韓國言論學報, 60(5) 2016.10. 353~407쪽 353

장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성:

100만 건 기사의 정보원과 주제로 본 신문 26년*

박대민**
(한국언론진흥재단 선임연구위원)

언론학계에서는 의제설정 연구를 비롯하여 적지 않은 시계열 내용 분석 연구가 이루어졌다. 그동안 시계열 방법론 측면에서는 많은 발전이 있었지만 수작업에 의존하는 전통적 내용 분석 방법으로는 대규모 내용 분석에 난관이 많았다. 이 연구는 본격적인 장기 시계열 연구를 수행하기 위해 자연어 처리와 의미연결망 분석이 결합된 뉴스 빅데이터 분석을 활용할 것을 제안한다. 또한 26년치(1990~2015) 8개 중앙지(〈경향신문〉, 〈국민일보〉, 〈동아일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레신문〉, 〈한국일보〉)의 정치와 사회면 기사 약 100만 건에 대해 분석했다. 기사는 한국언론진흥재단 뉴스 빅데이터 시스템인 '빅카인즈'를 활용하여 수집하고 자연어 처리한 뒤 기사의 정보원과 인용문 주제 중심으로 의미연결망 분석을 실시하여, 매체별로 정보원과 주제의 시계열적 변화를 살펴보았다. 분석 결과, 사회면 주제를 제외하면 중요도 최상위권 정보원과 주제의 매체 간 차이는 크지 않았던 반면, 시계열적으로는 2000년 전후로 가장 중요한 정보원과 인용문 주제가 전면적으로 변화하는 양상을 보였다. 기사당 정보원 수와 기사당 인용문 주제 수는 매체별로 다소 차이는 있지만 대체로 하락했다. 이 연구는 뉴스 빅데이터 분석을 활용해 수집된 온라인 기사 전수에 대해 지속적으로 모니터링하면서 자동화된 장기 시계열 내용 분석 데이터를 축적할 수 있을 뿐만 아니라, 이를 바탕으로 경제지표 등 다양한 시계열 데이터와 정교한 비교연구를 할 수 있는 토대를 마련했다는 데 의의를 갖는다.

핵심어: 자동화된 장기 시계열 내용 분석, 뉴스 자연어 처리, 뉴스 의미연결망 분석, 뉴스 빅데이터 분석, 빅카인즈

---

* 이 연구는 2016년 4월 한국언론진흥재단 뉴스 빅데이터 분석 리포트 〈News Big Data Analytics & Insights〉 1권 1호에 발표된 "신문 26년: 뉴스 빅데이터 시각화로 본 신문 보도의 역사"를 대폭 수정, 보완한 것입니다. 이 연구는 2016년 서울대학교 언론정보연구소 연구기금의 지원을 받았습니다.

** dmpark@kpf.or.kr; heathe1@snu.ac.kr

1. 문제 제기

뉴스 보도는 대통령 지지도에 영향을 주는가? 범죄 보도는 사회를 더 위험하게 바라보게 하는가? 뉴스는 주식시장의 거품이나 공황을 유도하는가? 매체의제는 공중의제에 영향을 주는 것인가, 아니면 받는 것인가? 이러한 언론학의 핵심적 연구 주제는 근본적으로 시계열 내용 분석 연구를 통해 규명되어야 할 것들이다. 우선 뉴스 보도에 대한 내용 분석이 필요하며, 다음으로 이를 여러 시점에서 반복해 분석해야 하고, 그렇게 얻은 결과물을 서로 비교하거나 대통령 지지도 조사, 여론조사, 주가를 비롯한 다양한 경제지표 등 다양한 또 다른 시계열 데이터와 비교해야 한다. 매체의 효과론적 인과성을 밝히기 위해서, 또는 매체현상이 사회에 미칠 영향을 예측하기 위해서 시계열 내용 분석은 필수적이다.

실제로 그동안 언론학계에서 시계열 내용 분석 연구는 적지 않게 이루어졌다. 초기에는 다양한 의사시계열적 방법이 동원됐지만 최근 들어 벡터자기회귀(VAR: Vector auto regression) 모형 등 고급 시계열 분석 방법을 적용한 연구도 적지 않게 눈에 띈다. 그러나 내용 분석 차원에서는 난관이 적지 않았다. 수작업에 의존하는 전통적 내용 분석 방법으로는 대규모의 내용 분석이 어려웠던 것이다. 예컨대 대통령 지지도나 GDP(gross domestic product) 등은 수많은 요인에 의한 영향을 집합적으로 반영하는 데 반해, 내용 분석의 양은 시점별로 때로는 수십 건에 지나지 않은 경우도 있었다. 즉, 본격적으로 장기 시계열 연구를 수행하기 위해 다른 종류의 대규모 시계열 데이터와 함께 비교 분석하기에는 뉴스 내용 분석의 양 자체가 지나치게 적었던 것이다.

그러나 최근 뉴스 자료가 아카이브화되고 자연어 처리와 의미연결망 분석 등을 자동화된 내용 분석에 어렵지 않게 활용할 수 있는 도구들이 제공되면서, 컴퓨터를 활용해 많은 양의 뉴스 기사를 분석하는 것에 대한 관심이 높아졌다. 이 연구에서는, 특히 자연어 처리와 의미연결망 분석이 결합된 자동화된 뉴스 내용 분석 방법을 뉴스 빅데이터 분석 방법(news big data analytics)으로 부르고자 한다. 이 연구에서는 뉴스 기사에 대한 장기 시계열 내용 분석에 뉴스 빅데이터 분석을 활용하는 방안을 모색할 것이다. 구체적으로는 기존의 뉴스 시계열 내용 분석의 성과와 한계를 짚어본 뒤 뉴스 빅데이터 특성을 탐색하고 이러한 이해를 바탕으로 장기 시계열 내용 분석의 한 축으로서 뉴스 빅데이터 분석을 제안할 것이다.

더 나아가 이 연구에서는 주요 중앙일간지 8개의 정치, 사회 지면 26년 치 기사 약 100만 건에 대한 정보원과 인용문 주제를 시행적으로 분석한다. 이러한 대규모 분석을 위해서는 최근 사회과학계에서 주목받고 있는 컴퓨터 보조 내용 분석(computerized content analysis)

내지 자동화된 텍스트 분석 기법을 활용할 것이다. 특히, 자연어 처리(natural language processing, 이하 NLP)와 의미연결망 분석(semantic network analysis), 그리고 시각화(visualization)를 결합한 뉴스 빅데이터 분석(news big data analysis)을 활용하여 대량의 기사 내용에 대한 자동화된 시계열 분석을 시도한다. 비록 시행연구이지만, 뉴스 빅데이터 분석을 통한 자동화된 시계열 내용 분석의 가능성을 확인하는 것과 함께, 신문사들을 중심으로 지난 26년간의 언론 지형의 변화를 연도별, 매체별, 지면별로 정보원과 인용문 주제 중심으로 개략적이나마 파악할 수 있을 것으로 기대한다.

2. 시계열 내용 분석에 대한 기존 문헌 검토

장기 시계열 내용 분석은 방법론적으로 두 측면에서 접근해야 한다. 첫째, 시계열 분석 측면이다. 시계열 분석 측면에서는 뉴스 데이터 특유의 시계열 변수 특성을 고려한 VAR이 제안됐다(이완수, 2009; 장병희 · 강형구 · 정일권 · 이혜진, 2008). 둘째, 내용 분석 측면이다. 장기 시계열 내용 분석을 하게 되면 데이터가 방대해지기 때문에 자동화된 내용 분석(automated content analysis)이 필연적으로 요청된다. 이 연구에서는 장기 시계열 내용 분석에 적합한 자동화된 내용 분석 방법으로서 NLP와 의미연결망 분석을 활용한 뉴스 빅데이터 분석을 제안한다.

1) 시계열 분석 측면에서의 발전: 의제설정 연구를 중심으로

언론학에서 대표적인 시계열 내용 분석 연구로는 의제설정(agenda setting) 연구를 들 수 있다(McCombs & Shaw, 1972; 반현 · 맥콤스, 2007). 1970년대 초 시작된 의제설정 연구는 기본적으로 한 매체의제가 시간을 두고 공공의제나 다른 매체의제에 영향을 주었는지를 판정하는 시계열 연구이다(이완수, 2009). 그럼에도 시계열 방법론을 본격적으로 활용한 의제설정 연구는 많지 않다.

국내에서는 많은 연구가 횡단 분석(cross-sectional analysis)을 활용한다. 이 경우 교차분석, t 검정 등 차이 판정을 통해 매체가 고유한 의제설정 역할을 하는지를 분석한다(김경희, 2008; 장하용, 2011; 하승태 · 조의현, 2008). 일반적으로 횡단 분석은 영향의 선후관계를 파악할 수 없다. 때문에 연구 설계상에서 이를 파악할 수 있는 장치를 담기도 한다. 예컨대 공중의제의 조사 분석 과정에서 설문지에 기사 노출 여부를 물음으로써 매체의제

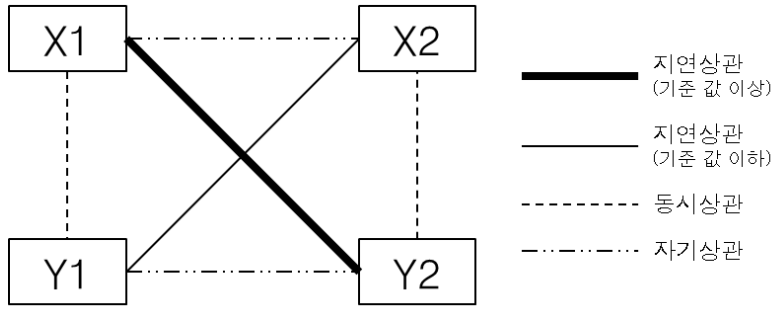


그림 1. 교차지연상관을 이용한 시계열 분석

와 공중의제의 선후관계를 파악할 수도 있다(김영욱 등, 2015). 실험설계에서 인과성을 활용하기도 한다. 즉, 내용 차이가 있는 기사를 독립변인으로 처치하고 피험자에게 기사를 읽게 한 뒤 의제설정 효과를 확인한다(이건호 · 유찬윤 · 맥콤스, 2007). 이상의 연구는 분석 기간이 대체로 짧다. 그러나 수년간의 기사를 한 기간으로 보고 매체 간 비교연구를 수행한 사례도 있다(정일권, 2010). 이상의 연구와 비교했을 때 둘 이상의 시기에 걸쳐서 분석하는 경향 분석(trend analysis)은 시차의 영향을 보다 분명하게 파악할 수 있다(하승태 · 조희연, 2008).

보다 전형적인 의제설정 연구는 매체와 공중 또는 매체 간의 의제나 속성을 현저성(salience)에 따라 순위화하고 이를 스피어만의 로(Spearman's rho) 등을 활용해 순위상관관계(rank order correlation)를 분석하고, 더 나아가 순위상관관계를 활용한 교차지연상관분석(cross-lagged correlation analysis)을 통해 서로 다른 두 매체의제 또는 매체의제 및 공공의제를 비교하는 방식이다(구교태, 2003; 반현 · 최원석 · 신성혜, 2004; 이건호, 2006; 이동훈, 2007; 이승희 · 송진, 2014; 임종섭, 2011; 최진호 · 한동섭, 2011). 예컨대 〈그림 1〉처럼 두 매체 X와 Y를 두 시기 1과 2에 대해 분석한다고 했을 때 'X1-X2', 'Y1-Y2'와 같은 자기상관(autocorrelation), 'X1-Y1', 'X2-Y2'와 같은 동시상관(synchronous correlation), 'X1-Y2'과 'X2-Y1'과 같은 교차지연상관(cross-lagged correlation)을 계산하여 상관관계의 유의미성을 파악한다.[1] 만일 자기상관, 동시상관이 유의미하지 않고, Y에 대한 X의 지연상관(X1-Y2)이 유의미하고 X에 대한 Y의 지연상관(X2-Y1)이 유의미하지 않다면, X가 Y에 영향을 준 것이다.

그러나 의제나 속성의 출현 빈도나 현저성, 보다 일반화해서 뉴스 등장 단어의 중요도

1) 유의미성을 판단하는 기준 값으로는 흔히 로젤-캠벨 기준 값(Rozelle-Campbell baseline)을 활용한다(이승희 · 송진, 2014; 임종섭, 2011).

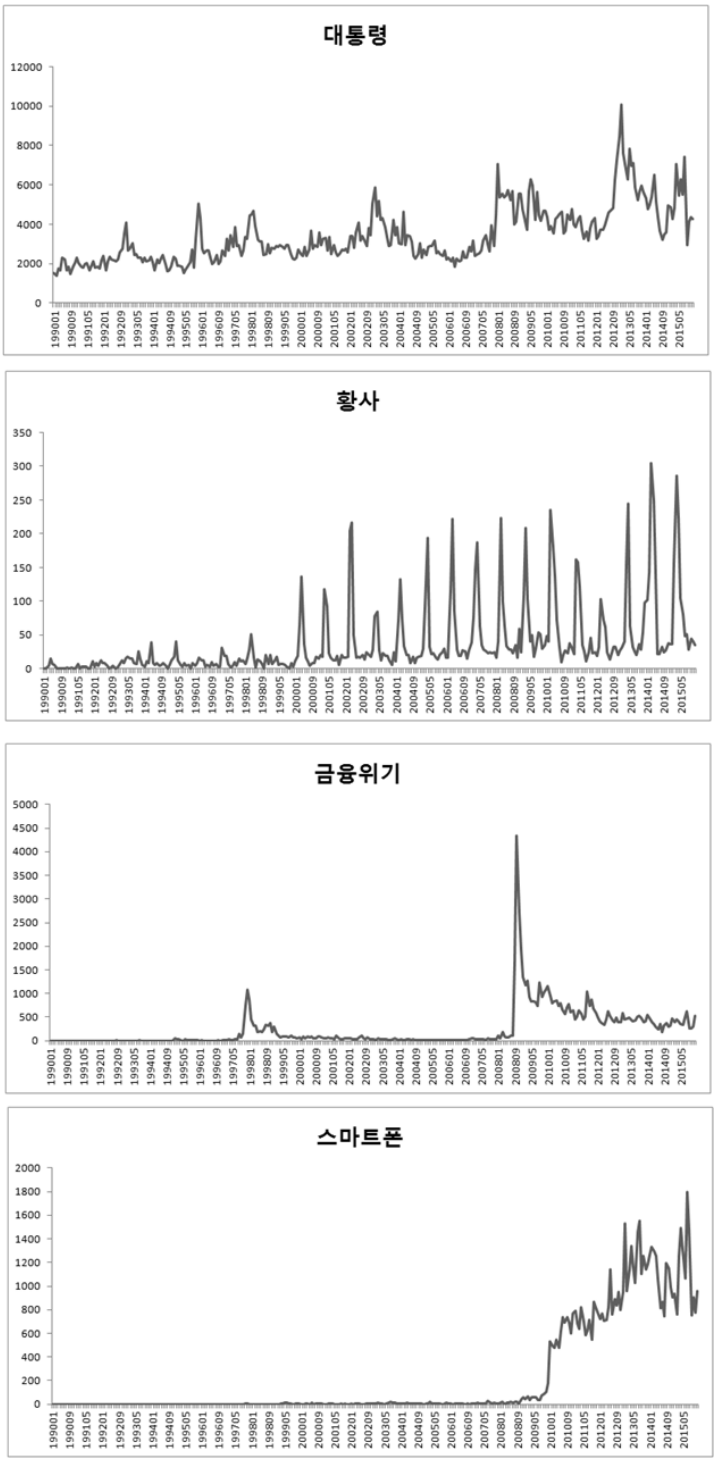


그림 2. 검색어별 26년간 월별 기사출현 빈도 추이

와 같은 뉴스 빅데이터는 교차지연상관 분석에는 적합하지 않은 고유의 시계열 자료(time-series data) 특성을 갖고 있다(이완수, 2009; 조진섭 · 손영숙 · 성병찬, 2016; 한광종, 2014).

첫째, 뉴스 빅데이터는 시계열 데이터이다. 즉, 자료들이 시차에 따라 서로 완전 독립이 아니며 과거 자료가 현재 자료에 영향을 준다. 예컨대 어제 중요한 의제는 다음날에도 중요하게 등장할 가능성이 높다.

둘째, 시차에 따른 데이터 간의 관계가 비선형적(nonlinear)이다. 예컨대 〈그림 2〉는 1990년 1월부터 2015년 12월까지 한국언론진흥재단의 뉴스 빅데이터 분석 시스템 '빅카인즈'에 축적된 7대 중앙지(〈경향신문〉, 〈국민일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레신문〉, 〈한국일보〉)의 검색어별 기사출현 빈도를 나타낸 그래프이다. '대통령'과 같은 경우 중요도(기사출현 빈도)가 점차로 상승하는 추세성분(trend component)을 가지며 선형 추세모형(linear trend model)으로 설명될 수 있다. 반면 '황사'의 중요도는 매년 겨울~봄 사이에 중시되는 계절성분(seasonal component)을 가지며 선형 계절추세모형(linear and seasonal trend model)에 따르는 것으로 보인다, '금융위기'와 같은 단어는 경제주기에 따른 순환성분(cyclical component)을 갖는 것으로 보인다. '스마트폰'과 같은 단어는 2009년 11월에 '아이폰 출시'라는 개입(intervention) 이후 중요도가 도약하게 된다.

셋째, 따라서 뉴스 빅데이터는 많은 경우 비정상적(nonstationary) 특성을 갖는다. 또한 NLP 성능 오류로 인한 결측이나 오측 문제로 이상값이 있을 수 있다. 때문에 본격적인 시계열 분석에 앞서 이상값이 있는지 확인하고 이를 해결해야 하며, 자기상관성을 검정하여 차분(differencing), 계절차분, 로그차분 등을 통해 데이터를 안정화해야 한다.

이러한 뉴스 빅데이터의 특성을 감안해 언론학계에서도 뉴스 내용 분석 데이터에 대해 자기회귀이동평균모형(ARMA: Auto-regression moving average), 자기회귀누적평균모형(ARIMA: Autoregressive integrated moving average), VAR모형, 개입시계열 분석(interrupted time series analysis) 등 다양한 시계열 분석 방법을 적용한 연구가 나오고 있다(이완수, 2007, 2009, 2015; 이완수 · 노성종, 2008, 2011; 이완수 · 심재철, 2007; 이완수 · 심재철 · 박양수, 2007; 이준웅, 2005; 장병희 등, 2008; 하승태, 2008, 2012; 한수연 · 윤석민, 2016; 홍원식, 2007).

2) 빅데이터 관점에서 본 내용 분석 측면의 한계

앞서 살펴본 대로 뉴스 기사의 장기 시계열 내용 분석은 시계열 분석 측면에서는 적지 않은 방법론적 개선이 이루어졌다. 반면 내용 분석 측면에서는 아직 난제가 적지 않다.

가장 큰 문제는 시간과 관련된다. 시계열 데이터는 보통 연 단위나 월 단위로 데이터가 축적되며, 분기 단위나 일 단위, 때로는 초 단위로도 수집될 수 있다. 온라인 뉴스는 입력 및 수정 시간이 초 단위로 기록되기도 하지만, 일반적으로 일 단위로 수집된다. 즉, 기사 내용은 주로 하루 동안 일어난 일을 다룬다. 흔히 언론학의 시계열 내용 분석 연구는 동종, 또는 이종 매체에서 둘 이상의 뉴스 내용 분석 데이터를 분석해 서로 비교하거나, 여론조사 결과나 각종 경제지표 등 다른 종류의 시계열 데이터와 비교하는 식으로 이루어진다. 이때 뉴스 내용 분석 시계열 데이터는 다른 시계열 데이터의 분석 기간과 시간 단위, 시차 등을 일치시켜야 한다. 예컨대 일 단위로 수집된 뉴스 내용 분석 데이터를 연 단위로 수집된 경제 시계열 데이터와 비교하거나, 시차가 일정하지 않은 지지도 조사 데이터 등과 비교할 경우 기사를 월 단위나 연 단위로 묶어서 분석해야 할 수도 있다(이완수 · 노성종, 2008, 2011; 이완수 · 심재철, 2007; 이완수 등, 2007; 이준웅, 2005; 장병희 등, 2008; 하승태, 2008, 2012; 홍원식, 2007). 때로는 날짜를 표본추출해서 선정한 다음 해당 날짜의 기사만 분석할 수도 있다(한수연 · 윤석민, 2016).

