뉴스 기사의 빅데이터 분석 방법으로서 뉴스정보원연결망분석

뉴스 기사의 빅데이터 분석 방법으로서 뉴스정보원연결망분석*

박 대 민**
(서울대학교 언론정보학과 박사수료)

뉴스 기사가 빅데이터가 되고 있다. 이는 저널리즘 연구에서 연구방법을 빅데이터 분석과 접목해야 하며, 빅데이터 분석에 의해 이론적 함의를 도출해야 할 필요성을 시사한다. 본 연구에서는 빅데이터화한 기사의 분석방법으로서 사회연결망분석을 활용한 뉴스정보원연결망분석(NSNA)을 제안하고 이를 프로그램으로 제작한 뒤 시행연구를 실시했다. 뉴스정보원연결망분석은 비정형데이터인 기사를 정보원과 인용문을 중심으로 정형화한 뒤, 연결정도중앙성 값에 따라 가중치를 부여하여, 주요 정보원과 주요 의제를 보여주며, 소속별로 서로 다른 주장들도 대비하여 보여줄 수 있다. 본 연구에서는 시행연구로 카인즈에서 20개 매체의 6개월 분량 기사를 크롤링한 뒤 '뉴타운'이라는 검색어로 나온 2,239개 기사를 분석했다. 그 결과 전통적인 정보원 연구에서 주목한 공식 정보원이 아닌, 시장의 경계에서 컨설턴트처럼 경제체계와 생활세계를 중개하는 정보원이 논쟁에 중요한 역할을 한다는 점을 알 수 있었다.

Keywords: 뉴스정보원연결망분석(NSNA), 빅데이터 분석, 사회연결망분석, 시각화, 정보원 편향성, 뉴타운

* 본 연구는 2012년 11월 사이버커뮤니케이션학회에서 발표된 논문 '컴퓨터 보조 담론분석으로서 정보원연결망 분석'을 대폭 수정한 것입니다. 프로토타입 프로그램 개발에 도움을 준 원인호 서울대학교 언론정보학과 석사과정과 김기남 아주대학교 미디어학과 석사과정생, 그리고 심사를 해주신 익명의 심사위원 세 분께 감사드립니다.

** heathe1@snu.ac.kr, heathe0@gmail.com

1. 문제제기

미디어 현상에서 다채널, 다매체, 매체 융합을 넘어서 이제 빅데이터(big data)가 화두가 되고 있다. 사실 이러한 변화는 다음과 같은 면에서 저널리즘 연구에 큰 도전이다. 첫째, 연구대상의 성격이 달라졌다. 둘째, 그에 따라 기존 연구방법의 한계에 부딪히고 있다. 셋째, 이에 따라 도출되는 이론의 성격도 달라져야 한다. 예컨대 과거에는 한국 언론계가 중앙일간지와 지상파 방송 중심으로 구조화됐다. 반면 현재는 뉴스 포털을 통해 제공되는 수많은 온라인매체의 비중을 무시할 수 없다. 설사 저널리즘 영역을 전통적인 범주로 한정하더라도, 2011년 현재 온·오프라인 매체 2,800여 개의 매체에 소속된 24,000여 명의 취재기자가 많게는 하루에 10건의 기사를 쏟아내고 있다(한국언론진흥재단, 2012). 또 이용자들은 포털을 통해서 뉴스 아카이브에 구축된 수억 건의 기사를 검색하고, 이를 카페나 블로그, 소셜네트워크서비스(social network service, SNS) 등을 통해 공유하고 복제하면서 이용한다.

이러한 상황에서 주요 일간지나 지상파 방송사만을 대상으로 많아야 수천 건의 기사 표본을 분석하는 전통적인 내용분석방법은 현재 한국 언론 현상의 일부만을 드러낼 뿐 전체적인 조망에서는 제한적일 수밖에 없다. 담론분석방법으로도 전형적인 몇몇 기사를 살펴보는 것만으로는 텍스트가 위치한 사회적 맥락이나 담론 속 논증 구조를 드러내기가 훨씬 더 어려울 수밖에 없다. 기존 연구방법의 타당성을 부정하는 것은 아니지만, 새로운 현상을 전체적으로 그려주고 이를 더 잘 설명할 새로운 이론을 도출하는데 있어 보다 최적화된 방법을 고민할 필요가 있는 것이다.

본 연구는 빅데이터가 이러한 변화된 저널리즘 환경을 포착하는 가장 중요한 개념 중 하나라고 주장한다. 즉 첫째, 기사는 빅데이터화했으며, 둘째, 때문에 기존의 방법론을 빅데이터 분석과 접목시키는 작업이 필요하며, 셋째, 빅데이터 방법에 기초하여 이론을 도출하는 작업이 저널리즘 연구에 필요하다고 본다.

이러한 문제의식 위에서 본 연구는 우선 빅데이터화한 자료를 처리하는 방법에 대해 모색하는데 초점을 두고 저널리즘에 관한 영역 지식(domain knowledge)을 기반으로 사회연결망분석(social network analysis, SNA)를 활용하여 빅데이터화한 기사를 축약(reduction)하고 시각화(visualization)하는 방법으로 NSNA(news source network analysis, NSNA)을 제안한다. NSNA는 ① 기사를 자연어 처리(natural language processing, NLP)하여 인용문 등을 추출하고, ② 개인실명, 집단, 익명 등 여러 유형의 정보원을 추출해서 이를 인용문과 매칭하며, ③ 공동 인용 여부에 따라 정보원을 연결망으로 시각화하고, ④

SNA를 통해 정보원에 가중치를 부여하고, ⑤ 정보원의 가중치를 바탕으로 대량의 기사를 중요도 순서로 검토하게 도와준다.

본 논문은 일차적으로 방법론 제안에 목적을 둔다. 때문에 깊이 있는 이론적 함의를 당장 도출하는 데는 무리가 있다. 다만 본 연구에서는 NSNA를 활용한 시행연구(pilot study)를 통해 기존 정보원 편향성 연구에 대한 제한점을 확인하도록 한다. 이를 위해 본 논문에서는 한국언론진흥재단의 뉴스 기사 아카이브 카인즈(kinds. or. kr) 데이터를 크롤링으로 수집한 뒤, '뉴타운'이라는 키워드로 검색한 기사를 대상으로 연결망 프로그램인 UCINET과 자체 제작한 NSNA 프로토타입 프로그램을 활용해 시행연구를 실시한다.

2. 기사 분석에서 정보원의 중요성과 정보원 편향성 연구의 제한점

NSNA는 정보원을 중심으로 전체 기사를 축약하여 시각화한다. 이런 방법이 타당하려면 우선 기사에서 정보원이 핵심적인 중요성을 가져야 한다. 여기서는 언론의 기사 작성 관행에서 정보원, 특히 개인실명정보원의 중요도를 이론적으로 살펴본다. 이어 NSNA가 방법론으로서 이론에 기여할 가능성을 탐색하기 위하여, 정보원 연구에서 가장 널리 수행되어 온 편향성 연구의 한계를 NSNA의 관점에서 지적하도록 한다.

1) 언론의 현실 구성에서 정보원의 중요성

뉴스는 현실 자체가 아니라 언론이 재구성한 현실을 기술한 것이다. 일찍이 리프만은 실제 현실과 머릿속 그림이 다르며 둘 사이를 언론이 매개한다고 지적했다(Lippmann, 1922/2013). 언론의 현실 매개에서 가장 중요한 방법으로 정보원 활용이 꼽혀왔다. 이 때 정보원은 뉴스에 등장하거나 인용된 사람, 또는 기사의 배경정보를 제보한 사람으로 언론인이 직접 목격하거나 인터뷰한 인물이다(Gans, 1979). 넓은 의미에서 정보원은 인물뿐만 아니라 문서 등 모든 형태의 정보 출처도 포함된다(임영호, 이현주, 2001; 한동섭, 유승현, 2008). 정리하면 정보원은 인물 정보원과 비인물 정보원으로, 인물 정보원은 고위 관리나 기업 임원 등 기득권 정보원과 시민 등 소수집단 정보원과 뉴스의 중심이 되는 쟁점 정보원과 그렇지 않은 부차적 정보원, 실명정보원과 익명정보원 등으로 구분할 수 있다(한동섭, 유승현, 2008; 한동섭, 임종수, 2002). 비인물 정보원은 인터뷰(전화, 대면),

공개연설, 기자회견, 연설 원고, 단체에서 발간 다양한 종류의 문서, 보도자료, 현재의 화제에 대한 기존 이야기, 통신사 뉴스, 기자 자신의 취재 메모 등이 있다(Bell, 1991). 인물 정보원은 많은 정보를 보도자료 등 비인물 정보원의 형태로 제공한다. 따라서 비인물 정보원은 궁극적으로는 인물 정보원이며 다만 직접 인용 여부에 의해 구분된다.

뉴스가 정보원의 입을 통해 말해지며 언론사는 정보원을 통해 원하는 바를 전달한다는 점은 오래 전부터 많은 연구자가 지적해왔다. 예컨대 시갈은 뉴스가 실제 발생한 일이 아니라 누군가가, 즉 정보원이 발생했다고 말해준 내용을 기술한 것이라고 지적했다(Sigal, 1973). 셧슨도 뉴스의 생산과정은 정보원에서 시작하며 저널리즘이 역사의 초고라면 정보원은 저널리즘의 초고라고 말했다(Schudson, 1978). 갠스 역시 뉴스는 정보원으로부터 독자에게 전달되는 정보로, 관료적이고 상업적인 조직에 고용된 전문가인 기자들이 정보원으로부터 입수한 것을 독자들에게 맞도록 요약하고 정제해서 변화시킨 것이라고 설명했다(Gans, 1979).

2) 정보원 편향성 연구의 제한점과 NSNA의 함의

언론의 정보원 연구 중 가장 널리 진행된 것이 정보원의 다양성과 편향성에 관한 연구이다. 미국에서 1970년대 초부터 진행된 이 연구들은 공통적으로 중앙지와 지방지, 신문・방송・잡지 등 매체 유형(Sigal, 1973; Brown et al., 1987; Gans, 1979; Soloski, 1989; Berkowitz, 1987)과 국가 안보 기사, 테러리즘 뉴스, 산업 뉴스 등 주제를 막론하고(Hallin et al., 1993; Atwater & Green, 1988; Chang, 1999), 정부 고위관계자나 기업 임원 등 공식적 정보원에 대한 언론의 의존도가 높다고 지적한다(Glasgow University Media Group, 1980; 김용진, 2004). 즉 기사에서 엘리트 정보원이 사건의 정의자가 되며(Manning, 2001), 이 때문에 보도가 권력에 대해 편향된다고 지적한다. 한국에서도 인력 부족과 지면 제한, 권위에 대한 의존과 함께 출입처에 대한 의존 등을 이유로 정보원 활용의 편향성이 나타났다(장호순, 2001).

그런데 대부분의 정보원 편향성 연구는 중요도를 출현 빈도로 측정하여 그에 따라 정부 고위 관료나 거물 정치인, 기업 고위 임원 등 기득권을 가진 정보원이 중요하다고 주장한다. 그러나 중요도는 빈도 이외에도 다양한 방식으로 평가될 수 있다. 예컨대 기사의 중요도는 1면에 배치했는지, 1면은 아니더라도 각 면의 톱기사인지, 기사 분량은 어떤지, 관련 기사나 후속 기사가 있는지, 몇 명의 기자를 기사 작성에 투입했는지, 얼마나 많은 사실이 포함됐는지, 해당 기사를 다룬 언론이 중요한 매체인지 등에 따라 다양

하게 평가될 수 있다.

