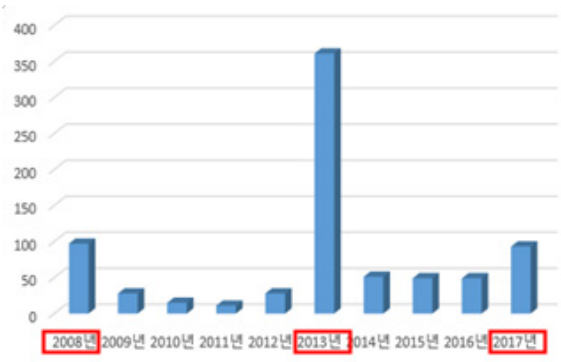
세 번째로는 1-mode matrix 기능을 활용하여 생성한 행렬데이터를 네트워크 분석 및 가시화 소프트웨어인 Ucinet을 활용하여 키워드 간의 연결중심성을 도출하여 네트워크 구조를 파악하였으며 Ucinet 시각화기능인 Netdraw를 사용하여 분석한 네트워크 관계를 시각화하였다.

3. 분석결과

3.1 수집데이터 연도별 분석

본 연구에서 조선일보를 대상으로 수집한 880

건의 이슈 중 전처리과정을 통해 정제된 분석대상은 총 782건이다. 이를 연도별로 나열한 결과 다음 <그림 1>과 같이 나타났다. 연도별로 나타난 특징은 2008년, 2013년, 2017년의 해당 기간에 대통령기록물 관련 이슈가 집중적으로 생산되었음을 알수있었다. 이를 분석한 결과 해당 기간은 정권교체 이후 후임 대통령의 부임년도와 일치함을 알 수 있었다. 이는 대통령기록물 관련 보도는 정권교체 이후 언론을 통해 대중에게 노출되고 있음을 추론할 수 있었다.



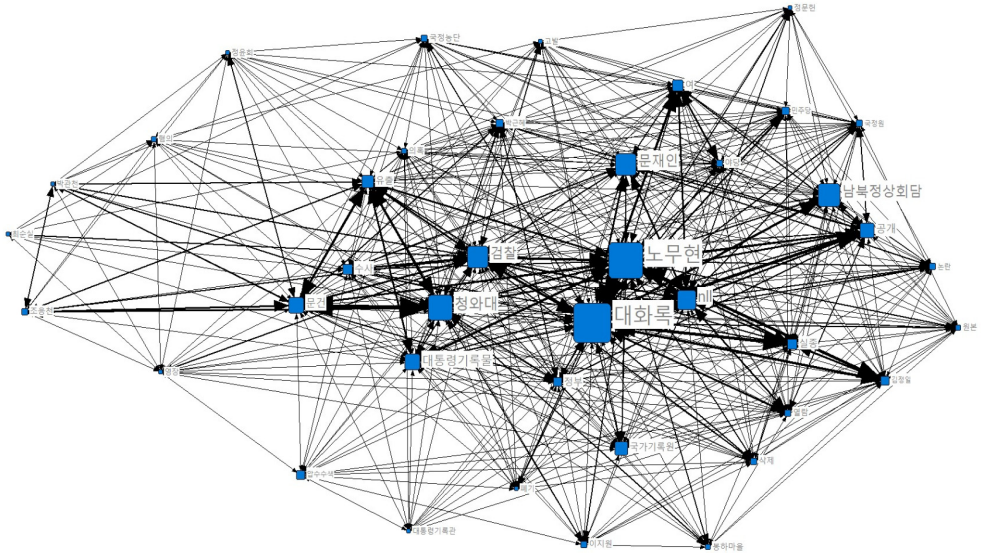
<그림 1> 대통령기록물 관련 연도별 이슈 (조선일보)