이 경우 정확한 시계열적 상관성을 분석하기 어려울 수 있다. 즉, 어떤 날짜 기사의 내용이 언제 어떤 다른 월별, 분기별, 연도별 수집 데이터에 영향을 주었는지 정확히 파악하기 힘들다. 뉴스 내용 분석 데이터가 일 단위이기 때문에 일 단위로 수집된 다른 시계열 데이터가 데이터 간 비교에 가장 좋다. 예컨대 일 단위로 분석한 뉴스 내용 분석 데이터를 일 단위로 수집된 종합주가지수와 비교하는 것이 이상적이다. 그러나 많은 경우 뉴스 데이터를 주, 월, 분기, 연 단위 등으로 바꾸어 분석한다. 뉴스 내용 분석 데이터 간에 비교할 경우에는 비교적 문제가 적지만, 다른 경우에는 해석을 제한적으로 해야 한다. 일 단위 데이터 간에 비교할 수 있더라도 문제는 있다. 일 단위로 뉴스 내용 분석을 하면 데이터가 수작업으로 분석하기에는 너무 많아지기 때문이다. 이때 분석 데이터의 수는 다음 공식에 따라 계산해 볼 수 있다.

$$\text{분석 데이터 수} = (\text{분석 기간} \div \text{시간 단위}) \times \text{분석 매체 수} \times \text{분석 지면 수} \times \text{분석 주제 수} \times \text{분석 범주 수}$$

예컨대 26년 치 기사를 연 단위로 8개 매체, 정치면 및 사회면에 대해 기사 전수를 정보원과 주제 측면에서 분석한다고 하면 내용 분석으로 총 832개의 데이터를 추출하면 된다. 그러나 이를 월 단위로 분석한다고 하면, '832 × 12'로 총 9,984개 데이터를 산출해야 한다.

타당한 시계열 분석을 위해 무조건 많은 양의 데이터가 반드시 필요한 것은 아니다.[2] 그러나 분석 기간이 수년 단위인 반면 기사가 일 단위의 내용을 다루면서 일 단위로 작성되는 뉴스를 분석하는 장기 시계열 내용 분석의 경우, 분석 데이터의 양이 방대하기 때문에 생기는 문제를 피하기 어렵다. 물론 분석 기간을 짧게 하는 대신 시간 단위도 짧게 해서 충분한 양의 시계열 데이터를 확보할 수도 있다(장병희 등, 2008). 그러나 분석 기간이 너무 짧다면 장기 시계열 분석이라고 보기에는 무리가 있다. 게다가 여론이 소수의 주요 매체에 크게 의존하던 과거와 달리, 수많은 매체가 포털이나 SNS 등 플랫폼을 통해 소비되는 현재의 상황에서는 뉴스의 양이 방대해지는 문제, 즉 뉴스의 빅데이터화는 분석상 피할 수 없는 문제다(박대민, 2013). 소수의 주요 매체만 분석하고 생산량으로 보나 이용량으로 보나 압도적 다수인 다른 매체를 제외할 때 분석의 타당성 문제가 제기될 수 있다.

예컨대 〈조선일보〉와 〈한겨레신문〉의 일 200여 개 기사에서 기사를 표집해 분석한 것보다는, 이를 제외한 140여 개 매체의 일 평균 2~3만 건 기사 전수를 분석하는 것이 더 타당할 수 있다. 이는 특히 대통령 지지도나 종합주가지수와 같이 정치, 경제, 사회 전반의 이슈에 대한 영향력이 반영된 시계열 데이터와의 비교에서는 더욱 문제시될 수 있다.

또한 기존에는 수십 년에 걸친 장기 시계열 분석을 위해, 시간이나 매체 외에도 지면이나 주제,[3] 분석 부분(예컨대 제목)을 한정하여 코딩의 수고를 더는 경우가 적지 않다. 주제도 현저성을 기준으로 1면 기사나 저녁뉴스, 기사 헤드라인만 분석하거나 검색어로 한정하기도 한다(이완수, 2015; 이완수 · 심재철, 2007; 이완수 · 노성종, 2011; 이완수 등, 2007; 장병희 등, 2008; 홍원식, 2007). 이는 매체와 기사와 주제를 사실상 비체계적 할당표집한 셈이다. 그러나 분석 기간과 시간 단위가 길수록 사실상 비체계적 할당표집 방식으로 수집된 소수의 기사를 내용 분석하는 것은 아무리 현저성이 상대적으로 높은 매체나 지면, 주제를 선별한다고 해도, 시장과 사회에 대한 포괄적인 시계열 데이터와 비교할 때 분석의 타당성이 높다고 하기는 어려워진다. 가능하면 일 단위로 최대한 많은 수의 기사를 내용 분석하는 것이 타당성을 높이는 방법이다.

---

2) 예컨대 개입시계열 분석을 위한 구간별 회귀 분석에서는 개입 전후로 12개 시점의 데이터가 필요한 것으로 알려져 있다(Wagner, Soumerai, Zhang, & Ross-Degnan, 2002).

3) 예컨대 기사검색 시 속성(attribute)에 해당하는 검색어를 여러 개 넣고 해당 검색어를 모두 또는 복수로 포함하는 기사만 선정할 수 있다.

표 1. 국내 언론학에서 주요 시계열 내용분석 연구의 비교

저자	발간 연도	시계열 분석	분석 기간	시계열 사례 수	분석 매체	표집 방식	분석 기사 수	시간 단위	내용분석 항목	비교 시계열 데이터
이완수 · 박양수	2016	회귀분석	1998년 12월~ 2014년 12월	48	〈조선일보〉, 〈동아일보〉, KBS, SBS	지면 한정(신문), 프로그램 한정(방송), 주제 한정, 헤드라인 한정	미기재	분기	논조지수	소비자동향지수 기업경기실사지수 경기동행지수
홍원식	2007	ARMA	1963년 1월~ 2002년 12월	480	〈NYT〉	기사 유형 한정, 주제 한정	447	월	지지도 보도건수, 보도된 지지도, 프레임	Gallup 대통령 지지도
하승태	2008	ARIMA	2006년 6월 21일~ 2007년 1월 10일	26	조인스닷컴	검색어 한정	1,177	주	기사 보도건수, 사진 보도건수, 사진 보도 톤	조인스닷컴 대선후보 지지도
하승태	2012	ARIMA	2007년 1월 15일~ 2008년 3월 16일	27	〈USA Today〉	검색어 한정	735	조사 연동	기사 보도건수	USA Today-Gallup 대선후보 지지도
이준웅	2005	ARIMA	1998년 1월~ 2002년 3월	63	〈조선일보〉, 〈중앙일보〉, 〈한겨레〉	기사 유형 한정, 주제 한정	1,081	월	프레임 보도건수	통일 정책 지지도
한수연 · 윤석민	2016	개입분석	2008~2010년, 2012~2014년	90	KBS, MBC, SBS	프로그램 한정, 연도별, 요일별, 주별 표집	2,328	일	아이템 수, 아이템 보도시간, 기사 제시 방식, 취재원 수, 사운드바이트 길이, 연성화 정도, 심층성, 앵커 의견 개입 정도	내용분석 데이터 간 비교

표 1. 계속

저자	발간 연도	시계열 분석	분석 기간	시계열 사례 수	분석 매체	표집 방식	분석 기사 수	시간 단위	내용분석 항목	비교 시계열 데이터
이완수 · 심재철	2007	VAR	1998년 12월~ 2005년 12월	85	〈조선일보〉, 〈동아일보〉, KBS, SBS	지면 한정(신문), 프로그램 한정(방송), 주제 한정, 헤드라인 한정	2,520	월	논조지수	대통령 지지도
이완수 · 심재철 · 박양수	2007									소비자 기대지수
이완수 · 노성종	2008									경기선행지수, 소비자평가지수, 소비자기대지수
장병희 · 강형구 · 정일권 · 이혜진	2008	VAR	2006년 2월 20일~ 2007년 8월 22일	77	KBS, MBC, SBS	프로그램 한정, 주제 한정, 헤드라인 한정	821(1차), 1,079(2차)	주	경선 후보별 프라이밍, 한나라당 프라이밍	경선 후보 지지율
이완수 · 노성종	2011	VAR	1998년 12월~ 2007년 12월	109	〈조선일보〉, 〈동아일보〉, KBS, SBS	지면 한정(신문), 프로그램 한정(방송), 주제 한정, 헤드라인 한정	3,047	월	논조지수	주가지수, 소매판매액지수

하지만 일 단위로, 또는 많은 매체로부터 수집된 기사 전수를 분석한다고 하면 이를 수작업으로 하기는 사실상 불가능하기 때문에, 장기 시계열 내용 분석에는 컴퓨터를 활용한 자동화된 내용 분석이 필연적으로 요청된다.

〈표 1〉은 국내 언론학에서 시계열 내용 분석을 활용한 주요 연구를 분석 대상과 방법 측면에서 비교한 것이다. 시계열 방법론은 다양한 방법이 시도되면서 점진적인 발전을 이루어왔음을 확인할 수 있다. 그러나 내용 분석 측면에서는 시계열 사례 수보다 분석 기사 수가 대부분 부족하다. 즉, 한 시점에서 분석된 기사는 시간 단위가 분기 단위나 월 단위, 주 단위, 일 단위든 상관없이 제시된 연구들 기준으로는 평균 25건 정도에 불과하다. 비록 매체, 지면과 주제를 한정한 것이라고 하더라도 이 정도의 기사량은 분기는 물론 하루 생산 기사량과 비교해도 극히 일부에 지나지 않는다. 반면 대통령 지지율이나 각종 경제지표 등 비교한 일부 시계열 데이터는 매 시점 전수에 준하는 데이터로부터 추출되어 비교 데이터 간 괴리가 크다.

또한 시계열 사례 수는 어느 정도 충분하더라고 하더라도 분석 기간 자체는 2년 미만으로 짧은 경우가 적지 않다(장병희 등, 2008; 하승태, 2008, 2012; 한수연 · 윤석민, 2016). 또 설사 분석 기간이 길더라도 시간 단위도 함께 길어지면서 시간 단위보다 시점별 분석 기사가 너무 적거나, 기사 내용을 일부 매체와 주제에 한정하고 제목과 논조, 단순보도 건수 등으로 다소 과하게 축약해 볼 수밖에 없다(이완수 · 박양수, 2016; 이완수 · 심재철, 2007; 이완수 등, 2007; 이완수 · 노성종, 2008, 2011; 이준웅, 2005; 홍원식, 2007).

요컨대 언론학에서 뉴스 내용에 대한 시계열 연구를 활성화하기 위해서는 경제지표와 같은 전수 또는 전수를 대표하는 표집으로부터 수십 년간 수집된 다른 시계열 데이터와 같은 수준의 장기 시계열 뉴스 데이터를 체계적으로 생산하고 비교하기 위한 방법이 요청된다. 이는 뉴스가 빅데이터화된 상황에서 자동화된 내용 분석을 활용하여 일 단위 장기 시계열 뉴스 빅데이터를 축적하는 것이 최선의 방법이라고 할 수 있다.

3) NLP와 의미연결망 분석을 활용한 자동화된 내용 분석

국내에서 뉴스 기사를 비롯해 소셜 미디어나 인터넷 게시판 등에 대한 자동화된 내용 분석은 크게 두 가지 방식으로 시도됐다. 우선 NLP를 활용한 방식이 있다. 주로 어절 분석이나 형태소 분석, 공기 분석(cooccurence analysis) 등을 활용해 분석의 기초가 되는 말뭉치를 구축하고, 단어의 빈도를 살펴보거나 문서 간 유사도를 비교하고 식으로 수행됐다(강남준 · 이종영 · 오지연, 2008; 강남준 · 김영희, 2010; 이귀혜 · 강남준 · 이종영, 2008;

이창환 · 심정미 · 윤애선, 2005). 이밖에 감성 분석(sentiment analysis)이나 평판 분석(opinion mining) 등을 수행하는 경우도 있다(최수진, 2014). 다만 영어가 아닌 한국어, 그리고 중립성을 중요한 저널리즘 가치로서 지향하는 한편, 정치, 경제, 사회, 문화 등 다양한 주제를 다루는 뉴스 기사에서는 성능이 충분히 구현되지 않는다는 난점이 있다. 또한 NLP의 경우 형태소 분석기의 활용 여부, 형태소 분석기의 성능, 형태소 분석과 구문 분석을 넘어서 개체명 인식이나 의미 분석 등이 포함되는지 여부 등에 따라 분석 품질이 달라질 수 있음을 염두에 두어야 한다(박대민, 2016).

또한 NLP를 수행한 뒤 단순 빈도를 파악할 경우 적절한 순위화가 이루어지지 않을 수 있다. 흔히 모든 문서에 많이 등장하는 불용어(stopword)나 검색어 등의 단어의 빈도가 높을 수 있다. 이러한 점을 해결하기 위해 흔히 언론학계에서는 수작업으로 불용어나 의미 없는 단어를 제거한다. 보다 체계적인 방식으로는 TF-IDF(term frequency-inverted document frequency)[4] 알고리즘과 같은 간단한 토픽모델링 기법을 사용할 수 있다(박대민, 2016). 이밖에 마르코프 체인 몬테카를로 방법(Markov chain Monte Carlo methods), 의사결정 트리(decision tree), SVM(support vector machine), LDA(latent Dirichlet allocation), 신경망(neural network), word2vec 등 다양한 토픽모델링(topic modeling) 기법을 활용해 뉴스를 분석한 연구가 나오고 있다(감미아 · 송민, 2012; 강범일 · 송민 · 조화순, 2013; 박종희, 2016, 2014; 박종희 · 박은정 · 조동준, 2015; 배정환 · 손지은 · 송민, 2013; 송태민 · 송주영, 2016; 안주영 · 안규빈 · 송민, 2016; 정영미 · 김용광, 2008; 정효정 · 배정환 · 홍수린 · 박찬웅 · 송민, 2016; 진설아 · 허고은 · 정유경 · 송민, 2013).

그러나 언론학 연구에서 널리 사용되어온 순위화 방법은 의미연결망 분석이다. 의미연결망 분석은 단어를 결점(node)으로 기사나 단락, 제목에서의 공동출현을 연결(edge)로 정의하여 사회연결망 분석(social network analysis) 기법을 응용하여 각종 중앙성(centrality)을 계산해 단어를 순위화(ranking)하거나 연결강도(tie strength)를 파악해 분석하는 것이 전형적이다(강명구, 2000; 남인용 · 박한우, 2007; 박대민, 2013, 2015). 이밖에 크기(size), 밀도(density), 응집성(cohesion), 중심화(centralization), 평균경로길이(average path length), 군집계수(clustering coefficient), 연결정도지수(degree exponent) 등을 통해 구조적 속성을 파악하고 그 의미를 해석하는 경우도 있다(김숙 · 박성희, 2009; 박대민, 2014, 2015; Park, Kim, & On, 2016). 둘 이상의 군집이나 연결망을 비교하기도

4) 문서에서 핵심어를 추출할 때 가장 널리 쓰이는 기본 알고리즘으로 어떤 단어가 전체 문서에 등장하는 빈도보다 특정 문서에 등장하는 빈도가 두드러지게 높을 경우 해당 단어를 그 특정 문서를 대표하는 단어로 제시한다.

한다(김해원 · 전채남, 2014; 장하용, 2001). 이러한 경우 각 지표를 저널리즘 관행과 저널리즘 가치 측면에서 어떻게 해석할 것인가가 중요하다.

사실 빈도 및 각종 중앙성 간의 순위상관관계는 매우 높다(박지영 · 김태호 · 박한우, 2013). 때문에 단순히 순위 자체만 도출하고자 한다면 빈도나 토픽모델링을 이용해 파악하면 될 수도 있다. 오히려 의미연결망의 결점이나 연결을 의미론적으로 모호하지 않게 명확히 정의하고, 중앙성은 가장 단순하고 해석이 명확한 절대적 연결정도 중앙성 정도만 파악하는 것이 나을 수도 있다.

일반적으로 토픽모델링 기법은 별도의 프로그래밍 작업이 필요하며 수학적으로 복잡한 반면, 의미연결망 분석은 Ucinet이나 Netminer, NodeXL 등 분석 도구가 있으며 수학적으로도 상대적으로 쉽기 때문에 해석이 편하다. 무엇보다 의미소 간의 관계를 시각화를 통해 직관적으로 파악할 수 있다는 점은 토픽모델링 기법이 갖지 못하는 장점이라고 할 수 있다.

최근에는 NLP와 의미연결망 분석을 혼합한 하이브리드 방식을 활용하는 사례가 늘고 있다. 즉, NLP를 통해 결점에 해당하는 의미소들(tokens)을 기사나 단락, 문장 등에서 자동으로 추출하고 문서 공동출현이나 유사도에 기초해 연결을 정의하여 의미연결망을 구성한 뒤 분석하는 것이다. 국내에서는 KrKwic(Korean key words in context) 패키지가 공개되면서 하이브리드 방식의 의미연결망 연구가 시작됐다(박한우 · 레이데스도르프, 2004). 최근 성능이 검증된 형태소 분석기를 활용하여 주로 명사나 형용사 등을 중심으로 연결망 분석 패키지 프로그램을 활용해 분석하는 사례가 늘고 있다(박지영 등, 2013; 안정윤 · 이종혁, 2015; 정효정 등, 2016).

의미연결망 분석은 의제설정 연구에도 활용될 수 있다. 즉, 의제와 속성이 의미연결망으로 구성된다고 보고, 메시지 생산 집단 간, 동종 매체 간, 또는 이종 매체 간 의미연결망을 비교하는 것이다(박한우 · 바넷 · 김장현 · 김태견 · 남윤재, 2011; 박지영 등, 2013; 안정윤 · 이종혁, 2015; 정효정 등, 2016; 최진호 · 한동섭, 2011). 최근에는 이를 네트워크 의제설정이라고 명명하기도 한다(안정윤 · 이종혁, 2015; Guo & McCombs, 2011, August). 분석 시기를 1기, 2기 등으로 나누고 단순한 수준에서 시계열 연구를 시도한 사례도 있다(박대민, 2015).

4) 연구 문제

정리하면 뉴스 기사의 장기 시계열 내용 분석은 필요성에도 불구하고 내용 분석을 수작업에 의존하기 때문에 제한적으로 이루어져 왔다. 그러나 자연어 처리나 의미연결망 분석 등을 활용한 자동화된 내용 분석, 또는 뉴스 빅데이터 분석을 통해 수집된 텍스트

데이터 전수에 대한 장기 시계열 내용 분석 데이터를 추출할 수 있게 됐다. 이 연구에서는 우선 뉴스 빅데이터 분석을 활용한 장기 시계열 내용 분석 방법론을 제안할 것이다. 이에 대한 연구 문제는 아래와 같다.

- 연구 문제 1: 뉴스 빅데이터 분석을 활용한 뉴스 기사의 장기 시계열 내용 분석은 어떻게 수행할 수 있는가?

이어 이러한 분석 방법을 실제 데이터에 적용한 시행연구를 실시한다. 구체적으로는 뉴스 빅데이터 분석 시스템 '빅카인즈'에서 추출한 전국종합지 8개 매체(〈경향신문〉, 〈국민일보〉, 〈동아일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레〉, 〈한국일보〉), 26년치(1990~2015) 정치면과 사회면 기사 100만 건에서 정보원과 인용문 주제를 검토할 것이다. 시행연구에 대한 연구 문제는 다음과 같다.

- 연구 문제 2-1: 지면별로 기사 수, 정보원 수, 인용문 주제 수는 시간(연도)에 따라 어떻게 증가하거나 감소하는가? (기술통계)
- 연구 문제 2-2: 지면별 기사 수, 정보원 수, 인용문 주제 수의 연도별 증감은 매체별로 어떤 차이가 있는가? (기술통계)

- 연구 문제 3-1: 지면별로 기사당 정보원 수와 기사당 주제 수는 시간(연도)에 따라 어떻게 증감하는가? (기술통계)
- 연구 문제 3-2: 지면별 기사당 정보원 수와 기사당 주제 수의 연도별 증감은 매체별로 어떤 차이가 있는가? (기술통계)

- 연구 문제 4-1: 지면별로 가장 중요한 정보원은 연도별로 어떻게 변화하는가? (정보원 분석)
- 연구 문제 4-2: 지면별로 가장 중요한 정보원은 매체별로 어떤 차이가 있는가? (정보원 분석)

- 연구 문제 5-1: 지면별로 가장 중요한 인용문 주제는 연도별로 어떻게 변화하는가? (인용문 주제 분석)
- 연구 문제 5-2: 지면별로 가장 중요한 인용문 주제는 매체별로 어떤 차이가 있는가? (인용문 주제 분석)

3. 뉴스 빅데이터 분석 활용 장기 시계열 내용 분석 방법의 제안

토픽모델링을 포함하여 NLP와 의미연결망 분석을 혼합하여 활용한다면, 약간의 데이터 정제(data cleansing) 과정을 제외하면 거의 완전히 자동화된 내용 분석이 가능하다. 즉, 표집된 기사가 아닌 수집된 모든 기사에 대해 메타데이터로 축적되어 있거나 NLP 데이터로 산출된 모든 속성에 대해서는 의미연결망 분석을 통해 순위화하고 속성 간의 관계를 파악할 수 있다. 이 경우 대규모 뉴스 기사에 대한 분석도 가능하기 때문에 이러한 분석을 뉴스 빅데이터 분석이라고 부를 수 있을 것이다(박대민, 2013; 박대민 · 백영민 · 김선호, 2015).