특히 기사가 사회 문제를 다룬다는 점에서, 단순히 많은 기사에서 인용된 정보원보다 논쟁적인 기사에서 많이 다루어진 정보원이 더 중요할 수 있다. 예를 들어 할리우드 스타가 기자회견을 열어서 5개 매체에서 5개의 기사를 작성했다고 하자. 그리고 어떤 정치평론가가 논쟁이 되는 사건에 대해 5명의 정보원과 1개의 기사에서 동시에 인용됐다고 하자. 빈도로 보면 할리우드 스타는 5이고 정치평론가는 1로, 할리우드 스타가 더 중요한 것처럼 보인다. 그러나 실제로 5개 매체가 인용한 할리우드 스타의 발언은 대동소이할 것이다. 반면 기사에 동일한 내용이 반복 인용되는 경우는 없기 때문에, 정치평론가를 포함하여 논쟁적 기사에 함께 게재된 5명의 정보원은 하나의 사건에 대해 보다 다양한 목소리를 내주고 있다고 볼 수 있다. 이처럼 연예계 스타처럼 많이 인용되지만 기자간담회를 통해 혼자만 발언하는 정보원이나 대기업 회장처럼 많이 다루어졌지만 인용이 되지 않은 인물보다, 평론가처럼 다른 논객들과 함께 지면상에서 논쟁을 펼치는 정보원이 어떤 면에서는 더 중요할 수 있는 것이다.

또한 주요 논객 정보원과 함께 인용되면서 근거 제시 또는 반박을 하는 정보원의 중요도도 생각할 수 있다. 이밖에 비록 적게 인용되더라도, 관련 기사에서 어떤 주제의 색다른 측면에 대해 말하는 정보원의 중요도 역시 연구 목적에 따라 고려할 수 있다. 이렇게 정보원의 중요도를 다양하게 정의할 수 있다면, 빈도분석에만 의존하는 것은 설명 범위가 제한적이라고 볼 수 있다.

NSNA는 정보원의 중요도를 다양하게 체계적으로 측정할 수 있는 방법을 제공한다. SNA에서는 여러 중앙성(centrality) 지표에 의존하여 결점(node)의 중요성을 다양하게 평가한다(손동원, 2002). NSNA에서는 정보원이 결점이 되고, 공동 인용 여부에 따라 결점 간 연결선(edge)이 그려진다. 이렇게 하면 연결정도중앙성(degree centrality)는 함께 인용된 정보원 수를 의미한다. 연결정도중앙성이 높은 정보원은 많은 정보원이 인용된 기사에서 많이 인용된 정보원, 즉 더 논쟁적인 기사에서 더 많이 인용된 정보원이다. 이런 정보원으로는 공식 정보원 외에 평론가, 여론조사기관 담당자, 증권가 애널리스트 등을 들 수 있다. 한편 위세중앙성(prestige centrality)은 연결정도중앙성이 높은 정보원과 함께 인용될수록 중요한 정보원으로 파악된다. NSNA에서는 위세중앙성으로 연결정도중앙성이 높은 정보원을 뒷받침하거나 반박한 정보원을 찾을 수 있다. 정보원의 중요도를 연결망 지표로 파악하는 접근은 정보원의 중요도가 함께 인용된 정보원에 의해 영향을 받는다는 점을 함축한다.

3. 뉴스 기사의 빅데이터 분석 방법으로서 NSNA

SNA는 빅데이터를 시각화하고 가중치를 부여하는 방법으로 널리 활용된다. 여기서는 뉴스 기사를 빅데이터로 규정하고, 뉴스를 SNA를 통해 분석하는 NSNA에 대해 설명하도록 한다.

1) 뉴스 기사의 빅데이터 특성

(1) 빅데이터의 개념

빅데이터는 엄밀한 학술용어는 아니다. 학계 전체에 통용되는 정의도 아직 없다. 그럼에도 불구하고, 빅데이터에 대한 각계의 논의를 살펴보면 빅데이터의 특징을 크게 데이터 측면과 데이터 저장, 관리, 처리의 측면에서 제시하는 것으로 보인다.

예컨대 전형적인 데이터베이스 소프트웨어 도구로 포착하고, 저장하고, 관리하고, 분석할(capture, store, manage, analyze) 수 없는 크기의 데이터 집합이라는 정의는 데이터 측면에 주목한 것이다(MGI, 2011. 5., 1쪽). 데이터 측면에서 빅데이터는 흔히 3V라고 해서, 크기(volume), 속도(velocity), 다양성(variety), 측면에서 기존 자료와 구별되는 것으로 이해된다. 우선 빅데이터는 저장 용량으로 따지면 흔히 10테라바이트 이상으로 간주되지만 컴퓨터 성능 향상에 따라 기준은 계속 높아지고 있다. 기록물이나 처리량, 파일 수가 많은 경우도 빅데이터라고 할 수 있다. 또한 빅데이터는 출처, 유형 등도 다양하다. 예컨대 파일 형태도 로그 파일, 수치, 텍스트, 오디오, 비디오 등 정형자료(structured data), 반정형자료(semistructured data), 비정형자료(unstructured data)가 뒤섞여있다. 속도 측면에서는 생성주기나 유통주기가 짧고, 수집과 분석이 실시간(real time)으로 이뤄진다(Russom, 2011, 6~7쪽). 이러한 측면에서 빅데이터 분석(big data analytics)의 핵심 기술은 대규모 데이터를 인식 가능한 수준으로 정리해주고 유형(pattern)을 보여주는(visualize) 축약(reducing)이다(Manovich, 2011). 이상의 정의는 사회적으로 널리 통용되는 반면 다소 모호하다. 즉 용량이 얼마나 커야 하는지, 개체수가 얼마나 많아야 하는지 기준이 없다. 또 비정형데이터라고 해서 빅데이터는 아닌데다가, 오디오 자료나 비디오 자료의 분석은 어렵다. 속도 역시, 처리가 얼마나 빨라야 하는지 불분명하다.

데이터 관리의 특성을 강조한 정의는 주로 기술적 측면에 주목하기 때문에 빅데이터가 아닌 것과의 구분이 비교적 명확한 편이다. 여기서 빅데이터란 대용량 비정형 데이터

에 최적화된 다양한 데이터베이스 기술을 활용하여 데이터를 수집, 저장, 관리, 처리하는 자료를 의미한다(함유근, 채승병, 2012, 72~88쪽). 즉 빅데이터 기술을 적용하면 더 효율적으로 관리할 수 있는 데이터가 빅데이터라는 것이다. 널리 쓰이는 빅데이터 기술로는 오픈소스(open source) 빅데이터 시스템 프레임워크인 하둡(Hadoop), 하둡과 연결된 파일 시스템으로 소수의 고성능 서버 대신 저사양 개인용 컴퓨터 여러 대를 병렬로 연결하여 처리하는 시스템인 분산 파일 시스템(Distributed File System, DFS), 데이터베이스 형식으로는 비정형데이터를 수집 저장하는 NoSQL(Not Only Structured Query Language) 등이 있다. 전체 데이터를 작은 단위의 맵(map)으로 쪼개어 처리한 다음, 각각의 결과를 합하는(reduce) 방법인 맵리듀스(MapReduce)도 널리 쓰인다.

데이터 측면보다 데이터 관리 측면이 더 명쾌하기는 하지만, 전자의 중요성을 간과할 수는 없다. 특히 비정형 빅데이터는 더욱 그렇다. 텍스트의 예를 들면 전체 텍스트에서 목적에 따라 NLP 등을 해야 할 핵심 요소, 예컨대 인명 등을 지정하고, 그 결과로 나오는 방대한 NLP 결과물에 어떤 식으로 관계를 설정하고, 그 관계에 따라 어떤 식으로 가중치를 부여하고, 가중치에 따라 결과물을 어떻게 축약하고, 축약한 것을 어떻게 시각화할지를 고민하는 과정이 선행되어야 한다. 이 과정을 수행하는데 필요한 컴퓨터 연산력을 빅데이터 시스템으로 효율화하는 것은 필수적이지만, 데이터를 어떻게 전처리할지 미리 정해두지 않는다면 무의미할 것이다. 이러한 전처리 과정을 일괄적으로 수행해주는 기법 중에 대표적인 것이 SNA이다. SNA 자체가 빅데이터 기술은 아님에도 불구하고, 빅데이터 분석에서 널리 사용되는 이유도 이 때문이다.

빅데이터 분석은 여러 장점을 갖는다. 우선 빅데이터 분석은 알고리즘(algorithm)에 기초하여 컴퓨터로 분석하기 때문에 일단 시스템을 구축한 이후에는 시간과 인력, 즉 비용을 절감할 수 있다. 또 전집을 대상으로 분석하기 때문에 표본오차가 없다. 일부 결측값을 비교적 정확히 추정할 수도 있다. 휴먼 코딩을 하지 않으므로 코더 간의 신뢰도 문제도 없다. 대규모 전집 데이터를 활용하기 때문에 다양한 관계를 파악할 수 있으며, 그 결과 2종 오류(type II error)를 범할 가능성이 낮아진다. 알고리즘 자체가 잘못되어 체계적인 오류가 발생할 수도 있는데 이 경우 알고리즘만 수정하면 전체 데이터를 비교적 쉽고 정확하게 수정할 수 있다. 또 데이터 추동적(data-driven)이기 때문에 해석에서 연구자의 편견이 개입할 가능성이 줄어든다.

하지만 이는 이상적인 경우로 단점도 적지 않다. 우선 빅데이터 시스템을 개발하고 구축하는데 오히려 더 많은 비용이 들 수 있다. 무엇보다 프로그램 개발자에 의존해야 한다. 자료 성격이 달라져도 프로그램을 다시 만들어야 한다. 또 현실적으로 빅데이터를

직접 만드는 기관이 아니면 자료에 대한 접근성이 떨어져서 여전히 표집 과정이 필요하다(Boyd & Crawford, 2012). 저작권 문제나 개인정보보호 문제도 크다. 특히 이 문제는 빅데이터에 너무 많은 정보가 담긴 있는데다가, 빅데이터 분석으로 주민등록번호처럼 수집되지 않은 정보도 상당히 정확하게 추정하기 때문에 심각하다(Acquisti & Gross, 2009). 이로 인해 자료를 제한적으로 수집하게 될 경우의 편향(bias)에 유의해야 한다. 분석 결과가 데이터 특수성을 지나치게 반영하는 과적합(overfiting) 문제로 일반화가 어려울 수도 있다(Linoff & Berry, 2011). 분석 결과물이 너무 많아서 해석이 어려울 수 있으며, 결국 이론적 또는 직관적으로 이미 아는 결론만 내릴 수 있다. 끝으로 알고리즘 자체가 이미 이론에 기초해 만들어지기 때문에 빅데이터 분석이 완전히 자료 추동적이라고 볼 수 없다. 또 알고리즘이 다중적으로 해석될 수 있다는 점을 간과하여 잘못된 결론을 내릴 수 있다.

(2) 뉴스 기사의 빅데이터 특성

기사는 빅데이터화하고 있다. 우선 양적인 면에서 추정을 위해 카인즈에 축적된 기사 자료에서 시작해보자. 한국언론진흥재단에 따르면 카인즈에는 1990년 1월 1일부터 2013년 9월 30일까지 신문, 주간지, 방송, 인터넷신문 등 66개 매체에서 작성된 약 2900만 건의 기사가 축적되어 있다. 날짜, 매체, 면종, 장르 등 기사 관련 정보는 전통적인 SQL 방식으로 정형화되어 저장되어 있지만, 기사 본문은 비정형 형태 그대로 저장되어 있다. 구체적으로는 자료는 기사 수에 해당하는 28,507,321개의 xml 파일과 관련 사진에 해당하는 2,963,140개의 jpg, bmp, gif 파일 등 총 31,472,227개의 파일로 저장된다. 그 용량은 총 360GB로 빅데이터치고는 작다.

