커뮤니케이션 이론
2018년 14권 3호 5~57쪽
https://doi.org/10.20879/ct.2018.14.3.005

유사도, 기사 공동 출현, 정보원에 기초한 뉴스 문장연결망 분석 알고리즘 제안*

박대민 한국언론진흥재단 선임연구위원**
서봉원 서울대학교 융합과학기술대학원 교수
김성현 서울대학교 융합과학기술대학원 석사
유재연 서울대학교 융합과학기술대학원 박사과정
송정우 서울대학교 융합과학기술대학원 석사과정

이 논문은 뉴스를 단어 중심의 의미연결망 분석을 통해 연구할 때 나타나는 한계를 지적하고, 문장 수준의 의미연결망 분석을 하는 방법으로서 뉴스 문장연결망 분석 방법을 제안했다. 이 연구에서는 특히 뉴스 인용문 중심의 분석 프로그램인 쿼트넷을 만들어 시행연구를 실시했다. 인용문 중심 뉴스 문장연결망은 개인실명 직접인용문을 결점으로, 동일 정보원 발언 여부, 기사 공동 출현 여부, 자카드 유사도를 결합한 관련도를 연결로 하는 의미연결망을 뜻한다. 뉴스 문장연결망은 의미경로를 가지며, 이를 활용해 중심문장, 요약문장, 상술문장 등을 정의할 수 있다. 이 연구에서는 〈빅카인즈〉에서 1990년 1월 1일부터 2016년 4월 30일까지 '인공지능'으로 검색된 기사 2,337개의 인용문 5,046개에 대해 뉴스 문장연결망 분석을 실시했다. 분석 결과, 유사도 계수를 0.333으로 했을 때 고립자(isolated node)를 제외하고 3,742개 결점 6,708개의 연결로 이루어진 문장연결망이 도출됐으며 알파고와 인공지능의 충격,

* 이 연구는 2018년 서울대학교 언론정보연구소 연구기금의 지원을 받았습니다

** heathe0@gmail.com; dmpark@kpf.or.kr

인공지능을 활용하는 여러 기술들을 적절히 분류하고 요약했다. 유사도 절삭 기준을 0.333에서 0.450로 높이면 고립자를 제외한 결점은 3,697개, 연결은 6,383개였다. 연결이 줄고 고립자가 늘어나면서 내용이 지나치게 세분화되어 묶이는 경향을 보였다. 뉴스가 중요한 사회 전반의 쟁점들에 대해 대중적인 논증을 다년간 축적하고 있다는 점을 고려할 때 IBM 프로젝트 디베이터와 같은 컴퓨터 논증의 기초 기술로서 사회과학자가 쉽게 활용할 수 있는 컴퓨터 논증 프로그램을 설계하는데도 기여할 수 있을 것이다.

KEYWORDS 문장 연결망 분석, 쿼트넷, 의미연결망 분석, 컴퓨터 논증, 뉴스 빅데이터 분석

1. 문제제기

소셜 미디어를 비롯한 각종 사용자 행동 데이터(user generated behavior data)를 대규모로 수집할 수 있게 되면서 관련 분석 방법론도 빠르게 발전하고 있다. 사회과학계에서도 사회 구조와 인간 행동 관련 빅데이터(big data)를 컴퓨터로 분석하는 컴퓨터 이용 사회과학(computational social science)이 주목을 받고 있다(Watts, 2013). 특히 사회학과 사회물리학 분야에서는 행위자(actor) 간 상호작용을 그래프 이론(graph theory)을 활용해 분석하는 사회연결망 분석(social network analysis)이 발전해왔다.

그러나 사회연결망 분석은 행위와 행위체계를 분석 할 뿐, 행위의 의미나 행위자들이 만드는 담론 및 기호체계는 제대로 분석하기 어렵다. 연결망 이론의 관점에서 이러한 기호체계 분석은 사회연결망 분석과 구분되는 의미연결망 분석(semantic network analysis)을 통해 이뤄질 수 있다.

사회연결망 분석이 사회과학분야 중 특히 상호작용을 다루는 사회학에서 주목을 받았다면 의미연결망 분석은 특히 의사소통을 다루는 커뮤니케이션학에서 빠르게 확산되고 있다(윤호영, 2018). 국내에서는 의미연결망 분석을 뉴스 저작권 위탁사인 한국언론진흥재단의 기사 아카이브와 연계한 뉴스 빅데이터 시스템 〈빅카인즈(BigKinds)〉에 접목해 공개함으로써 다양한 분야의 연구로 확산되고 있다. 그런데 기존의 의미연결망 분석은 대부분은 형태소, 단어, 어절 분석이나 분석의 편의를 위해 웹이나 기사처럼 문서 단위 분석이 이루어졌다. 그러나 사회 현상을 행위에 관한 미시 수준(micro level)과 구조에 관한 거시 수준(macro level), 그리고 둘을 매개하는 중위 수준(mezzo level)으로 이해하듯이 텍스트와 같은 기호 현상에서도 미시 수준인 단어와 거시 수준인 문서, 그리고 중위 수준인 문장을 생각할 수 있다. 사회학에서도 보다 정교한 분석을

위해 미시와 거시를 매개하는 중위 수준의 분석이 주목을 받았듯이 의미연결망 분석에서도 중위 수준인 문장 분석이 요청된다.

그동안 자동화된 텍스트 분석은 자연어처리와 의미연결망분석, 텍스트 마이닝, 토픽모델링, 빅데이터 분석 시스템과의 결합, 딥 러닝(deep learning) 기술의 접목 등을 통해 많은 발전을 이루었다. 이에 따라 언론학을 비롯한 사회과학분야에서도 점차 활용도가 높아지고 있다. 그럼에도 불구하고 대부분의 분석이 단어 수준에 치중됨으로써 맥락을 직관적으로 깊이 있게 파악하는데 어려움을 겪고 있다.

이 연구는 뉴스 기사를 대상으로 단어와 문서, 문장 등 다수준 의미연결망 분석(multi level semantic network analysis)을 위한 기초 작업으로서 뉴스 문장연결망 분석을 제안한다. 뉴스 문장 연결망에서는 의미 단위(token)에 해당하는 결점(node)은 문장으로, 의미소 간 관련도(relevance)에 따른 연결(edge)은 유사도(similarity)와 공동출현(co-occurrence) 등으로 정의한다. 이를 바탕으로 이 연구에서는 문장 간의 의미 맥락을 더 직관적으로 이해할 수 있도록 의미거리(semantic distance), 의미경로(semantic path), 핵심 문장, 요약 문장, 상술 문장 등을 수학적으로 정의할 것이다. 이어 이를 바탕으로 뉴스 문장 연결망 분석 프로그램 〈쿼트넷(QuoteNet)〉[1] 프로토타입 버전을 개발했다. 그리고 시행연구(pilot study)로서 〈빅카인즈〉에서 인공지능 관련 기사 데이터를 활용해 테스트해보았다. 뉴스 문장연결망 분석 방법은 단어, 문서 외에 문장 수준에서 다양하고 직관적인 분석을 가능하게 해줄 것으로 기대된다. 더 나아가 뉴스가 중요한 사회 전반의 쟁점들에 대해 대중적인 논증을 다년간 축적하고 있다는 점을 고려할 때 IBM 프로젝트 디베이터(IBM Project Debater)와 같은 컴퓨터 논증(computational argumentation)의 기초 기술로서 사회과학자가 쉽게 활용할 수 있는

1) https://goo.gl/2rc7iT

컴퓨터 논증을 설계하는데도 기여할 수 있을 것이다.

2. 기존 문헌 검토

1) 자동화된 텍스트 분석

전통적으로 사회과학 분야에서는 뉴스와 같은 텍스트 분석을 수작업에 의존해왔다. 그러나 텍스트가 양(volume), 속도(velocity), 다양성(variety) 측면에서 빅데이터화함에 따라 자동화된 텍스트 분석의 활용도가 높아지고 있다.

텍스트 분석은 일반적으로 수집, 정제, 분석, 해석 등 네 단계로 수행된다. 이 중 분석 부분을 좀 더 상세히 살펴보자. 수집되고 정제된 데이터는 벡터 공간 모형(vector space model, VSM)에 따라 가중 벡터 값을 갖는 데이터로 구조화된다(Salton, 1971). 대표적으로 용어 출현 빈도(term frequency)와 역문서 빈도(inverse document frequency)와 같은 수집 빈도(collection frequency)를 고려한 TF IDF(term frequency inverse document frequency)와 여기에 길이 정규화(length normalization)를 통해 문서 길이에 따른 차이를 보정한 방식이 쓰인다(김남규, 이동훈, 최호창, 2017; Jones, 1972; Lee, Chuang & Seamons, 1997; Salton, 1988).

벡터화된 의미데이터는 매우 높은 차원을 갖기 때문에 많은 경우 차원 축소(dimension reduction) 과정을 거친다. 대표적인 차원 축소 기법으로는 주성분 분석(principle component analysis, PCA), 특이값 분해(singular value decomposition, SVD), 음수 미포함 행렬 분해(non negative matrix factorization, NMF) 등이 있다(Lee, Chuang & Seamons, 1997; Hotelling, 1933; Stewart, 1993). 이어 차원 축소와 함께, 또는 별도로 동시출현(co occurrence) 및 유사

성(similarity)를 바탕으로 단어나 문서 등 의미 단위 간 관계를 파악하기도 한다(Lee & Seun, 1999; Turney & Patrick, 2010).

분석 단계에서는 크게 빈도분석(frequency analysis), 군집화(clustering), 분류(classification)와 이를 통한 순위화(ranking) 과정이 수행된다. 방대한 텍스트 데이터의 빈도분석에 대한 전체적인 결과는 흔히 워드 클라우드(word cloud)를 비롯한 각종 그래프로 시각화(visualization)되어 제시된다. 한편 빈도를 시간에 따라 제시하면 추세분석(trend analysis)과 같은 자동화된 시계열 내용분석이 된다(박대민, 2016; Aiden & Michel, 2014).

다양한 수준의 의미 단위들을 토픽 모델링(topic modeling) 기법을 활용해 군집화할 수도 있다. 과거에는 잠재 의미 분석(latent semantic analysis, LSA), 확률적 잠재 의미분석(probabilistic LSA, pLSA)이 많이 활용됐다(Deerwester, et al., 1990; Hofmann, 1999). 최근에는 잠재 디리클레 할당(latent Dirichlet allocation, LDA), Word2Vec 등이 주목을 받고 있다(Blei et al., 2003; Mikolov, et al., 2013). Word2Vec 역시 문장, 문단, 문서로 분석 수준을 확장한 Seq2seq, Paragraph2vec, Doc2vec 등으로 확장됐다(Le & Mikolov, 2014; Sutskever, et al., 2014).