뉴스 빅데이터 분석은 자동화된 시계열 내용 분석에서 중요하다. 자동화된 시계열 내용 분석은 크게 자동화된 내용 분석과 시계열 분석으로 나눌 수 있다. 이 연구에서는 특히 방법론적 난점이 큰 내용 분석 측면에 초점을 맞추도록 한다.

뉴스 빅데이터 분석 절차는 크게 자료 수집, NLP, 의미연결망 분석, 해석 내지 담론 분석으로 나누어볼 수 있다. 이 연구는 특히 NLP 중심의 뉴스 빅데이터 분석 시스템인 한국언론진흥재단의 '빅카인즈'를 활용해 분석한다. 또한 해석 또는 담론 분석은 상당부분 연구자의 통찰에 의존하며 아직 자동화될 수 없는 부분이 적지 않다. 따라서 이 연구는 자료 수집과 담론 분석보다는 NLP와 의미연결망 분석에 초점을 두어 기술할 것이다.

이 절에서는 우선 기존 내용 분석과의 절차적 차별점을 염두에 두고 분석상에 염두에 두어야 할 뉴스 빅데이터의 주요 특성을 살펴볼 것이다. 이어 장기 시계열 내용 분석에 맞는 내용 분석 방법으로서 뉴스 빅데이터 분석의 절차를 간략하게 기술할 것이다.

1) 뉴스 빅데이터의 특성

뉴스 빅데이터 분석은 빅데이터를 뉴스로 가공하는 데이터 저널리즘과는 정반대의 과정이다. 즉, 뉴스 빅데이터는 사전에 입력된 메타데이터 외에 비정형상태의 뉴스를 NLP 데이터, 의미연결망 분석 데이터 등으로 정형화한 것이다. 뉴스 빅데이터는 그 양이 방대할 뿐만 아니라 다양한 종류의 데이터가 혼재되어 있고 본질적으로 데이터의 편향, 결측, 오측 등 다양한 오류 가능성과 시간에 따라 변화하는 동적 특징을 갖는 등 분석에 어려움이 있다. 그러나 뉴스 빅데이터는 완전히 무질서한 데이터가 아니라 일종의 복잡계(complex system) 특성을 갖는 데이터로 효율적 분석을 가능하게 하는 특성을 갖고 있다.

(1) 데이터 편향

뉴스 빅데이터 분석은 기존 내용 분석과 달리 표집 과정이 없다. 그럼에도 불구하고 뉴스 빅데이터는 전집은 아니며 정확히 말해 DB에 수집된 전체 데이터를 의미한다. 이 때 뉴스 빅데이터는 수집 및 DB 구축 단계에서 데이터 편향이 있을 수 있다.[5] DB 수집 단계의 데이터 편향은 표집으로 해결할 수 없다. 표집이 DB 데이터 내에서 이루어지기 때문이다. DB상의 편향성을 표집으로 조정하려면 DB화 이전의 전집에 대한 정확한 정보가 있어야 하는데, DB화 이전의 자료에 대해서는 분포는커녕 간단한 기술통계조차도 확보할 수 없는 경우가 많다. 오히려 빅데이터 분석이 DB의 편향성을 확연하게 부각시켜서 이를 보정할 수 있게 도움을 줌으로써, 사후적으로나마 데이터의 편향성을 줄일 수 있다. 사실 표집은 데이터의 편향성을 해소하기 위한 것이 아니라 분석상의 편의 때문에 이루어지는 경우가 많다.

DB 축적 데이터의 편향 가능성 때문에 데이터에 대한 깊은 영역지식(domain knowledge)을 바탕으로 데이터 특성을 면밀히 파악하고 타당한 연구 설계를 해야 한다. 예컨대 이 연구에서 활용할 '빅카인즈'의 매체 분류는 종이신문이 아닌 언론사닷컴을 기준으로 한다. 즉, 〈경향신문〉으로 분류된 기사는 종이신문 외에도 닷컴에서 제공되고 있는 자매지 기사를 모두 포함한다.[6] 따라서 '빅카인즈'를 활용한 매체 간 비교연구는 '신문'이 아닌 '신문사' 간 비교연구가 된다. 또 '빅카인즈'에서는 전국중앙지 중 〈조선일보〉, 〈중앙일보〉, 〈동아일보〉[7]와 KBS, MBC, SBS 등 지상파 3사의 데이터도 빠져 있다는 점도 연구 설계 단계에서 고려사항이 될 수 있다.

(2) 성능과 오류 보정

뉴스 빅데이터 분석은 수작업에 의한 코딩 대신 NLP를 활용한다. 때문에 코더 간 신뢰도를 측정할 필요는 없지만, 대신 NLP 도구에 대한 성능을 파악해야 한다.

고난도 분석에 들어갈수록 완벽한 NLP 도구를 찾기는 어렵기 때문에 꽤 우수한 NLP

---

5) 사실 연구자가 아날로그 자료로부터 직접 데이터를 입력하지 않고 포털이든 사이트든 공공 DB든 이미 디지털화된 데이터를 분석한다면 데이터 수집 단계의 편향성 문제를 염두에 두어야 한다.

6) '빅카인즈'는 기본적으로 과거 '카인즈' 데이터를 활용하기 때문에 과거 '카인즈'에서 데이터를 표집한 경우도 마찬가지 난점을 갖고 있다.

7) 이 연구에서 분석을 수행한 2016년 3월에는 〈동아일보〉 데이터가 '빅카인즈'에 포함되어 있었지만 2016년 10월 현재는 제외됐다. 이 연구에서는 〈동아일보〉 데이터를 연구 목적으로 한정해 분석에 포함시켰다.

데이터라고 하더라도 20% 정도의 인식 오류를 갖는 경우가 많다. 오류 종류는 크게 두 가지인데, 하나는 데이터가 있는데도 불구하고 추출되지 않은 결측 문제이고 다른 하나는 데이터를 추출했지만 값을 잘못 추출한 오측이다. 이는 각각 재현율 문제와 정확도 문제에 대응한다.

따라서 뉴스 빅데이터 분석을 위해서는 주요 기능들에 대한 NLP 인식 성능이 공개된 도구를 사용하고 자체 개발한 경우에는 NLP 성능을 명시하고 분석을 해야 한다. 가능하면 주요 알고리즘도 파악하면 더 좋다. 이를 바탕으로 연구자는 추출된 NLP 데이터를 활용할 수 있을 정도로 NLP 도구의 성능이 충분한지, 충분하지 않다면 그에 대해 어떻게 적절하게 대응할지를 판단할 수 있다. 또한 일반적으로 NLP 성능은 특정 데이터 집합에만 높게 나오는 경우도 적지 않으므로 소규모 데이터를 기준으로 원 자료와 비교하면서 실제 성능을 판단해 볼 필요도 있다.

NLP 데이터가 많지 않을 경우 수작업에 의한 정제가 가능하다. 이러한 정제 작업은 일반적으로 오측 문제를 해결한다. 그러나 재현율이 낮으면 원 자료를 일일이 확인해야 하는데 이 경우 사실상 자동화한 내용 분석이라고 할 수 없다. 게다가 데이터가 방대할 경우 재현율이 높아도 현실적으로 수작업에 의한 정제 작업이 불가능하다.

따라서 데이터의 오류를 가능한 통계적으로 파악하고 제거하는 방향을 모색할 필요가 있다. 우선 의미 있는 속성 내지 개체(entities)들이 제대로 추출됐는지, 그리고 의미 관계가 충분히 드러났는지 파악하기 위해 분석할 데이터의 분포가 일반적으로 알려진 데이터 분포와 유사한지 살펴보아야 한다. NLP 데이터의 경우 많은 경우 빈도든 연결정도 중앙성이든 멱함수 분포를 갖는 경우가 많다(박대민, 2014). 또 시계열적으로 매체 간 추이를 비교하는 등의 방식으로 이상점을 찾아내고 이를 해결할 수 있다. 만일 특별한 이유 없이 특정 매체만 기사 수가 크게 적다거나 특정 해에만 기사 수가 급감한다면 결측을 의심할 수 있다. 이밖에 오류에 상관없도록 데이터를 표준화(normalization)하는 방안도 고려할 수 있다. 예컨대 정보원 수보다 기사당 정보원 수가 기사 수집상의 결측 오류를 줄여줄 수 있다. 대수의 법칙(law of large numbers)에 근거해 데이터가 일정 이상으로 많다면 오류가 상쇄될 수도 있다. 즉, 매체별로 분석하기에 데이터가 충분해 보이지 않는다면 매체를 통합하거나 시간 단위를 늘림으로서 데이터를 어느 정도는 늘릴 수 있다. 이밖에 알고리즘이나 분석 과정은 복잡할수록 오류가 커질 수 있으므로 분석 과정을 가능한 간결하게 설계할 필요도 있다. 분석 자료를 공개하는 방안도 검토될 수 있다.

(3) 차원 축소

차원 축소(dimension reduction)는 의미연결망에서 결점과 연결을 분석하기 쉽게 단순 명료하게 정의하는 문제와 관련된다. 일반적으로 의미연결망은 사회연결망보다 결점과 연결이 많아지기 쉽기 때문에 결점과 연결의 수를 줄일 수 있는 차원 축소가 특히 필요하다. 차원 축소가 충분히 이루어지지 않으면, 의미연결망이 너무 많은 결점으로 구성되거나 모든 결점이 연결된 완전연결망(complete network)에 가깝게 된다. 반대로 차원 축소가 너무 과하면, 결점이 너무 적거나 결점들이 서로 충분히 연결되지 않은 성긴 연결망(sparse network)이 된다. 완전연결망이나 성긴 연결망은 오류가 있는 연결망은 아닐 수도 있지만 적절히 순위화가 되지 않기 때문에 전수를 분석해야 해서 사실상 뉴스 빅데이터 분석을 활용하는 의미가 없어지게 된다. 따라서 결점이 너무 많으면 결점을 다중 분류하여 하위 유목으로 분석하고, 너무 적으면 상위 유목으로 종합할 수 있다. 예컨대 결점을 단어, 명사, 고유명사, 인명으로 구체화할수록 그 수는 줄어들고 해석도 명확해질 수 있다. 또 연결이 너무 많이 나타나면 유사도든 공동출현이든 관련도(relevance)의 임계값을 높이는 방식으로 연결을 줄여갈 수 있다. 예컨대 기사 공동출현보다는 문장 공동출현이 의미관계를 더 엄격하게 정의한 것이라고 볼 수 있다.

(4) 의미연결망의 성장

뉴스 빅데이터 분석에서 활용하는 뉴스 의미연결망의 가장 큰 특징은 시간에 따라 성장하는 연결망이라는 점이다. 예컨대 결점을 정보원으로, 연결을 기사 공동출현으로 정의한 뉴스 정보원 연결망의 경우, 정보원은 어떤 시기에는 등장했다가 다른 시기에는 등장하지 않을 수 있다. 또 두 정보원은 어떤 시기에는 같은 기사에 함께 인용됐다가 다른 시기에는 그렇지 않을 수 있다. 따라서 시차가 있는 뉴스 정보원 연결망은 결점이나 연결이 다를 수 있다. 또 시간 단위를 길게 잡을수록 결점과 연결이 점점 누적적으로 많아지게 된다. 그리고 일정 규모 이상으로 커지게 되면 연결정도 중앙성별 결점 수가 멱함수(power law) 분포를 유지하게 된다. 이는 일종의 의미론적 상전이(phase transition)이다. 저널리즘 관점에서 해석하면 일정 기간 이후에는 중요한 정보원은 대부분 등장하고 관련된 정보원은 같은 기사에 한번이라도 인용되지만 관련되지 않은 정보원은 시간이 지나도 연결될 가능성이 거의 없으며, 중심(hub)을 점하는 정보원은 시간 단위를 늘려도 그 중요도를 유지하는 반면 주변부에 있는 정보원이 중심이 되기는 대단히 어렵다는 것을 뜻한다.

뉴스 의미연결망이 성장하는 연결망이라는 사실은 뉴스 의미연결망 분석을 위해서는 빅데이터를 분석해야만 한다는 것을 의미한다. 의미연결망이 충분히 성장할 정도로 분

그림 3. 성장하는 뉴스 정보원 연결망

석 데이터가 많지 않다면 실제로 중요한 결점이나 연결이 누락될 수 있다.

〈그림 3〉은 차세대융합기술원이 개발한 뉴스 빅데이터 시스템인 '뉴스소스 베타'를 이용하여 '창조경제'를 검색어로 2013년 6월 1일부터 1주, 2주, 4주 단위로 뉴스 정보원 연결망(연결정도중앙성 0 이하는 제외)을 같은 배율로 시각화한 것이다. 1주 연결망의 경우 결점이 적고 연결이 성긴 상태이지만, 4주 연결망의 경우 결점이 많아지고 연결도 점차 중심이 드러나는 척도 없는 연결망의 모습으로 바뀌는 것을 알 수 있다.

(5) 두터운 꼬리 분포

일정 성능을 만족하는 NLP 도구를 통해 충분히 많은 데이터가 적절하게 차원 축소되어 추출되면, 뉴스 의미연결망은 매체 유형이나 주제, 기간, 오류 포함 여부에 관계없이 척도 없는 연결망, 정확히 말해 두터운 꼬리(fat tailed) 형태의 연결망 특성을 갖게 된다. 즉, 연결정도 중앙성 값에 따른 연결정도지수 γ 값이 발산하지 않고 특정 값으로 수렴한다.[8] 뉴스 정보원 연결망의 경우 연결정도지수 값은 '1.6 ± 0.2'으로 수렴한다(강병남,

8) 한 결점 i의 연결선 수를 뜻하는 연결정도(degree)를 k_i라고 할 때, 연결정도 k_i의 분포함수를 연결정도분포함수를 $P_d(k)$라고 하고, 이를 연결정도가 k인 결점 수를 총 결점 수 N으로 나눈 양으로 정의하면 척도 없는 연결망에서 $P_d(k)$는 k가 클 때 멱함수인 $k^{-\gamma}$에 근사하며 이를 γ 기준으로 변환하면 아래와 같은 공식을 도출할 수 있다(강병남, 2010; 박대민, 2014; Barabási & Albert, 1999).

$$\gamma \sim -\frac{\ln(P_d(k))}{\ln k}$$

참고로 연결정도지수 값에 따른 분포 특징은 다음과 같다. $\gamma < 0$: 푸아송 분포(Poisson distribution), $0 < \gamma < 2$: 두터운 꼬리, $2 < \gamma < 3$: 척도 없는 연결망(scale free network), 매우 작은 세상

2010; 박대민, 2014; Park et al., 2016).

물론 뉴스 의미연결망은 저널리즘 관행이나 의제에 따라서 다양한 형태로 그려질 수 있다. 예컨대 뉴스 정보원 연결망의 경우 권위주의 언론 관행이나 매우 전문적인 분야일 경우 특정 정보원을 극단적으로 집중 인용하게 되면서 별 모양의 연결망이 될 수도 있다. 또 의제가 충분히 성숙하지 않은 상태로 간헐적으로 보도될 경우 기간을 늘리고 데이터를 늘려도 성긴 연결망이 나타날 수 있다. 그러나 충분히 논쟁적인 주제인 경우, γ는 ① 발산하지 않으며 수렴하며 ② 같은 취재 관행을 가지면 연결정도지수 값이 유사하게 된다. 이를 뒤집어 말해 뉴스 정보원 연결망의 경우 γ가 1.6 안팎으로 수렴하는지를 살펴봄으로써, ① 의미연결망의 결점과 연결이 적절한 수준에서 축약됐는지(결점과 연결의 정의 측면), ② 주제를 충분히 많은 정보원이 다루고 있으며 충분히 성장한 의제인지(의미연결망의 성장 측면), ③ 기사가 정보원 인용 등을 통해 사실성을 성취하고자 하며, 마감시간이나 기자의 자율적 취재 등 객관주의 관행을 따르는지(저널리즘 관행 측면)를 어느 정도 파악할 수 있다. 또한 뉴스 의미연결망이 두터운 꼬리 분포를 갖기 때문에 연결정도가 낮은 정보원이나 주제 등을 체계적으로 분석에서 대거 제거할 수 있으며, 최상위권 정보원과 주제를 분석할 경우에도 전체적인 의미연결망의 주된 내용을 파악할 수 있다고 말할 수 있다. 또한 뉴스 의미연결망에 의해 결점을 순위화할 경우 결점 간의 중요도 차이가 등간이 아니며, 1~2위의 격차가 2~3위의 격차보다 훨씬 큰 극단적인 서열 차이를 갖는다는 점을 시사한다.

2) 자동화된 장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석 절차

일반적인 뉴스 빅데이터 분석 절차는 〈그림 4〉처럼 자료수집, 자연어 처리, 뉴스의미연결망 분석, 해석으로 이루어진다. 1단계인 자료수집 단계에서는 수집할 데이터를 정의하고 DB를 설계하기 위한 데이터 모델링 단계가 포함될 수 있다. 또한 자료수집은 수작업이나 크롤러를 활용한 자동 수집, 또는 아카이브를 통한 다운로드나 협약을 통한 데이터 이전 등 다양한 방식으로 이뤄질 수 있다. 2단계의 자연어 처리는 크게 형태소 분석, 구문 분석, 개체명 인식, 의미 분석, 담론 분석 등을 수행하게 된다. 자연어 처리를 통해서 결점과 연결에 관한 데이터를 얻게 되면 이를 활용해 의미연결망 분석을 할 수 있게 된다. 3단계 의미연결망 분석은 개체명, 단어, 문장, 기사, 매체 등 다양한 수준에서 수행

현상(ultra small world), $3 < \gamma$: 척도 없는 연결망, 작은 세상 현상.

그림 4. 뉴스 빅데이터 분석 절차

할 수 있다. 또 앞서 언급했듯이, 연결망의 구조나 결점에 대한 중앙성 등을 분석한다. 그리고 결점 간의 관계를 보다 직관적으로 파악하기 위해 시각화를 수행한다. 끝으로 의미연결망 분석을 통해 순위화된 주요 결점을 유형별로 분석하거나 연결망 구조를 보고 해석할 수 있다. 대표적으로 개체명 중 정보원이나 주제, 인용문 등을 분석할 수 있다.

이 연구에서 활용할 '빅카인즈'와 같은 뉴스 빅데이터 분석 시스템을 이용하면 뉴스 빅데이터 분석에서 자료수집 및 자연어 처리 등의 과정을 상당 부분 생략할 수 있다. 다만 이 경우 DB와 데이터의 특성에 대한 이해가 필수적이다.

앞서 간략하게 언급했듯이, 뉴스 빅데이터 분석을 자동화된 장기 시계열 내용 분석에 접목시킬 경우 전통적 내용 분석을 이용한 시계열 분석과는 몇 가지 차이가 있게 된다. 첫째, 수작업에 의한 코딩 대신 자연어 처리를 수행하므로 코더 간 신뢰도 대신 자연어 처리 성능을 공개할 필요가 있다. 둘째, 표집한 기사 대신 DB에 쌓인 기사 전수, 또는 훨씬 많은 양의 기사를 분석한 데이터를 다루게 된다. 이 때문에 시각화 등이 거의 필수일 수 있다. 셋째, 의미연결망 분석을 해 보면 뉴스 빅데이터가 시기별, 매체별, 주제별로 그 크기나 속성이 매우 다른, 즉 비모수적이고 비선형적이며 명목적이고 순위차가 극심한 데이터라는 점을 알 수 있다. 따라서 뉴스 빅데이터를 다른 종류의 시계열 데이터 간 비교할 때는 데이터 간 특성 차이에 유념하여 모형을 제안해야 한다.