하지만 카인즈에 축적되어 있지 않은 국내 전 매체로 범위를 확대하고 기간을 더 늘리면 기사의 수와 크기는 크게 늘어난다. 2012년도 한국언론연감에 따르면 2011년 기준 매체 수는 2831개, 기자 수는 24,553명에 달한다(한국언론진흥재단, 2013, 124쪽). 카인즈에서 매체 당 평균 기사 수가 약 432,000개, 용량이 5.5GB이므로, 단순 추정하면 기사 수는 12억여 개, 전체 용량은 1.5TB를 훌쩍 넘을 수 있다.[1] 게다가 이는 텍스트와 사진만 기준으로 삼은 것이고 방송의 오디오와 비디오가 추가되면 용량은 급증한다. 더 나아가 기사는 텍스트, 사진, 오디오, 비디오 등 다양한 형태의 파일로 구성된다. 또 의

1) 상당수의 매체는 2000년대 이후 설립된 온라인 매체 등일 수 있지만, 이들 온라인 매체는 일반 매체보다 더 많은 기사를 생산해 내기도 한다. 실제로 한 온라인 매체는 기자 1명이 하루에 기사 10건씩 송고하도록 할당해놓고 있다.

미 있는 기사 분석을 위해서는 텍스트만 해도 문장 단위로 살펴봐야 한다. 즉 정확한 숫자는 알 수 없지만 기사 당 문장 수가 10건 안팎이라고 한다면, 분석할 수 있는 전체 기사의 문장 수만 따져도 전체 매체 기준으로 120억 건이 넘을 수도 있는 셈이다.

또한 기사는 비정형데이터로 자료의 수집과 관리 및 분석이 까다롭다. 먼저 수집 측면에서 뉴스 기사 아카이브에 축적된 자료를 제외하면, 기사 자료는 언론사별로 따로 저장되어 있다. 이는 언론사별로 기사를 자동으로 다운받는다면, 즉 크롤링한다면 각각 다른 방식으로 해야 한다는 것을 뜻한다. 게다가 주요 언론사는 자사 홈페이지에 크롤링 방지 기술을 적용하고 있어 이를 우회해야 한다. 이 경우 저작권 문제도 해결해야 한다.

관리 및 분석도 까다롭다. 텍스트만 생각할 경우에도 고도의 NLP 기술이 필요하다. 예컨대 본 논문이 제안하는 NSNA의 경우 정보원 이름, 소속, 직함과 문장의 식별 및 매칭 등의 문제가 발생한다.[2] 더 나아가 정보원에 대해 NSNA의 각종 연결망 지표들을 구하고, 이를 문장과 매칭해서 시각화하는 작업을 신속하게 하기 위해서는 빅데이터 기술이 활용되는 것이 불가피하다. 속도 측면에서 보면, 기사는 속보성이 중시되기 때문에 수시로 생산되고 소비된다. 물론 연구를 위해서는 속보성이 크게 중요하지는 않지만, 여론에 민감한 정책 기관이나 기업의 경우 빅데이터 기술을 활용할 필요가 있다.

사실 뉴스 기사는 다른 전통적인 미디어의 비정형데이터보다 빅데이터 분석 측면에서 몇 가지 유리한 점이 있다. 첫째, 책이나 그림 등은 아직은 대부분 일단은 아날로그 형태로 제작된 뒤에 별도로 디지털화되는 반면, 뉴스 기사의 텍스트나 사진, 영상은 처음부터 끝까지 디지털 형태로 생산되고 유통되는 것이 이제 더 일반적이 됐다. 물론 컴퓨터 게임처럼 처음부터 끝까지 디지털화된 형태로만 존재하는 데이터도 있지만 이는 전통적인 미디어 자료는 아니다. 또 신문의 최종 생산물은 종이 형태로도 존재하지만, 현재는 주요 매체마저 종이신문을 포기할 계획을 세울 정도로 디지털 형태가 보편화됐다. 카인즈처럼 여러 매체의 기사가 함께 저장된 뉴스 기사 아카이브도 존재한다. 이 경우 크롤링이 아니라 자료 제공 협력을 통해 기사 빅데이터를 일괄적으로 제공받을 수도 있다. 더 나아가 이미 다양한 자료를 부분적으로는 정형화해뒀기 때문에 분석이 그만큼 용이

2) 정보원의 이름과 문장을 매칭하는 경우만 예를 들면, 우선 기사 내에서 문장들을 서로 분리해 내고, 문장 가운데 인용문을 추려내고, 인용문에서 정보원의 이름과 소속과 직함을 찾아내야 한다. 이 때 정보원의 이름이 인칭대명사나 '성(姓) + 직함'으로 처리되어 있다면, 이 대명사나 성이 지칭하는 정보원의 이름을 다른 문장에서 찾아낼 수 있어야 한다. 이 때 같은 '성'(姓)을 가졌지만 이름이 다른 정보원이 둘 이상 있다면, 한 인용문에 나온 '성'(姓)을 어떤 이름과 매칭시킬지를 파악해야 한다. 게다가 기사 작성 관행이 매체는 물론 주제에 따라서도 조금씩 다르므로 기사 작성 관행에 대한 폭넓은 영역 지식을 활용한 휴리스틱(heuristic)을 활용해야 한다.

〈표 1〉 뉴스 기사에 빅데이터 분석 기술을 적용한 주요 서비스 사례

주요 기능	주요 사례
검색	네이버 뉴스 검색, 카인즈
가중치 부여	구글 뉴스, 뉴스메이트
요약	썸리(Summly), 뉴스썸머
클러스터링	네이버 뉴스클러스터링
선별	플립보드(Flipboard), 트위터 뉴스봇(Tweeter Newsbot)
오피니언 마이닝	다음소프트 소셜매트릭스
시각화	연합뉴스 인터랙티브
공유	페이스북(Facebook), 트위터(Twitter)
편집	위키트리

하다.

둘째, 기사는 많은 경우 중복되므로 축약 효과가 크다. 예컨대 대통령의 해외순방은 모든 언론사에게 공통으로 주어진 사실이다. 또 언론사들은 기자회견이나 보도자료 등으로 기사화할 내용을 공유한다. 때문에 한편으로는 대동소이한 내용의 기사들을 하나의 대표 기사로 대신할 수 있고 다른 한편으로는 여러 기사들을 중복되지 않은 사실 위주로 종합한 종합 기사를 만들 수도 있다. 즉 중복된 여러 기사를 축약한 대표 기사나 종합 기사를 만드는 것이 의미를 갖는다. 실제로 관련 기사를 묶어서 중요도 순으로 보여주는 구글 뉴스와 같은 뉴스 어그리게이션(news aggregation) 서비스 등은 다양한 측면에서 기사 축약 및 종합을 시도한 사례로 볼 수 있다.

셋째, 기사는 비정형데이터이지만 비교적 형식을 갖췄기 때문에 NLP에 용이하다. 즉 기사는 문법에도 맞을 뿐만 아니라, 제목(title), 전문(lead), 본문(body), 기자이름(byline) 등의 형식을 갖춘다. 날짜나 매체는 물론, 지면 종류로 주제도 파악할 수 있으며, 인용 방식 등 기사 작성 관행 등도 최소한 매체별로는 통일되어 있다. 이러한 영역 지식을 활용한다면 다른 비문 텍스트에 비해 NLP의 속도나 정확도도 높아질 수 있다.

이러한 이유로 기사에 대한 빅데이터 분석 기술은 업계에서는 이미 본격 활용되고 있다. 예컨대 포털사이트에서 여러 매체의 뉴스를 취합하는 뉴스 어그리게이션, 원하는 기사를 찾는 뉴스 검색, 기사에 대한 가중치 부여, 기사의 요약, 같은 분야 또는 의제의 기사끼리 묶어주는 클러스터링(clustering), 필요한 기사를 선별하는 뉴스 큐레이션(news curation), 오피니언 마이닝(opinion mining) 또는 감성분석(sentiment analysis) 등을 통한 가치 발굴, 기사의 시각화 등 빅데이터화된 기사에 최적화된 기술이 활용된다. 특히 SNS와 연계해서 빅데이터화된 기사의 가중치를 부여하거나 오피니언 마이닝을 하고,

SNS나 위키엔진 등을 활용하여 사용자 간의 기사 공유나 참여를 통해 추가로 사용자 생성 빅데이터를 만드는 서비스가 적지 않게 나와 있다.

2) 뉴스 기사의 빅데이터 분석 방법으로서 NSNA

(1) 뉴스 기사에 대한 컴퓨터 보조 분석 방법 현황

기사 분석에서 컴퓨터를 활용한 분석이 없지는 않다. 기사 분석에 제한된 방법은 아니지만, 양적 연구에서는 컴퓨터 이용 내용분석(computerized content analysis, CCA)이, 질적 연구에서의 컴퓨터 보조 질적 데이터 분석 소프트웨어(computer assisted qualitative data analysis software, CAQDAS)가 활용되어 왔다.

먼저 내용분석에서는 이미 컴퓨터가 널리 활용되어왔지만 주로 분석 단계에서 통계 프로그램을 이용하는데 한정됐다. 하지만 최근에는 텍스트 데이터의 수집과 코딩 단계에도 컴퓨터를 활용하는 CCA가 주목을 받고 있다. CCA는 말뭉치분석(corpus analysis), 형태소분석(morphological analysis), 텍스트 마이닝(text mining) 등 어휘분석(lexical analysis)이 주를 이룬다. 이 때 표본 추출 과정은 불필요할 수 있으며 기술적 분석이 중심이 된다(Rourke et. al., 2001).

국내 언론학계에서도 컴퓨터 이용 분석을 사용한 연구들은 점차 늘어나고 있는데 형태소 분석 또는 말뭉치 분석을 활용한 연구가 주를 이룬다. 예컨대 노무현 대통령 기자회견문의 수사학적 특징을 살펴본 연구(이귀혜 등, 2008), 신문기사 표절을 형태소 매칭 기법을 통해 살펴본 연구(강남준 등, 2008), 독립신문 논설의 저자 식별 연구(강남준 등, 2010), 대한매일신보의 국문 논설의 형용사와 부사의 활용빈도를 살펴본 연구(윤상길 등, 2011), 대한매일신보의 국문 논설에서 언론 개념의 변화를 공동 빈출 어휘를 통해 살펴본 연구(김영희 등, 2011) 등이 있다. 언론학계의 연구는 아니지만, 신문 기사의 표제와 본문에 쓰인 형태소 유형별 비율 살펴본 연구도 있다(송경화, 강범모, 2006). 하지만 이러한 연구는 단순한 빈도분석에 바탕을 둔다는 한계를 갖는다. 최근에는 진일보한 텍스트 마이닝 기법을 활용하여 관련 어휘를 묶어주는 시도도 눈에 띈다. 예컨대 감미아와 송민은 텍스트 마이닝 기법을 이용하여 특정 기간의 3개 매체 6개 분야 기사의 단어에 대해 단순빈도분석 외에도 군집분석(clustering), 분류분석(classification)을 실시하여 신문의 논조를 분석한다(감미아, 송민, 2012).