분류는 기본적으로 문서 분류(document classification)에 사용된다. 문서 분류 방법으로는 나이브 베이즈(Naïve Bayes)나 의사결정 트리(decision tree), SVM(support vector machine) 등을 이용한 기계학습 방법이 널리 활용된다(Joachims, 1998; Lewis & Ringuette, 1994; Vapnik & Kotz, 1982). 한 문서를 여러 범주로 다중 분류할 수도 있다. 〈빅카인즈〉 역시 SVM을 기본으로 기사 지면을 다중 분류한다(SaltLux, 2015). 텍스트에 담긴 호불호나 의견을 파악하는 감성분석(sentiment analysis)이나 평판분석(opinion mining) 등도 분류의 일종으로 볼 수도 있다(Pang & Lee, 2008)

뉴스는 지면 분류, 자동 부착, 요약, 연관어 추천, 기사 군집화, 중복 판정, 감성분석 등을 통한 뉴스의 가치 판단, 자동 편집, 추천 등 텍스트 분석이 활발하게 활용되고 있다(Bharat, et al., 2009; Newman, et al., 2006; Curtiss, et al., 2009; McKeown, et al., 2002; Park, et al. 2009). 이와 함께 텍스트 분석을 자연어 생성(natural language generation) 기술과 접목한 로봇저널리즘(robot journalism)도 주목을 받았다(김동환, 이준환, 2015).

자동화된 텍스트 분석은 많은 발전에도 불구하고 아직까지 사회과학 분야에서 활용하는 데에 있어 몇 가지 난점이 있다. 첫째는 맥락 파악의 어려움, 둘째는 담론분석 수준의 깊이 있는 분석의 어려움이다. 이는 모두 분석 단위가 단어 수준에 치중되어 있기 때문에 발생하는 문제로 문장 수준의 분석을 통해 해결할 수 있을 것으로 기대된다.

사실 현재 기술적으로는 단어, 문장, 문단, 문서 등 다양한 수준의 분석이 가능하다. 하지만 사회과학 분야에서는 단어 수준의 분석에 치중되어 있다. 이제 사회과학 분야에서도 LDA와 같은 수준 높은 단어 수준 분석을 활용하는 사례가 적지 않다(예컨대 진설아 등, 2013). 그러나 단어 수준 분석은 맥락을 파악하기가 어렵다. LDA의 경우 군집화된 단어들을 다시 하나의 의제로 묶는 라벨링(labeling) 이슈가 발생한다. 정확한 라벨링을 위해서는 다시 영역 지식 전문가가 개입하거나, 아니면 연구자가 원래 문서를 들여다보면서 군집화된 단어들이 한 문장에서 사용되는 등 명백히 관련된 사례를 최소한 하나 이상 수작업으로 찾아야 한다. 더 큰 문제는 많은 경우 연구자들이 군집화된 단어들을 보고 개인의 상식과 통찰력에 기초해 라벨링을 하고 해석한다는 것이다. 그 결과 실제로는 존재하지 않는 맥락을 제시하거나, 너무 일반적인 수준에서 맥락을 이해하고 만다.

한편 사회과학 분야의 자동화된 분석은 파이썬(python)이나 R의 형태소분석 라이브러리 정도만을 활용하는 경우가 많다. 그 결과 분석의

깊이가 형태소분석과 구문분석에서 그치는 경우가 많다. 정교한 개체명 인식 단계만 가도 사전 구축에 어려움을 겪고 있다. 의미중의성 해소나 대용어 해소 등 의미분석은 시도조차 안 하는 경우가 많은데, 이는 결국 분석의 재현율이나 순위화에 영향을 준다. 이를 고려하는 경우에도 사전 정제 작업 과정에서 사실상 수작업으로 진행하는 사례가 많다. 이 어려움 때문에 자동화된 분석을 대규모 데이터 분석에 활용하지 못하거나 자동화된 분석이 수작업에 의한 분석에 비해 크게 효율적이라고 생각하지 못하는 경우가 많다.

게다가 의미분석에서 이미 난점에 봉착하고 있지만, 정작 언론학 등 사회과학 분야의 텍스트 분석은 더 높은 수준인 담론분석, 그것도 비판적 담론분석을 주로 수행한다. 예컨대 기본적으로 뉴스의 논조나 뉘앙스, 맥락 등은 파악해야 한다. 이를 위해 텍스트 분석에서는 감성분석이나 평판분석을 활용하고 있다. 그러나 한국어 감성분석과 평판분석은 특정 제품군의 상품평 등 영역을 한정한 경우가 아니면 아직 충분한 성능을 구현하고 있지 못하다. 특히 뉴스의 경우, 사회 전 분야를 장기간에 걸쳐서 다루기 때문에 감성분석과 평판분석이 더욱 어렵다.

감성분석과 평판분석을 문서 단위, 즉 기사 수준에서 수행하면 문제는 더 커진다. 기사는 원칙적으로 중립적으로 작성되기 때문이다. 게다가 저널리즘 관행을 고려하면, 중립은 사실상 찬성인 경우가 많으며, 전체적인 의견 기후(opinion climate)가 찬성 쪽으로 경도돼 있는 경우 반대는 양이 적더라도 큰 의미를 갖는 경우가 많다(박대민, 2015). 즉 감성분석과 평판분석은 문서 단위가 아니라 문장 단위에서 이루어져야 한다. 더 나아가 일종의 해상도 이슈도 있을 수 있다. 즉 기사에 담긴 치열한 사회적 논쟁을 살펴보면, 어떤 의제에 일반적으로는 같은 입장인 두 진영이 각론에서는 서로 간 대립하는 경우도 있다. 어떤 두 정당이 모두 진보적이라고 해도 대립할 수 있는 셈이다. 더 나아가 의견 역시 '호/불호/중립'이 아니라 정치는 '찬성/반대', 경제는 '매도/매수', 문화는 '호/불호' 등

으로 세분화될 수 있다. 예컨대 보호무역 정책이 잘못됐다 생각하면서도 관련 제조업 주식을 매수할 수도 있다. 결국 주제나 조직, 사람, 시기에 따라서 감성과 의견을 세분화해야 할 수 있다.

문장 수준의 분석은 이러한 문제를 상당 부분 해소한다. 단어와 달리 완전한 문장은 그 자체만으로 의미를 정확히 파악할 수 있다. 관련도에 따라 군집화된 문장 역시 맥락을 보다 직관적으로 파악할 수 있다. 특히 뉴스의 인용문은 발언자와 조직, 주제가 연결되어 있다. 더 나아가 단순히 관련된 문장을 군집화하는 것이 아니라 문장을 관련도에 따라 순차적으로 제시할 수 있다면 담론이 해체되고 모순적으로 전개되는 상황도 파악할 수 있을 것이다. 예컨대 뉴타운 담론이 투기화된 신도시 담론을 대체하고 서민 주택 공급 및 서민 주거 환경 개선이라는 사회적 목표 달성을 위해 정당화됐지만, 결국 신도시 담론처럼 투기 조장으로 비판 받고 대체되는 논의 과정을 이해할 수도 있다(박대민, 2014a).

2) 의미연결망 분석

넓게 보면 의미연결망 분석 역시 자동화된 텍스트 분석의 일부이다. 그러나 다른 한편으로는 사회 연결망 분석 방법으로부터 영향력도 적지 않다[2]. 그러나 사회연결망 분석은 행위와 행위체계를 대상으로 한다(Sowa, 2000). 반면 의미연결망 분석은 의미와 기호체계를 다룬다. 기호체계는 행위체계와 긴밀한 관계를 갖고 있지만, 담론이 행위로 환원되지는 않는

---

2) 사회 연결망 분석 연구는 소시오메트리(sociomatrix)를 시작으로, 6단계 분리 법칙(six degrees of separation), 약한 연결의 힘(the strength of weak tie), 하버드 학파의 집합이론(set theory)과 그래프 이론(graph theory)의 활용, 구조적 공백(structural hole), 각종 중앙성 지표(centrality), 작은 세계(small world) 현상, 척도 없는 연결망(scale free network), 스며들기(percolation), 3단계 영향 규칙(three degrees of influence rule) 등으로 발전했다(Moreno, 1937; Travers, 1967; Granovetter, 1973; White, 1963; Burt, 1992; Freeman, 1979; Watts & Strogatz, 1998; Barabási & Albert, 1999; Callaway, et al., 2000; Christakis & Fowler, 2009).

다. 행위와 행위체계와 구분되는 담론과 기호체계의 중요성은 소통 행위(communicative action), 언어학적 전회(linguistic turn), 담론구성체(discursive construction) 등의 개념으로도 강조됐다(박대민, 2015; Habermas, 1985; Rorty, 1992).

의미연결망 분석에 사회연결망의 분석 방법을 활용할 수 있지만 몇 가지 차이점이 있다. 사회연결망의 결점와 연결은 각각 행위자(actor)와 상호작용(interaction)이다. 반면 의미연결망에서 결점은 형태소, 품사, 단어, 문장, 문단, 문서, 매체 등 의미 단위가 되며, 연결은 의미론적인 관련도이다. 또 의미연결망은 사회연결망에 비해 결점과 연결이 훨씬 많다. 예컨대 원고지 10매 분량의 기사 1개에도 30개가 넘는 문장과 450 500개의 단어가 사용된다. 이러한 기사를 수 만개 이상 분석한다면, 결점만 해도 수백 만 개를 쉽게 넘어설 수 있다. 따라서 의미연결망 분석은 앞서 말한 것처럼 적절한 차원감소를 비롯한 자동화된 분석이 필수적이다. 실제로 의미연결망 분석은 결점 및 연결 정보를 자동으로 추출하기 위해 다양한 자연어처리 기법을 활용한다.

기존의 의미연결망 분석은 형태소, 단어, 어절 수준이나 웹, 기사처럼 문서 수준의 분석이 이루어졌다. 그러나 사회 현상을 행위와 구조를 중위 수준 분석으로 보다 정확히 이해 수 있는 것처럼, 텍스트와 같은 기호 현상 역시 중위 수준을 분석하는 것이 유용할 수 있다. 이 때 문장은 형태소, 단어, 어절과 같은 미시 수준과 기사, 매체, 언론계 등 거시 수준에 비해 상대적으로 중위 수준이 된다. 형태소나 품사 등은 형식언어학적 단위로 탈맥락적이고 문서는 맥락을 이해할 수는 있지만 하나의 의미론적 기본 단위로 간주하기에는 너무 복합적인 내용을 담고 있다. 반면 문장, 특히 단문은 하나의 의미를 완결된 형태로 제시하여 단순하면서 직관적인 의미 전달이 가능한 텍스트의 기본 단위이다(정태석, 2002; Vater, 2001).

의미연결망 분석은 다양한 분야에서 활용되어 왔다. 과거에는 문헌

정보학에서 연구 동향을 파악하기 위해 의미연결망과 사회연결망을 혼합한 성격의 지식연결망 분석이 가장 많이 이뤄졌다. 이러한 지식연결망으로는 공동인용 연결망(cocitation network) 분석, 동시인용 분석(cocitation analysis), 공동연구 연결망(collaboration network) 분석, 공저자 연결망(coauthorship network) 분석, 동시단어분석(coword analysis) 등이 있다(Barabâsi et al., 2002; Callon et al., 1986; Garfield, 1964; Garfield & Merton, 1979; Kretschmer, 1994; Moody, 2004; Newman, 2001; Newman, 2004; White & Griffith, 1981; White & McCain, 1998; Zhao & Strotmann, 2008).