4. 분석 사례: 100만 건 기사의 정보원과 주제 분석으로 본 신문 26년

여기서는 제안된 뉴스 빅데이터 이용 시계열 내용 분석을 뉴스 빅데이터 분석 측면을 중심으로 시행적으로 적용해 보도록 한다. 구체적으로는 1990년부터 2015년까지 주요 중앙지 8개 매체의 정치면과 사회면 기사 약 100만 건을 정보원과 인용문 주제 측면에서 간략하게 분석해 볼 것이다. 저널리즘 이론이나 역사, 매체 간 비교연구 관점에서 보다 면밀한 해석은 별도의 연구를 통해 수행할 필요가 있다.

1) 연구대상

시행연구의 분석 대상은 다음과 같다.

① 검색어: 없음(모든 기사)
② 분석 매체: 〈경향신문〉, 〈국민일보〉, 〈동아일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레신문〉, 〈한국일보〉[9][10]
③ 분석 지면: 정치, 사회[11]
④ 분석 기사: 정치면 505,002건, 사회면 483,053건 등 총 988,055건
⑤ 분석 기간: 26년(1990년 1월 1일~2015년 12월 31일)
⑥ 분석 항목: 정보원(중복 제거)[12], 인용문 주제(중복 제거)[13]
⑦ 분석 의미연결망: 정보원연결망, 인용문 주제연결망

9) 〈문화일보〉는 1990년부터 1995년까지 기사가 없다.
10) 정확하게는 종이신문 외에 해당 언론사 닷컴에 게재된 인터넷 신문, 잡지, 방송 등 자매매체 기사가 모두 포함된다.
11) 자료 수집 당시인 2016년 3월 21일엔 자동 지면 분류가 되지 않은 상태였다. 따라서 구 '카인즈'에서 수작업에 의해 지면 분류된 기사만 분석했다. 대체로 지면 기사는 수작업으로 지면 분류가 돼 있으며, 온라인 전용 기사는 지면 분류가 되어 있지 않는 경향이 있다. 2016년 4월 4일 현재 '빅카인즈'는 자동 지면 분류기를 통해 수집기사 전수를 추가 분류해둔 상태다. 8개 매체에서 자동분류된 기사 수는 정치면 1,250,441건, 사회면 1,901,222건 등이다. 이 연구의 분석 기사 수는 정치면의 경우 자동 분류된 기사 수의 40.39%, 사회면은 25.41% 수준이다. 개발사에 따르면 '빅카인즈'의 정치면과 사회면과 그 하위 지면에 대한 자동지면분류기 성능은 F1점수 기준 53~88% 수준으로 일부 세부 지면 분류를 제외하면 대체로 80% 안팎의 성능을 보인다. 이 연구는 전체 기사 중 수작업으로 지면 분류된 30% 정도의 기사를 분석대상으로 했지만, 추후 자동 지면 분류한 기사 전수를 분석할 수도 있다.
12) 여기서 정보원은 직접인용문으로 발언이 인용된 개인 실명 인물 또는 기관으로 정의했다. 익명 정보원은 제외했다.
13) 인용문 주제는 인용문당 3개씩 자동 추출했다.

'빅카인즈'는 전문가 버전[14] 을 통해 메타데이터와 NLP 데이터를 다운로드 받을 수 있다. 다만 연구 실행시점에서 아직 전문가 버전이 공개된 상태가 아니었고 수집할 데이터도 워낙 많기 때문에 데이터는 '빅카인즈' 개발사를 통해 별도로 엑셀 파일 형태로 받았다. 이 파일은 인용문 정보를 포함한 인용문 파일로, 파일은 연도별, 매체별, 지면별로 생성했다. 인용문 파일에서 유의미한 데이터 값이 없는 경우를 제외한 데이터 유형은 〈표 2〉와 같으며 그 엑셀 파일의 예시는 〈그림 5〉와 같다.

표 2. 데이터 유형

열 영문명	설명	비고
INFOSRC	정보원	이름 + 소속 + 직함 / 직업명
INFOSRC_NAME	이름	
INFOSRC_ORG	소속	
INFOSRC_POS	직함	
STN_CONTENT	인용문 본문	
SEN_ID	인용문 ID	매체, 날짜, 정보원 정보 포함
ART_ID	기사 ID	매체, 날짜 정보 포함
ART_DATE	날짜	연, 월, 일
ART_PROVIDER	매체명	
ART_CATEGORY1	지면분류 1	다중분류
ART_CATEGORY2	지면분류 2	다중분류
ART_CATEGORY3	지면분류 3	다중분류
ART_TAG_1	기사 태그 1	3개 추천
ART_TAG_2	기사 태그 2	3개 추천
ART_TAG_3	기사 태그 3	3개 추천
SNT_TAG_1	인용문 태그 1	3개 추천
SNT_TAG_2	인용문 태그 2	3개 추천
SNT_TAG_3	인용문 태그 3	3개 추천
ART_HEADLINE	기사제목	
ART_BYLINE	기자명	
NEWS_LINK	원문링크	빅카인즈 내 원문기사 인링크

14) http://tools.kinds.or.kr/adam/login.do

그림 5. 인용문 관련 NLP 다운로드 엑셀 파일

2) 연구 방법

이 연구에서는 뉴스 NLP는 '빅카인즈'를 활용했다. 연결망 분석 및 시각화는 별도 프로그램과 사이트를 개발해 이용했다.

(1) 메타데이터 수집 및 뉴스 NLP

'빅카인즈'는 개발사인 '솔트룩스'의 LEA(Language Engineering & Analysis)와 ADAM(Advanced Data Analysis Management)을 기반으로 두는 NLP에 특화된 뉴스 빅데이터 분석 시스템이다. LEA의 상세한 알고리즘은 개발사의 기업 기밀로 알려져 있지 않다. 대략적으로 주요 알고리즘을 살펴보면 형태소 분석과 개체명 인식, 감성 분석 등에는 SSVM (structured support vector machine)[15] 및 그 변용 알고리즘을, 구문 분석에는 단어를 트리 구조로 나타내어 특정 언어의 문법이 아닌 기계학습을 통해 분석하는 그래프 기반 방법(graph based dependency parsing) 등 통계적 기계학습 알고리즘을 기초로, 여기에 규칙 기반(rule based) 방식을 결합한 혼종(hybrid) 방식을 활용한다.

이 연구에서는 특히 형태소 분석을 거쳐 개체명 인식, 인용문 추출, 인용문 주제 추출을 통해 산출된 NLP 데이터를 활용한다. 각각의 추출 성능은 계속 개선 중으로, 이 연구에서 데이터를 수집하기 전에 개발사 측에서 밝힌 성능은 다음과 같다.

우선 형태소 분석은 44개 모든 형태소를 분석하며, F1 점수(F1 score) 기준 98%의 인식 성능을 제공한다.[16] 인명, 조직명, 직업/직위, 지역명 등에 대한 개체명 인식 성능은 〈표 3〉과 같다.[17]

---

15) https://en.wikipedia.org/wiki/Structured_support_vector_machine

16) 2015년 3월 기준, 〈솔트룩스 LEA 기술백서〉.

17) 2015년 11월 4일 기준, 〈분석엔진 품질 평가서 버전 1.2〉.

표 3. 개체명 인식 성능

개체명 종류	재현율	정확도	F1
인명(PS)	81.64	89.78	85.51
조직명(OG)	87.69	90.27	88.96
직업/직위(OC)	77.11	88.98	82.62
지역명(LC)	92.98	90.27	93.82
PLO + OC	84.85	90.92	87.72

F1 점수는 인식 성능 평가에서 가장 널리 쓰이는 지수로 그 공식은 다음과 같다.

$$F = \frac{2 \times 재현율 \times 정확도}{재현율 + 정확도}$$

인용문 추출 성능을 살펴보면 '빅카인즈'는 인용문을 쌍따옴표를 기준으로 추출한다. 즉, 한 문장에 쌍따옴표로 묶인 인용이 두 개가 있다면 이를 2개의 인용문으로 추출한다. 인용문 추출 성능은 정확도 기준으로 82.26%이지만 재현율이 보고되지 않은 상태이다. 다시 말해 기사에 있는 인용문이 얼마나 추출되는지 확실하지 않다.[18]

인용문 주제 추출에는 토픽랭크(TR, TopicRank)[19] 알고리즘 등이 적용됐다. 기사 단위 태그 추출 결과에 대해 사용자 만족도 기준 80% 이상의 성능을 구현한다(Berlocher, Lee, & Kim, 2008, July).

18) 무엇보다 인용문 추출 기능에는 의미 중의성 해소 기능이 포함되어야 한다. 즉, 해당 인용문의 정보원을 함께 추출해야 하는데, 이 정보원이 대명사나 '성 + 직함' 등으로 축약되어 나타난다면 해당 지칭어가 누구를 말하는지를 파악해주어야 한다.

19) 토픽랭크 알고리즘은 주제어 후보 추출(candidate extraction), 단어 군집화(word clustering), 순위화 등 3단계로 이루어진다. 후보 추출 알고리즘은 아래와 같은 공식에 의해 추출된다.

$$TR(K, w) = (df(K, w)/df(w))^{\alpha}(-p(w)\log(p(w)))^{\beta}$$

여기서 K는 사용자가 입력한 검색어, w는 각 주제어, $df(K, w)$는 K와 w가 동시에 출현한 빈도, $df(w)$는 w가 출현한 빈도, $p(w)$는 w가 문서에 출현할 확률, α와 β는 가중치 조정 값을 뜻한다. 수식에서 '$df(K, w)/df(w)$'는 K와 w가 해당 문서에서 얼마나 자주 출현했는지를 보는 공기어 분석을, '$-p(w)\log(p(w))$'는 주제어가 전체 문서에서 출현할 확률보다 특정 문서에서 얼마나 높은 확률로 출현했는지로 순위화하는 *TF-IDF* 알고리즘을 반영한다. 단어 군집화는 각 주제어들을 문서별로 갖는 *TR*값을 벡터로 표현한 뒤, 두 주제어 간의 코사인 유사도(cosine similarity)를 계산해 유사한 것끼리 묶는 과정이다. 순위화는 군집 간에는 군집 내 포함된 주제어의 *TR*합에 따라, 군집 내에서는 주제어별 *TR*에 따라 순위를 부여하는 방식으로 이뤄진다.

(2) 뉴스 의미연결망 분석

이 연구에서는 앞서 설명한 개체명 수준의 뉴스 의미연결망 분석, 즉 뉴스 정보원연결망 분석과 뉴스 주제 연결망 분석을 수행한다. 기존의 연결망 분석 프로그램으로는 메모리 문제로 많게는 수만 건의 결점을 분석하기 어려운 경우가 있다. 때문에 이 연구에서는 연결정도 중앙성 계산을 위해 앞의 〈그림 5〉와 같은 인용문 파일을 입력하고 결점과 연결에 해당하는 열(column)을 하나씩 지정하면 에지 리스트(edge list) 형태의 파일을 내부적으로 생성해 자동으로 연결정도 중앙성을 구해주는 연결정도 중앙성 분석기(DegreeAnalyzer)를 개발자와 함께 자체적으로 만들어 활용했다. 그 형태와 출력 데이터 예는 〈그림 6〉과 같다.

정보원 연결망의 결점 정보는 인용문 파일의 'INFOSRC_NAME' 열에서, 연결정보는 'ART_ID' 열에서 뽑은 에지 리스트를 바로 생성할 수 있다. 하지만 주제 연결망의 경우, 디그리 애널라이저를 쓰기 전에 인용문 파일에서 연결 정보를 담은 'SEN_ID'와 결점 정보를 담은 'SNT_TAG_1', 'SNT_TAG_2', 'SNT_TAG_3' 등 3개열로 구성된 링크드 리스트(linked list)를 추출하고 이를 에지 리스트로 변환해주는 과정이 필요하다. 따라서 링크드 리스트를 에지 리스트로 변환해주는 프로그램(EdgeListConverter)을 추가 개발했다. 즉, 주제의 연결정도 중앙성은 에지 리스트 컨버터로 인용문 파일을 에지 리스트 파일로 바꾸고, 이를 다시 디그리 애널라이저에 입력해 구한다. 디그리 애널라이저와 에지 리스트 컨버터는 파일 단위뿐만 아니라 폴더 단위로 입출력이 이루어져 수천 개의 파일을 한 번에 분석할 수 있게 설계했다.[20]

이밖에 연도별, 매체별, 지면별로 정보원과 주제에 대해 연결망을 그릴 경우 도출되는 의미연결망이 매우 많다. 따라서 뉴스 빅데이터를 효과적으로 보여주기 위한 사이트

	A	B	C	D	E	F
1	INFOSRC	DEGREE_IN	YEAR	ART_PROV	ART_CATEGORY	
2	박근혜 대	55	2016	경향신문	정치	
3	최고위원	40	2016	경향신문	정치	
4	국방부	38	2016	경향신문	정치	
5	더불어민주	22	2016	경향신문	정치	
6	한민구 국	21	2016	경향신문	정치	
7	홍용표 통	20	2016	경향신문	정치	
8	안철수 의	19	2016	경향신문	정치	
9	기자 간담	18	2016	경향신문	정치	
10	김무성 대	18	2016	경향신문	정치	

그림 6. 디그리 애널라이저와 출력된 엑셀 데이터

20) 연구자가 프로그램 기획을, 강남용 카이스트 연구원이 개발을 맡았다.

그림 7. 신문 26년 시각화 사이트

(http://112.175.24.115/~kpf/)를 별도로 만들었다.[21] 〈그림 7〉은 사이트 초기 화면 예시이다. 사이트에서는 경제지 데이터와 2016년 1~3월 데이터, 경제와 문화 지면 데이터, 기관 연결망 데이터를 포함해 총 3,888개의 의미연결망을 시각화한다. 사이트는 특정 결점을 선택할 경우 주제어 중심 연결망을 포함하거나, 연결망을 확대 및 축소, 이동할 수 있는 기능을 제공한다.

(3) 기술통계 및 시계열 분석

'빅카인즈'에서는 연도별, 매체별 기술통계 자료를 추출할 수 있다. 또 정보원과 인용문 주제에 대해 중앙성 분석을 하는 과정에서 중복이 제거된 정보원과 주제의 리스트를 얻을 수 있으므로 각각의 기술통계값을 알 수 있다. 시계열 분석에는 Excel 2007을 활용했다(조신섭, 손영숙, 성병찬, 2016; 한광종, 2014).

21) 연구자가 사이트 기획을, 데이터 시각화 전문기업 비주얼다이브 및 퍼넥스가 개발을 맡았다.

3) 분석 결과

(1) 기술통계

연도별, 매체별, 지면별 기사 수와 정보원 수, 인용문 주제 수[22]는 〈부록 Ⅰ-1〉에서 〈부록 Ⅰ-6〉까지와 같다. 결측이나 미분류 기사가 있긴 하지만, 기사나 정보원, 주제, 또는 전체 매체, 지면별, 매체별 추세가 모두 비슷해 대체적인 추이를 파악할 수 있다.

기사 수가 증가하면서 정보원 수와 주제 수도 늘어나는 추세이다. 이는 무엇보다 인터넷 등장 등으로 기사가 늘면서 절대적 다양성은 더 커진 결과로 보인다. 그러나 정보원 수와 주제 수는 기사가 증가한 만큼 크게 늘지는 않았다. 매체의 증가와 기사 수의 결측 등을 감안해 정보원 수와 주제 수를 표준화한 값의 추이, 즉 기사당 정보원 수와 기사당 주제 수의 연도별 매체별 추이는 매체에 따라 다소 차이가 있지만 〈부록 Ⅱ-1〉에서 〈부록 Ⅱ-4〉와 같이 되레 감소하고 있다.

기사당 정보원 수와 기사당 주제 수에 관한 26년 치 시계열 모형을 추정하기 위해 우선 선형회귀모형을 가정하고 분석한 결과, 잔차의 표준편차를 통해 파악한 이상값의 판별, *F*검정, 더빈왓슨 검정을 통한 오차항 자기상관 판정, 예측력(결정계수) 등의 측면에서 소수의 매체를 제외하면 적절한 모형이 도출되지 않았다. 즉, 시계열 그래프는 증감이 있는 형태였다.

따라서 비선형 회귀분석을 통해 시계열 모형을 다시 추정해 보았다. 가장 많은 모형을 간결하게 예측할 것으로 보이는 모형은 〈표 4〉와 같은 3차 다항식으로 표현될 수 있다.[23] 〈문화일보〉는 1996~2015년 연 단위 자료 20개 중 1996년을 이상치로 판단하여 제외한 19개 자료로 회귀식을 산출했으며, 다른 매체는 1990~2015년까지 각각 26개 자료를 통해 회귀식을 산출했다.

정치면의 경우 기사당 정보원 수 기준으로는 대부분 매체가 1995년에 정점을 찍고 하락하는 추세였다. 〈경향신문〉, 〈국민일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레신문〉 등 6개 매체는 2013~2015년 중에 기사당 정보원 수가 가장 작았다. 기사당 주제

---

22) 같은 매체, 같은 연도, 같은 지면 내에선 중복이 제거됐지만, 같은 연도의 다른 매체 간엔 중복이 제거되지 않았다. 즉, 같은 해, 같은 매체, 같은 지면에 대통령이 10번 등장했다면 이는 중복으로 보고 정보원 1명으로 간주하지만, 다른 해나 다른 매체, 또는 다른 지면에 대통령이 등장했다면 각각 1명으로 놓고 연도별 또는 매체별 정보원 수를 합산했다.

23) 대체로 6차 다항식 형태의 회귀식이 결정계수 값은 가장 컸지만 너무 복잡하고 특정매체에만 맞아 과적합 문제가 있을 것으로 보인다. 다만 이 연구는 시행연구로 예측을 목적으로 하지 않으므로 각 회귀모형의 최적화(optimization)를 수행하지는 않았으며 3차 다항식의 예시만 제시했다.

수로 보면 정점은 주로 1995~2006년 사이에 있었으며, 반면 2013~2015년에 최하인 경우가 5개 매체(〈국민일보〉, 〈문화일보〉, 〈서울신문〉, 〈세계일보〉, 〈한겨레신문〉)였다. 사회면의 경우에도 최댓값은 기사당 정보원은 1996~2006년 사이, 기사당 주제는 2003~2007년 사이에 있는 매체가 많았다. 반면 2013~2015년에는 최솟값을 기록한 경우가 많다.