질적 연구에서도 컴퓨터를 활용하지 않는 것은 아니다. 워드프로세서 활용은 수작업으로 이뤄진 데이터 분류와 저장과 검색을 자동화한 선구적인 사례다(최희경, 2008,

128). 하지만 워드프로세서는 모든 문서 작업을 위한 범용 프로그램이다. 질적 데이터 관리에 적합한 본격적인 CAQDAS는 1980년대 이후에 등장하기 시작했다.[3]

국내 언론학계에서는 CAQDAS를 사용한 사례가 거의 없다. 이는 CAQDAS가 적지 않은 한계를 갖기 때문이다. 대표적인 CAQDAS 프로그램인 NVivo를 기준으로 살펴보자. NVivo는 세부 내용 파악에 유용하고, 체계적 데이터 관리에 유용하며 편향되거나 과장된 결과를 내놓을 가능성이 낮다는 장점을 갖는 반면, 여전히 코딩 작업에 너무 많은 시간과 노력이 투입되며, 그 결과 이론 개발에 소홀해질 수 있고, 속성 부여와 코딩과정에서 비일관성이 남아 있어서 양적 자료 등 다른 자료와의 비교를 어렵게 만든다(최희경, 2008). 이런 문제는 분석할 기사가 많아질수록 더욱 심각해진다. 이 때문에 질적 자료를 양적 자료와 연관시켜 분석하거나 대규모 조사연구를 할 때, CAQDAS를 활용하는 경우는 많지 않다(Bolden & Moscarola, 2000).

요컨대 CCA는 빅데이터에 활용할 수 있지만 심층적인 분석은 어렵다. 단순 빈도를 넘어서 어휘의 관계도 살펴보는 연구도 나왔지만, 어휘 분석만으로는 불충분하다. 기사는 텍스트이며 기본 단위는 문장이기 때문이다.[4] 따라서 전체 기사는 문장을 중심으로 구성되는 복합논증적 구조로서 파악될 수 있어야 한다. CAQDAS는 이러한 면에서는 강점을 갖지만, 문제는 대규모 데이터의 분석에 적합하지 않다.

(2) NSN의 개념

본 연구에서는 데이터 전처리 측면에서 NSNA를 활용한다. 앞서 살펴봤듯이, SNA는 빅데이터 기술 자체는 아니지만 빅데이터 분석 기법으로 널리 응용된다. 본 논문에서는 SNA를 기사 분석에 적용해본다. 빅데이터 분석 기법으로서 NSNA는 CCA와 CAQDAS의 한계를 극복할 대안이 될 수 있다. 단어 빈도에 주로 의존하는 CCA와 달리 NSNA는

---

3) CAQDAS는 크게 에스노그라프(Ethnograph)와 같은 코드 추출 프로그램(code-and-retrieve program), 하이퍼리서치(HyperRESEARCH)와 같은 규칙 기반 이론 개발 체계(rule-based theory building system), AQUAD와 같은 논리 기반 체계(logic-based systems), NUD*IST와 같은 색인 기반 접근(index based approach), ATALS/ti와 같은 개념적 네트워크 개발(conceptual network builder) 등으로 분류하기도 한다(Richards & Richards, 1994).

4) 본 논문에서는 기사를 텍스트언어학의 관점의 텍스트로 간주한다. 대체로 좁은 의미에서 텍스트는 응결성, 응집성, 의도성, 용인성, 정보성, 상황성, 상호텍스트성 등의 텍스트성을 충족시키는 의사소통 출현체로 정의된다(de Beaugrande & Dressler, 1981, Vater, 2001/2006, 39쪽에서 재인용). 광의의 의미로는 영상물이 함께 있는 텍스트도 텍스트에 포함된다. 기사는 협의의 텍스트 정의에도 부합하는 명백한 텍스트이다.

매체 · 기사 · 문장 · 의제 · 정보원 등 분석 기사의 다양한 요소 간에 관계를 설정하고 가중치를 부여할 수 있다. 또 시각화를 통해 정보원 간, 문장 간, 기사 간의 관계 및 그 구체적인 내용을 살펴보는 것도 손쉽게 가능해지므로, 기사의 복합논증 구조를 이해하는 데 도움을 줄 것으로 기대한다.

NSNA를 설명하기에 앞서 NSN의 개념을 먼저 살펴보자. NSN은 같은 기사에 두 정보원이 직접인용문으로 함께 인용됐을 경우 이 정보원들 간에 서로 의미론적인 관계가 있는 것으로 보고 간접적으로 만드는 준연결망(quasi network)을 뜻한다(박대민, 2011). 이 때 NSN은 관계의 방향이 양방향인, 즉 사실상 방향이 없는 쌍방향 연결망(un-directed graph)이다.

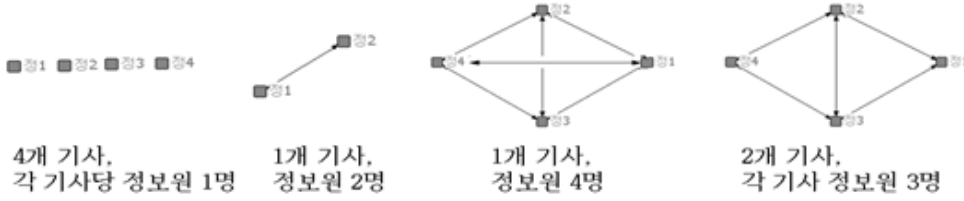
〈그림 1〉을 예로 들면, 우선 기사 1에서 정보원 A와 정보원 B, C는 같은 기사에 인용되었으므로, 설사 같은 입장은 아닐지 몰라도 서로 의미론적으로 관련된다. 다음으로 정보원 A는 동일 인물이므로 매개로 기사 1과 기사 2가 연결된다. 이러한 두 의미론적 관계에 따라 기사 1의 정보원 B, C와 기사 2의 정보원 D, E가 의미론적으로 연결된다.[5]

〈그림 1〉 뉴스정보원연결망 개념도

5) 언뜻 생각하면 정보원 A가 서로 다른 기사에 전혀 다른 주제에 대해 인용될 수 있을 것처럼 보인다. 이는 특히 정보원 A가 여러 기관에 소속된 경우에 그렇다. 하지만 동일인물이라고 할지라도 소속을 달리 인용하면 기자들은 이들을 사실상 다른 인물로 취급한다. 예컨대 어떤 정보원이 교수이자 기업인이라면 해당 정보원은 교수로 인용될 때는 대학 출입기자가 학계에 대해, 기업인으로 인용될 때는 산업부 기자가 업계에 대해 취재하는 것이다. 이런 관행을 반영해 소속이 다르게 인용된 동일 정보원은 NSN에서 필요할 경우 서로 다른 결점으로 나타낼 수 있다. 게다가 기사 1과 기사 2가 비슷한 기간에 게재된 기사라면, 기자들은 기사 1에서나 기사 2에서나 정보원 A에게 소속과 직함에 걸맞은 주제에 대해 이야기해줄 것으로 기대하며 취재한다. 예컨대 기자들은 언론학 교수에게는 언론에 대한 전문적인 견해를 기대하고 인용한다. 이러한 논리에 따라, 기사 1 · 2, 정보원 A · B · C · D · E 및 그들의 인용문은 서로 의미론적으로 관련된다.

〈표 2〉 뉴스정보원연결망 행렬 표

	기사1	기사2	기사3	기사4	…
정보원1	1	0	1	0	…
정보원2	1	0	0	0	…
정보원3	0	1	0	0	…
정보원4	0	0	1	0	…
…	…	…	…	…	…



〈그림 2〉 정보원연결망 그래프의 기본 형태

NSN은 기본적으로 결점과 연결선으로 구성된다. 결점은 정보원에 해당하며 연결선은 정보원 간 공동인용을 나타낸다. 즉 연결선으로 연결된 결점은 한 기사에서 공동 인용되었음을 뜻한다. NSN의 행렬표는 '정보원×기사'의 2원 자료(2 mode data)를 '정보원 × 정보원'의 1원 자료(1 mode data)로 변환해 만들어진다. 2원 자료의 행렬은 〈표 2〉와 같다. 이 행렬은 기사 1에서는 정보원 1과 정보원 2가 동시 인용됐으며 기사 2에서는 정보원 3만, 기사 3에서는 정보원 1과 정보원 4, 기사 4에서는 아무도 인용되지 않았음을 나타난다.

NSN의 기본적인 그래프 형태 예시는 〈그림 2〉[6]와 같다. 4명의 정보원이 모두 다른 기사에 인용됐다면 NSN은 4개의 결점이 연결되지 않은 채 도시된 그래프로 나타난다. 1개의 기사에 2명의 정보원이 인용됐다면 NSN은 2개의 결점이 하나의 선으로 연결된 형태가 된다. 4명의 정보원이 모두 같은 기사에 인용되었다면, 4개의 결점이 완전히 연결된 NSN으로 그려진다. 반면 2개의 기사에 3명의 정보원이 인용됐고 이중 2명이 겹친다면, 공동 인용되지 않은 정보원 간에는 연결이 없고, 공동 인용된 3명의 정보원은 선

6) 〈그림 2〉에서 화살표는 두 정보원이 서로 의미론적 관계를 주고받는다는 것을 의미한다. 흔히 연결망에서 단방향 화살표는 한 결점이 다른 결점으로 방향 있는 관계를 맺을 때(예컨대 편지를 주는 사람에게서 받는 사람에게로 화살표가 감), 양방향 화살표는 두 결점이 상호적인 관계를 맺을 때(예컨대 친구 관계) 사용된다.

〈표 3〉 정보원연결망의 분석대상, 구성단위, 의미단위

분석대상	구성단위	의미단위
검색어	전체 연결망	소재
기사뭉치	구성집단	의제
기사	파당	주제
인용문	결점	명제

으로 연결된 NSN이 제시된다. 이상에서 하나의 파당(clique)이 하나의 기사에 대응된다는 점을 알 수 있다.

한편 NSN에서 하나의 구성집단(component)은 하나 이상의 파당이 연결된 형태로 만들어진다. 예컨대 〈그림 2〉의 네 번째 NSN은 2개의 기사로 이뤄진 구성집단이다. 두 기사는 정보원 2와 정보원 3에 의해 의미론적으로 관련되어 하나의 어떤 의제(agenda)를 이룬다고 말할 수 있다. 예컨대 박근혜 대통령의 인용문이 담긴 두 기사는 어떤 하나의 의제를 구성한다. 박근혜 대통령이 한 기사에서는 정치를, 다른 기사에서는 경제를 다룰 수 있다. 그러나 두 기사는 우선 박근혜 대통령을 다룬다는 점에서 하나의 의제를 구성한다. 또 검색어와 분석기간을 통해 추가로 기사의 범위를 제한하면, 박근혜 대통령이 관련되어 있으면서 해당 검색어와 관련된 의제가 NSN의 구성집단 형태로 시각화될 수 있다. 예컨대 '경제'라는 검색어로 제한하면, 박근혜 대통령의 인용문이 포함됐더라도 '경제'라는 검색어가 없는 정치 관련 기사는 제외된다. 이 때 검색어는 하나의 소재(topic)로 볼 수 있다. 한편 NSN에서 하나의 구성집단은 각종 군집분석을 통해 보다 작은 구성집단으로 나눠질 수도 있다. 이는 하나의 큰 의제가 세부적인 의제들로 분화될 수 있다는 점을 시사한다. 더 나아가 전체 연결망은 서로 연결되지 않은 복수의 구성집단으로 이뤄질 수 있다. 이는 하나의 소재가 다양한 큰 의제들로 구성되어 있다는 점을 함축한다.