언론학 분야에서는 Kr-Kwic 프로그램이 도입된 이후 다양한 텍스트에 대한 의미연결망 분석이 활성화됐다(Park & Leydesdorff, 2004). 최근에는 자연어처리 기술을 결합해 결점을 자동으로 추출하고 사회연결망 분석 기법을 활용해 순위화하는 자동화된 의미연결망이 댓글, 블로그, 소셜 미디어, 뉴스 등의 분석에 활용되고 있다(예컨대 박대민, 2016; 박지영, 김태호, 박한우, 2013; 채영길, 유용민, 2017 등).

기존의 뉴스 의미연결망 분석은 대부분 기사나 제목에서 추출된 형태소 내지 명사나 형용사와 같은 품사를 분석 단위로 삼고 있다. 예컨대 기사 속 인용된 정보원을 분석하는 뉴스 정보원 연결망 분석의 경우 사실상 인용문의 중요도를 정보원의 중요도로 환원한다(박대민, 2013). 즉 중요 정보원의 발언을 중요 인용문으로 가정한다. 이는 어느 정도 타당하지만, 그럼에도 불구하고 주요 인물의 발언이 모두 중요한 것은 아니며, 주변적 인물의 발언이 모두 중요하지 않은 것은 아니다. 이를 해결하기 위해서는 개체명 수준이 아닌 문장 수준에서 바로 분석을 하는 것이 필요하다.

분석 수준은 재현율(recall)과 정확도(precision)의 트레이드 오프(trade off) 문제와도 관련된다. 즉 검색어를 포함한 기사에서 인용문을 추출할 경우, 기사는 해당 검색어와 부분적으로나마 관련되지만 인용문

은 관련이 없을 수 있다. 반대로 검색어를 포함한 인용문이 있는 경우만 추출할 경우, 검색어를 포함하지 않지만 해당 검색어와 관련된 인용문이 있을 수 있다. 예컨대 질의어를 '감세'로 하고 기사를 추출할 때, 어떤 기사는 전적으로 감세에 관한 기사일 수 있지만, 어떤 기사는 '대선'의 하위 의제로서 '감세'를 부분적으로만 다룰 수도 있다. 그 결과 '감세'가 포함된 기사의 인용문 가운데 '감세'가 아닌 '대선'의 다른 하위 의제, 예컨대 '안보'에만 관련된 인용문이 있을 수 있다. 즉 인용문의 재현율은 높지만 정확도는 떨어진다. 이를 해결하기 위해 '감세'라는 단어가 포함된 인용문만 필터링(filtering)한다면, 이번에는 정확도는 100%에 가깝게 되지만, 재현율은 크게 떨어지게 된다. 즉 문장 안에 '감세'라는 단어는 포함되지 않지만 '감세'에 관해 논한 인용문이 제외된다. 뉴스 문장연결망 분석의 과제는 예컨대 '감세'를 포함하지 않은 인용문을 검색 결과로 제시하는 한편, '감세'와 관련되지 않은 인용문의 순위를 낮추는 것일 수 있다.

뉴스 문장연결망은 한 문장이 등장할 때, 어떤 문장이 관련도에 따라 연쇄적으로 활성화될 수 있는지, 즉 의미경로에 대한 정보를 제공한다. 의미경로 개념은 자연수로 주어지는 두 문장 간 의미거리 개념을 내포한다. 즉 일정 거리 이내의 문장은 인지적으로 다 관련된 문장일 수 있지만, 근사적으로 의미거리가 가까운 것은 더 관련되고, 먼 것은 덜 관련된다. 의미거리를 결정하는 관련도는 문장 간 유사도만으로 측정할 수 없다. 예컨대 기사 공동 출현 문장들은 전혀 다른 문장임에도 관련될 수 있다. 이 연구에서 뉴스 문장연결망의 연결은 관련도로, 관련도는 유사도와 함께 기사공동출현, 일정 기간 내 동일 정보원 발언 여부로 정의하고자 한다.

3) 뉴스 빅데이터 분석

뉴스 빅데이터 분석은 자연어처리와 의미연결망 분석을 결합해 뉴스 기사를 대규모로 분석하는 연구로 볼 수 있다(박대민, 2013; 2015). 특히 국내에서는 뉴스 저작권 대행사이자 공공기관인 한국언론진흥재단이 수

십 개 주요 매체로부터 수십 년간 기사를 수집해 자연어처리한 뉴스 빅데이터 분석 시스템을 활용한 연구가 진행되고 있다.

공개된 뉴스 빅데이터 분석 시스템은 2013년 차세대융합기술원이 한국언론진흥재단의 카인즈 데이터를 처리한 개발한 〈뉴스소스 베타〉가 시작이다(차세대융합기술원, 2013). 이어 한국언론진흥재단은 〈뉴스소스 베타〉를 바탕으로 〈빅카인즈〉를 기획해 2016년 공개했다[3]. 해외에서는 수작업으로 메타데이터를 부착하는 구조화된 저널리즘(structured journalism)을 시작으로, 〈뉴스소스 베타〉와 유사한 뉴스 빅데이터 분석 시스템인 IBM이 2015년 〈IBM 왓슨 뉴스 익스플로어 베타〉[4]를 내놓았으며 이후 이를 인공지능 시스템인 IBM 익스플로어에 반영했다(김선호 등, 2015.12).

표 1. 〈빅카인즈〉의 인식 성능

entities	recall	precision	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
인용문		82.26	

〈빅카인즈〉는 기사 게재일자, 매체, 기사 제목과 같은 언론사가 수작업으로 제공하는 메타데이터 외에 개체명 인식, 지면 분류, 인용문 추출 등 자연어처리 기술을 바탕으로 자동 부착(auto tagging)된 정보원 인명, 기관명, 직함, 지면, 인용문 본문, 기자명, URL 등을 xls이나 CSV 파일 형태로 다운로드 받을 수 있다. 2016년 초기 버전 〈빅카인즈〉의 자연어처리 인식 성능은 F1 점수 기준으로 아래와 같다(솔트룩스,

3) http://www.kinds.or.kr/

4) http://news explorer.mybluemix.net/

2015.11.). 다만 인용문 성능은 재현율이 공개돼 있지 않았다.

〈빅카인즈〉가 본격적으로 연구자들에게 공개된 것이 불과 2년 남짓함에도 불구하고 뉴스 빅데이터 연구는 비교적 빠르게 확산되고 있다. 언론학계는 물론 인문사회 분야와 이공계 분야를 막론하고 〈빅카인즈〉를 활용한 뉴스 빅데이터 분석이 광범위하게 시도됐다(곽재현, 홍지숙, 2018; 권혜진 등, 2017; 김대진 등, 2018; 김민준 등, 2017; 김봉제, 2018; 김재욱, 김한수, 2018; 김종성, 2017; 김혜원, 이정욱, 2018; 맹미선, 2017; 박대민, 2016; 박이수, 2017; 박현정 등, 2017; 박희봉, 이민화, 2016; 배진수, 2016; 성미애, 2017; 손기준 등, 2015; 신사임 등, 2017; 이은별 등, 2017; 이정석, 2017; 조현채, 박철용, 2018; 최모나 등, 2016; 최영지, 2017; 최준희 등, 2017; 최충익, 김철민 , 2017). 이밖에 뉴스 빅데이터 분석을 활용한 기획 기사[5] 작성, KISDI의 ICT 지수 개발 및 트렌드 보고서 발행[6], 뉴스 미디어 스타트업 공모전[7] 등도 진행 중이다.

〈빅카인즈〉는 44개 언론사의 기사를 수집해 개체명 인식, 인용문 분석, 의미분석 등 비교적 높은 수준의 자연어처리를 사전에 수행한 뒤 사회과학자들에게도 익숙한 형태인 엑셀 파일로 제공한다는 장점이 있다. 그러나 여전히 몇 가지 개선점은 있다. 무엇보다 의미연결망 분석 기능을 제공하지만 체계적인 분석을 하지 않아 분석에 큰 혼선을 준다. 예컨대 결점의 수가 너무 많을 때 결점의 수를 절삭(cut-off)하게 되는데, 이 때 절삭 기준이 임의적인지, 빈도 수나 연결정도 중앙성 값에 따라 상위권만 표시한 것인지, 아니면 최근 일자 순인지 등을 파악하기 어렵다. 전체 결

5) 2018년 7월 5일 현재 네이버 뉴스에서 '뉴스 빅데이터'로 검색한 결과 581건의 기사가 나왔다. https://goo.gl/ULFoQ8

6) KISDI 보도자료: https://bit.ly/2MN4WFe

7) https://bit.ly/2KzgSOf

점 수가 몇 개인데 실제 표현된 결점 수는 몇 개인지 등의 정보도 누락돼 있다. 다양하면서도 높은 수준의 분석을 위해서, 즉 주제 분석이나 인용문 분석을 보다 원활하게 하기 위해서, 기사 단위가 아니라 인용문 단위에서 주제를 추출하고 인용문의 id를 부여할 필요도 있다. 주제를 결점으로 하고 기사 공동 출현 기준으로 연결하면, 결점은 너무 많고 결점 간 의미론적 관계도 낮은 것끼리 연결되게 된다. 즉 연결의 정확도가 너무 떨어진다. 그러나 인용문 기준으로 주제를 연결하면 이러한 문제가 해소된다. 같은 인용문에 등장한 주제들은 한 문장에서 다뤄질 만큼 매우 밀접한 관련성을 갖기 때문이다. 인용문 수준에서 연결 정보를 추출하려면 인용문 id를 부여하는 것이 필요하다. 또한 인용문 id를 부여하면, 인용문을 하나의 결점으로 보고 뉴스 문장 연결망 분석을 할 수도 있다. 사실 이는 이미 〈빅카인즈〉 내부적으로는 구현된 기능이고 사용자에게 제공되고 있지만 않을 뿐이어서 운영자 측의 정책 변화와 간단한 시스템 디자인 변경만으로 충분히 성능 개선이 가능하다. 정리해자면 〈빅카인즈〉는 현재 제공하고 있는 인물, 기관, 주제 등 개체명 분석 기능 외에 문장 분석 등 다양한 수준의 자동화된 담론분석을 지원할 잠재력을 갖고 있지만, 현재로서는 이를 충분히 구현하고 있지 않다.

4) IBM의 디베이터

뉴스 기사는 기본적으로 사실과 의견을 담고 있다. 그리고 스트레이트 기사를 제외하고 논설은 물론 피처 기사 역시 대부분 어떤 하나의 주장을 담고 있다. 예컨대 정치 뉴스의 경우 인물, 조직, 정책에 대한 지지와 반대, 경제 뉴스의 경우 궁극적으로 자산에 대한 매수와 매도의 의견, 문화 뉴스의 경우 어떤 문화콘텐츠를 좋아하느냐 싫어하느냐와 같은 취향을 담는다. 즉 뉴스 기사는 대부분 사실을 근거로 펼치는 논증(argumentation)이다.

뉴스 기사는 논증으로서 논문이나 보고서 등과 다른 특징을 갖는다.