표 4. 지면별, 매체별 기사당 정보원 수와 기사당 주제 수의 회귀식과 R^2

지면	매체	유형	회귀식	R^2	최댓값 연도	최솟값 연도
정치	〈경향〉	기사당 정보원 수	y = 6E - 05x^3 - 0.0038x^2 + 0.0484x + 0.6678	0.6709	1996	2015
	〈국민〉		y = 2E - 05x^3 - 0.0019x^2 + 0.0274x + 0.746	0.7019	1996	2013
	〈동아〉		y = 2E - 05x^3 - 0.0007x^2 + 0.0037x + 0.8085	0.0376	1996	2003
	〈문화〉		y = 7E - 05x^3 - 0.0065x^2 + 0.1388x - 0.0029	0.7308	2004	2013
	〈서울〉		y = -7E - 05x^3 + 0.0015x^2 + 0.0013x + 0.7283	0.5619	1996	2015
	〈세계〉		y = -3E - 05x^3 + 0.0004x^2 - 0.0074x + 0.8421	0.6701	1996	2015
	〈한겨레〉		y = 6E - 06x^3 - 0.0011x^2 + 0.0096x + 0.7728	0.678	1996	2013
	〈한국〉		y = 2E - 05x^3 - 0.0011x^2 + 0.0117x + 0.7518	0.261	1996	2010
	〈경향〉	기사당 주제 수	y = 0.0002x^3 - 0.0113x^2 + 0.0908x + 4.6543	0.6744	1996	2008
	〈국민〉		y = -0.0003x^3 + 0.0045x^2 + 0.0365x + 4.7292	0.6464	2005	2013
	〈동아〉		y = -0.0003x^3 + 0.0122x^2 - 0.1805x + 5.383	0.3008	1996	2003
	〈문화〉		y = 0.0005x^3 - 0.0439x^2 + 0.9851x - 1.0708	0.8106	2004	2013
	〈서울〉		y = -0.0008x^3 + 0.0244x^2 - 0.1635x + 4.8789	0.6285	2004	2015
	〈세계〉		y = -0.0005x^3 + 0.0189x^2 - 0.2199x + 5.482	0.6021	1996	2015
	〈한겨레〉		y = -1E - 05x^3 - 0.0057x^2 + 0.0476x + 5.1032	0.7679	1990	2015
	〈한국〉		y = -0.0006x^3 + 0.0212x^2 - 0.193x + 5.1587	0.4004	2007	1998
사회	〈경향〉	기사당 정보원 수	y = 2E - 05x^3 - 0.0019x^2 + 0.0244x + 0.7806	0.6856	1996	2008
	〈국민〉		y = 6E - 05x^3 - 0.0032x^2 + 0.0422x + 0.7699	0.7396	2000	2013
	〈동아〉		y = -4E - 06x^3 + 0.0003x^2 - 0.0032x + 0.8736	0.1524	2006	1999
	〈문화〉		y = 7E - 05x^3 + 0.006x^2 + 0.1221x + 0.1857	0.748	2004	2013
	〈서울〉		y = -0.0002x^3 + 0.0045x^2 - 0.0287x + 0.8359	0.8085	2004	2015
	〈세계〉		y = -2E - 05x^3 - 1E - 05x^2 + 0.0024x + 0.866	0.8033	1995	2015
	〈한겨레〉		y = -2E - 05x^3 - 0.0003x^2 + 0.0022x + 0.9142	0.7884	1990	2013
	〈한국〉		y = -3E - 05x^3 + 0.0013x^2 - 0.0145x + 0.8371	0.0777	2004	1994
	〈경향〉	기사당 주제 수	y = 0.0002x^3 - 0.0118x^2 + 0.1439x + 3.8691	0.577	2003	2008
	〈국민〉		y = -0.0002x^3 + 0.0022x^2 + 0.0571x + 4.0626	0.6727	2006	2013
	〈동아〉		y = -0.0003x^3 + 0.0135x^2 - 0.1303x + 4.4764	0.568	2010	1992
	〈문화〉		y = -1E-06x^3 - 0.0164x^2 + 0.5007x + 1.048	0.6739	2005	2013
	〈서울〉		y = -0.001x^3 + 0.0303x^2 - 0.2189x + 4.5451	0.8296	2004	2015
	〈세계〉		y = -0.0005x^3 + 0.0141x^2 - 0.0907x + 4.3588	0.8542	2004	2015
	〈한겨레〉		y = -4E - 05x^3 - 0.0068x^2 + 0.1229x + 4.0596	0.8037	2004	2013
	〈한국〉		y = -0.0004x^3 + 0.0123x^2 - 0.0823x + 4.3827	0.5938	2007	2015

26년간 기사는 매체와 지면에 따라서는 많게는 23배가량[24] 급증했다. 이는 종이신문의 지면 수도 늘어났을 뿐만 아니라, 인터넷 신문, 방송 등 자매매체도 다수 생겨났기 때문으로 보인다.

정리하면 중앙지 중심의 매체는 1990년대 중반부터 2005년대 중반까지 기사 하나에 많은 정보원과 주제를 담았다. 반면 최근 3년간에는 정보원과 주제 기준으로 기사 간 중복이 많았다. 매체별로는 〈한겨레신문〉의 경우, 꾸준히 하락하는 추세였다. 반면 〈동아일보〉는 대체로 유지되는 것처럼 보이지만 이는 수집된 데이터 자체의 차이일 가능성이 크다. 즉, 〈동아일보〉는 기사 수가 상대적으로 적고 기사당 정보원 수와 기사당 주제 수가 유지되는 경향이 있는데, 이는 아마도 〈동아일보〉만 데이터 자체에 닷컴 기사 등이 빠져 있기 때문으로 추정된다.

(2) 정보원 분석

① 정치면

중요도 1위 정보원을 기준으로 볼 때 정치면의 경우 2003년을 기점으로 정당 대변인의 시대와 대통령의 시대로 갈린다. 1990년대엔 특히 야당 대변인이 중시됐다. 예컨대 새정치국민회의가 창당된 해인 1995년 〈세계일보〉에선 '박지원 대변인'이 162명의 정보원과 공동인용됐다. 참고로 같은 해 같은 매체에서 '김영삼 대통령'은 42명의 정보원과 공동인용됐을 뿐이다. 이밖에도 '이규택 대변인', '손학규 대변인', '박범진 대변인'이 김영삼 대통령보다 중요도가 높았으며, '안성열 대변인', '구창림 대변인', '안성열 대변인' 등이 20위권 안에 포진했을 정도로 대변인이 중시됐다.

노무현 임기 첫해인 2003년 이후엔 대통령의 중요도가 단연 높다. 2003년 〈서울신문〉은 '노무현 대통령'을 271명의 정보원과 공동인용했는데 이는 전 시기 모든 매체의 정치면을 통틀어 가장 많다. 당시 중요도 2위는 한나라당으로 87명의 정보원과 인용됐다. 대통령별로 살펴보면 2008년 〈한겨레신문〉은 '이명박 대통령'을 224명과, 2015년 〈경향신문〉은 '박근혜 대통령'을 209명과 인용했다. 〈부록 Ⅲ-1〉은 연도별 매체별로 정치면에서 가장 중요한 정보원을 나타낸 것이다.

24) 〈세계일보〉 사회면 1990년과 2015년 비교.

② 사회면

사회면은 2000년 전후로 중요 정보원이 바뀌었다. 〈부록 Ⅲ-2〉는 연도별 매체별로 사회면에서 가장 중요한 정보원을 나타낸 것이다. 2000년 이전엔 국회의원 등 정치인의 비중이 높았다. 이 시기엔 사회면의 상당 부분이 정치면에 실릴만한 주제를 다뤘기 때문이다. 특히, 검찰 출입기자가 정치인에 대한 각종 수사를 취재하고 이에 관한 정치인의 멘트를 받아 싣는 경우가 많았던 것으로 보인다. 예컨대 1994년 '장석화 의원'은 주사파에 대한 공안 수사나 '김말용 의원 뇌물 수수' 등과 관련해 비판적 논의를 진행했다. 1990년 〈한겨레신문〉은 '유준상 의원'을, 〈동아일보〉는 '김우석 의원'을 각각 79명, 62명의 정보원과 인용했는데, 이때 두 신문은 사회면에서 국정감사를 지면 중계했다. 2000년 이후엔 교육부가 가장 중요한 정보원으로 부상했다. 예컨대 2007년 〈경향신문〉은 '교육부'를 93명의 정보원과 함께 인용하는데, 특목고나 사교육, 대입, 등록금 등 다양한 주제를 언급하고 있다. 교육부는 '교육부 관계자'처럼 기관정보원으로 보도되는 경우가 많았다.

(3) 인용문 주제 분석

① 정치면

〈부록 Ⅲ-2〉는 지난 26년간 중앙지 정치면에서 가장 중요하게 나온 주제를 연도별, 매체별로 나타낸 표이다. 정치면의 인용문 주제로는 '미국'이 가장 많았다. 26년간 8개 매체에서 총 64회[25] 로 1위를 했다. 특히, 2001~2009년 중 2007년[26] 을 제외한 8년 동안 많았다. '미국' 외에도 '소련', '일본', '북한' 등 주변국의 거론이 많았던 점으로 미뤄볼 때 정치면에서 한국의 지정학적 요인에 대한 고려가 많다는 점을 알 수 있다.

정당, 특히 보수정당도 중요한 화두로 제시됐다. '민자당'-'신한국당'-'한나라당'-'새누리당'으로 이어지는 보수정당이 1위를 차지한 수를 합하면 총 71회나 됐다. 반면 '민주당'은 23회 1위를 기록했다. 보수정당에 치중된 경향은 1990년대에 강했다. 앞서 야당 대변인이 중시된 것으로 나온 정보원 분석 결과와 모순된다고 생각할 수 있지만, 이는 야당 대변인 역시 보수정당에 대한 비판을 주로 가했기 때문이다.

그밖에 1997년 외환위기 시절 '한보', 1999년 '내각제', 2013년 진보지 중심으로 '국정원' 등이 눈에 띄는 주제였다. 특히, 2013년 〈경향신문〉에서 '국정원'은 644개 주제와 공동인용됐을 만큼 쟁점이 됐다.

---

25) 〈부록 Ⅲ-1〉부터 〈부록 Ⅲ-4〉까지 각 표의 전체 칸수가 총 202개이므로 어떤 정보원이나 주제가 모든 매체, 모든 연도에서 1위할 경우 최대 202회 1위를 할 수 있다.

26) 대선이 있던 2007년에는 '한나라당'이 중요도 1위였다.

② 사회면

〈부록 Ⅲ-4〉는 지난 26년간 중앙지 정치면에서 가장 중요한 것으로 나온 주제를 연도별, 매체별로 나타낸 표이다. 사회면 주제는 정치면이나 경제면보다 주제가 다양했다. 주제를 매체 특성, 연도 특성, 출입처, 문체 특성 측면에 따라 살펴보자.

우선 사회면에선 매체별 특성이 비교적 뚜렷이 나타났다. 즉, 〈한겨레신문〉은 '노동자'와 관련해 다양한 주제를 다룬 반면, 〈서울신문〉은 '공무원'을 중요한 주제로 봤다. 〈국민일보〉는 2010년대 이후 '환자들'이 중요한 주제였던 경우가 상대적으로 많았는데 이는 사회면에 의료기사를 다뤘기 때문으로 보인다.

다음으로 연도별로 살펴보면, 1997년 '한보' 파산, 1999~2000년 외환위기 이후 '구조조정', 2015년 '메르스' 사태가 중시됐다.[27] 특히, 2015년 〈세계일보〉에서는 '메르스'를 636개의 주제와 함께 언급해 공동인용 주제 수가 가장 많았다.

사회면 특성상 출입처 측면에선 '수사', '피고인', '피해자' 등 경찰이나 검찰 관련 주제가 많았다. '서울', '교육부', '서울대' 등 서울시나 교육 담당 출입기자가 다뤘을 만한 주제도 중시됐다.

문체상으로는 '가능성'이란 단어가 다양한 주제와 함께 등장했다. 이러한 현상은 특히 2000년 이후 두드러졌다. '가능성'이란 단어는 "유독물질 취급 도중 휴·폐업한 업체의 경우, 장기 보관 과정에서 탱크 손상 등에 의한 누출사고 발생 가능성을 배제할 수 없다"(정수근 대구환경운동연합 생태보존국장, 〈문화일보〉, 2013, 1, 15)와 같이 전문가나 책임자의 전망이나 우려, 예상 등을 담을 때 많이 사용됐다.

5. 결론 및 함의

의제설정 연구와 같은 사례에서 보듯이 언론학에서 시계열 내용 분석은 주요 방법론으로 널리 사용되어왔으며 많은 발전이 있었다. 그러나 전통적인 내용 분석에서는 수작업으로 기사나 문서를 코딩하기 때문에 장기적이고 세분화된 시계열 연구를 수행하는 데 한계가 있었다.

27) 참고로 2014년 세월호와 같이 여러 주제어가 복합적으로 얽힌 경우는 중요도 1위 주제어 분석만으로는 잘 나타나지 않는다. 즉, '빅카인즈' NLP 특성상 세월호는 '세월호' 외에도 '세월호_참사', '세월호_사고', '세월호_특별법', '세월호_사건', '세월호_유가족' 등으로 나뉜다. 또 이 주제는 '청와대', '교육부' 등 평년에도 중시되는 주제어와도 연관되며, '유가족', '위원회' 등 다른 관련 주제어와도 연관된다.

이러한 난점은 자동화된 내용 분석, 특히 NLP와 의미연결망 분석이 결합된 뉴스 빅데이터 분석을 통해 극복할 수 있다. 이 연구는 뉴스 빅데이터 분석에서 고려할 점을 명시하고 간단한 분석 절차를 제안했다. 이어 시행연구를 통해 실제로 8개 매체의 정치 및 사회면 약 100만 건의 기사를 분석해 봄으로써 분석 가능성을 보여주었으며, 연 단위의 개략적인 시계열 분석을 통해 대체적 매체 지형의 변화를 파악했다.

구체적으로는 8개 매체 중앙지의 26년 치 정치면과 사회면 기사에 대해 NLP와 의미연결망 분석을 결합한 뉴스 빅데이터 분석을 실시했다. 이를 통해 분석 기사 수나 시계열 데이터 수, 내용 분석 결과 수의 한계를 극복할 수 있었다. 즉, 분석 기사가 연도별, 매체별로 수천 건에 달하고 전체 분석 기사 수가 100만 건에 육박한다고 해도 데이터가 수집된 이후에는 수분 내에 분석이 가능하다. 시간 단위는 1년으로 삼았지만, 분석 목적에 따라서는 주, 분기, 월, 일 등 다양한 시간 단위의 분석도 가능하다. 1일 단위의 경우 시계열 데이터는 1990년 1월 1일부터 2015년 12월 31일까지 9,496개나 되겠지만 기계적 분석에서는 제한이 없다. 이는 일 단위 이상의 다른 다양한 시간 단위의 시계열 데이터와 비교 분석이 가능함을 시사한다. 흔히 뉴스 포털에서는 기사 업로드 시간을 초 단위까지 제시하는데, 이 경우 방송 시청률 데이터처럼 초 단위 분석도 가능하다. 이 경우 주식 매매와 같은 초 단위 시계열 데이터와 뉴스 데이터를 비교할 수도 있을 것이다.

한편 시간에 따라 매체별로 정보원 수, 인용문 주제 수가 늘어나기는 했지만, 이는 기사가 급격하게 늘어났기 때문이다. 기사당 정보원 수와 기사당 인용문 주제 수로 표준화하면 매체마다 추이가 다르기는 했고 시간에 따라 다소 상승하기도 했지만 대부분 장기적으로 하락하는 점을 확인했다. 이 연구에서는 그 원인에 대한 면밀한 분석을 제시하지 않았으며 이는 추후 언론사적 관점에서 분석이 필요할 것이다. 중요한 점은 GDP와 같은 경제지표와 마찬가지로 기사의 내용에 대한 장기 시계열 지표를 쉽게 만들 수 있다는 점이다. 기사당 정보원 수와 기사당 인용문 수는 내용적 측면에서 가장 간단한 품질 지표가 될 수도 있을 것이다.

한편 정보원과 인용문 주제를 연도별, 매체별로 검토한 결과, 정치면의 경우 정보원과 인용문 주제에 관한 한 매체 간 차이는 크지 않았으며, 시간에 따라서는 2000년 안팎으로 지면 차이 없이 크게 두 시기로 나누어지는 것을 확인할 수 있었다. 사회면에서도 정보원의 경우 비슷한 현상이 발견됐다. 비록 사회면에서 매체별, 연도별로 인용문 주제의 다양성은 나타났지만 이는 일반적으로 매체 간 비교연구가 갖는 함의를 부분적으로 반증한다고 볼 수 있다. 예컨대 정치 관련 기사에서 보수지와 진보지 간의 차이는 비록 주제별 차이나 같은 주제에 대한 논조의 차이는 있을 수 있지만 중요도 최상위 정보원

의 의존도나 의제 측면에서는 생각만큼 크지 않을 수 있다. 그러나 이를 매체 간 차이가 일반적으로 없다고 성급히 해석해서는 안 될 것이다. 왜냐하면 매체 간 차이는 중요도는 떨어지지만 두터운 꼬리를 형성하고 있는 압도적 다수의 정보원이나 주제에서 두드러질 수 있기 때문이다. 즉, 각 매체가 고유하게 발굴한 정보원과 그들이 던진 독특한 의제를 통해서 매체 간 차이가 형성될 가능성도 있다. 요컨대 보다 세부적인 매체 간 비교를 통한 종합적 결론은 후속 연구를 통해 내려야 할 것이다.

분석상의 여러 제한점은 언급할 필요가 있다. 우선 이 연구는 수작업으로 지면 분류된 기사만을 대상으로 분석했다. 그러나 온라인 기사 중에서는 상당수가 지면 분류 없이 DB에 저장되는 경우가 많다. 대체로 DB에 축적된 기사 전수의 30% 안팎의 기사만 분석한 셈이다. 이 경우 자동지면 분류기를 활용해 완전 분류하여 분석할 수 있다. 다만 NLP 도구에 오류가 있을 수 있다는 점은 감안해야 한다. 즉, 오류를 포함하는 NLP된 전수를 분석한 결과와 사실상 정답인 수작업으로 지면분류된 기사 간에는 트레이드 오프(trade off)가 있을 수 있다.

NLP 측면에서는 인용문 추출 성능을 정확하게 확인할 수 없었다는 한계가 있다. 의미연결망의 분포는 대체로 두터운 꼬리 형태의 전형적인 멱함수 분포를 보이는 것으로 보이지만, 중요도 상위권의 정보원이나 주제의 연결정도 중앙성 값이 기사 수보다 다소 작은 한편, 중요도가 낮은 정보원과 주제의 수는 다소 많은 것으로 보인다. 그 원인으로는 첫째, '세월호'와 '세월호_침몰'처럼 주요 정보원이나 주제가 실제로는 하나로 볼 수 있는데 NLP 상에서 둘로 분할되면서 연결이 줄어들었을 가능성, 둘째, 인용문의 재현율이 낮아서 주제도 덜 추출되고 그에 따라 주제 간 연결이 줄었을 가능성이 있다. 보다 정확한 분석을 위해서는 NLP 성능을 보다 정확히 파악하고 개선할 필요가 있다.

이 연구는 자료 수집을 '빅카인즈'에 의존했기 때문에 8개 매체에 대해서만 분석했다. 〈조선일보〉와 〈중앙일보〉 등 빠진 중앙일간지를 비롯해 〈연합뉴스〉와 같은 뉴스 통신사 기사, 지상파 방송사의 기사를 추가로 분석해 비교하는 것도 의미가 있을 것이다. 또 8개 매체의 기사에 대해 종이신문 기사와 자매지 기사를 나눠서 분석하지 않았다. 매체 간 비교를 보다 세세하게 하기 위해서는 자매지를 구분한 분석도 필요하다. 지면 역시 정치, 사회 등 2개면만 분석했는데, 경제, 문화, 국제 등 다른 지면으로 확장해 분석해 볼 필요도 있다.

추가로 이 연구는 기사에 대해 단어, 특히 개체명 수준, 개체명 중에서도 정보원과 인용문 주제에 대해 분석했다. 추후 분석에서는 장소나 직함, 수치 등 보다 다양한 개체명 분석과 함께, 인용문을 비롯한 다양한 문장 단위의 분석을 추가하여 복합논증 분석과 같

은 담론 분석에 가까운 분석을 수행하면 더 의미가 있을 것이다. 또한 이 연구는 분류 체계나 현저성에 따라 한정하지 않고 모든 정보원과 주제에 대해 분석할 가능성을 제시했다는 점에서 의제의 복합적 성격을 드러내는 동시에 일반화 가능성도 높였다. 이밖에 이 연구는 최상위 정보원과 주제에 대한 분석 결과만 제시했지만, 기사 내용에 대한 보다 세부적인 분석도 가능할 것이다. 예컨대 정보원이나 인용문 주제들을 군집화하고 이들의 특성과 동적인 변화 과정을 살펴본다거나, 주요 개체명들에 대한 장기적인 중요도의 변화나 개체명 간 상관성을 살펴본다면, 의제설정의 동역학에 대한 보다 깊은 이해를 할 수 있을 것이다.

여러 한계에도 불구하고 이 연구는 뉴스 빅데이터 분석을 활용하여 수백만 건의 기사를 대상으로 자동화된 장기 시계열 내용 분석을 수행할 수 있는 가능성을 열었다. 장기 시계열 내용 분석은 분석 양이 방대하여 하나의 논문에 그 내용을 전부 담기는 어렵다. 따라서 데이터를 시각화하거나 데이터를 공개하는 것, 즉 오픈 데이터(open data)가 연구의 반복가능성(replicability)을 위해 중요하다. 이 연구에서도 연도별, 매체별, 지면별 정보원연결망과 주제연결망을 볼 수 있는 시각화 사이트를 함께 제작했으며, 모든 순위의 전체 정보원과 주제 리스트 및 가중치가 담긴 파일도 추후 공개하여 다른 연구자들도 확인하거나 활용할 수 있게 할 예정이다.

■ 참고문헌

감미아 · 송 민 (2012). 텍스트 마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석. 〈지능정보연구〉, 18권 3호, 53-77.

강남준 · 김영희 (2010). 형태주석 전산 말뭉치를 활용한 〈독립신문〉 논설의 저자 연구. 〈한국언론학회 학술대회 발표논문집〉, 115-116.

강남준 · 이종영 · 오지연 (2008). 신문기사의 표절 가능성 여부 판정에 관한 연구: 컴퓨터를 활용한 형태소 매칭기법을 중심으로. 〈한국언론학보〉, 52권 1호, 437-466.