정리하면 NSN는 정보원이 인용문을 통해 제시하는 명제(proposition), 인용문과 그 밖의 문장들로 구성되는 한 기사가 제시하는 한 주제(theme), 여러 기사들로 이뤄진 기사뭉치로 대변되는 한 의제, 그리고 그러한 의제들로 구성되어 있고 연구에서는 검색어 입력을 통해 만들어지는 어떤 소재를 시각화하고 있다. 즉 인용문 또는 명제는 정보원으로, 기사 또는 주제는 파당으로, 기사뭉치 또는 의제는 구성집단으로, 검색어가 아우르는 기사 전체 또는 소재는 구성집단의 뭉치인 전체 연결망으로 시각화된다. NSNA는 NSN 분석을 통해 기사의 명제, 주제, 의제, 소재를 분석 대상으로 삼는다. 전체연결망과 주요구성집단의 예는 뒤의 연구결과 〈그림 4〉로 제시했다.

(3) NSNA의 개요

NSNA는 내용분석이나 담론분석에서 흔히 진행되는 기사의 수집, 분류, 가중치 부여, 의견 분류 등의 사전 작업을 NLP을 활용해 컴퓨터로 자동화할 수 있고, 이를 통계적인 표집이 아닌 SNA, 시각화 등을 통해 축약시켜준다.

NSNA에는 기자가 뉴스가치를 판단하고 기사를 작성하는 관행에 대한 영역 지식이 반영되어 있다. NSNA는 집단정보원이나 익명정보원에 적용하거나 수치나 장소 등을 포함한 문장, 또는 간접인용문 등의 분석에도 적용할 수 있지만, 기본적으로는 개인실명정보원과 그 직접인용문을 중심으로 기사를 분석하는데 적합하다. 이는 우선 한국 언론이 객관주의 저널리즘 관행을 따라 특히 개인 실명 직접인용문의 사용을 통해 사실성을 부여하는 관행을 반영한다. 기사에서 사실성을 부여하는 방법으로는 인용문, 수치, 사례 제시가 대표적이며, 이 가운데 인용문이 특히 중시된다(van Dijk, 1988; 송용회, 2005; 박재영, 2006; 남궁은정, 강태완, 2006). 한국 언론은 사설이나 칼럼과 같은 의견뉴스보다 사실뉴스를 훨씬 많이 생산하며, 기자나 언론사의 의견은 직접 드러내기보다는 수치나 인용문을 선택하는 등 게이트키핑 과정을 통해 간접적으로 제시된다.

또한 NSNA는 뉴스가치가 높은 기사일수록 더 많은 사실을 담는 관행, 특히 사회적으로 논쟁이 되는 사건일수록 각계의 더 많은 목소리를 담으려 하는 관행을 반영한다. 어떤 소재에 대해 더 많은 정보원으로부터 인용을 한다는 것은 더 많은 기자가, 더 많은 시간을 들여, 더 많은 사람을 만나고, 더 많은 기사를 쓴다는 것을 의미하는데 이는 언론이 해당 사건의 뉴스가치를 높게 봤기 때문이다. 또 NSNA는 사실이라도 단순히 보도자료를 기술한 기사를 평가절하한다. 사실 이런 기사는 가독성도 떨어지고, 기자들도 중시하지 않으며, 무엇보다 정보원 인용도 많지 않다. 덧붙여 기자들은 정치, 사회 영역을 경제, 문화 영역보다, 경제, 문화 중에서는 경제를 더 중시하는 경향이 있는데, 이 역시 인용문 활용 관행에 반영된다. 즉 정치, 사회 분야 기사의 정보원 활용이 가장 많고, 그 다음이 경제이며, 문화는 인용문 활용이 비교적 적다. NSNA에서도 이러한 저널리즘 관행이 반영돼 있다.

NSNA에서는 정보원의 중요도를 단순히 기사 등장 빈도만으로 평가하지 않고, 다른 정보원과 어떤 관계를 맺고 있는지에 따라 다양하게 평가한다. 특히 한 정보원의 연결정도중앙성은 해당 정보원이 얼마나 많은 정보원과 함께 인용되었는지를 의미한다. 즉 모든 기사가 아니라 여러 정보원이 각자의 의견을 제시하는 논쟁의 장이 되는 기사에서 등장해야 중요한 정보원인 것이다. 2장에서 살펴본 대로 빈도와 연결정도중앙성 모두 어떤 중요도를 나타내기는 하지만, 논쟁적인 기사를 추려내고 다양한 의견을 대조 검토하

는 담론분석에서는 빈도보다는 NSNA로 도출된 연결정도중앙성으로 가중치를 부여하는 편이 더 타당한 것이다.

특정한 저널리즘 관행이 잘 반영돼 있다는 점 때문에 오히려 NSNA을 수행할 때 유의점이 있을 수 있다. 우선 문화 분야 기사를 분석하거나, 사설, 리뷰, 미담, 인터뷰 기사 등을 분석할 때 한계가 있다. 경제 기사를 분석할 때는 인용문 외에 수치도 중시해야 한다. 너무 전문적이거나 뉴스가치가 낮거나 이데올로기적 편향 등의 이유로 기사화되지 않는 주제 또한 분석이 어려울 수 있다. 때로는 매체 특성도 고려해야 한다. 기사 길이를 보면 잡지가 가장 길고 신문이 그 다음, TV가 가장 짧다. 즉 중요도가 아니라 매체 특성 때문에 잡지의 직접인용문의 개수가 과도하게 많을 수 있다. 이때는 연구 목적에 따라 매체나 기간을 동일하게 통제하거나 기사 수 등의 가중치를 고려해야 한다. 분석 기간도 적절히 정해야 한다. 기간이 너무 짧으면 대상 기사가 너무 적어서 분석하기에는 정보원이 불충분하고 유의미한 관계가 발견되지 않을 수 있다. 반대로 기간이 너무 길면 정보원이 너무 많아져서 실제로는 중요하지 않은 정보원이나 유의미하지 않은 관계가 중시될 수 있다. 따라서 다른 방법론과 마찬가지로 NSNA를 수행할 때 역시 연구목적에 따라 주제, 장르, 매체, 기간을 적절히 설정해야 한다.

(4) NSNA의 절차

NSNA는 자료수집, 정형화, 분석 및 시각화 등 크게 3단계로 진행된다. 먼저 기사 자료를 수집한 뒤, 비정형자료인 기사를 NLP 등을 거쳐 정형자료로 변환한다. 이 때 날짜, 매체, 기사, 제목, 기사 본문, 개별 문장, 인용문 및 수치 문장, 정보원 이름・소속・직함, 소속 분류 등이 코딩된다. 다음으로 정보원을 중심으로 SNA를 활용하여 NSN을 그리고 연구 목적에 맞는 중앙성 값 등을 구한 다음 그 결과를 바탕으로 가중치에 따라 해석을 진행한다.

기사 자료는 뉴스 포털 사이트, 각 언론사 뉴스 사이트, 뉴스 기사 아카이브 등에서 크롤링을 통해 모을 수 있다. 각 기관과 직접 연락하여 자료 제공 협력을 받을 수도 있다. 기사 본문을 정형화하기 위해서는 컴퓨터를 활용한 NLP가 활용된다.[7] NLP 과정을 간략히 기술하면, ① 기사의 식별 및 매칭, ② 문장의 식별 및 매칭, ③ 정보원의 이름・소속・직함 식별 및 매칭[8] 등의 단계가 있다.[9] 또 프로토타입 프로그램에서는 구

7) 기사의 NLP 과정에는 공학적인 NLP 알고리즘 방식 외에도 기사 작성 관행과 관련된 영역 지식에 기초한 휴리스틱 방식도 중요하게 활용될 수 있다.

8) 각 인용문에서 정보원의 이름・소속・직함을 찾는데 대명사 등으로 대체되어 있을 경우 다른 문장

현하지 않았지만, 유사 기사나 유사 문장을 클러스터링해서 매칭할 수도 있다.

분석을 통해서는 기초적인 기술통계 값, NSN 그래프, 중앙성 값 등이 도출된다. 중앙성 값을 바탕으로 정보원에 가중치를 부여하고 이 가중치에 따라 정보원과 기사를 시각화할 수 있다. 기초적인 기술통계 값으로는 연구자가 입력한 분석 기간, 검색어, 검색매체 등과 전체정보원 수, 전체인용문 수, 전체기사 수, 전체의제 수, 주요정보원 수, 주요의제 수 등 기본적인 기술분석 결과가 된다. 이 때 주요정보원과 주요의제의 수는 결점 중 고립자(isolated)와 종속자(pendant)를 제외하여 파악한다. 고립자를 소거하면 1명의 정보원만 인용된 기사가, 종속자를 소거하만 2명의 정보원만 인용된 기사가 제외된다.

NSN 그래프는 전체 NSN(global NSN), 주요 NSN(main NSN), 자아중심 NSN(ego-centric NSN)으로 그려질 수 있다. 전체 NSN은 검색어로 선별된 기사의 전체 담론구조를 축약하여 시각화한다. 고립자와 종속자를 제거하여 주요 NSN으로 축약할 수도 있다. 전체 NSN의 구성집단은 전체 기사의 담론을 구성하는 세부 의제들을 보여준다. 가장 큰 구성집단은 가장 중심적인 의제이다. 다른 구성집단도 고립자와 종속자를 제거한 후 남아 있다면 주목할 만한 의제를 다룬다고 볼 수 있다. 특정 정보원(자아)의 자아중심연결망은 자아와 공동 인용된 타자 정보원을 연결망으로 보여준다. 연구목적에 따라 자아중심연결망도 활용할 수 있다.

가중치 부여는 SNA의 중앙성 값을 활용한다. SNA에서는 연결망 내에서 결점의 구조적 중요성을 연결정도중앙성 등 각종 중앙성으로 파악한다. NSNA에서도 이를 활용하여 중요도를 다양하게 정의할 수 있다. 이 가운데 연결정도중앙성은 가장 기본적이고 대표적이다. NSN에서 연결정도중앙성은 분석기사 전체에서 해당 정보원과 함께 인용된 정보원 수를 뜻한다. 연결정도중앙성이 높은 정보원은 다수의 정보원이 인용된 논쟁적인 기사에 인용되었을 뿐만 아니라, 여러 기사에 인용될 정도로 활발하게 논쟁에 참여하는 정보원이다. 또한 연결정도중앙성이 높은 정보원은 해당 검색어의 여러 의제를 다뤘을 가능성이 크다. 즉 해당 정보원은 다양한 분야에 의견을 제시해줄 만한 전문성을 갖추고 있다. 따라서 연결정도중앙성이 높은 정보원의 발언은 그렇지 않은 정보원보다 논쟁에서 중요성이 더 높다고 볼 수 있다. 중요한 정보원을 NSN에서 연결정도중앙성이 높은 정보원으로 정의할 경우, 정보원의 중요도를 계산하는 식은 아래와 같이 한 결점의

에 있는 정보원의 이름 · 소속 · 직함을 찾아 해당 인용문에 매칭한다.