우선 뉴스 기사의 논증은 학자나 전문가가 아니라 대중을 대상으로 하므로 쉽게 서술된다. 다음으로 뉴스 기사는 사회 현상 전 분야에 대한 논증을 담고 있다. 또한 뉴스 기사는 사소한 것이 아니라 주목할 만한 사회 현상에 대한 논증을 담고 있다.

그런데 객관주의 저널리즘에서 기자는 자신의 의견을 그대로 제시할 수 없다. 따라서 뉴스 기사에서는 주장이 인용문을 통해 담긴다. 즉 기자는 정보원의 입을 통해 주장을 담는 한편, 정보원을 인용함으로써 사실성을 확보한다(박대민, 2015). 또한 불편부당성을 이념적으로 지향하기 때문에, 그리고 하나의 주장만 실었을 때 위험성을 회피하기 위해서라도, 하나의 주장을 강조할 때도 최소한 하나 이상의 반론을 함께 담으려고 한다.

이는 뉴스 기사를 문장 수준에서 연결하는 뉴스 문장 연결망이 일종의 자동 생성된 논증의 시각화, 즉 논증 지도(argumentation map) 내지 논증 그림(argumentation diagram)이 될 수 있음을 시사한다(Carr, 2003). 그것도 사회적으로 중요한 수많은 주제에 대해 장기간에 걸쳐서 전개된 논증의 내용을 담고 있을 가능성이 있다. 즉 일종의 토론 기계를 만든다면, 뉴스 문장 연결망 분석이 유용하게 활용될 가능성이 있다.

실제로 컴퓨터공학에서는 인공지능 기술과 연계해 이러한 토론기계를 만들고자 하는 시도가 있어왔다. 컴퓨터 논증 연구가 그것이다[8]. 2018년 6월에는 IBM의 하이파(Haifa)라는 이스라엘 소재 개발팀이 프로젝트 디베이터라는 인공지능 기반 토론기계 개발 계획을 6년째 진행 중이며 소규모로 공개 테스트를 수행했다고 밝혔다[9].

IBM 디베이터는 복잡한 주제에 대한 논쟁에 전면적으로 참여할 수 있는 시스템을 목표로 한다. 디베이터를 구성하는 기술로 IBM은 크게 ① 논증 발굴(argument mining), ② 입장 분류(stance classification)

---

8) https://www.doc.ic.ac.uk/~ft/argumentation.html

9) https://www.research.ibm.com/artificial intelligence/project debater/

및 감성분석(sentiment analysis), ③ 딥 뉴럴넷(deep neural nets, DNNs)과 약한 지도학습(weak supervision), ④ 자연어처리, ⑤ 텍스트 음성 변환 시스템(text to speech systems, TTS) 등을 제시했다. 논증 발굴은 복수의 문서에서 주장과 근거, 반대 의견을 찾아내고, 논증들을 연결하고, 논증의 질을 평가하고 새로운 주장을 합성하는 등의 과정을 통해 논증 코퍼스를 만드는 과정이다. 입장 분류 및 감성분석은 논증의 입장을 파악해 분류해 시스템이 어떤 입장을 결정하면 그러한 입장에 따라 찬성 또는 반대의 주장을 근거와 함께 전개할 수 있게 하는 기술이다. 딥 러닝과 자연어처리는 논증 구조를 찾아내고 논증의 품질이나 풍부함을 심화시키기 위해, 그리고 음성인식 기능을 개선하기 위해 활용된다. 텍스트 음성 변환 시스템은 디베이터가 구성된 논증을 음성으로 발화할 수 있게 한다. 끝으로 벤치마크 데이터 세트(benchmark datasets)는 성능을 비교, 개선하기 위해 축적하는 데이터다.

이러한 기술을 편의상 재구성해보면, ① 주어진 문서에서 논증 코퍼스를 만드는 과정, ② 논증의 입장에 따라 일관된 견해를 제시하는 과정, ③ 새로운 논증을 구성하는 과정, ④ 논증의 난이도나 문체 등을 조정하는 과정, ⑤ 음성을 인식하고 합성하는 인터페이스, ⑥ 논증의 평가와 시스템 개선 과정 등으로 나눠볼 수 있다. 이 가운데 토론기계로서 가장 핵심적인 기술은 1, 2, 3이다.

뉴스 문장 연결망은 이 가운데 1과 2의 과정과 관련된다. 뉴스 문장 연결망 분석에서는 논증 탐지를 인용문 식별로 갈음한다. 즉 일종의 규칙 기반으로 논증을 탐지한다. 비록 기사 안의 다른 부분 중 일부는 논증으로 간주할 수도 아닐 수 있지만, 모든 인용문은 논증으로 간주할 수 있다. 즉 뉴스 기사에서 논증 탐지의 재현율은 약간 떨어질 수 있지만, 정확도는 인용문을 제대로 식별하기만 했다면 100%이다. 사실과 의견은 구분하지 않는데, 실제로 기사에서 이 둘을 구분하는 것은 인간으로서도 쉬운 일이 아니다. 논증 코퍼스는 인용문 연결망 형태로 제시된다. 뒤에서 제

시할 뉴스 문장 연결망 분석을 활용하면 문장이 연결망에서 차지하는 위치에 따라, 문장을 핵심, 요약, 상술 등으로 태깅할 수 있다. 또 문장 간의 관계 역시 의미경로나 의미군집 등으로 태깅할 수 있다. 입장은 다층적으로 군집화된다. 즉 하나의 군집을 통해 찬반을 함께 제시하면서 입장을 제시할 수도 있고, 조직이나 인물에 따라 좀 더 일관되고 세분화된 입장으로 제시할 수도 있다.

IBM 디베이터와 비교하면 뉴스 문장 연결망 분석은 이러한 코퍼스를 만드는 과정이 매우 직관적이고 단순하다. 즉 뒤에서 살펴보겠지만 논증 탐지는 인용문으로, 논증 간 연결은 문서 공동출현이나 동일 정보원의 최근 발언 여부, 그리고 문장 유사도로 처리할 수 있다. 다만 이 연구에서 제시하는 뉴스 문장 연결망 분석 알고리즘과 그를 구현한 프로토타입 프로그램은 문장 간의 연결을 매끄럽게 하는 스타일 측면이나 논조, 인터페이스 측면은 빠져있다. 그러나 IBM이 목표로 하는 디베이터가 인간과 자연스럽게 토론하는 논객을 만드는 것일 수도 있지만 현 단계에서는 대외적으로 공개되지 않았고, 사회과학자, 특히 언론학자로 많은 기사를 입장별로 묶어서 때로는 요약적으로, 때로는 상세하게 검토하는 데에는 뉴스 문장 연결망 분석 프로토타입이 부족하지 않은 성능을 제공한다. 또한 뉴스 특성상 다방면에 대중적인 논증을 제시하는 데에는 오히려 더 적합할 수도 있다.

표 2. IBM 디베이터에 적용되는 주요 기술

논증 발굴	관련 문서(relevant documents)에서 주장 탐지(Detecting claims)
	관련 문서의 근거 탐지(Detecting evidence)
	주장 반박(Negating claims)
	새로운 주장 합성(Synthesizing novel claims)
	코퍼스 전체에서 주장 탐지(Detecting claims throughout a corpus)
	코퍼스 전반의 주장 탐지(claim detection) 개선
	논증의 질 평가(Assessing argumentation quality)
	복수 텍스트 내 논증 연결(Relating arguments across texts)
입장 분류 및 감성분석	전문가 의견 입장의 결정(Determining expert opinion stance)
	주장 입장의 결정(Determining claim stance)
	입장 분류 및 감성 분석(Stance Classification and Sentiment Analysis)
	주장 입장 분류의 개선(Improving claim stance classification)
	감성 문구 분류(Classifying sentiment of phrases)
	감성 숙어 분류(Classifying sentiment of idioms)
딥 뉴럴넷과 약한 지도학습	논증의 점수화(Scoring arguments)
	자동 음성 인식(Automatic Speech Recognition, ASR), 이해, 출력
	문구 분절의 예측(Predicting phrase breaks)
	단어 및 구문의 강조(Emphasizing words and phrases)
	음성 패턴(speech patterns) 개선
	유사 문장 식별(Identifying similar sentences)
	논증 발굴(argument mining) 개선
	전체 코퍼스 내 주장 검색(Searching for claims)
	개념 추상성(concept abstractness) 결정
자연어 처리 알고리즘	관련 문서 식별(Identifying related documents)
	논증 구조 탐지(Detecting argumentative structures)
	자동 음성 인식(Automatic Speech Recognition, ASR), 이해, 출력
	유사한 문장 식별
텍스트 음성 변환 시스템	문구 분절의 예측
	단어 및 구문의 강조
	음성 패턴 개선
벤치 마크 데이터 세트	주석 요소(Annotated argument elements)
	복합어 관련성(Multi word term relatedness)
	위키화(Wikification)
	개념 관련성(Concept relatedness)
	관용 표현의 감성(Sentiment of idiomatic expressions)
	토론 연설(Debate speeches)
	듣기를 위한 논증적인 내용 (Argumentative content for listening comprehension)

3. 뉴스 문장연결망 분석 모형의 제안

이 연구는 뉴스 문장연결망을 수학적으로 모형화하고, 이 모형을 실제 데이터에 적용해 테스트한 뒤 모형을 정교화하고자 한다. 이 절에서는 뉴스 문장연결망 분석의 핵심 개념과 수학적 모델을 정의한다.

1) 인용문 중심의 뉴스 문장연결망 분석

이 연구에서는 문장, 특히 인용문의 뉴스 문장연결망 분석(news sentence network analysis)을 제안한다. 여기서는 동일 정보원 발언 여부와 기사 공동 출현, 그리고 유사도 기반의 혼합(hybrid) 방식을 이용한다. 인용문은 한국이나 미국처럼 객관주의 저널리즘 관행을 따르는 언론에서 가장 중요한 사실성 관행 중 하나이다(Van Dijk, 1988). 뉴스 문장연결망 분석은 특히 논쟁적 대화, 즉 토론에 대한 말뭉치를 제공할 수 있다. 우선 뉴스 기사는 기본적으로 논쟁적이며, 정치, 경제, 사회, 문화, 국제 등 다양한 범주의 사회문제를 다루고 다양한 의견과 의견의 근거가 되는 사실을 담고 있다(이준웅, 2010; Protess, 1992). 또한 언론은 흔히 후속 보도나 매체 간 의제설정(intermedia agenda setting)을 한다. 즉 중요한 의제에 대해 관련 기사를 연쇄적으로 작성한다(Roberts & McCombs, 1994). 뉴스 문장연결망 분석은 기사를 통한 사회적 논쟁의 맥락에 부합하도록 자동으로 연결망을 구성하고 연관성과 화제성에 따라 문장의 경로와 순위를 제공할 수 있어야 한다.

2) 뉴스 문장연결망의 정의

뉴스 문장연결망은 기사에서 쌍따옴표 안에 포함되는 개인실명직접인용문을 결점으로 한다. 연결은 정보원 발언, 기사 공동 출현, 유사도를 결합한 관련도를 기준으로 부여한다.