3.2 핵심키워드와 언어네트워크 분석

본 연구는 분석대상의 핵심키워드 선정을 위해 앞서 논의된 방법론을 적용하여 상위 40개의 핵심키워드를 선정하였다. 아래의 <표 2>는 조선일보의 상위 40개의 키워드와 출현빈도를 나타낸 표이다. 이를 살펴보면 100회 이상의 빈도수가 나타난 키워드는 '대화록', '노무현', '청와대', '남북정상회담', '문재인', '검찰', 'nll' 이다. 이를 통해 남북정상회담의 대화록과 관련한 사안이 대통령기록물 관련 이슈의 중점 사건이 되어왔음을 추론할 수 있었다. 상위 40개의 키워드에는 '노무현', '문재인', '김정일', '박근혜', '조응천', '박관천', '최순실', 등 인물명이 포함되어 있었다. 이는 대통령기록물 관련 이슈의 중심에 자리한 인물의 구성을 파악할 수 있게 하였다. 핵심키워드에 포함된 기관의 경우 '청와대', '검찰', '국가 기록원', '여당', '야당', '국정원', '봉하마을' 등이 나타났으며, 대통령기록물 이슈와 밀접한 관련이 있는 기관은 어떻게 구성되어 있는지 파악할 수 있었다. 다음으로 매체별 핵심키워드의 빈도분석을 통해 도출할 수 없었던 키워드 간의 관계를 파악하기 위해 분석대상의 네트워크 구조와 특성을 알아보았다. 빈도분석을 통해 도출된 핵심키워드 40개를 대상으로 동시출현 매트릭스데이터를 도출하여 핵심키워드 간의 연결중심성을 파악하였다. 연결중심성의 경우 수치가 높게 나올수록 네트워크 내의 가장 많은 노드들 간의 직접적인 연결관계를 가지게 되며, 네트워크의 중심에 위치하게 된다(권호천, 2017). 연결중심성의 경우 '대화록', '노무현', '검찰', '청와대', 'nll' 순으로 높게 나타났으며, 이와 같은 순위는 핵심키워드의 빈도별 수치와 비교적 평행하게 나타나고 있음을 알 수 있었다. 그러나 '남북정상회담', '김정일', '문재인', '국가 기록원' 의 키워드의 경우 핵심키워드의 빈도순위에 비해 연결중심성이 낮게 측정되었다. 이는 빈도별 분석에서 추론하였던 대통령기록물 관련 이슈의 인물, 사건, 기관의 영역에서의 비중이 핵심키워드의 출현빈도를 통해 분석한 결과 만큼 크지 않았음을 나타내고 있었다. 이와 반대로 '문건', '공개', '유출', '실종' 의 키워드는 연결중심성이 높게 나타났으며, 이를 통해 빈도에서의 추론과 달리 이들 키워드가 실질적으로 네트워크 중심에 더욱 가까이 자리한 키워드임을 알 수 있었다. 다음의 <그림 2>는 대통령기록물 관련 이슈에 대한 네트워크를 그린 지도이다. 네트워크표현을 위해 네트워크 분석 및 가시화 소프트웨어인 Ucinet의 Netdraw를 활용하여 네트워크 분석 및 시각화를 진행하였다.

<표 2> 핵심키워드 및 네트워크 분석 결과

순위	빈도 상위 40개 키워드	연결중심성 상위 40개 키워드	순위	빈도 상위 40개 키워드	연결중심성 상위 40개 키워드
1	대화록(246)	대화록(0.227)	21	논란(34)	민주당(0.034)
2	노무현(225)	노무현(0.160)	22	국정원(33)	조응천(0.033)
3	청와대(151)	검찰(0.132)	23	이지원(32)	박근혜(0.032)
4	남북정상회담(140)	청와대(0.122)	24	조응천(31)	삭제(0.029)
5	문재인(132)	nll(0.109)	25	열람(31)	국가기록원(0.028)
6	검찰(129)	문건(0.097)	26	야당(29)	박관천(0.025)
7	nll(113)	공개(0.086)	27	국정농단(28)	원본(0.022)
8	대통령기록물(94)	유출(0.083)	28	삭제(27)	의혹(0.021)
9	문건(90)	대통령기록물(0.070)	29	봉하마을(25)	국정농단(0.021)
10	공개(82)	실종(0.066)	30	의혹(22)	고발(0.020)
11	국가기록원(76)	김정일(0.059)	31	혐의(20)	이지원(0.019)
12	유출(69)	여당(0.059)	32	원본(19)	영장(0.018)
13	여당(59)	남북정상회담(0.054)	33	박관천(18)	혐의(0.018)
14	실종(51)	문재인(0.054)	34	고발(18)	정문헌(0.016)
15	수사(51)	정부(0.051)	35	영장(18)	정윤회(0.016)
16	정부(49)	수사(0.049)	36	최순실(17)	폐기(0.016)
17	압수수색(43)	야당(0.037)	37	대통령기록관(17)	압수수색(0.014)
18	김정일(42)	논란(0.036)	38	정문헌(16)	봉하마을(0.013)
19	민주당(40)	열람(0.034)	39	폐기(16)	최순실(0.010)
20	박근혜(40)	국정원(0.034)	40	정윤회(15)	대통령기록관(0.006)