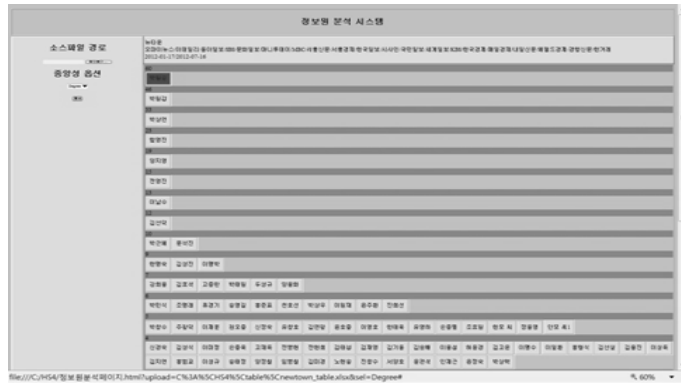
9) 카인즈와 같은 기사 아카이브의 자료를 이용하면 기사를 매체 별, 날짜 별로 구분해서 저장하는 작업은 이미 진행되어 있을 수 있다. 하지만 각 기사의 문장들을 각각 분리해 저장하고, 각 문장이 인용문인지 식별하고, 정보원의 식별 및 매칭 과정 등은 대부분 별도로 NLP를 수행해야 한다.

절대적 연결정도중앙성[10]을 구하는 공식에서 간단히 도출된다.

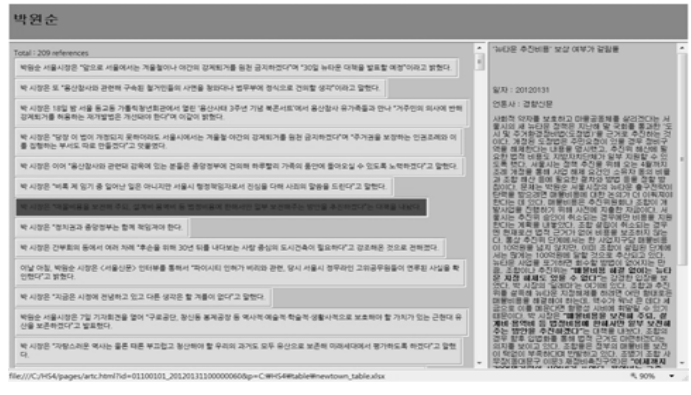
정보원의 중요도 = 여러 기사에서 함께 인용된 정보원 수

연결정도중앙성에 기초한 가중치에 따라 정보원을 순서대로 제시하고 각 정보원의 인용문을 보여줄 수 있다. 〈그림 3〉은 프로토타입 프로그램으로 구현한 시각화 예시이다. 검색어, 매체 명, 검색기간과 함께 연결정도중앙성 값이 큰 순서대로 정보원의 이름을 보여준다. 특정 정보원을 클릭하면 새 창에서 해당 정보원이 언급한 인용문들을 볼 수 있고, 다시 인용문을 클릭하면 인용문이 포함된 기사가 나타난다.

중요도 순으로 정보원 나열



선택한 정보원의 인용문과 기사 시각화



〈그림 3〉 정보원, 인용문, 기사의 시각화 예

10) 연결정도중앙성의 절대적 중심성은 한 결점에 직접 연결된 결점의 절대적 수로 측정된다. 상대적 중심성은 절대적 중심성을 전체 가능한 결점의 수로 나눈 값이다(손동원, 2002, 97~100쪽).

4. 시행연구

1) 연구문제 및 연구방법

여기서는 실제로 NSNA 프로토타입 프로그램을 활용해 기사를 시험적으로 분석해보도록 한다. 이를 통해 NSNA를 통해 분석할 수 있는 내용이 무엇인지 살펴본다. 본 논문에서는 이론적 함의를 도출하기 위한 본격적인 분석을 하지는 않지만, NSNA를 통해 앞에서 살펴본 정보원 편향성 연구의 한계를 부분적으로나마 확인해보도록 한다.

기사 수집은 크롤링 방식으로 할 수도 있고 자료를 가진 언론사나 기사 아카이브 구축 기관으로부터 이전받을 수도 있다. 본 시행연구를 위해서는 카인즈에서 2012년 1월 17일~2012년 7월 16일까지 6개월치 분량 20개 매체[11]의 619,328개 기사를 크롤링 방법으로 수집했다. 전체 기사를 대상으로 '뉴타운'을 검색어로 하여 '뉴타운'이라는 단어가 포함된 기사 수는 2,239개였다. 이 가운데 매체별 개인실명직접인용문 포함 기사 수는 경향신문이 87개, 국민일보 44개, 내일신문 8개, 동아일보 101개, 문화일보 87개, 서울신문 82개, 세계일보 33개, 한겨레신문 81개, 한국일보 62개, KBS 1개, MBC 28개, SBS 5개, 매일경제신문·매경이코노미·MBN 694개, 서울경제신문 162개, 한국경제신문 395개, 헤럴드경제신문 247개, 머니투데이 11개, 오마이뉴스 7개, 이데일리 30개, 시사인 6개 등으로 총 1844개였다.[12] 크롤링한 데이터는 html 형식으로 저장됐다. 연구자가 프로그램 명령어와 함께 검색어와 매체 기간 등을 입력하면 프로토타입 프로그램은 파일을 분석해 해당 검색어가 포함된 기사들의 날짜, 매체, 기사, 정보원 이름, 인용문, 소속 및 직함, 기사 원문 등으로 정형화해준다. 또한 초보적인 NLP을 수행하여 기사에서 정보원(이름+소속+직함 세트)과 인용문을 추출하고, 정보원과 인용문과 기사를 매칭했다. 이밖에 '정보원×기사'의 행렬을 구성하고 각 정보원의 연결정도중앙성을 계산하여 액셀 파일 형태로 결과물을 산출했다. 프로그램을 구동하여 얻은 결과물을 오류를 수정해 살펴본 결과,[13] 개인실명정보원 수는 321명(중복 제거)이었으며 총 인용

11) 카인즈에서는 주간경향 기사를 경향신문 기사와, 매경이코노미와 MBN의 기사를 매일경제신문 기사와 함께 분류해놓았다. 따라서 각각의 매체를 분리하면 총 23개 매체가 된다

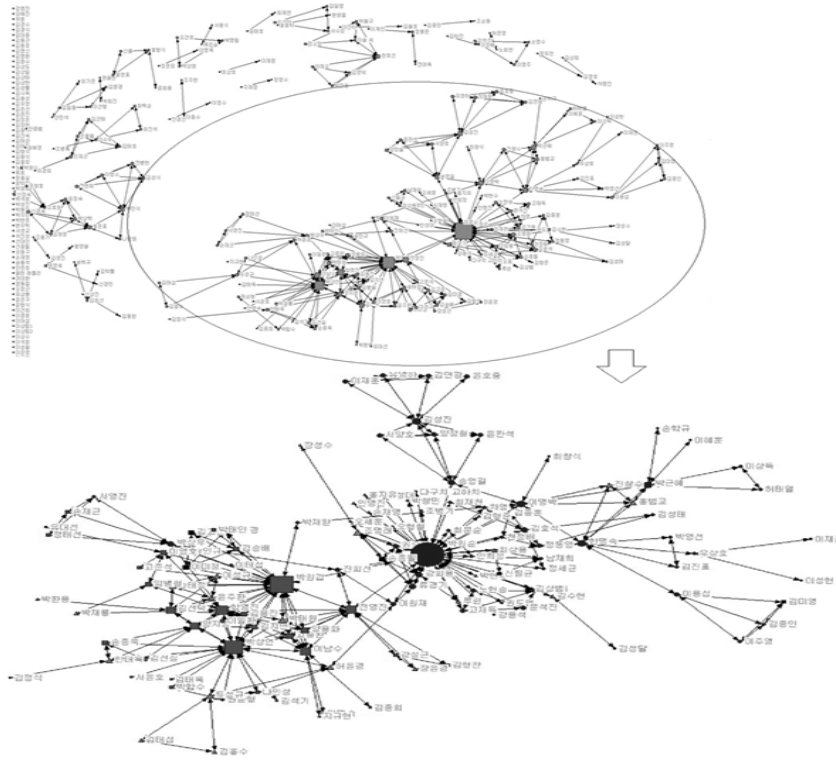
12) 머니투데이, 오마이뉴스 등은 '뉴타운'이라는 검색어가 포함된 기사는 더 많지만, 개인실명직접인용문을 활용한 기사는 크게 적다. KBS는 기사 형식의 차이로 파일럿 프로그램의 인용문 수집 성능이 떨어졌다.

13) 이름에 대한 NLP는 오류가 많은 편이다. 프로토타입 프로그램 역시 NLP가 완벽하지 않아 연구자가 직접 오류를 수정해야 한다. 현재 한국정보화진흥원 빅데이터 활용 스마트서비스 시범사업의

문 수는 1,100개였다. 또한 프로그램은 앞서 〈그림 3〉에서 본 것처럼 정보원을 연결정도중앙성에 따라 배열하고 각 정보원의 직접인용문을 제시해준다. 전체 NSN을 연결망 그래프로 보기 위한 시각화 프로그램으로는 UCINET 6에 포함된 Netdraw를 사용했다. 또한 정보원의 소속 코딩은 정치, 경제, 사회, 문화, 국제로 나누어 연구자가 수동으로 진행했다.

2) 연구결과

뉴타운 관련 기사에서 도출된 NSN의 전체 연결망과 주요구성집단(main component), 즉 주요주제연결망은 〈그림 4〉와 같이 시각화된다. 결점의 크기는 연결정도중앙성 값에 비례한다.



〈그림 4〉 뉴타운 기사의 뉴스정보원연결망: 전체연결망(위)와 주요구성집단

일환으로 차세대융합기술원, 서울대학교 산학협력단, 한국언론진흥재단 등이 협력하여 기능을 개선한 베타 버전을 개발 중으로 2013년 내로 카인즈를 통해 웹앱 형태로 공개할 예정이다.

〈표 4〉 연결정도중앙성에 따른 뉴타운 관련 정보원

연결정도	이름
60	박원순
46	박원갑
33	박상언
23	함영진
19	양지영
15	전영진
13	이남수
12	김선덕
10	박근혜, 문석진
9	한명숙, 이명박
8	김성진
7	강희용, 김효석, 고종완, 양용화, 박태원, 두성규
6	박민식, 천호선, 홍준표, 박상우, 이원재, 류경기, 송영길, 진희선, 윤주환, 조명래
5	정동영, 권오중, 이영호, 한태욱, 손종필
4	김미경, 김성식, 김연광, 서양호, 양정철, 유영하, 윤관석, 윤호중, 이명수, 이상득, 이용섭, 인재근, 전병헌, 전현희, 고재득, 김기동, 노현송, 이재훈, 손종욱, 김선실, 이일환, 김용진, 김지연, 임병철, 김승배, 이성규, 김태섭, 송태정, 진창수, 허윤경, 홍범교, 홍형식, 김고운, 김재영, 박상혁, 신경숙, 윤정숙, 이미정
3	김종인, 김종훈, 김태호, 민홍철, 서영진, 우상호, 이재오, 이주영, 정몽준, 천정배, 최재천, 차영, 남재희, 김근태, 강성근, 김형찬, 권순형, 나인성, 손재근, 이수우, 정태선, 최상용, 박재룡, 신필균, 김미영, 김영덕, 박진권, 박창수, 박철규, 신정숙, 유창호, 이솔, 정소망, 조효원, 장윤경
2	권영세, 김문경, 김성태, 김일웅, 김진표, 나경원, 박영선, 이계안, 정세균, 허태열, 홍영표, 권도엽, 박재완, 김일영, 안희정, 오형철, 최문순, 최성태, 최진석, 박태인, 송인규, 고준석, 김희선, 박합수, 서수정, 이선형, 최백순, 김수현, 박명림, 신율, 김흥수, 오건호, 윤희웅, 이명옥, 정준모, 루쉰, 박상훈, 김영주, 다구치 고하치, 손영수, 이영주
1	김철호, 강용석, 김동완, 김상희, 김영호, 김종민, 노회찬, 박희진, 손학규, 신경민, 신상진, 안종범, 원혜영, 유경희, 이성헌, 이혜훈, 전여옥, 정두언, 조순용, 차명진, 최재연, 허준영, 황영철, 김상범, 김정식, 김학진, 윤태호, 이재문, 주재영, 권철현, 오세훈, 최창식1, 박창민, 서용식, 예은실, 이정배, 김석기, 김종희, 김학렬, 서윤호, 송혁규, 이태섭, 진경선, 안명숙, 안민석, 최상호, 김태욱, 유대선, 이영수, 이진하, 조병기, 조병록, 이종수, 박환용, 김형준, 박현수, 변창흠, 손재명, 조주현, 지규현, 한성대, 홍종학, 장성수, 김영진, 인명진, 조용기, 정영수, 홍지유, 김성달, 이재근, 최기준
0	노영희, 박주민, 김명신, 김상도, 김상현, 문학진, 박래학, 박용진, 서영교, 선대인, 이상수, 이의엽, 이정희, 이준석, 정송학, 조윤선, 김경식, 김동호, 김진숙, 김황식, 박성진, 김동근, 김문수, 김성렬, 김용성, 김우영, 김종석, 김태균, 김형식, 류훈, 문충실, 성동구, 안준호, 임인구, 임철수, 차성수, 최창식, 추재엽, 이지송, 이지형, 김대중, 강영진, 곽훈, 김상태, 김시욱, 앨런 래플리, 이건희, 이상림2, 이정화, 임종구, 정연주, 한주희, 허명수, 김정현, 박경희, 석진성, 송용석, 이대섭, 조세헌, 김은선, 김은진, 김정은, 선종필, 여대환, 이춘우, 임성환, 장경철, 정병래, 박석훈, 변우택, 이석훈, 정종근, 한정호, 정영균, 이성철, 이은영, 임석재, 김경수, 박현호, 손재영, 오정근, 이경훈, 김상일, 전효찬, 정상협, 장성건, 윤형식, 김흥모, 이은홍, 최호철, 강혜진, 김명환, 김태진, 염동걸, 윤은구, 정미영, 주영철, 허먼 칸, 유상봉, 조희경, 이상림1