우선 동일 정보원의 발언 행렬은 '인용문×기사'의 무방향 2원 행렬

(undirected 2 mode matrix)로 제시된다. 이를 전치행렬과 행렬곱셈하면 동일 정보원의 발언에 따른 인용문 행렬이 도출된다. 이를 Qr라고 하자. 기사를 ai, 인용문을 qj, 정보원을 sk라고 할 때 기사 a1 에 정보원 s1 의 발언 q1과 s2의 발언 q2, q3가, 기사 a2에 s2의 발언 q4, s3의 발언 q5가 공동인용 됐다면 Qr은 〈그림 1〉과 같다.

그림 1. 기사 정보원 인용문 행렬

			a_1			a_2	
			s_1	s_2		s_2	s_3
			q_1	q_2	q_3	q_4	q_5
a_1	s_1	q_1	1				
	s_2	q_2		1	1		
		q_3		1	1		
a_2	s_2	q_4				1	
	s_3	q_5					1

그런데 인용문은 여러 날짜에 걸쳐 수집될 수 있다. 같은 날짜의 동일 정보원 발언은 관련성이 높을 것이다. 그러나 그러한 인용문은 너무 적다. 사실 관련성이 높은 다른 동일 정보원 발언이 다른 날짜에서도 나타날 수 있다. 그러나 날짜 범위가 너무 길어지면 동일 정보원의 발언이라고 해도 관련성은 떨어지게 될 것이다. 이 연구에서는 2일 이내, 즉 인접한 날짜의 동일 정보원 발언인 경우 관련성이 있다고 가정했다. 프로그램에서는 연구자가 날짜 범위 값을 지정할 수 있게 만들었다.

다음으로 인용문의 기사 공동 출현 행렬 역시 '인용문×기사'의 무방향 2원 행렬로 나타난다. 이를 앞서의 경우와 같이 행렬연산하면 기사 공동 출현에 의한 인용문 행렬이 도출된다. 이를 행렬 QA라 하자. 기사 ai에 인용문 q1, q2, q3가, 기사 a2에 인용문 q4, q5가 공동인용됐다면 행렬 QA는 〈그림 2〉와 같다.

그림 2. 기사 공동 출현 기준 인용문 행렬

		a_1			a_2	
		q_1	q_2	q_3	q_4	q_5
a_1	q_1	1				
	q_2		1	1		
	q_3		1	1		
a_2	q_4				1	
	q_5					1

끝으로 유사도 행렬은 인용문을 형태소분석하고 품사 부착을 한 뒤 명사 기준으로 벡터화한 다음, 자카드 유사도(Jaccard similarity)를 구한다. 참고로 두 벡터집합 A와 B를 비교할 때, 자카드 유사도 공식은 아래와 같다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \qquad (1)$$

문장 간 유사도 값은 크게 세 유형이 있을 수 있다. 첫째, 완전 일치이다. 즉 비교한 두 문장은 동일하다. 이 경우 두 문장은 하나의 결점으로 합친다. 둘째, 완전 불일치이다. 이 경우 두 문장은 별도의 결점으로 간주하며, 유사도에 의한 관련도도 없기 때문에 유사도 기준으로 두 문장 간 연결도 부여하지 않는다. 셋째, 유사도 값이 완전 일치와 완전 불일치 사이에 있는 경우다. 유사도가 높다는 것은 그만큼 비슷한 내용을 다뤘다고 볼 수 있기 때문에 관련도가 높고 유사도가 낮으면 그 반대로 볼 수 있다. 따라서 완전 일치는 아니지만 일정 값 이상의 유사도가 있는 두 문장은 관련도가 있다고 간주하고 연결을 부여할 수 있다.

자카드 유사도의 경우 값이 0이면 완전 불일치, 1이면 완전 일치를 뜻한다. 따라서 자카드 유사도가 1인 경우 인용문을 중복으로 간주해 하나의 결점으로 합치고, 0이면 별개의 결점으로 간주하고, 일정 값 이상이면 연결을 부여한다. 만일 인용문 q3와 q5의 명사벡터가 많이 겹치고 q2

가 q1, q3과 약간, 그리고 이보다 q1가 q5와 좀 더 유사하다면 유사도 QS 행렬은 예컨대 〈그림 3〉과 같을 수 있다.

그림 3. 인용문 유사도 행렬

	q_1	q_2	q_3	q_4	q_5
q_1	1.00	0.20			0.43
q_2	0.20	1.00	0.34		
q_3		0.34	1.00		0.89
q_4				1.00	
q_5	0.43		0.89		1.00

행렬 Qr과 QA의 원소 값은 0 또는 1인 이원(binary) 값인 반면, 행렬 QS의 원소 값은 0과 1 사이의 실수이다. 따라서 행렬연산을 위해 QA의 유사도를 절삭 값(threshold)에 따라 0과 1의 이원 값으로 변환하여 유사도 행렬 QS′를 〈그림 4〉와 같이 최종 도출한다. 여기서는 0.450로 가정했다. 프로그램에서 절삭 값은 연구자가 임의로 정할 수 있도록 개발했다.

그림 4. 변환된 인용문 유사도 행렬

	q_1	q_2	q_3	q_4	q_5
q_1	1				
q_2		1			
q_3			1		1
q_4				1	
q_5			1		1

최종적인 관련도 기준 인용문 행렬 Q는 다음과 같은 행렬연산으로 구한다.

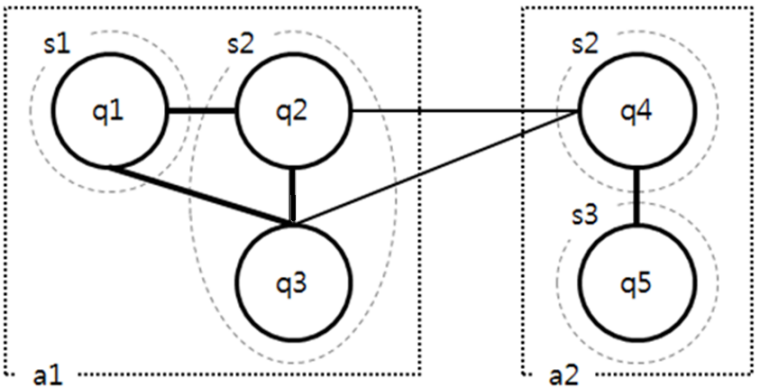
$$Q = Qr + QA + QS' \qquad (2)$$

〈그림 1〉에서 〈그림 4〉까지의 사례에 해당하는 행렬을 덧셈해 최종 도출한 관련도 행렬은 〈그림 5〉와 같다. 관련도 행렬은 서로 다른 종류의 행렬을 합산해 도출한 것이다. 따라서 관련도 행렬의 각 원소의 값은 '0'은 그대로, 1' 이상이면 '1'로 절삭해 기입한다. '0'은 동일 정보원 발언, 기사 공동 출현, 유사도 측면에서 관련도가 없음을, '1'은 최소한 하나 이상의 측면에서 관련도가 있음을 뜻한다. 이에 따른 뉴스 문장연결망은 〈그림 6〉과 같이 나타낼 수 있다. 굵은 선(q1 q2 q3, q4 q5)은 기사 공동 출현, 가는 선(q2 q3 q4)은 동일 정보원 발언, 복선(q3 q5)은 유사도에 의한 연결을 나타낸다.

그림 5. 관련도 행렬

	q_1	q_2	q_3	q_4	q_5
q_1	1				
q_2		1	1		
q_3		1	1		1
q_4				1	
q_5			1		1

그림 6. 뉴스 문장연결망 예시



3) 뉴스 문장연결망에서 중심, 요약, 상술의 정의

문장은 순서가 바뀌면 의미 전달이 어색해진다. 이는 문장이 의미경로와 의미거리를 갖고 있음을 시사한다. 또한 모든 기사는 소위 '야마'라고 하는, 한 문장으로 기술되는 중심문장을 갖는다(박창섭, 2010). 또 다소 긴 기사는 중심문장과 관련되면서 '꼭지'로 구분되는 요약 문장으로 나타낼 수도 있다. 이 밖에 중심문장과 요약 문장은 아니지만 내용을 부연하는 상술 문장들이 있다.

이 절에서는 뉴스 문장연결망 분석을 통해 의미군집, 의미경로, 의미거리, 중심문장, 요약, 상술을 수학적으로 정의하고자 한다. 첫째, 의미군집은 전체 뉴스 문장연결망을 구성하면서, 서로 연결되지 않은 작은 연결망들이다. 각 의미군집은 검색어를 공통된 주제로 다루기는 하지만, 그 검색어를 중심에 놓고 다양하고 심도 깊게 다룰 수도 있고 다른 의제의 하위 주제로 언급하거나 단편적으로 간단히 다룰 수도 있다. 전자의 경우 문장 연결망에서 주요 의미군집으로 나타날 것이다.

둘째, 문장 연결망에서 의미거리는 두 문장 간의 최단경로(shortest path)이다. 최단경로를 설명하면, 우선 두 결점 v와 v′ 간의 경로 P는 {v1, v2, …, vn}, v=v1, v′=vn로 나타낼 수 있다. 이 때1≤i≤n인 i에 대해 vi과 vi+1는 인접한다. 두 결점 vi, vj 간의 거리함수를 f(ei,j)라고 할 때, 최단경로는 모든 가능한 n에 대해 아래의 거리함수 f를 최소화하는 경로이다.

$$\sum_{i=1}^{n-1} f(e_{i,i+1}) \qquad (3)$$

셋째, 의미경로는 한 문장에서 시작해 거리가 1인 문장들을 연쇄적으로 이어감으로써 만들어진다. 의미경로상의 문장들은 의미거리가 먼 문장 간에는 관련도가 있을 수도 없을 수도 있지만 최소한 인접한 문장

간에는 관련도가 있는 일종의 가족유사성(family resemblance)을 갖는다. 즉 인접 문장은 같은 기사에 등장하거나, 같은 정보원이 발언하거나, 유사한 문장이다.

넷째, 중심문장은 문장 연결망의 중심으로 연결정도 중앙성, 즉 연결수로 구한다. 이는 중심문장을 밀접하게 관련된 문장, 즉 인접 문장이 가장 많은 문장으로 정의한다는 뜻이다. 문장연결망 그래프를 결점과 연결에 대한 함수 G=(V, E)라고 하면, 연결정도 중앙성 공식은 아래와 같다.

$$C_D(v) = \deg(v) \qquad (4)$$

다섯째, 요약 문장은 요약경로 상의 문장으로 정의한다. 요약경로는 문장연결망의 지름(diameter)으로 찾을 수 있다. 지름은 가장 긴 최단경로이다. 특히 이 연구에서는 지름 중 중심문장을 지나는 경우를 요약경로로 간주한다. 보통 중심문장은 지름의 양 끝이 아닌 결점에 위치한다. 만일 요약 문장을 제시할 때 중심문장을 가장 먼저 제시한다면, 요약경로는 중심문장에서 2개 이상의 의미경로로 제시될 수 있다. 또한 문장 연결망에서 지름은 보통 여러 개가 있다. 때문에 이 연구에서는 문장별로 연결정도 중앙성을 계산한 뒤 해당 지름에 속한 문장의 연결정도 중앙성 합산 값이 큰 지름을 더 중요한 요약경로로 간주한다. 정리하면 요약 문장은 중심문장에서 시작하여 지름 끝에 있는 문장으로 요약경로에 따라 순서대로 제시한다.