강명구 (2000). 정치뉴스에 나타난 한국 정치권력구조의 네트워크 분석: '동시출현빈도'의 타당도 검증. 〈언론정보연구〉, 37권, 93-130.

강범일 · 송 민 · 조화순 (2013). 토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구. 〈한국문헌정보학회지〉, 47권 4호, 315-334.

강병남 (2010). 〈복잡계 네트워크 과학: 21세기의 정보과학〉. 서울: 집문당.

구교태 (2003). 매체 간(intermedia)과 매체 내(intramedia) 의제분석을 통한 뉴스획일화 연구: 2000년 미국 대통령 선거운동에 관한 뉴스보도를 중심으로. 〈한국언론정보학보〉, 21권, 7-34.

김경희 (2008). 포털 뉴스의 의제설정과 뉴스가치: 포털 뉴스와 인쇄신문의 비교 분석. 〈한국언론학보〉, 52권 3호, 28-52.

김 숙 · 박성희 (2009). 블로그에 나타난 언론인의 사회적 관계망에 대한 탐색적 연구: 중앙일보 기자 블로그를 중심으로. 〈사이버커뮤니케이션학보〉, 26권 3호, 5-42.

김해원 · 전채남 (2014). 빅데이터를 활용한 콘텐츠 제작방안에 관한 탐색적 연구: TV홈쇼핑을 중심으로. 〈사이버커뮤니케이션학보〉, 31권 3호, 5-51.

남인용 · 박한우 (2007). 대권 예비후보자 관련 신문기사의 네트워크 분석과 홍보전략. 〈한국정당학회보〉, 6권 1호, 79-107.

박대민 (2013). 뉴스 기사의 빅데이터 분석 방법으로서 뉴스정보원연결망분석. 〈한국언론학보〉, 57권 6호, 234-262.

박대민 (2014). 뉴스 정보원 인용에서의 폭발성과 언론의 편향성. 〈커뮤니케이션 이론〉, 10권 1호, 295-324.

박대민 (2015). 사실기사의 직접인용에 대한 이중의 타당성 문제의 검토: 〈동아일보〉와 〈한겨레신문〉의 4대강 추진 논란 기사에 대한 뉴스 정보원 연결망 및 인용문 분석. 〈한국언론학보〉, 59권 5호, 121-151.

박대민 (2016). 뉴스 기사의 자연어처리. 〈커뮤니케이션 이론〉, 12권 1호, 4-52.

박대민 · 백영민 · 김선호 (2015). 〈뉴스 빅데이터 분석시스템 연구〉. 서울: 한국언론진흥재단.

박종희 (2014). 베이지안 사회과학 방법론이란 무엇인가?. 〈평화연구〉, 22권 1호, 481-529.

박종희 (2016). 세월호 참사 1년 동안의 언론보도를 통해 드러난 언론매체의 정치적 경도. 〈한국정치학회보〉, 50권 1호, 239-269.

박종희 · 박은정 · 조동준 (2015). 북한 신년사(1946-2015)에 대한 자동화된 텍스트 분석. 〈한국정치학회보〉, 49권 2호, 27-61.

박지영 · 김태호 · 박한우 (2013). 의미연결망 분석을 통한 셀러브리티의 SNS 메시지 탐구: 아이돌의 미투데이 메시지를 중심으로. 〈방송통신연구〉, 82권, 36-74.

박한우 · 레이데스도르프 (2004). 한국어의 내용분석을 위한 KrKwic 프로그램의 이해와 적용: Daum.net에서 제공된 지역혁신에 관한 뉴스를 대상으로. 〈Journal of the Korean Data Analysis Society〉, 6권 5호, 1377-1387.

박한우 · 바넷 · 김장현 · 김태견 · 남윤재 (2011). 〈인터넷 소셜미디어 개론: 이론과 사례〉. 경산: 영남대학교출판부.

반 현 · 맥스웰 맥콤스 (2007). 의제설정 이론의 재고찰: 5단계 진화 모델을 중심으로. 〈커뮤니케이션 이론〉, 3권 2호, 7-53.

반 현 · 최원석 · 신성혜 (2004). 뉴스의 속성과 2차 의제설정 효과 연구: 위도 핵폐기장 보도를 중심으로. 〈한국언론정보학보〉, 25권, 65-102.

배정환 · 손지은 · 송 민 (2013). 텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석. 〈지능정보연구〉, 19권 3호, 141-156.

송태민 · 송주영 (2016). 〈R을 활용한 소셜 빅데이터 연구방법론〉. 서울: 한나래출판사.

안정윤 · 이종혁 (2015). '네트워크 의제설정'의 출현: 뉴스 매체와 온라인 게시판 간 이슈 속성 네트워크의 유사성 분석. 〈한국언론학보〉, 59권 3호, 365-394.

안주영 · 안규빈 · 송 민 (2016). 텍스트 마이닝을 이용한 매체별 에볼라 주제 분석. 〈한국문헌정보학회지〉, 50권 2호, 289-307.

이건호 (2006). 한국 인터넷 매체들의 상호 의제 설정 효과: 8개 온라인 신문의 내용 분석을 중심으로. 〈한국언론학보〉, 50권 4호, 200-227.

이건호 · 유찬윤 · 맥스웰 맥콤스 (2007). 환경 문제의 2차 의제설정효과: 지구 온난화 이슈 내 서로 다른 속성을 중심으로. 〈한국언론학보〉, 51권 2호, 153-179.

이귀혜 · 강남준 · 이종영 (2008). 탄핵 시기 노무현 대통령의 수사학 연구: 컴퓨터 언어 분석 기법을 중심으로. 〈한국언론학보〉, 52권 5호, 25-55.

이동훈 (2007). 뉴스수용자에 대한 포털뉴스의 의제설정효과 연구: 북핵보도 관련 종이신문의 의제전이과정을 중심으로. 〈한국언론학보〉, 51권 3호, 328-357.

이승희 · 송 진 (2014). 재난보도에 나타난 소셜 미디어와 방송 뉴스의 매체 간 의제설정. 〈한국언론학보〉, 58권 6호, 7-39.

이완수 (2007). 한국 경제뉴스의 속성(attributes) 프레임효과 연구. 〈언론과 사회〉, 15권 1호, 86-122.

이완수 (2009). 의제설정이론에서 그랜저 인과관계 모형의 방법론적 타당성 연구. 〈커뮤니케이션 이론〉, 5권 2호, 54-100.

이완수 (2015). 정부 구조변동에 따른 경제커뮤니케이션 효과의 비대칭성. 〈미디어 경제와 문화〉, 13권 3호, 7-44.

이완수 · 노성종 (2008). '무엇'에서 '언제'로: 벡터자기회귀모형을 통한 경제현실, 경제보도, 경제인식 간상호영향의 시간차 탐구. 〈한국언론학보〉, 52권 5호, 320-345.

이완수 · 노성종 (2011). 경기 국면에 따른 경제커뮤니케이션 효과의 비대칭성: 경제보도, 주가, 소비행위 간 효과의 위계, 속도, 강도에 관한 시계열 분석. 〈한국방송학보〉, 25권 3호, 302-348.

이완수 · 박양수 (2016). 경제 정보에 대한 비대칭적 반응: 경제뉴스에 대한 경제 주체의 심리와 행위. 〈한국언론학보〉, 60권 1호, 165-201.

이완수 · 심재철 (2007). 집합적 경제보도와 국가적 경제상황 및 국민적 경제인식이 대통령 지지도에 미치는 영향에 관한 시계열 분석. 〈한국방송학보〉, 21권 2호, 506-545.

이완수 · 심재철 · 박양수 (2007). 경제뉴스, 경제상황, 소비자 기대심리 그리고 소비행위의 상호 속성 의제설정 관계에 대한 시계열 분석. 〈한국언론학보〉, 51권 4호, 280-307.

이준웅 (2005). 갈등적 사안에 대한 여론 변화를 설명하기 위한 프레이밍 모형 검증 연구: 정부의 통일 정책에 대한 뉴스 프레임의 형성과 해석적 프레임의 구성을 중심으로. 〈한국언론학보〉, 49권 1호, 133-162.

이창환 · 심정미 · 윤애선 (2005). 언어적 특성을 이용한 '심리학적 한국어 글분석 프로그램(KLIWC)' 개발 과정에 대한 고찰. 〈인지과학〉, 16권 2호, 93-121.

임종섭 (2011). 매체 관심도와 기사 부각도에서 분석한 대형 뉴스사이트 간 의제설정효과. 〈미디어 경제와 문화〉, 9권 3호, 57-94.

장병희 · 강형구 · 정일권 · 이혜진 (2008). 대통령 후보 경선 관련 방송 뉴스 보도와 후보자 지지도간 시계열적 관련성 분석: 2007년 한나라당 대통령 후보 경선을 중심으로. 〈한국방송

학보〉, 22권 4호, 355-400.
장하용 (2011). 매체 간 경쟁의 심화에 따른 안내적 저널리즘의 약화: 중앙종합언론의 보도에 대한 실증적 분석. 〈한국언론정보학보〉, 56권, 48-70.
정영미 · 김용광 (2008). 사건중심 뉴스기사 자동요약을 위한 사건탐지 기법에 관한 연구. 〈정보관리학회지〉, 25권 4호, 227-243.
정일권 (2010). 사회면 기사 분석(1998년~2009년)을 통해 본 뉴스 미디어의 현실구성. 〈한국언론정보학보〉, 50권, 143-163.
정효정 · 배정환 · 홍수린 · 박찬웅 · 송 민 (2016). 정치적 이념에 따른 트위터 공간에서의 집단 간 의견차이 분석. 〈한국언론학보〉, 60권 2호, 269-302.
조진섭 · 손영숙 · 성병찬 (2016). 〈SAS/ETS를 이용한 시계열분석〉. 서울: 율곡출판사.
진설아 · 허고은 · 정유경 · 송 민 (2013). 트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구. 〈정보관리학회지〉, 30권 1호, 285-302.
최수진 (2014). 한류에 대한 미 · 중 언론보도 프레임 및 정서적 톤 분석. 〈한국언론학보〉, 58권 2호, 505-532.
최진호 · 한동섭 (2011). 정치인 트위터와 신문 · 방송뉴스의 의제 상관성에 관한 연구. 〈언론과학연구〉, 11권 2호, 501-532.
하승태 (2008). 지지율 조사 보도에 따른 유력 대선 후보별 뉴스 보도의 분석: 후보 인용(sound-bite)과 보도사진을 중심으로. 〈한국언론학보〉, 52권 5호, 346-366.
하승태 (2012). 선거여론조사와 후보별 보도량 분석: USA Today의 미대선 경선 보도를 중심으로. 〈언론학연구〉, 16권 2호, 115-140.
하승태 · 조의현 (2008). 중요한 사회적 의제(MIP)에 대한 공적 합의: 1991~2006년의 갤럽데이터 분석. 〈한국언론정보학보〉, 41권, 41-74.
한광종 (2014). 〈Excel 활용 미래예측과 시계열분석〉. 서울: 백산출판사.
한수연 · 윤석민 (2016). 종합편성채널 출범이 지상파 방송 뉴스에 미친 영향: 저녁종합뉴스 보도관행에 대한 개입시계열 분석. 〈한국방송학보〉, 30권 1호, 169-210.
홍원식 (2007). 대통령지지도와 언론의 관계를 통해 살펴본 여론의 순환적 형성에 관한 연구. 〈한국언론학보〉, 51권 6호, 33-61.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509-512.
Berlocher, I., Lee, K. I., & Kim, K. (2008, July). *TopicRank*: *Bringing insight to users*. Presented in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval(pp. 703-704). ACM, New York, NY.
Guo, L., & McCombs, M. (2011, August). *Toward the third level of agenda setting theory*: *A network agenda setting model*. Presented in AEJMC annual conference, St. Louis, MO.
McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176-187.
Park, D., Kim, G., & On, B. (2016). Understanding the network fundamentals of the news sources associated with a specific topic. *Information Sciences*, *372*, 32-52.

Wagner, A. K., Soumerai, S. B., Zhang, F., & Ross-Degnan, D. (2002). Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, *27*(4), 299-309.

최종 투고일 2016년 8월 12일
게재 확정일 2016년 9월 23일
논문 수정일 2016년 10월 6일

부록 Ⅰ-1. 정치면의 연도별 매체별 기사 수

연도	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉	연도별 총합
1990	1,091	1,031	1,208		1,124	1,011	1,246	1,084	7,795
1991	1,275	971	1,291		1,040	1,130	1,309	1,152	8,168
1992	1,479	1,113	1,503		1,225	1,295	1,818	1,269	9,702
1993	1,022	851	1,186		1,257	1,096	1,529	1,008	7,949
1994	1,409	1,128	1,301		1,508	1,301	1,757	808	9,212
1995	1,460	1,329	1,570		1,591	1,532	1,920	1,469	10,871
1996	1,485	1,469	1,486	83	1,714	1,756	1,886	1,624	11,503
1997	1,672	1,577	1,864	1,718	1,987	2,059	2,123	1,694	14,694
1998	1,124	1,167	1,100	1,187	1,095	1,450	2,044	1,437	10,604
1999	1,480	1,662	1,616	1,488	1,558	1,509	2,271	1,684	13,268
2000	2,034	2,858	2,413	1,785	2,672	2,153	3,041	2,020	18,976
2001	1,773	2,361	2,302	2,317	2,256	2,118	2,753	2,145	18,025
2002	2,499	2,671	2,663	2,586	2,901	2,899	3,612	2,388	22,219
2003	2,500	3,231	2,950	2,732	3,271	3,061	3,061	2,504	23,310
2004	1,933	2,692	2,521	2,058	2,599	2,559	2,523	2,363	19,248
2005	1,712	2,031	1,902	1,861	1,866	2,446	2,118	1,986	15,922
2006	1,707	1,935	2,237	2,169	2,063	2,167	2,223	1,949	16,450
2007	2,142	2,246	2,635	2,490	2,563	2,841	3,129	2,496	20,542
2008	5,731	3,067	2,428	2,337	2,730	3,265	4,680	2,568	26,806
2009	5,829	2,798	2,387	2,453	2,450	3,284	5,125	2,867	27,193
2010	5,859	3,048	2,770	2,785	2,505	3,408	5,482	3,804	29,661
2011	6,190	3,130	2,335	2,058	2,442	2,337	5,245	3,216	26,953
2012	7,845	3,406	2,811	2,983	2,886	3,996	6,166	3,229	33,322
2013	6,560	5,801	2,027	3,819	3,544	7,201	4,925	3,544	37,421
2014	5,629	2,620	1,790	3,499	4,212	5,966	4,218	2,813	30,747
2015	7,056	3,085	1,312	2,188	5,685	8,527	4,124	2,464	34,441
매체별 총합	80,496	59,278	51,608	44,596	60,744	72,367	80,328	55,585	505,002

부 록 Ⅰ-2. 정치면의 연도별 매체별 정보원 수

연도	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉	연도별 총합
1990	798	803	937		774	821	975	840	5,948
1991	968	791	1,085		777	828	1,024	935	6,408
1992	1,224	941	1,306		898	1,261	1,539	947	8,116
1993	803	610	870		847	859	1,237	737	5,963
1994	967	877	930		1,144	945	1,289	547	6,699
1995	1,251	1,149	1,307		1,403	1,398	1,442	1,310	9,260
1996	1,464	1,475	1,479	111	1,752	1,759	1,762	1,454	11,256
1997	1,353	1,300	1,478	1,390	1,536	1,568	1,527	1,312	11,464
1998	917	1,071	1,027	841	807	929	1,462	1,060	8,114
1999	1,049	1,345	1,151	1,190	1,276	1,149	1,612	1,227	9,999
2000	1,616	2,318	1,934	1,526	2,084	1,661	2,301	1,632	15,072
2001	1,287	1,875	1,755	1,716	1,717	1,458	1,788	1,619	13,215
2002	1,904	2,156	1,839	2,061	2,276	2,105	2,494	1,606	16,441
2003	1,981	2,165	2,006	2,108	2,223	2,185	2,106	1,864	16,638
2004	1,776	2,297	2,001	2,047	2,404	2,194	2,214	2,037	16,970
2005	1,339	1,706	1,345	1,689	1,631	1,802	1,725	1,459	12,696
2006	1,564	1,686	1,958	1,991	1,774	1,595	1,891	1,598	14,057
2007	1,639	1,757	2,193	1,962	1,930	1,799	2,006	1,772	15,058
2008	2,494	1,970	2,071	1,894	1,969	2,197	2,479	1,858	16,932
2009	2,480	1,786	1,873	1,640	1,792	1,978	2,324	1,935	15,808
2010	2,826	2,020	2,215	1,634	1,962	1,946	2,784	2,390	17,777
2011	2,688	1,953	1,860	1,461	1,819	1,588	2,343	2,086	15,798
2012	3,584	2,214	2,291	1,886	2,213	2,332	2,841	2,506	19,867
2013	3,109	2,586	1,564	1,745	2,178	3,324	2,090	2,306	18,902
2014	2,922	1,528	1,485	1,712	2,310	3,155	2,226	2,084	17,422
2015	2,861	1,645	1,080	1,143	2,112	3,383	1,838	1,632	15,694
매체별 총합	46,864	42,024	41,040	31,747	43,608	46,219	49,319	40,753	341,574

부 록 Ⅰ-3. 정치면의 연도별 매체별 주제 수

연도	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉	연도별 총합
1990	5,473	5,421	6,355		5,191	5,435	7,151	5,253	40,279
1991	6,019	4,672	6,390		4,604	5,680	6,432	5,324	39,121
1992	7,418	5,330	7,528		5,367	6,792	9,596	6,591	48,622
1993	4,798	3,801	5,742		5,573	4,930	7,867	4,542	37,253
1994	5,994	5,151	5,721		7,176	5,492	8,838	3,682	42,054
1995	7,032	6,583	7,448		8,648	8,114	9,010	7,444	54,279
1996	7,859	8,112	7,892	593	9,588	9,553	10,159	8,242	61,998
1997	8,023	8,434	8,194	8,262	9,292	9,591	10,167	7,974	69,937
1998	4,741	5,638	5,011	4,794	4,555	5,669	9,089	6,156	45,653
1999	6,866	8,643	7,153	7,205	7,701	7,012	11,255	8,099	63,934
2000	8,349	14,117	11,182	8,445	11,778	8,990	14,385	9,058	86,304
2001	7,910	13,323	11,291	12,794	10,900	9,576	13,589	10,602	89,985
2002	10,968	13,880	11,546	13,414	14,126	14,070	17,096	10,390	105,490
2003	11,025	16,890	12,161	13,966	16,202	14,057	15,308	12,464	112,073
2004	9,480	15,392	10,926	12,053	14,793	13,603	14,049	12,722	103,018
2005	7,468	11,667	7,917	10,704	10,118	11,424	10,916	10,538	80,752
2006	8,188	10,827	10,477	11,642	10,186	10,531	11,140	10,427	83,418
2007	9,629	12,379	13,672	13,163	12,565	12,862	12,907	14,025	101,202
2008	15,816	12,836	11,857	12,242	13,928	13,952	15,031	13,801	109,463
2009	16,485	12,479	11,076	12,054	12,108	14,028	15,263	14,542	108,035
2010	18,319	13,621	12,877	11,782	12,673	14,898	16,134	18,118	118,422
2011	19,324	13,723	10,850	9,772	12,257	9,745	15,225	15,657	106,553
2012	23,358	13,421	12,416	12,081	13,198	13,227	15,507	16,116	119,324
2013	22,263	19,121	9,139	13,527	15,474	25,304	11,338	17,600	133,766
2014	18,775	10,376	7,692	12,608	14,934	24,274	12,347	12,887	113,893
2015	22,672	11,905	6,117	7,818	14,684	27,165	12,696	10,675	113,732
매체별 합계	294,252	277,742	238,630	208,919	277,619	305,974	312,495	272,929	2,188,560