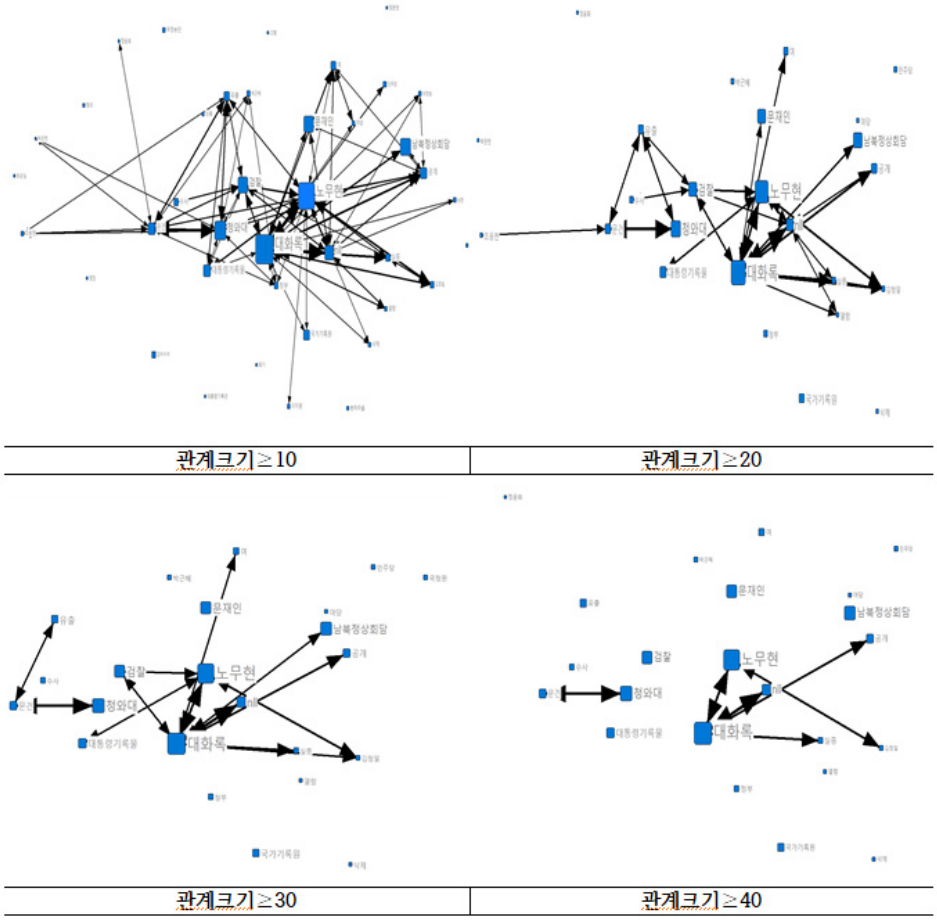
<그림 2> 대통령기록물 관련 이슈 네트워크 지도

대통령기록물 관련 이슈의 전체 네트워크는 총 40개의 노드와 총 936개의 링크로 연결되어 나타났다. 이 중 '노무현', '대화록', '문재인', '청와대', '검찰', 'nll' 의 노드가 크고 진한 것을 확인할 수 있다. 이는 네트워크 노드의 크기가 클수록 연결 정도가 높게 나타나

며, 동시 출현빈도가 높은 노드들은 서로 인접하게 나타남을 파악할 수 있었다. 이와 같은 키워드들은 네트워크의 형성에 기여하며, 노드 간의 핵심적인 허브 역할을 맡고 있다. 노드 간의 연결된 링크의 경우 연결선의 두께에 따라 키워드 간의 연계 빈도를 나타낸다. 따라서 연결선의 두께가 두꺼울수록 연계 빈도 역시 높게 나타난다. <그림 2>를 보면 허브 역할을 맡은 '노무현', '대화록', '검찰', '청와대', 와 같은 노드들을 중심으로 두껍고 얇은 링크가 복잡하게 연결되어 있음을 볼 수 있다. 이처럼 네트워크 관계망의 밀도가 높게 표현되면 노드 간의 관계 파악이 어렵기에 이를 파악하기 위해 관계 크기에 따른 네트워크의 재분석이 필요하다.