끝으로 상술 문장은 지름에 속하지는 않지만, 최단 경로 상에 속한 어떤 요약 문장과 완전 연결망, 즉 의미파당(semantic clique)을 구성하며 연결된 문장들이 있을 수 있다. 이는 해당 요약 문장을 상술하는 상술 문장으로 정의할 수 있다.

4. 뉴스 문장연결망 분석 프로그램 '쿼트넷'의 개발

이상을 바탕으로 연구진은 뉴스 문장연결망 분석 프로토타입 프로그램 〈쿼트넷〉을 만들었다. 링크된 파일에는 이 연구의 분석 결과도 예시로 포함돼 있다. 프로그램은 파이썬(python)으로 제작됐다. 형태소분석을 위한 konlpy, 연결망 분석을 위한 networkx를 비롯해 xlrd, click, tqudm과 같은 라이브러리가 활용됐다. 〈쿼트넷〉은 〈빅 카인즈〉에서 얻은 1개 이상의 엑셀 파일 형태의 메타데이터를 입력하면, 인용문을 형태소분석한 결과를 필요한 메타데이터와 함께 CSV 파일 형태로 변환해 제공해준다. 또 기사 공동출현, 동일 정보원 발언, 자카드 유사도에 따라 연결된 문장 정보를 제공해준다. 이를 별도의 연결망 분석 프로그램을 활용하여 연결정도 중앙성이나 경로 등을 구하면, 중심문장, 요약 문장, 상술 문장 등을 파악하고 시각화할 수 있다. 〈쿼트넷〉은 중심, 요약, 상술의 문장을 효과적으로 검토하기 위해 문장을 일종의 트리 구조로 보여주는 시각화 기능도 갖추었다(〈그림 7〉부터 〈그림 12〉까지 참조). 간단히 기능을 설명하면 먼저 서로 다른 군집에 속하면서 연결정도 중앙성 값이 높은 인용문 상위 10개를 중심문장으로 제시한다. 이어 중심문장 중 하나를 선택하면 중심문장이 포함된 의미경로를 제시한다. 끝으로 의미경로상의 문장 중 하나를 클릭하면 해당 문장을 포함하는 상술 문장을 보여준다. 또 연결망의 분포 등을 파악하기 쉽도록 연결망분석 프로그램 Gephi에서 활용할 수 있는 파일도 생성한다. 필요한 경우 산출된 엑셀 파일을 활용해 익숙한 UCINET이나 NETMINER에서 시각화할 수도 있다. 사용자는 〈표 3〉, 〈표 4〉와 같이 분석 항목과 관련도를 결정하는 계수 값을 지정할 수 있다. 각 계수의 디폴트 값은 영역지식(domain knowledge)을 활용해 임의로 책정했다. 명령어 옵션과 변수 옵션에 대한 파이썬 명령어는 아래와 같다.

표 3. 명령어 옵션

옵션	설명	디폴트 값
i	정보원이 있는 열의 번호	0
q	인용문이 있는 열의 번호	7
a	기사 ID가 있는 열의 번호	9
d	기사 날짜가 있는 열의 번호	10
quote id	인용문 ID가 있는 열의 번호	8
c	카테고리가 있는 열들의 번호	12, 13, 14
p	매체명이 있는 열의 번호	11
s	출력 디렉토리 접미사	_output

표 4. 변수 옵션

옵션	설명	디폴트 값
i		2 (이틀)
a	동일 기사인용문 사이 연결설정 안 함	False (설정 중)
s	유사도 역치 값	0.450
m	유사도 척도	Jaccard
n	동일 인용문 제거하지 않음	False (제거 중)
e	연결정도 절삭 값	1 (고립자 제고)
f	분석 기간 시작일	1990 01 01
t	분석 기간 종료일	2016 04 30

명령어 옵션: python xls2csv.py 〔명령어옵션〕〔옵션값〕〔폴더명〕
변수 옵션: python construct.py 〔변수옵션〕〔변수값〕〔폴더명〕

이 연구에서는 인용문을 결점으로, 관련도를 연결로 하는 뉴스 문장 연결망을 분석하기 위해 관련도를 기사 공동출현, 동일정보원 발언 여부, 자카드 유사도에 따라 정의해 연결을 부여했다. 또 뉴스 문장연결망에서 의미군집, 의미거리, 의미경로, 중심문장, 요약 문장, 상술 문장을 군집화, 맨하튼 거리, 연결정도 중앙성 등을 이용해 수학적으로 정의했다. 그

리고 이를 구현한 뉴스 문장 연결망 분석 프로그램 〈쿼트넷〉을 개발했다.

이 논문에서는 〈쿼트넷〉을 테스트하기 위한 시행연구를 실시했다. 분석 대상은 1990년 1월 1일부터 2016년 4월 30일까지 〈빅카인즈〉에서 '인공지능'으로 검색된 기사 2,337개에서 추출된 인용문 5,046개이다(기사 id 및 문장 id 기준으로 중복 없음). 정보원 수는 2,029명(중복 제거)이었다. 전국일간지, 경제지, 지역신문 등 〈빅카인즈〉에서 서비스하는 30개 매체(경남도민일보, 경남신문, 경상일보, 경향신문, 광주일보, 국민일보, 국제신문, 내일신문, 대전일보, 동아일보, 매일경제, 매일신문, 무등일보, 문화일보, 부산일보, 서울경제, 서울신문, 세계일보, 영남일보, 전남일보, 전북일보, 중도일보, 중부매일, 충북일보, 파이낸셜뉴스, 한겨레, 한국경제, 한국일보, 한라일보, 헤럴드경제)를 대상으로 분석했다. 다만 어떤 연도에는 특정 매체에 '인공지능'이 포함된 기사가 1개도 없는 경우도 있었다.

〈빅카인즈〉 데이터 중 뉴스 문장연결망 분석을 위해 활용한 데이터는 〈표 5〉와 같다. 정보원명(INFOSRC)은 개체명 인식기를 통해 인명, 기관명, 직함을 추출해 얻어지며 인용문 (STN_CONTENT)도 자연어처리를 통해 추출해야 한다. 인용문 성능은 재현율이 공개돼 있지 않기 때문에 이 연구에서는 여러 매체에서 다년간 데이터를 수집해 데이터 양을 늘림으로써 분석의 타당성을 최대한 보완했다.

표 5. 〈빅카인즈〉 데이터

정보		필드 이름	내용
결점	인용문	STN_ID	인용문 ID
연결	기사공동출현	ART_ID	기사ID
	일정 기간 내 동일정보원 발언	INFOSRC	정보원명
		ART_DATE	기사 게재일
	유사도	STN_CONTENT	인용문 본문

인용문 간 연결은 기사 공동출현 여부 외에 이틀 간 동일 정보원 발언인지 여부, 유사도로 부여했다. 유사도 절삭 값은 0(완전 불일치)과 1(완전 일치)를 제외한 전체 문장 간 유사도 값들의 중간 값(median)인 0.333 초과와 디폴트 값인 0.450 초과 등 2개로 설정했다. 0.450는 과거 카인즈 데이터를 단어 수준에서 군집화한 기존 연구를 참고해 제시한 값이다(신유현 등, 2013). 다만 이 값은 참고 값일 뿐 단어와 문장은 분석 수준이 다르기 때문에 사실상 임의로 정한 값으로 추가 분석을 통해 정확한 유사도 계수를 찾을 필요가 있다. 테스트를 위한 시행연구의 연구문제는 아래와 같다.

- 연구문제 1: 인공지능 관련 기사를 뉴스 문장연결망 분석했을 때 추출되는 요약 문장, 상술 문장, 상세문장은 중요도와 관련도에 따라 추출되고 배열되는가?

중요도와 관련도는 인간끼리도 서로 다르게 판단할 수 있다. 인간과 기계도 마찬가지다. 일반적으로 특정 주제에 대해 중요하지 않은 문장이나 관련되지 않은 연결을 판단할 때는 의견 차이가 적을 수 있다.

- 연구문제 2: 인공지능 관련 기사를 뉴스 문장연결망 분석했을 때 관련도를 적절하게 반영하는 유사도 계수 값은 얼마인가?

정확한 계수 값은 매체 유형이나 주제, 시기 등에 따라 달라질 수도 있고 그렇지 않을 수도 있다. 이 연구에서는 유사도 계수를 0.333과 0.450으로 놓고 차이를 살펴보도록 한다.

5. 테스트 결과

1) 유사도 계수가 0.333인 경우

분석 결과 고립자(isolated node)를 제외하고 3,742개 결점 6,708개의 연결로 이루어진 문장연결망이 도출됐다. 참고로 기사 공동출현에 의한 연결은 6,141, 동일정보원에 의한 연결은 6,623개, 유사도에 의한 연결은 1,814개가 생성됐다. 이들 합은 실제 연결 수보다 많다. 이는 복수의 요인에 의해 연결이 생성될 수 있는 데다가 완전 일치한 문장을 하나의 결점으로 합치면 연결 수가 줄기 때문이다. 이밖에 의미군집은 997개였다. 연결망의 지름은 18, 평균 군집 계수(average clustering coefficient)는 0.873, 평균 경로 길이(average path length)는 6.657이었다.

연결정도 중앙성이 높으면서 각 군집을 대표하는 중심문장 중 상위 10위권에 해당하는 문장은 〈그림 7〉과 같다. 문장 순서는 중요도 순서이다. 일단 각 문장의 내용은 뒤의 상술 문장과 함께 자세히 검토하겠다.