연결정도중앙성이 높은 순서, 즉 함께 인용된 정보원이 많은 순서로 정보원을 제시하면 〈표 4〉와 같다. 주요 정보원 10명을 순서대로 제시하면 박원순(서울시장), 박원갑(KB국민은행 부동산서비스사업단 수석부동산팀장), 박상언(유엔알컨설팅 대표), 함영진(부동산써브 실장), 양지영(리얼투데이 리서치자문팀장), 전영진(예스하우스 대표), 이남수(신한은행 부동산전략사업팀장), 김선덕(건설산업전략연구소장), 박근혜(새누리당 비상대책위원장), 문석진(서대문구청장) 등이다.

연결정도중앙성 6 이상인 정보원을 소속별로 파악하면 〈표 5〉와 같다. 한 기사에 6명의 정보원이 함께 인용되는 경우가 드문 점을 감안하면 이 정보원은 2개 이상 기사에 인용된 정보원으로 우선적으로 발언을 살펴볼 필요가 있다.

프로토타입 프로그램을 통해 각 정보원의 주요 발언을 쉽게 파악할 수 있다. 〈표 6〉은 박원순에 버금가는 주요 정보원인 박원갑의 주요 인용문만 제시한 것이다.

〈표 5〉 뉴타운 관련 소속별 주요 정보원

정당	박근혜, 한명숙, 김성진, 강희용, 박민식, 천호선, 홍준표
대통령/청와대	이명박
과학기술학계	윤주환,조명래
장관/행정부처	박상우,이원재
인문사회학계	박태원
지방자치단체	박원순, 송영길, 문석진, 류경기, 진희선
컨설팅/애널리스트	박원갑, 박상언, 함영진, 양지영, 전영진, 이남수, 김선덕, 고종완, 양용화
연구소	두성규

〈표 6〉 박원갑의 뉴타운 관련 주요 인용문

"서울시의 뉴타운 정비사업 신정책으로 뉴타운 옥석가리기가 빠르게 진행될 것", "주민들의 의사에 따라 사업방식을 결정하게 되면 상당 부분 조합을 해산시키거나 추진을 포기하는 곳이 많이 늘어날 것으로 예상된다"(문화일보, 2012. 1. 30).
"전체적으로 부동산 가격이 하향안정세를 보이고 있는 상황에서 뉴타운 지정이 많았던 강북 지역은 이번 뉴타운 출구전략으로 가격이 떨어질 개연성이 높다", "서울시가 도로와 인프라를 설치해도 시가 제기능을 할 수 있도록 충분히 지원해줘야 한다", "이런 후속작업이 없으면 뉴타운으로 예정됐다가 해제된 지역이 슬럼화할 소지가 있다"(동아일보, 2012. 2. 1.)
"지지부진한 재개발 등 뉴타운 사업을 더 이상 방치할 수 없다는 점에서 이번 조치는 시점 면에선 적절했다"(서울신문, 2012. 2. 1).
"뉴타운 사업은 겉으로 봐서는 공공성이 강하지만 깊이 들여다보면 이를 통해 수익을 내려는 재테크 사업", "민간성격이 강한 사업에 이와 관계없는 주민 세금을 지원하는 것은 이치에 맞지 않는다"(매일경제, 2012. 2. 4.)
"뉴타운 재검토는 전세난과 (직접적인) 관계가 없다", "뉴타운 해제가 (주택) 공급 부족으로 이어진다는 주장은 논리적으로 약하다", "뉴타운 개발 이후 오히려 (주택 수가 줄어) 인구밀도가 줄어든다는 연구결과도 나와 있지 않으냐"(서울신문, 2012. 2. 6).
"규제 해제로 팔고 싶었던 사람이 매물을 내놓을 수 있게 됐지만 뉴타운 해제가 거론되는 상황에서 매수를 감행하기는 어렵다", "매물이 늘어 지분가격이 더 내려갈 수 있다"(한국경제, 2012. 2. 17).

이와 같은 방식으로 인용문을 살펴보고 각 소속을 대표하는 주요정보원의 견해를 대조해보면, 우선 박원순 시장은 뉴타운 출구전략 및 관련 정책을 최초 제안한 뉴스메이커이다. 그는 기존의 대형 건설사와 주택 소유자 또는 부동산 투자자에게 이익이 돌아가는 뉴타운 정책을 장기 거주자나 세입자 등 거주자가 주도하는 방향으로 전환하고자 한다. 특히 그는 기존 뉴타운 정책을 사실상 지지한 정부를 비판하면서, 뉴타운 출구전략에 필요한 비용을 서울시는 물론 정부가 일부 부해야 한다고 주장한다. 다음으로 박원갑 팀장은 박 시장 다음으로 중요한 정보원이다. 그는 새로운 뉴타운 정책이 시장에 미치는 영향은 물론, 서울시 정책의 실효성을 다면적으로 평가한다. 이원재 국토해양부 주택정책관은 박 시장에 대해 전면 반대하는 입장으로 뉴타운 출구전략 비용을 정부 재정으로 지원하는 것에 대해 시장 논리를 들어 반박한다.

박근혜, 한명숙 등 정당인들은 총선 관련 기사에서 뉴타운 문제를 간략하게 언급한 수준에 불과하여 뉴타운 관련 기사에서 중요한 논객으로 등장하지는 않는 것으로 보인다. 반면 문석진 구청장은 서울시의 뉴타운 출구전략 발표 이후 뉴타운 정책 개발 테스크포스(TF)에 참여하면서 서울시, 정부와 주민이 모두 재원 부담을 해야 한다고 주장하는 등 현실적이고 구체적인 주장을 내놓고 있다. 손종필 서울풀뿌리시민사회단체네트워크 운영위원은 뉴타운 출구전략 및 그 후속 대책인 마을 만들기 사업에 대해 기대와 우려를 동시에 나타낸다.

뉴타운 기사에 대한 본격적인 담론분석과 이론적 함의 도출은 본 논문의 주된 목적은 아니다. 다만 앞서 검토한 정보원 편향성 연구와 관련하여 간략하게 살펴보자. 기존 편향성 연구는 언론사가 고위 관리나 대기업 임원 등 기득권층 정보원에 지나치게 의존한다고 지적한다. 그러나 사실 이러한 정보원은 기자회견 등을 통해 한꺼번에 많은 매체에 보도된다. 따라서 설사 인용의 빈도는 높아도 내용은 대동소이할 수밖에 없다. 게다가 대기업 회장 등 거물은 중요한 인물이지만 직접 인용되는 사례는 거의 없다. 반면 NSNA의 결과에 따르면, 최소한 뉴타운 분야에서는 기득권층에 가까운 공식적 정보원의 중요도에 의문을 제기할 수 있다. 주택시장과 생활세계 사이에 자리 잡은 컨설턴트와 같은 정보원들이 큰 역할을 하고 있는 것이다. NSNA에서 연결정도중앙성 값이 큰 컨설턴트들은 기득권층이라기보다는 부동산 업계에서 오랜 경험을 쌓아 독립한 소규모 연구소 대표나 금융기관 부장 등 중간관리자급에 불과하며, 이들이 가진 자본이나 권력, 학력 등도 기존 정보원 편향성 연구에서 지적한 공식정보원의 권위에 비할 바가 아니다. 이들은 아마도 투자설명회나 은행 창구 등을 통해 일반 투자자나 주민들과 자주 대면할 것이다. 언론은 주민의 의견을 직접 반영하는 것은 아닐지 모르지만, 시장의 경계에 자리 잡

은 이런 컨설턴트의 입을 통해 시민들의 의견을 정제된 형태로 대거 수렴하는 관행을 갖고 있는 것으로 보인다. 그렇다면 최소한 시장 영역 가운데 대중이 대거 관여하는 금융이나 부동산 분야에 대한 보도에서는 이들의 역할에 주목해야 할 필요가 있는 셈이다.[14]

5. 결론 및 제언

기사는 빅데이터가 됐다. 이는 저널리즘 연구에서 연구대상이 변화했고, 그에 따라 방법론과 이론이 변화해야 한다는 점을 시사한다. 저널리즘 연구자로서는 빅데이터화한 기사를 축약하는 기술이 중요하다. 본 연구에서는 빅데이터화한 기사를 축약하고 분석하는 방법으로서 SNA을 활용한 NSNA를 제안했다. 이를 바탕으로 프로토타입 프로그램을 제작했으며, 이 프로그램을 바탕으로 '뉴타운'을 검색어로 한 시행연구를 실시했다.

NSNA은 정보원과 그 인용문을 중심으로 기사를 분석한다. 그 절차는 자료수집, 정형화, 분석 및 시각화 등 크게 3단계로 나눠진다. 자료수집은 크롤링 방법을 쓸 수 있지만, 저작권 문제 등을 고려할 때 자료 보유 기관과 협력하여 데이터를 제공받는 것이 좋다. 본 연구에서는 크롤링 기법을 활용했다. 자료를 수집한 뒤에는 비정형자료인 기사를 NLP 등을 거쳐 정형자료로 변환한다. 이어 SNA를 활용하여 연결망과 중앙성 값 등을 도출하고 정보원을 클러스터링하고 인용문, 기사 등을 시각화한다. 끝으로 이를 바탕으로 주요 정보원, 그들의 주장, 주요 의제, 소속별 의견 대조 등을 해석한다.