3.3 관계 크기에 따른 언어네트워크 분석

노드 관계의 크기에 따라 네트워크를 재구성하여 대통령 관련 이슈에 대한 세밀한 네트워크 관계 분석을 실행하였다. <그림 3>은 네트워크 관계 크기별 네트워크 지도이다. 네트워크의 관계 크기에 따라 복잡하게 연결되어있는 노드와 링크 간의 연결 정도가 감소하는 것을 볼 수 있다. 관계 크기가 커질수록 연결 강도가 높은 노드들만 남게 된다. 다음 <그림 3>을 통해 알 수 있는 점은 다음과 같다. 관계 크기가 40을 넘었을 때 남게 된 키워드는 '노



<그림 3> 대통령기록물 관련 이슈의 관계크기 별 네트워크 지도

무현', '대화록', 'nll', '공개', '김정일', '실종', '문건', '청와대' 였다. 이는 조선일보에서의 대통령기록물 관련 이슈의 구성을 파악할 수 있는 명확한 네트워크 관계를 제시한다. 대통령기록물 관련 이슈의 중심의 인물은 '노무현',김정일'로 구성하였으며, 사건은 'nll', '대화록', '공개'와 관련한 사안이 주를 이루었다는 점을 파악할 수 있었다. 반면 '청와대'와 '문건'의 경우 서로 간의 연결 강도는 높았으나 이와 연결된 다른 노드들의 관계는 관계크기≥40 에서 보이지 않았다. 그러나 이들의 관계는 관계크기≥30과 관계크기≥20의 네트워크 지도 통해 명확히 나타난다. 먼저 관계크기≥30에서의 '문건'과'유출'의 링크가 연결되어 청와대의 문건 유출과 관련되었음을 파악할 수 있었으며, 관계크기≥20에서의 '조응천'과 '문건'의 연결, '유출'과'검찰'의 연결된 관계를 보았을 때 사건의 핵심 인물은 '조응천', '검찰'의 '수사'가 이루어지고 있음을 파악할 수 있었다.

4. 결론 및 후속연구

본 연구는 주요 언론사 중 하나인 조선일보에서 대통령기록물 관련 이슈에 대해 어떤 방식으로 이슈를 구성하며 이를 보도하는지 알아보기 위해 언어네트워크 분석기법을 활용하여 내용분석을 실행하였다. 먼저 수집대상인 조선일보의 대통령기록물 관련 이슈를 수집하였다. 이후 수집된 데이터의 전처리과정 및 키워드의 정제과정을 거쳐 핵심키워드를 도출하였다. 도출된 핵심키워드에 대한 분석을 통해 대통령기록물 관련 이슈의 사건, 인물, 기관을 파악하였다. 이후 핵심키워드 간의 동시출현행렬을 파악하여 네트워크 분석을 실행하였으며, 이를 바탕으로 키워드 간의 관계를 분석하였다. 분석결과 핵심키워드를 중심으로 한 대통령기록물 관련 이슈의 네트워크 관계 파악이 가능하였다. 하지만 복잡하게 얽혀있는 네트워크 지도를 통해 세밀한 분석을 진행하는 것은 어려운 일이었다. 따라서 핵심키워드 간의 형성된 네트워크 관계 크기에 따라 네트워크를 재구성하여 대통령 관련 이슈에 대한 세밀한 네트워크의 관계 분석을 실행하였다. 그 결과 복잡하게 연결되어있는 키워드 간의 네트워크 관계를 자세하게 파악할 수 있었다. 그러나 본 연구에서는 연구대상을 주요 언론사 중 하나인 조선일보만을 대상으로 자료를 수집 및 분석하였기 때문에 후속연구에서는 연구대상을 확장하여 각 미디어별 이슈분석 양상을 비교하는 과정이 수행되어야 한다. 또한 이슈분석을 통해 도출된 결과를 토대로 관련 분야에서의 연구가 얼마나 진행되었는지 살펴보는 것 역시 후속연구에 반영해야 할 사항으로 판단된다.

참고문헌

권호천 (2017). 사드(THAAD) 관련 신문기사의 의미네트워크 분석 - <조선일보>와 <한겨레신문> 기사를 중심으로. 언론정보연구, 54(2), 2017.5, 114-154.

이수상 (2014). 언어 네트워크 분석 방법을 활용한 학술논문의 내용분석. 정보관리학회지 31(4), 49-68.

이수상. (2017). 신문기사에 나타난 경주지진사건의 사회적 이슈분석. 한국도서관·정보학회지, 48(2), 53-72.

KISTI(2016), KnowledgeMatrix Plus ver.0.80 for supporting Scientometric Network Analysis, Department of Scientometric Research, Korea Institute of Science and Technology Information (KISTI)