그림 7. 인공지능 관련 기사의 중심문장

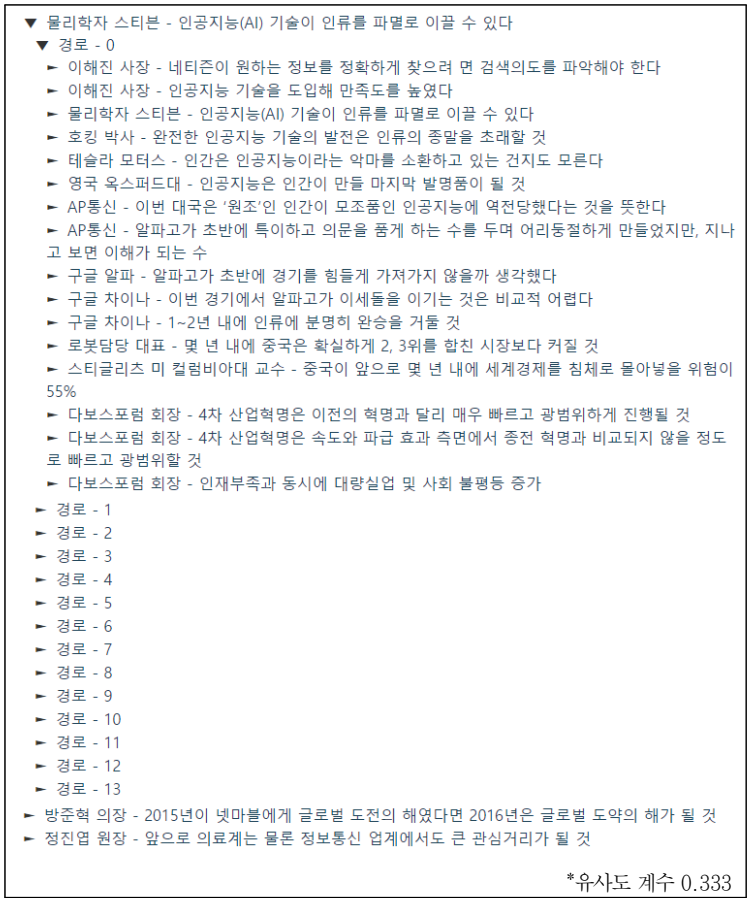
► 물리학자 스티븐 - 인공지능(AI) 기술이 인류를 파멸로 이끌 수 있다
► 방준혁 의장 - 2015년이 넷마블에게 글로벌 도전의 해였다면 2016년은 글로벌 도약의 해가 될 것
► 정진엽 원장 - 앞으로 의료계는 물론 정보통신 업계에서도 큰 관심거리가 될 것
► 맥아피 수석연구원 - 소비자들 입장에선 세계 최고 제품이나 서비스도 그렇게 비싸지 않은 가격에 이용할 수 있는데, 굳이 2위와 3위 제품을 써야 할 이유가 없지 않으냐
► 오준호 교수 - 지금은 한국 로봇산업 중흥의 기회로 정부가 장기적으로 육성.지원하려는 의지가 중요하다
► 이조원 단장 - 우리의 나노기술은 미국.일본에 비해 25% 수준이지만 반도체 공정기술만은 세계 최고 수준
► 현대모비스 - 이러한 경쟁력을 바탕으로 친환경 LED 헤드램프를 일반 차종까지 보급화시키는 한편 해외 완성차 업체에도 수출할 계획
► 국민회의 정호선 의원 - 114안내 서비스에 대한 국민들의 불만이 고조되는 상황에서 한국통신이 유료화에만 신경쓰는 것은 공사로서 취할 태도가 아니다
► 박현구 대표 - 자동차 최악의 '연료소비' 주범 찾았다
► 김원종 한국IBM 수석 부사장 - 센터는 파트너를 위한 공간으로, 고객 및 시장에 적극적이고 구체적인 솔루션을 갖고 다가서자는 부분이 가장 크다

*유사도 계수 0.333

다음으로 전체 문장연결망에서 가장 중요한 '물리학자 스티븐 (호킹)'의 인용문을 포함하는 의미경로 상의 요약 문장은 〈그림 8〉과 같다. 인공지능에 대한 우려를 담은 중심문장에서 시작할 경우, 아래쪽 문장은 가까운 두 문장은 중심문장과 비슷한 직접적 우려를, 다음 다섯 문장은 우려와 함께 '알파고 대국'에 대한 논의를 담았다. 이어진 세 문장은 중국 시장 관련 논의를, 그 다음 문장은 4차 산업혁명이 중심이 되는 논의를 담았다. 한편 위쪽의 문장은 이해진 사장의 2000년 보도인데, 인공지능에 대한 낙관론으로 중심문장에 대한 반론이다.

정리하면 의미거리가 먼 인용문 간은 논의가 다소 달라지지만 인접 인용문 간의 논의는 밀접하게 관련되며, 그러면서도 큰 틀에서도 알파고 대국 이후 인공지능의 위협과 가능성이라는 주제를 다루고 있음을 확인할 수 있다. 즉 요약 문장 간에 가족유사성을 확인할 수 있는 것이다.

그림 8. 인공지능 관련 기사의 요약 문장



이어 물리학자 스티븐의 인용문을 상술하는 문장을 살펴보면 〈그림 9〉와 같다.

그림 9. 인공지능 관련 기사의 상술 문장

▼ 물리학자 스티븐 - 인공지능(AI) 기술이 인류를 파멸로 이끌 수 있다
▼ 경로 - 0
► 이해진 사장 - 네티즌이 원하는 정보를 정확하게 찾으려 면 검색의도를 파악해야 한다
► 이해진 사장 - 인공지능 기술을 도입해 만족도를 높였다
▼ 물리학자 스티븐 - 인공지능(AI) 기술이 인류를 파멸로 이끌 수 있다
S I SK텔레콤 - 이런 기술이 축적되면 추후 데이터 분석 기술과 연계해 인공지능에도 접목될 수 있다
S I SBS 뉴스 - 인공 지능과 싸워도 제가 이길 수 있다고 생각한다
S I 환경기술연구소 상무 - 인공지능 기술의 발달로 영화에서나 상상할 수 있었던 기능들이 속속 제품에 구현되고 있다
S I 호킹 박사 - 완전한 인공지능 기술의 발전은 인류의 종말을 초래할 것
S I I 물리학자 스티븐 - 완전한 '인공지능'의 개발이 인류의 멸망을 불러올 수 있다
S I 중구 환경재단 - 인공지능(AI)은 인류의 미래에 대한 가장 위협적인 기술
S I 박홍준 투자본부장 - BT, NT, 인공지능 로봇 기반기술은 서로 융합될 수 있는 미래산업
I I 영국 BBC - 생각하는 로봇 개발을 위한 완전한 인공지능의 등장은 인류의 멸망을 가져올지 모른다
S I 스티븐 호킹 박사 - 인공지능은 인류 문명을 위협할 재앙을 불러올 수 있다
S I 호킹 박사 - 인공지능이 인류의 종말을 불러올 수도 있다
I A 영국 BBC방송 - 테러 위협에 맞서는 인터넷 기업들의 노력이 필요하지만, 개인의 자유와 사생활을 침해하지 않는 방법을 찾아야 하는 어려움이 따른다
S I IBM 회장 - 사람처럼 데이터를 이해하고 추론하는 인공지능기술이 필요하다
I A 영국 BBC방송 - 인간 능력에 필적하거나 이를 뛰어넘는 AI가 등장할 가능성에 두려움을 느낀다
S I 스티븐 호킹 박사 - 완전한 AI 기술의 발전은 인류의 종말을 초래할 수 있다
S I 이해진 사장 - 인공지능 기술을 도입해 만족도를 높였다
S I 네이처 - 인공지능 기술이 사람 수준의 능력에 도달할 수 있는 가능성을 보여줬다
S I 호킹 박사 - 인공지능이 인류 멸망 가져올 수 있다
S I IBM - 실생활에 사용할 수 있는 인공지능을 만들겠다.
I A 영국 BBC방송 - 생각하는 로봇 개발을 위한 완전한 AI의 등장은 인류의 멸망을 가져올지 모른다
I A 영국 BBC방송 - 인터넷이 '양날의 칼'과 같은 존재가 되고 있다
S I 물리학자 스티븐 - 인공지능의 발전은 인류의 종말을 가져올 수 있다
I I 물리학자 스티븐 - 지금까지의 초기 인공지능 기술은 유용성을 충분히 입증했다면서도 인간 능력에 필적하거나 이를 뛰어넘는 인공지능이 등장할 가능성에 두려움을 느낀다
I A 영국 BBC방송 - 지금까지의 AI 기술은 유용성을 충분히 입증했다
S I 이준표 대표이사 - 한국의 인공지능형 네비게이션 기술력을 전 세계에 알릴 수 있어 자랑스럽다

*유사도 계수 0.333

각 군집의 중심문장, 요약 문장, 상술 문장을 〈그림 7〉, 〈그림 8〉, 〈그림 9〉와 같은 방식으로 세부적으로 살펴봤다. 10개의 중심문장과 그에 속한 요약 문장과 상술 문장들은 상위 10위권 군집을 대표한다. 1위 군집은 주로 알파고 이후 인공지능의 위협을 다뤘다. 이 주요 군집(main component)의 크기(size), 즉 인용문 수는 594개였고, 정보원은 213명이었다. 이는 234개의 기사에서 추출됐다. 이어 2위 군집은 인공지능

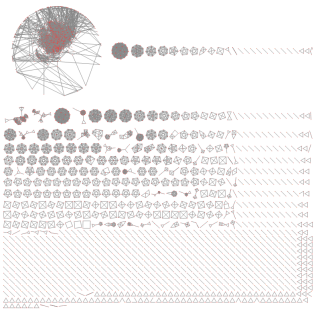
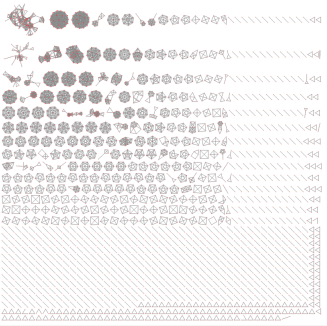
등을 활용하는 넷마블의 전략을, 3위와 4위 군집은 의료 기술과 일자리 감소를, 5위부터 7위까지 군집은 로봇 기술, 나노 기술, 자동차 분야 등에서의 인공지능 활용 등을 논의했다. 8위 군집은 1개의 기사로 이뤄져 있는데 주제가 혼재돼 있다. 구체적으로 보면 동아일보 1996년 국감 기사로 통신과학기술위에서 인공지능이 언급된다. 그런데 농림수산위, 건설교통위 내용이 통신과학기술위의 내용과 함께 언급된다. 때문에 인공지능과 관련이 낮은 인용문인데도 해당 기사에 많은 인용문이 등장해 가중치가 높아지면서 상위권 의미군집으로 제시됐다. 9위 군집은 현대모비스와 헬스캐어에 관련 인용문이 혼재돼 있다. 10위 군집은 1개의 기사로 이루어져 있는데 IBM 왓슨 서울 클라이언트 센터를 다룬다.

정리하면 제시된 10개 군집은 서로 구분되는 주제를 다루고 있다. 즉 군집화가 잘 됐다. 특히 1위 군집은 서로 다른 많은 기사의 많은 정보원과 인용문들에서도 관련도 높은 인용문끼리 적절하게 묶고 있음을 확인할 수 있었다. 한편 위 이하의 군집은 기사 수는 크게 줄어서 1위 군집보다 쉽게 요약, 상술된다. 다만 8위권 이하의 군집에서는 관련 없는 주제들이 함께 묶이기도 하는데, 이는 기사 자체가 예외적으로 관련도가 낮은 주제를 하나의 기사로 다루기 때문이다.

2) 유사도 계수가 0.450인 경우

한편 유사도 기준을 0.333에서 0.450로 높이면 연결이 줄고 고립자가 늘면서 고립자를 제외한 결점 수가 줄어든다. 실제로 유사도 계수를 디폴트 값인 0.450로 했을 때 고립자를 제외한 결점은 3,697개, 연결은 6,383개로 나타났다. 의미군집은 더 파편화되면서 1,120개로 늘었고, 최대 군집의 크기가 작아지면서 연결망의 지름은 9로 짧아졌다. 각각의 군집이 작아진 한편 구집 수가 많아지면서 평균 군집화 계수는 0.936로 높아지고, 평균 경로 길이는 2.164로 짧아졌다. 유사도 기준이 0.333인 경우와 0.450인 경우의 전체 연결망을 시각화해 비교하면 〈그림 10〉과 같다.