부 록 Ⅰ-4. 사회면의 연도별 매체별 기사 수

연도	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉	연도별 총합
1990	539	562	791		514	516	888	517	4,327
1991	764	621	954		662	664	1,072	763	5,500
1992	622	733	828		741	904	1,305	679	5,812
1993	906	991	883		916	1,016	1,442	779	6,933
1994	1,014	1,015	1,064		906	983	1,458	596	7,036
1995	1,209	1,197	1,476		1,142	1,403	1,807	1,131	9,365
1996	1,369	1,268	1,509	107	1,144	1,382	1,818	1,236	9,833
1997	1,466	1,275	1,703	1,504	1,250	1,520	1,731	1,558	12,007
1998	1,269	1,325	1,227	1,025	988	981	1,870	1,416	10,101
1999	1,701	1,989	1,958	1,479	1,589	1,225	2,524	1,841	14,306
2000	1,574	2,035	1,982	1,039	1,922	1,253	2,231	1,534	13,570
2001	2,040	2,286	2,605	1,897	2,115	1,543	2,647	1,878	17,011
2002	2,169	2,204	2,221	1,851	2,085	1,602	2,394	1,892	16,418
2003	2,735	2,659	2,459	2,201	2,829	2,083	2,726	2,460	20,152
2004	2,367	2,485	2,437	1,828	2,552	2,081	2,776	2,131	18,657
2005	2,960	2,294	2,268	1,921	2,310	2,516	3,102	2,282	19,653
2006	2,566	2,298	2,180	2,141	2,165	1,958	2,533	1,826	17,667
2007	3,034	1,947	2,561	2,116	2,242	2,118	3,344	1,994	19,356
2008	6,477	4,090	2,338	1,666	2,620	2,566	5,298	1,924	26,979
2009	7,072	4,223	2,845	2,015	3,170	3,561	6,366	2,634	31,886
2010	5,771	3,655	2,591	1,868	2,614	3,139	5,606	2,656	27,900
2011	6,570	4,420	2,549	1,392	2,869	2,510	5,719	3,051	29,080
2012	6,190	3,266	2,143	1,891	2,838	3,513	5,363	2,863	28,067
2013	5,792	7,030	1,381	3,080	3,294	7,160	4,227	3,231	35,195
2014	6,013	3,949	1,508	2,462	6,082	8,027	4,290	3,327	35,658
2015	7,214	2,669	1,139	1,500	8,823	11,813	4,159	3,267	40,584
매체별 총합	81,403	62,486	47,600	34,983	60,382	68,037	78,696	49,466	483,053

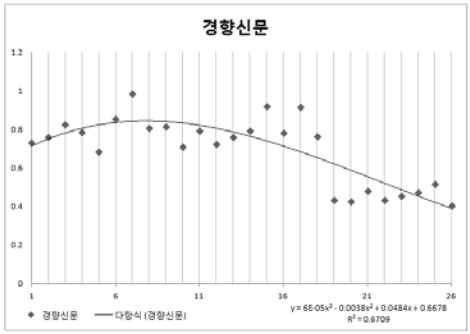
부 록 Ⅰ-5. 사회면의 연도별 매체별 정보원 수

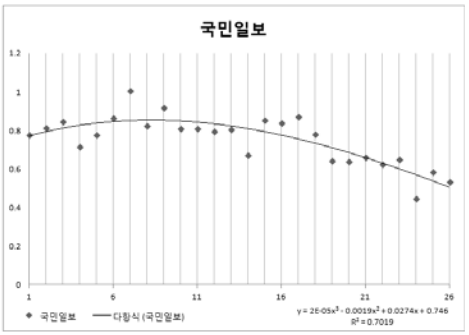
연도	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉	연도별 총합
1990	470	479	676		426	476	878	460	3,865
1991	659	541	870		514	546	945	582	4,657
1992	490	619	681		540	767	1,243	539	4,879
1993	724	808	749		693	855	1,283	620	5,732
1994	818	882	856		729	853	1,324	412	5,874
1995	1,000	1,092	1,355		962	1,331	1,529	908	8,177
1996	1,274	1,193	1,462	135	1,034	1,257	1,635	1,099	9,089
1997	1,267	1,238	1,528	1,321	1,117	1,366	1,556	1,268	10,661
1998	964	1,260	1,086	785	703	748	1,515	1,071	8,132
1999	1,430	1,751	1,524	1,294	1,374	1,039	2,091	1,426	11,929
2000	1,299	1,990	1,764	997	1,557	1,092	1,953	1,281	11,933
2001	1,688	2,061	2,092	1,681	1,830	1,323	2,352	1,443	14,470
2002	1,875	1,906	1,871	1,556	1,698	1,421	2,099	1,425	13,851
2003	2,485	2,231	2,078	1,920	2,362	1,749	2,423	2,029	17,277
2004	2,151	2,215	2,178	1,783	2,432	1,908	2,524	1,898	17,089
2005	2,378	2,008	1,891	1,729	2,127	1,974	2,522	1,701	16,330
2006	2,196	2,061	2,121	1,917	1,847	1,592	2,287	1,409	15,430
2007	2,525	1,757	2,406	1,815	2,083	1,624	3,033	1,664	16,907
2008	3,456	2,808	2,130	1,425	2,289	1,997	3,310	1,641	19,056
2009	3,838	2,967	2,472	1,533	2,360	2,356	3,416	2,049	20,991
2010	3,389	2,707	2,339	1,195	1,973	2,034	3,005	1,987	18,629
2011	3,715	3,194	2,268	1,186	2,020	1,912	3,063	2,329	19,687
2012	3,505	2,487	1,925	1,290	2,032	2,430	2,903	2,344	18,916
2013	3,536	4,197	1,274	1,427	2,329	4,749	2,064	2,521	22,097
2014	3,650	2,660	1,386	1,369	3,058	4,909	2,286	2,689	22,007
2015	3,894	1,928	1,050	889	2,917	6,248	2,261	2,389	21,576
매체별 총합	54,676	49,040	42,032	27,247	43,006	48,556	55,500	39,184	359,241

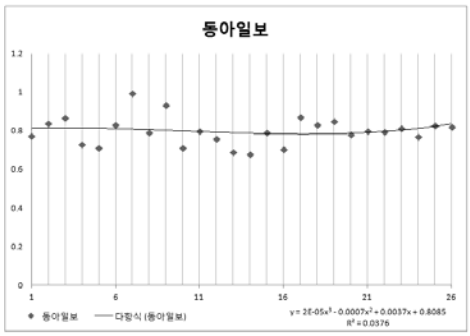
부록 Ⅰ-6. 사회면의 연도별 매체별 주제 수

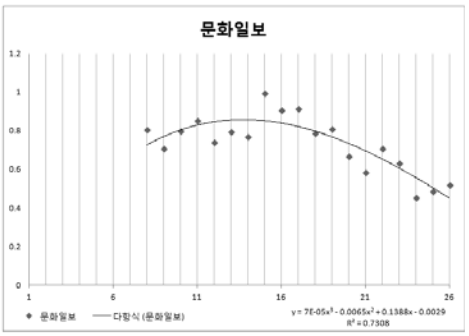
연도	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉	연도별 총합
1990	2,355	2,212	3,522		2,141	2,262	4,062	2,337	18,891
1991	3,192	2,813	4,128		2,855	2,848	4,585	3,134	23,555
1992	2,558	3,124	3,180		2,860	3,627	5,503	2,771	23,623
1993	3,773	4,281	3,822		3,686	4,195	6,410	3,283	29,450
1994	4,111	4,383	4,252		3,926	4,097	6,496	2,594	29,859
1995	4,982	5,124	6,397		5,157	6,241	7,645	4,824	40,370
1996	5,998	5,895	6,445	588	5,131	6,010	7,996	5,526	43,589
1997	6,218	5,785	6,899	6,323	5,584	6,421	7,815	6,680	51,725
1998	5,023	5,870	4,943	4,128	4,028	3,871	8,033	5,747	41,643
1999	7,135	8,900	7,670	6,549	7,041	5,258	11,023	7,782	61,358
2000	6,617	9,442	8,628	4,884	8,161	5,272	9,987	6,907	59,898
2001	9,035	10,757	11,457	9,062	9,461	7,001	12,155	8,786	77,714
2002	9,461	9,863	9,609	8,042	8,986	7,278	10,615	8,460	72,314
2003	12,790	12,606	10,737	10,016	13,238	9,732	12,825	11,408	93,352
2004	10,531	12,149	10,669	9,063	12,758	10,005	13,758	10,384	89,317
2005	12,250	11,457	10,181	9,882	11,335	11,373	14,213	10,558	91,249
2006	11,465	11,653	10,410	10,563	10,156	9,169	11,665	8,595	83,676
2007	12,775	9,825	12,177	10,112	10,918	8,829	14,308	9,935	88,879
2008	16,886	16,786	10,821	8,331	12,324	10,439	16,142	9,271	101,000
2009	20,139	17,246	12,704	9,267	13,549	13,999	17,962	11,716	116,582
2010	18,392	15,506	12,464	6,953	11,638	12,441	16,217	11,400	105,011
2011	19,980	18,039	11,500	6,768	12,007	10,193	15,420	12,642	106,549
2012	18,570	12,912	9,128	7,214	11,950	11,605	14,612	12,531	98,522
2013	19,749	22,619	6,302	8,215	13,187	23,593	8,705	13,895	116,265
2014	20,093	14,184	6,837	8,099	16,685	26,313	10,604	14,019	116,834
2015	22,975	10,050	5,174	5,177	16,789	32,001	11,042	12,835	116,043
매체별 합계	287,053	263,481	210,056	149,236	235,551	254,073	279,798	218,020	1,897,268