본 논문에서 제안한 NSNA는 기본적으로 개인실명정보원과 그 인용문을 중심으로 다양한 방식으로 가중치를 부여해 축약한다. 이러한 방식은 국내 언론의 취재 및 기사 작성 관행을 반영한다. 하지만 인용문을 사실성 기제로 적극 활용하는 통신사나 중앙일간지의 정치 사회 관련 기사에 더 적합한 측면이 있다. 그럼에도 매체 특성과 지면 특성 등을 반영하여 수치나 익명정보원 등의 사용 수준을 조율한다면, NSNA는 다양한 매체와 주제의 분석에도 충분히 도움을 줄 수 있다.

---

14) NSNA를 활용해 다른 분야의 검색어를 분석한 결과에서도 흔히 알려진 기득권층 공식정보원보다는 중간 지식인으로 불릴만한 정보원의 중요도가 높게 나타났다. 예컨대 '총선'에서는 여론조사기관 담당자가, '금융위기'에 대해서는 각 증권사 애널리스트가 중시된다. 이는 취재관행을 생각해볼 때 불가피한 것으로 보인다. 즉 기자는 마감시간에 쫓기면서 수많은 기사를 작성하기 위해서 다양한 주제에 대해 언제든 말해줄 수 있는 정보원을 선호한다. 거물급 정보원은 취재에 성공하면 좋지만 말을 아끼기 때문에 인용하기 편한 정보원이 아니다. 이런 정보원으로는 평론가, 애널리스트, 컨설턴트, 여론조사전문가, 홍보 담당자, 대변인, 교수 등이 있다.

기사에서 중요도의 판단은 다양할 수 있다. 본 연구에서는 연결정도중앙성만을 따졌다. 그 결과 논쟁적인 기사에서 다양한 의견을 제시하는 인물과 그의 발언을 중요한 것으로 간주된다. 하지만 NSNA에서도 연결정도중앙성 외에 다양한 방식으로 의미 있는 결점들을 찾는 방법이 존재한다. 만일 이러한 방법을 연구에서 찾는 다른 정보로 해석할 수 있다면 빅데이터 분석을 더 광범위하게 활용할 수 있을 것이다. 단, 이 때 어떤 값이 다중적으로 해석 될 수 있다는 점에 주의해야 한다. 예컨대 인접중앙성이 높은 정보원은 종종 논쟁의 중심이 되는 정보원을 찾아주지만, 항상 그렇지만은 않다. 이는 연결정도중앙성이 높으면 다른 모든 중앙성 값도 높게 나오는 경향과 무관하지 않다. 때문에 연결정도중앙성 이외의 값을 분석에 활용하고자 한다면, 이러한 점을 어떻게 통제할지 고려해야 한다.

NSNA는 한글로 작성된 기사에만 적용했다. 하지만 다른 언어 기사에도 유사한 분석을 진행할 수 있을 것으로 본다. NLP를 제외하고 그래프를 그리고 가중치를 부여하는 방식만 놓고 보면, NSNA는 인용문을 중요한 사실성 기제로 활용하는 영미권 언론의 기사 분석에도 타당하다.

끝으로 무엇보다 질적 연구자가 빅데이터 분석을 적극 활용할 필요가 있다. 빅데이터 분석은 질적 연구자에게 시간과 인력 절감 측면에서 큰 도움이 될 수 있을 뿐만 아니라, 질적 연구자의 영역 지식이 빅데이터의 전치리 과정에 결정적인 기여를 할 수 있다. 질적 연구자는 프로그램 기획에 적극 참여하여 텍스트와 그 분석방법에 대한 깊이 있는 영역 지식을 제공하면서 휴리스틱 측면을 대폭 개선시켜준다. 사실 알고리즘만으로는 분석의 깊이가 부족하고, NLP의 정확도도 떨어지기 때문에 실제 연구에 적용하는데 한계가 있다.

물론 빅데이터 분석은 프로그래밍 작업이 불가피하다. 자료의 특성과 연구목적에 따라 프로그램이 설계 · 수정되기 때문이다. 이 때문에 질적 연구자가 빅데이터 분석에 어려움을 느낄 수도 있다. 하지만 NSNA에서 보듯이, 빅데이터 분석은 표집이나 코딩을 최소화하고, 담론 구조를 부분이 아닌 전체로서 보여주는 큰 장점이 있다. 다만 분석 편의를 위해 자료를 중요한 것부터 보여주고, 연결망 그래프로 제시할 뿐이다. 사실 질적 연구자의 영역 지식은 비정형 빅데이터 분석의 초석이다. 이런 면에서 빅데이터 분석은 인문사회와 과학기술 간 융합연구 성격이 강하며 그 효과도 크다고 하겠다.

■ 참고문헌

강남준 · 이종영 · 최운호 (2010). 독립신문 논설의 형태 주석 말뭉치를 활용한 논설 저자 판별 연구. 『한국사전학』, 15권, 73~101.

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

김영희 · 윤상길 · 최운호 (2011). 대한매일신보 국문 논설의 언론 관련 개념 분석. 『한국언론학보』, 55권 2호, 77~102.

김용진 (2004). 신문 뉴스 인용문의 담화 기능: 미국 신문의 9 · 11 사건 보도를 중심으로. 『담화와 인지』, 11권 2호, 19~42.

남궁은정 · 강태완 (2006). 신문 인용 보도의 텍스트 구조. 『스피치와 커뮤니케이션』, 6권, 7~44.

박대민 (2011). 정보원 활용을 통한 신문사의 타당성 요구 응답 전략: 동아일보와 한겨레신문의 4대강 관련 정보원 연결망 분석. 『제 15회 전국 언론학 대학원생 컨퍼런스 자료집』. 77.

박재영 (2006). 뉴스 평가 지수 개발을 위한 신문 1면 머리기사 분석. 『한국의 뉴스 미디어』. 서울: 한국언론재단. 147~220.

손동원 (2002). 『사회 네트워크 분석』. 서울: 경문사.

송경화 · 강범모(2006). 신문 기사의 언어 사용 양상. 『인지과학』, 17권 4호, 255~269.

송용회 (2005). 한국 종합일간지 기사의 사실성 입증 기제에 관한 연구: 조선일보와 한겨레신문 사회면을 중심으로. 『한국언론학보』, 49권 3호, 80~104.

윤상길 · 김영희 · 최운호 (2011). 대한매일신보 국문논설 번역문체 판별의 어휘적 준거에 대한 탐색적 연구. 『언론정보연구』, 48권 1호, 188~228.

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

이재경 · 김진미(2000). 한국 신문 기사의 정보원 사용 관행. 『한국언론학회 2000 봄철 정기 학술 발표대회 논문집』, 293~307.

임영호 · 이현주 (2001). 신문기사에 나타난 정보원의 권력 분포: 1949~1999년 동아 기사의 내용분석. 『언론과학연구』, 1권 1호, 300~330.

장호순 (2001). 한국 신문의 정보원과 취재경로 분석. 『한국언론학회 가을철 정기학술대회 논문집』, 179~193.

최희경 (2008). 질적 자료 분석 소프트웨어(NVivo2)의 유용성과 한계: 전통적 분석방법과 Nvivo2 분석방법의 비교. 『정책분석평가학회보』, 18권 1호, 123~151.

한동섭 · 유승현 (2008). 언론보도에서 나타난 익명정보원에 관한 연구: '미국산 쇠고기 수입 논란'을 중심으로. 『언론과학연구』, 8권 4호, 702~739.

한동섭 · 임종수 (2002). 미디어의 뉴스원 활용과 헤게모니 투쟁에 대한 고찰. 『2001년 한국언론학회 가을철 정기학술대회 논문집』. 27~56.

함유근 · 채승병 (2012). 『빅데이터, 경영을 바꾸다』. 서울: 삼성경제연구소.

한국언론진흥재단 (2013). 『2012 한국언론연감』. 서울: 한국언론진흥재단.

Acquisti, A., & Gross, R. (2009). Predicting Social Security numbers from public data. Paper presented at the Proceedings of the National academy of sciences.

Atwater, T., & Green, N. F. (1988). News sources in network coverage of international terrorism. *Journalism Quarterly*, *65*(4), 967～971.

Bell, A. (1991). *The Language of News Media*. Oxford: Blackwell.

Berkowitz, D. (1987). TV news sources and news channels: a study in agenda-building. *Journalism Quarterly*, *64*(2/3), 508～513.

Bolden, R., & Moscarola, J. (2000). Bridging the quantitative-qualitative divide: the lexical approach to textual data Analysis. *Social Science Computer Review*. *18*(4), 450～460.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society*, *15*(5), 662～679.

Brown, J. D., Bybee, C. R., Weardem, S. T., & Straughan, D. M. (1987). Invisible power: newspaper news sources and the limits of diversity. *Journalism Quarterly*, *64*(1), 45～54.

Chang, K. (1999). Auto trade policy and the press: auto elite as a source of the media agenda. *Journalism and Mass Communication Quarterly*, *76*, 312～324.

Gans, H. (1979). *Deciding Whats News*. Austin: University of Texas Press.

Glasgow University Media Group (1980). *More Bad News*. London: Routledge.

Hallin, D. C., Manoff, R. K., & Weddle, J. K. (1993). Sourcing patterns of national security reporters. *Journalism Quarterly*. *70*(4), 754～766.

Linoff, G. S., & Berry, M. J. (2011). *Data Mining Techniques*: *For Marketing*, *Sales*, *and Customer Relationship Management*. John Wiley & Sons.

Lippmann, W. (1922). *Public Opinion*. 이동근 옮김 (2013). 『여론』. 서울: 아카넷.

Manning, P. (2001). *News and News Sources*: *A Critical Introduction*. London: SAGE publications.

Manovich, L. (2011). Trending: the promises and the challenges of big social data. In Gold, M. K.(ed.). *Debates in the Digital Humanities* (pp. 460～475). University of Minnesota Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. (2011. 5). *Big Data*: *The Next Frontier for Innovation*, *Competition*, *and Productivity*. McKinsey Global Institute.

Richards, T. J., & Richards, L. (1994). Using computers in qualitative research. *Methods*, *1*.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Methodological Issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education*, *12*(1), 8～22.

Russom, P. (2011). *Big Data Analytics*: *An Overview in 20 Tweets*. The Data Warehousing Institute.

Schudson, M. (1978). *Discovering the News*: *A Social History of American Newspapers*. NY: Basic Books.

Sigal, L. V. (1973). *Reporters and Officials*: *the Organization and Politics of Newsmaking*. Lexington, Mass: DC Heath.

Soloski, J. (1989). Sources and channels of local news. *Journalism Quarterly*. *66*(4). 864～870.

Van Dijk, T. A. (1988). *News as Discourse*. NJ: Lawrence Erlbaum.
Vater, H. (2001). *Einfuhrung in die Textlinguistik: Struktur und Verstehen von Texten*. 이성만 옮김(2006).『텍스트의 구조와 이해: 텍스트언어학의 새 지평』. 서울: 배재대 출판부.

네이버 뉴스 news. naver. com
네이버 뉴스클러스터링 news. search. naver. com/newscluster
뉴스메이트 www. newsmate. kr
뉴스 썸머 play. google. com/store/apps/details?id=com. albo7. newsummer
다음소프트 소셜매트릭스 insight. some. co. kr/campaign. html
위키트리 www. wikitree. co. kr
연합뉴스 인터랙티브 www. yonhapnews. co. kr/medialabs/index. html
카인즈 www. kinds. or. kr
Google News news. google. com
Facebook www. facebook. com
Flipboard www. flipboard. com
Summly summly. com/index. html
Twitter twitter. com
Twitter Newsbot twitter. com/news_kor

최초 투고일 2013년 8월 7일
게재 확정일 2013년 11월 15일
논문 수정일 2013년 11월 25일