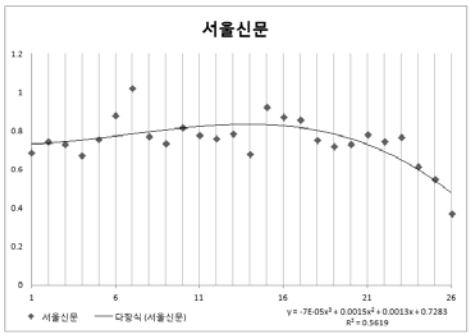
부록 II-1. 정치면의 매체별 기사당 정보원 수

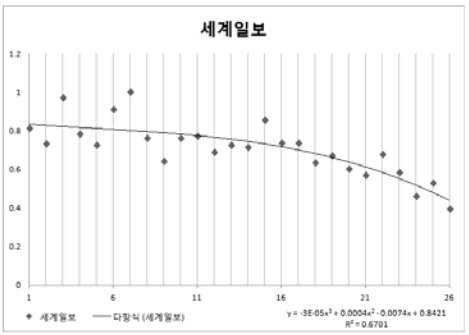


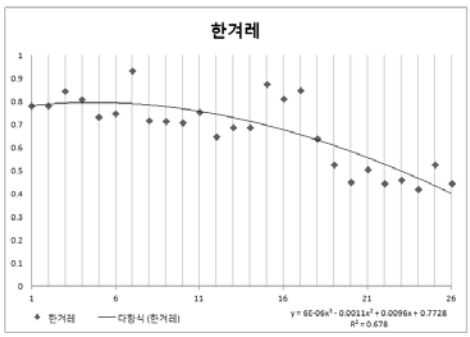


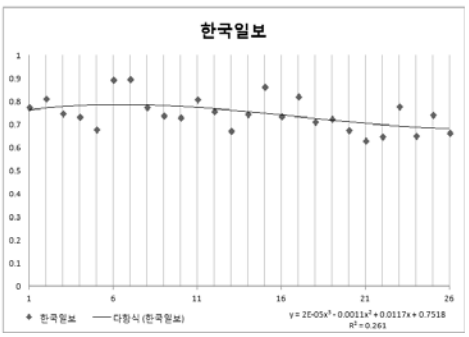




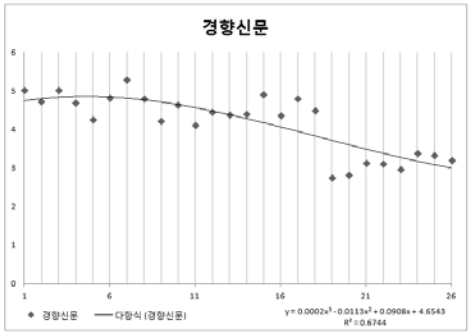


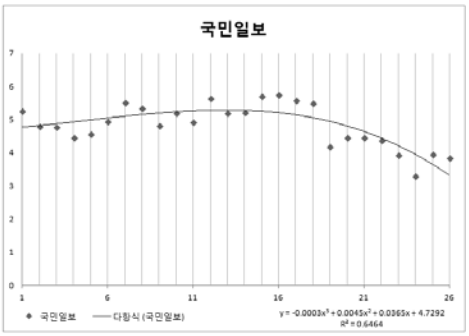


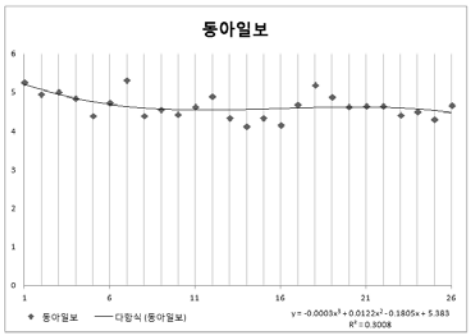


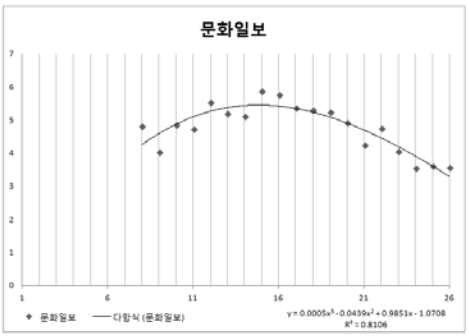


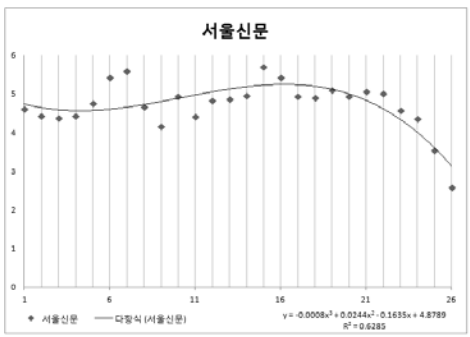
부록 II-2. 정치면의 매체별 기사당 주제 수

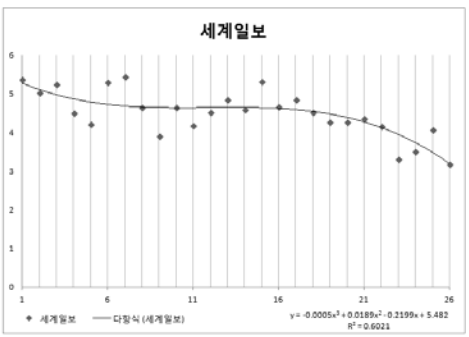


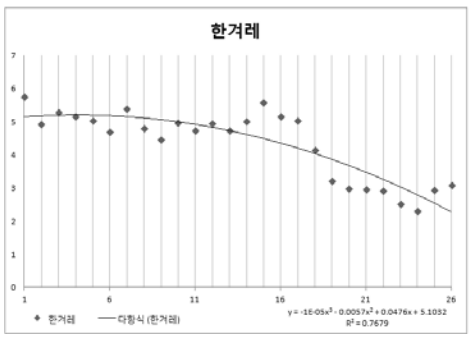


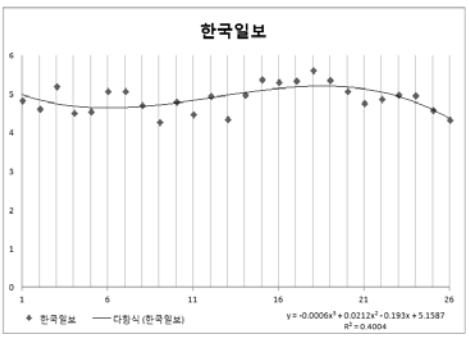




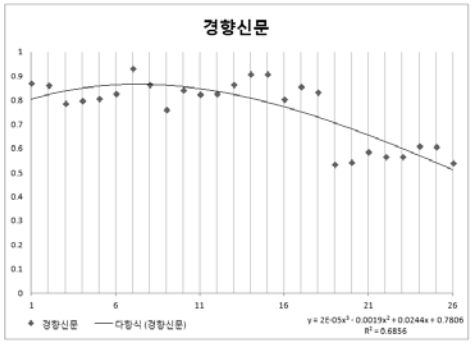


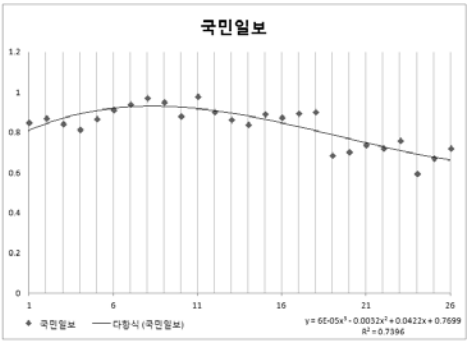


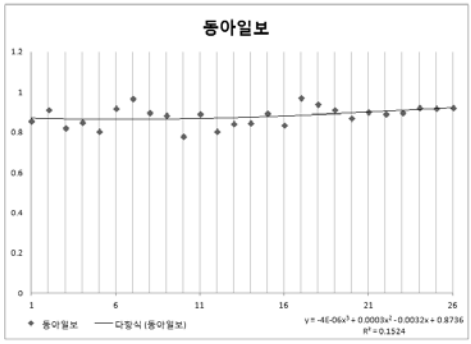


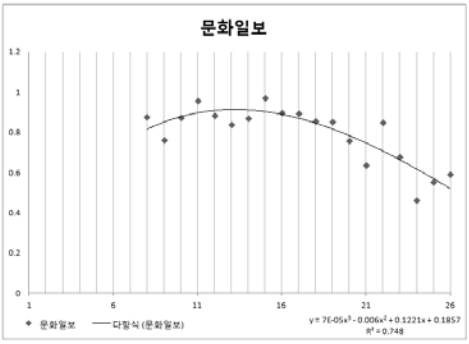


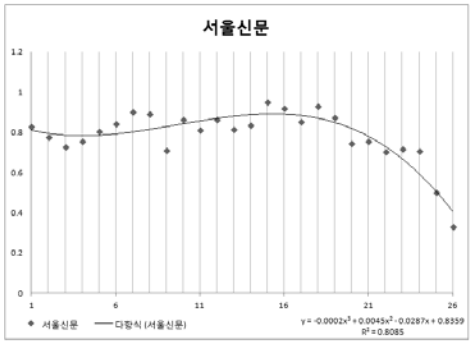
부록 II-3. 사회면의 매체별 기사당 정보원 수

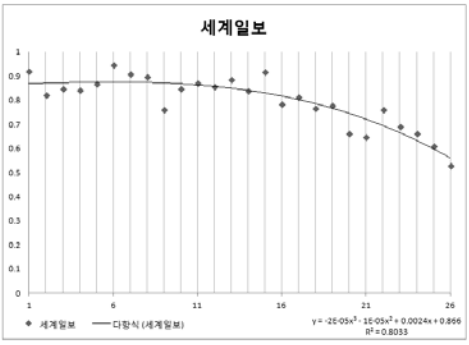


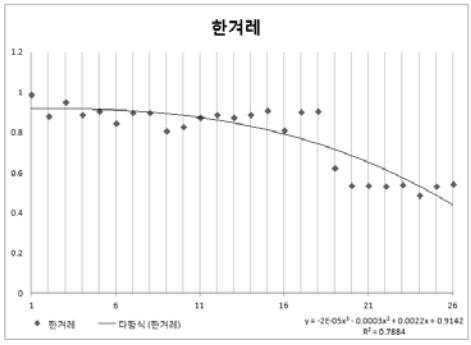


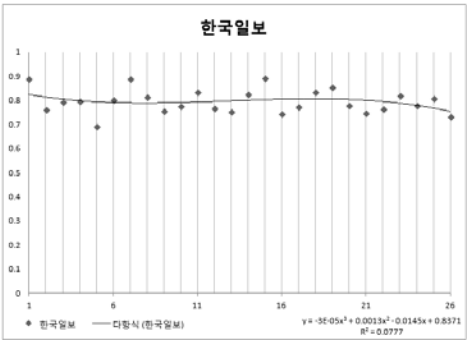




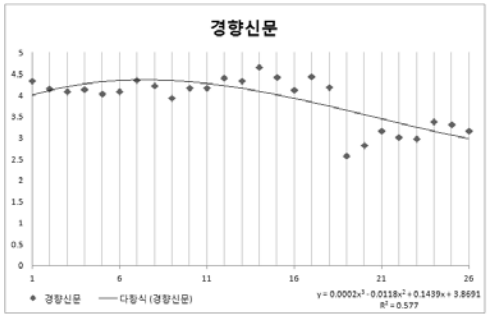


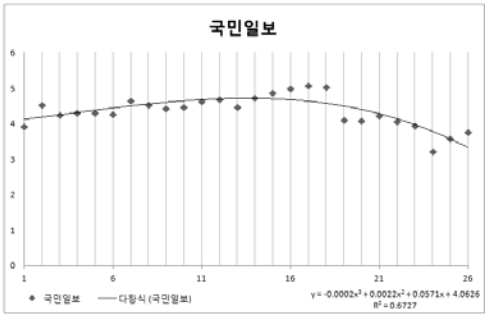


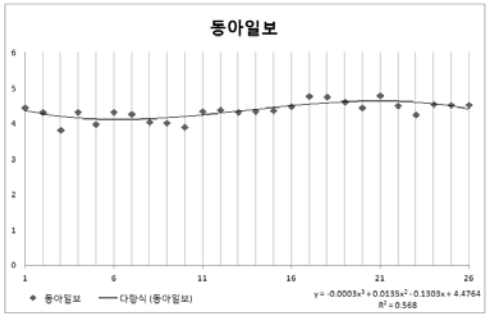


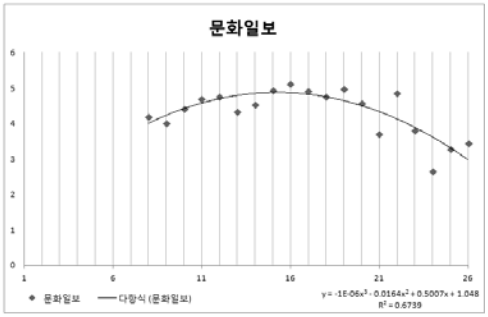


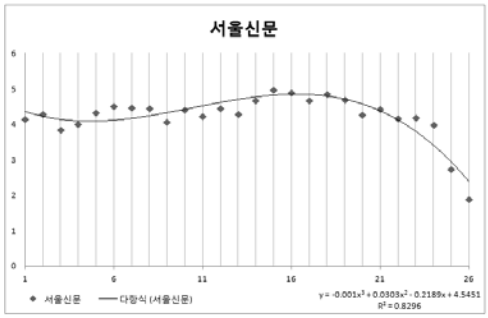
부록 II-4. 사회면의 매체별 기사당 주제 수

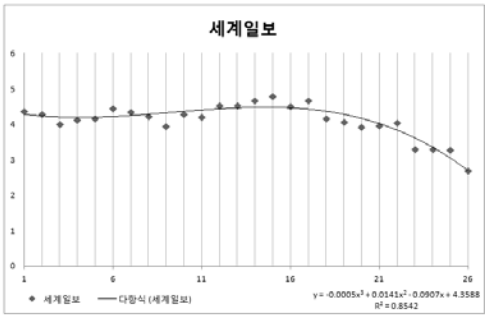


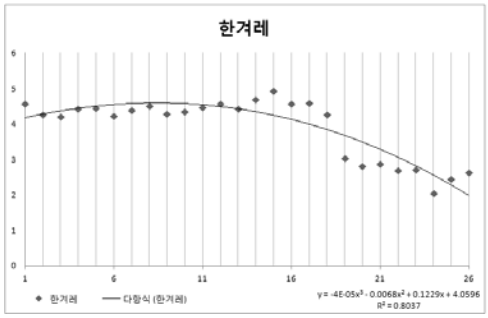


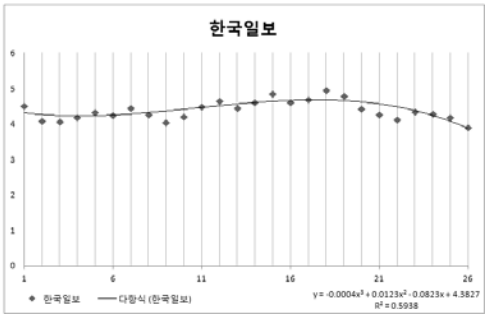












부 록 III-1. 정치면에서 가장 중요한 정보원

	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉
1990	**김태식 대변인**	김대중 총재	정대철 의원	null	김동영 총무	권노갑 의원	김충조 의원, 이수인 의원	박실 의원
1991	김대중 총재	김대중 총재	**박희태 대변인**	null	*노태우 대통령*	김대중 총재	**박상천 대변인**	**박상천 대변인**
1992	김대중 대표	김대중 대표	민자당	null	**박희태 대변인**	정해남 후보	김대중 대표	황인성 후보
1993	**박지원 대변인**	민주당	**박지원 대변인**	null	*김영삼 대통령*	이기택 대표	*김영삼 대통령*	황인성 후보
1994	**박지원 대변인**	문정수 총장	**박지원 대변인**	null	**박지원 대변인**	**박지원 대변인**	민주당	민주당
1995	**박지원 대변인**	**박지원 대변인**	**박지원 대변인**	null	**박지원 대변인**	**박지원 대변인**	**박지원 대변인**	민주당 의원
1996	국민회의 김대중 총재	국민회의	강삼재 사무총장	null	강삼재 사무총장	국민회의 김대중 총재	강삼재 사무총장	국민회의
1997	**정동영 대변인**	**정동영 대변인**	**정동영 대변인**	**정동영 대변인**	*김영삼 대통령*	**정동영 대변인**	**정동영 대변인**	**정동영 대변인**
1998	한나라당	**김철 대변인**	한나라당	한나라당	**맹형규 대변인**	한나라당	*김대중 대통령*	한나라당
1999	이회창 총재	이회창 총재	*김대중 대통령*	*김대중 대통령*	*김대중 대통령*	**안택수 대변인**	*김대중 대통령*	*김대중 대통령*
2000	한나라당	*김대중 대통령*	*김대중 대통령*	한나라당	*김대중 대통령*	한나라당 이회창 총재	*김대중 대통령*	민주당
2001	민주당	**장광근 수석부대변인**	한나라당	한나라당	한나라당	민주당	**권철현 대변인**	한나라당
2002	한나라당	**남경필 대변인**	**한나라당 남경필 대변인**	**남경필 대변인**	한나라당	**남경필 대변인**	**남경필 대변인**	민주당
2003	정부 당국자	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*
2004	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*
2005	*노무현 대통령, 청와대*	*노무현 대통령*	청와대	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*
2006	*노무현 대통령*	**우상호 대변인**	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*	*노무현 대통령*
2007	**천호선 청와대 대변인**	강재섭 대표	*노무현 대통령*	청와대	*노무현 대통령*	**나경원 대변인**	**천호선 청와대 대변인**	*노무현 대통령*
2008	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*

부록 III-1. 계속

	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉
2009	*이명박 대통령*	청와대	*이명박 대통령*	청와대	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*
2010	*이명박 대통령*	박지원 원내대표	민주당	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*	*이명박 대통령*
2011	*이명박 대통령*	청와대	*이명박 대통령*	김진표 민주당 원내대표	*이명박 대통령*	민주당	청와대	*이명박 대통령*
2012	*이명박 대통령*	**박용진 대변인**[a]	새누리당	명지대 교수	새누리당	민주당	새누리당[a]	새누리당
2013	*박근혜 대통령*	*박근혜 대통령*	청와대	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*
2014	*박근혜 대통령*	*박근혜 대통령*	새누리당	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*
2015	*박근혜 대통령*	새누리당 김무성 대표	*박근혜 대통령*	김무성 새누리당 대표	*박근혜 대통령*	*박근혜 대통령*	*박근혜 대통령*	김무성 새누리당 대표

주 1: '대변인'은 볼드체, '대통령'은 이탤릭체, 'null'은 밑줄.
주 2: 'null'은 데이터가 없는 경우, 'null*'은 중요도가 10 이하로 유의미하지 않은 경우.
[a] NLP 오류로 '한다'와 같은 불용어가 1위로 나왔을 때 차상위 단어를 표시한 경우.

부록 III-2. 정치면에서 가장 중요한 주제

	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉
1990	*민자당*[a)]	*민자당*	*민자당*	null	소련	소련	*민자당*	소련
1991	소련[a)]	*민자당*[a)]	**미국**	null	소련	한국[a)]	*민자당*	소련
1992	*민자당*	*민자당*	*민자당*	null	*민자당*	*민자당*	*민자당*	*민자당*
1993	김영삼	김영삼	*민자당*[a)]	null	**미국**	일본[a)]	**미국**[a)]	*민자당*[a)]
1994	**미국**	**미국**	**미국**	null	**미국**	**미국**[a)]	**미국**[a)]	**미국**
1995	*민자당*[a)]	*민자당*[a)]	*민자당*[a)]	null	한국[a)]	*민자당*[a)]	*민자당*[a)]	*민자당*[a)]
1996	*신한국당*	*신한국당*	*신한국당*	자민련	*신한국당*[a)]	*신한국당*[a)]	*신한국당*	*신한국당*[a)]
1997	한보[a)]	한보[a)]	한보[a)]	한보[a)]	한보[a)]	*신한국당*[a)]	한보[a)]	한보[a)]
1998	*한나라당*	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]
1999	내각제[a)]	한국[a)]	**미국**[a)]	**미국**[a)]	한국[a)]	내각제[a)]	*한나라당*[a)]	내각제[a)]
2000	민주당[a)]	*한나라당*[a)]	**미국**[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	민주당[a)]
2001	**미국**[a)]	**미국**[a)]	**미국**	**미국**[a)]	**미국**[a)]	**미국**[a)]	**미국**[a)]	**미국**[a)]
2002	**미국**[a)]	민주당[a)]	**미국**	민주당[a)]	민주당	민주당[a)]	민주당[a)]	**미국**[a)]
2003	**미국**	**미국**[a)]	**미국**	**미국**[a)]	청와대	**미국**[a)]	**미국**[a)]	민주당[a)]
2004	**미국**[a)]	*한나라당*[a)]	**미국**[a)]	*한나라당*[a)]	*한나라당*	*한나라당*[a)]	*한나라당*[a)]	**미국**[a)]
2005	**미국**[a)]	일본[a)]	**미국**	**미국**[a)]	**미국**	**미국**[a)]	**미국**[a)]	**미국**[a)]
2006	**미국**	**미국**[a)]	**미국**	**미국**[a)]	*한나라당*	**미국**[a)]	**미국**[a)]	**미국**
2007	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]	*한나라당*	*한나라당*[a)]	*한나라당*[a)]	*한나라당*[a)]
2008	*한나라당*[a)]	**미국**[a)]	**미국**[a)]	*한나라당*[a)]	청와대	**미국**[a)]	**미국**[a)]	**미국**[a)]
2009	**미국**[a)]	**미국**[a)]	**미국**[a)]	민주당[a)]	**미국**	**미국**[a)]	**미국**[a)]	**미국**[a)]
2010	*한나라당*[a)]	중국[a)]	중국[a)]	중국[a)]	중국	*한나라당*[a)]	*한나라당*[a)]	민주당[a)]
2011	*한나라당*[a)]	*한나라당*[a)]	민주당[a)]	민주당[a)]	*한나라당*	**미국**[a)]	민주당[a)]	민주당[a)]
2012	*새누리당*[a)]	민주당[a)]	민주당[a)]	민주당	민주당[a)]	민주당[a)]	*새누리당*[a)]	*새누리당*[a)]
2013	국정원[a)]	민주당[a)]	**미국**[a)]	민주당[a)]	민주당[a)]	민주당[a)]	국정원	국정원[a)]
2014	청와대[a)]	*새누리당*[a)]	일본[a)]	**미국**[a)]	*새누리당*[a)]	청와대[a)]	청와대[a)]	일본[a)]
2015	*새누리당*[a)]	한국[a)]	한국[a)]	**미국**[a)]	청와대[a)]	청와대[a)]	청와대[a)]	**미국**[a)]

주 1: '미국'은 볼드체, '새누리당' 등 보수 정당은 이탤릭체, 'null'은 밑줄.
주 2: 'null'은 데이터가 없는 경우.
[a)] NLP 오류로 '한다'와 같은 불용어가 1위로 나왔을 때 차상위 단어를 표시한 경우.

부 록 III-3. 사회면에서 가장 중요한 정보원

	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉
1990	최병렬 공보처 장관	null*	김우석 의원	null	null*	서청원 의원	유준상 의원	야당 의원
1991	박석무 의원	김광일 의원	박희태 대변인	null	*윤형섭 교육부 장관*	*교육부*	박석무 의원	null*
1992	null*	대한 감사	null*	null	null*	null*	이해찬 의원	대한 감사
1993	null*	청와대	김영삼 대통령	null	이인제 노동부 장관	변정일 대변인, 이기택 대표	원혜영 의원	*교육부*
1994	민주당	null*	null*	null	박지원 대변인	박지원 대변인, 이기택 대표	장석화 의원	null*
1995	박지원 대변인	박지원 대변인	박지원 대변인	null	이부영 의원	박지원 대변인	김영진 의원	민주의원
1996	국민회의 정동영 의원	국민회의 이기문 의원	제정구 의원	null*	이양우 변호사	국민회의 이해찬 의원	국민회의	국민회의
1997	정동영 대변인	국민회의 이상수 의원	국민회의 조순형 의원	국민회의	조순형 의원	국민회의 조순형 의원	국민회의	국민회의
1998	*교육부*	김재천 의원	여당 의원	한나라당	null*	한나라당	김대중 대통령	한나라당
1999	야당 의원	김대중 대통령	이회창 총재	이부영 총무	이부영 총무	안택수 대변인	민주노총	한나라당
2000	*교육부*	김규한 교수	금감원	권철현 대변인	*교육부*	null*	민주노총	*교육부*
2001	*교육부*	장광근 수석 부대변인	*교육부*	한나라당	*교육부*	보건복지부	*교육부*	*교육부*
2002	노무현 대통령	남경필 대변인	*교육부*	한나라당	*교육부*	자민련 송광호 의원	*교육부*	*교육부*
2003	*교육부*	노무현 대통령	전교조	노무현 대통령	노무현 대통령	노무현 대통령	노무현 대통령	노무현 대통령
2004	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*
2005	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	노무현 대통령	*교육부*	*교육부*
2006	*교육부*	*교육부*	*교육부*	*교육부*	환경부	*교육부*	*교육부*	*교육부*
2007	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*	*교육부*
2008	이명박 대통령	*교과부*	*교과부*	청와대	이명박 대통령	이명박 대통령	*교과부*	*교과부*
2009	이명박 대통령	*교과부*	노동부	*교과부*	행안부	이명박 대통령 a)	노동부	*교과부*
2010	*교과부*	*교과부*	*교과부*	청와대	*교과부*	*교과부*	*교과부*	*교과부*
2011	*교과부*	복지부	*교과부*	서울대 교수	*교과부*	*교과부*	*교과부*	*교과부*

부록 III-3. 계속

	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉
2012	*교과부*	*교과부*	*교과부*	*교과부*	*교과부*	*교과부*	*교과부*	*교과부*
2013	*교육부*	복지부	**청와대**	*교육부*	*교육부*	*교육부*	박근혜 대통령	박근혜 대통령
2014	*교육부*	복지부	*교육부*	박근혜 대통령	*교육부*	박근혜 대통령	*교육부*	*교육부*
2015	*교육부*	*교육부*	박근혜 대통령	null*	*교육부*	대법원	대법원	*교육부*

주 1: 정당 및 정당인 등 정치인은 볼드체, '교육부'(교과부 포함)는 이탤릭체, null은 밑줄.
주 2: 'null'은 데이터가 없는 경우, 'null*'은 중요도가 10 이하로 유의미하지 않은 경우.
[a)] NLP 오류로 '한다'와 같은 불용어가 1위로 나왔을 때 차상위 단어를 표시한 경우.

부 록 III-4. 사회면에서 가장 중요한 주제

	〈경향〉	〈국민〉	〈동아〉	〈문화〉	〈서울〉	〈세계〉	〈한겨레〉	〈한국〉
1990	부동산	KBS[a], 근로자	부동산	null	피고인	근로자	**노동자**	부동산
1991	서울	서울	수서[a]	null	서울	서울[a]	**노동자**	피고인
1992	피고인	서울	근로자	null	서울	피고인	전교조	피고인[a]
1993	피고인	전교조[a]	피고인[a]	null	우리나라[a]	피고인[a]	**노동자**	교육부
1994	서울	*공무원*[a]	서울	null	우리나라[a]	우리나라[a]	*공무원*[a]	서울
1995	수사[a]	수사[a]	수사[a]	null	노씨[a], 특별법	노씨[a]	노씨[a]	노씨
1996	피고인	피고인[a]	피고인[a]	피고인	피고인[a]	피고인[a]	*공무원*	피고인
1997	한보	한보[a]	한보	한보	한보	한보[a]	한보	한보
1998	가능성	한나라당	구조조정[a]	구조조정	*공무원*	안기부[a]	구조조정[a]	가능성[a]
1999	구조조정[a]	수사[a]	수사[a]	서울[a], 수사	*공무원*[a]	수사[a]	구조조정[a]	한나라당[a]
2000	구조조정[a]	구조조정[a]	구조조정[a]	서울[a]	*공무원*[a]	가능성[a]	**노동자**[a]	서울[a]
2001	서울[a]	이씨[a]	언론사[a]	언론사[a]	*공무원*[a]	언론사[a]	언론사[a]	서울[a]
2002	*공무원*[a]	서울[a]	서울[a]	*공무원*[a]	위원회	청와대	**노동자**	가능성
2003	가능성[a]	서울[a]	전교조[a]	수사[a]	*공무원*	수사[a]	**노동자**[a]	가능성[a]
2004	가능성[a]	서울[a]	한국[a]	*공무원*[a]	*공무원*	어떤	*공무원*[a]	*공무원*[a]
2005	서울[a]	가능성[a]	서울대[a]	가능성[a]	*공무원*	가능성[a]	**노동자**[a]	서울대[a]
2006	교육부[a]	가능성[a]	서울[a], 한국	청와대[a]	*공무원*	*공무원*[a]	**노동자**[a]	외환은행[a]
2007	교육부[a]	교육부[a]	*공무원*[a]	교육부[a]	변호사	가능성[a]	**노동자**[a]	김씨[a]
2008	청와대[a]	전문의[a]	한국[a]	가능성[a]	*공무원*	*공무원*[a]	**노동자**[a]	청와대[a]
2009	*공무원*[a]	가능성[a]	가능성[a]	가능성[a]	*공무원*	*공무원*[a]	**노동자**[a]	가능성[a]
2010	청와대[a]	환자들[a]	학부모[a]	*공무원*[a]	*공무원*	가능성[a]	*공무원*[a]	가능성[a]
2011	가능성[a]	환자들[a]	가능성[a]	가능성[a]	가능성	더욱	종편[a]	가능성[a]
2012	가능성[a]	가능성[a]	가능성[a]	가능성[a]	*공무원*	피해자[a]	**노동자**[a]	가능성[a]
2013	국정원[a]	환자들[a]	가능성[a], 한국	가능성[a]	*공무원*[a]	피해자[a]	국정원	국정원[a]
2014	청와대[a]	환자들[a]	교육부[a], 자사고	가능성[a]	가능성[a]	매우[a]	청와대[a]	가능성[a]
2015	메르스[a]	메르스[a]	일자리[a]	메르스[a]	메르스[a]	메르스	**노동자**[a]	메르스

주 1: '노동자'는 볼드체, '공무원'은 이탤릭체, 'null'은 밑줄.
주 2: 'null'은 데이터가 없는 경우.
[a] NLP 오류로 '한다'와 같은 불용어가 1위로 나왔을 때 차상위 단어를 표시한 경우.

Automated Time Series Content Analysis with News Big Data Analytics:

Analyzing Sources and Quotes in One Million News Articles for 26 Years

Daemin Park
Senior Researcher, Korea Press Foundation

Time series content analysis in communication studies such as agenda setting theory is increasingly popular. There have been methodological advances in time series analysis. However, it is impossible to do content analysis for a large number of news articles with traditional manual techniques. This study suggests news big data analytics for automated time series content analysis in a long term, mixing natural language processing (NLP) and semantic network analysis of news. A pilot study focusing on news sources and quotes' topics is also conducted analyzing news about political or social issues. Around one million news articles for 26 years (1990～2015) are collected from 8 major nationwide Korean dailies including *Kyunghyang Shinmun*, *Kukmin Ilbo*, *Donga Ilbo*, *Munhwa Ilbo*, *Seoul Shinmun*, *Segye Ilbo*, *Hankyoreh*, and *Hankook Ilbo*. NLP with 'BigKinds', a news big data analysis database developed by Korea Press Foundation, and semantic network analysis with independent development tools are used. Studies showed that less differences among newspapers and complete time series changes between 1990s and 2000s were found in the most important sources and topics except for topics in society section. The number of sources and topics per article has decreased in general. Sophisticated automated times series content analysis in a long term enables researchers to monitor the press system as a whole and to compare other time series data such as many economic indexes.

Keywords: automated times series content analysis, natural language processing of news, semantic network analysis of news, news big data analytics, BigKinds