그림 10. 유사도 기준별 뉴스 문장 연결망 비교

유사도 기준	0.333	0.450
연결망 시각화		
분포 특성	고립자 제외 결점 수: 3,742 고립자 제외 연결 수: 6,708 고립자 제외 의미군집 수: 997 지름: 18 평균 군집계수: 0.873 평균 경로길이: 6.657	고립자 제외 결점 수: 3,697 고립자 제외 연결 수: 6,383 고립자 제외 의미군집 수: 1,120 지름: 9 평균 군집계수: 0.936 평균 경로길이: 2.164

중심문장은 〈그림 11〉과 같이 나타난다.

그림 11. 인공지능 관련 기사의 중심문장

► AFP통신 - 컴퓨터가 최고수와의 5번기 가운데 첫 승리를 낚았다
► 호킹 박사 - 인공지능이 인류 멸망 가져올 수 있다
► 방준혁 의장 - 2015년이 넷마블에게 글로벌 도전의 해였다면 2016년은 글로벌 도약의 해가 될 것
► 하성민 SK텔레콤 사장 - ICT노믹스로 5G 이통시장 선점할것
► 맥아피 수석연구원 - 기술이 발전하면서 이 기술을 활용하는 소수 슈퍼스타들이 부를 독차지 하고 있다
► 서울대 이준환 교수 - 현재 국내 언론시장에서 존재하고 있는 사람 기자가 해서는 안 되는 일 중 하나가 단순히 인터넷 클릭수를 높이기 위해 검색어 위주의 기사를 작성하는 일인데, 필요하다면 이런 일은 로봇 기자가 맡을 수도 있다
► 삼성동 코엑스 - 인공지능을 두려워하지 말고 새로운 변화를 사회와 경제 속에서 포용해야 한다
► 대국 후 - 정말 무엇과도 바꾸지 않을, 값어치를 매길 수 없는 1승
► 권영 도 박사 - 대기업 중에서 국내 로봇업체를 리드하는 기업이 아쉽다
► 이수영 센터장 - 2030년 정도면 어느정도의 지능을 갖춘 인공지능 컴퓨터나 인조인간을 만들어 낼 수 있을 것

*유사도 계수 0.450

AFP통신의 인용문을 포함하는 요약 문장은 〈그림 12〉와 같다. 유사도 계수가 0.333인 경우의 요약 문장보다 알파고 대국 기사로 보다 집

중돼 있다. 또한 미국 NBC 방송을 상술하는 문장은 〈그림 13〉과 같다.

그림 12. 인공지능 관련 기사의 요약 문장

▼ AFP통신 - 컴퓨터가 최고수와의 5번기 가운데 첫 승리를 낚았다
▼ 경로 - 0
► 미국 NBC방송 - 알파고의 승리로 AI의 한계를 넓혔다
► 신화통신 - 인간이 컴퓨터를 지배할 수 있는 마지막 게임으로 여겨진 바둑에서 처음으로 진 경기로 기록됐다
► AFP통신 - 컴퓨터가 최고수와의 5번기 가운데 첫 승리를 낚았다
► AFP통신 - 알파고가 '직관력을 갖춘 인공지능'으로서 세계를 까무러치게 할 신고식을 치른 셈
► AFP통신 - 인간 바둑 챔피언이 슈퍼컴퓨터를 상대로 놀라운 승리를 거뒀다
► 중국 신화통신 - 인간 바둑 챔피언이 3연패 끝에 마침내 인공지능을 이겼다
► 일본 마이니치 - 알파고가 전날까지는 이 9단을 압도했지만 이 날은 이 9단의 승부수에 알파고의 수가 갑자기 흐트러졌다
► 호킹 박사 - 인공지능이 인류 멸망 가져올 수 있다
► 방준혁 의장 - 2015년이 넷마블에게 글로벌 도전의 해였다면 2016년은 글로벌 도약의 해가 될 것
► 하성민 SK텔레콤 사장 - ICT노믹스로 5G 이통시장 선점할것
► 맥아피 수석연구원 - 기술이 발전하면서 이 기술을 활용하는 소수 슈퍼스타들이 부를 독차지 하고 있다
► 서울대 이준환 교수 - 현재 국내 언론시장에서 존재하고 있는 사람 기자가 해서는 안 되는 일 중 하나가 단순히 인터넷 클릭수를 높이기 위해 검색어 위주의 기사를 작성하는 일인데, 필요하다면 이런 일은 로봇기자가 맡을 수도 있다
► 삼성동 코엑스 - 인공지능을 두려워하지 말고 새로운 변화를 사회와 경제 속에서 포용해야 한다
► 대국 후 - 정말 무엇과도 바꾸지 않을, 값어치를 매길 수 없는 1승
► 권영 도 박사 - 대기업 중에서 국내 로봇업체를 리드하는 기업이 아쉽다
► 이수영 센터장 - 2030년 정도면 어느정도의 지능을 갖춘 인공지능 컴퓨터나 인조인간을 만들어 낼 수 있을 것

*유사도 계수 0.450

그림 13. 인공지능 관련 기사의 상술 문장

▼ AFP통신 - 컴퓨터가 최고수와의 5번기 가운데 첫 승리를 낚았다
▼ 경로 - 0
▼ 미국 NBC방송 - 알파고의 승리로 AI의 한계를 넓혔다
I A 신화통신 - 이세돌이 알파고에 '충격패'를 당했다
I A 신화통신 - 인간이 컴퓨터를 지배할 수 있는 마지막 게임으로 여겨진 바둑에서 처음으로 진 경기로 기록됐다
I A 영국 BBC방송 - 이세돌이 경기 도중 불안해하고 한숨을 쉬며 고개를 젓는 등의 행동을 보였다
I A AP통신 - 이세돌 9단의 패배는 한국인들은 물론 전 세계 바둑팬들에게는 충격적으로 받아들여지고 있다
I A AP통신 - 아직 4차례의 대국이 더 남았지만 가장 창의적이고 복잡한 게임으로 여겨지는 바둑에서 AI 알파고가 이세돌을 이긴 것은 매우 의미심장한 일
I A 영국 BBC방송 - 이세돌이 우위를 점하는 듯했으나 경기 종료 20분을 남기고 알파고가 난공불락의 리드를 했다

*유사도 계수 0.450

정리하면 전체 연결망에서 가장 중요한 중심문장을 포함하는 가장 큰 주요군집을 상술 문장까지 검토했을 때 알파고 대국에 집중돼 있음을

알 수 있었다. 여기에는 알파고의 승리와 패배가 모두 담겨 있다. 한편 인공지능의 위협에 대한 논의는 별도 군집으로 묶여서 두번째 중심문장으로 제시된다. 유사도 계수 값을 높였기 때문에 군집이 분할된 것으로 이해할 수 있다.

7. 결론

이 연구는 뉴스를 단어 중심의 의미연결망 분석을 통해 연구할 때 나타나는 한계를 극복하고자, 문장 수준의 의미연결망 분석으로서 뉴스 문장연결망 분석 방법을 제안했다. 구체적으로는 뉴스 문장연결망의 결점과 연결, 의미거리와 의미경로, 중심과 요약 및 상술 등의 개념을 정의했다. 또 뉴스 문장연결망 분석 프로토타입 프로그램인 〈퀴트넷〉을 개발했다. 끝으로 이를 이용해 인공지능 관련 기사에 대해 시행연구를 실시했다.

이 연구는 '인공지능'이라는 검색어를 포함한 기사의 인용문을 분석했다. 따라서 다른 주제에 대해서도 뉴스 문장 연결망 분석의 타당성을 검증할 필요가 있을 수 있다. 이 연구에서는 소개하지 않았지만, 〈퀴트넷〉 개발 이전에 넷마이너 등 기존 연결망 분석 솔루션을 활용해 '북한'이라는 검색어가 포함된 기사를, 〈퀴트넷〉 초기 버전을 활용해 '가습기 살균제'라는 검색어가 포함된 기사를, 〈퀴트넷〉을 활용해 '중국'이라는 검색어가 포함된 기사를 분석해 활용성을 확인했다(박대민, 2015; 박대민 등, 2016; 박대민, 2017).

형태소 수준의 연결망 분석은 의미 없는 결점이나 연결도 뉴스 의미연결망에 포함시켜 정교한 분석을 어렵게 했다. 너무 많은 결점과 연결이 있어서 이를 연구자가 자의적으로 빈도 기준으로 절삭하면서 사실상 데이터 기반의 성격을 잃어버리는 경우도 많다. 이를 보완한 것이 뉴스 정보원 연결망 분석이나 뉴스 주제 연결망 분석 등 개체명 수준의 의미연결

망 분석이었다. 그러나 개체명 수준의 의미연결망 분석만으로는 직관적인 분석이 어려웠다. 예컨대 주제 연결망 분석에서 주제들이 군집화될 때 그 의미를 직관적으로 파악하지 못해 공동출현한 두 주제어를 포함한 문장을 다시 찾아야 했다. 게다가 이 문장과 관련된 논의를 찾는 것은 더욱 막연한 일이었다. 뉴스 문장연결망 분석은 뉴스 정보원 연결망 분석과 뉴스 주제 연결망 분석, 뉴스 정보원-주제 연결망 분석 등 개체명 수준의 분석과 함께 맥락을 보다 직관적으로 파악하는데 도움을 줄 수 있다. 특히 최근 R이나 파이썬 등 프로그래밍 언어를 활용해 형태소 수준에서 자동화된 내용분석을 하는 사회과학자들이 늘어나고 있다. 문제는 이러한 분석을 위해 프로그래밍 언어를 배워야 한다는 점이다. 〈퀴트넷〉을 비롯한 의미 연결망 분석 프로그램은 프로그래밍을 할 줄 모르는 문화연구자라고 해도 자동화된 담론분석을 간편하게 수행할 수 있게 도움을 준다.

의미연결망 분석은 크게 결점, 연결, 방향성, 분포와 분석, 인식 성능 등 측면에서 발전시킬 수 있다. 첫째, 결점을 확장하는 방향으로 심화시킬 수 있다. 이 논문은 인용문만을 결점으로 간주해 분석했다. 그러나 기사에는 인용문만 있는 것이 아니다. 수치를 포함한 문장이나 장소와 관련된 문장 등도 있을 수 있다. 일반적으로 문장은 인물, 기관, 장소, 수치 등 문장이 포함하고 있는 개체명, 정치, 경제, 사회, 문화 등의 지면, 뉴스, 논문, 보고서, 블로그, 댓글, 소셜 네트워크와 같은 문서의 유형 등 내용 측면에서 다양하게 다중 분류할 수 있다. 이 때 문장의 5형식과 같은 형식 측면에서 다중 분류는 큰 의미가 없다. 의미연결망은 형식이 아니라 내용으로 관련된 결점을 연결하기 때문이다.

둘째, 관련도를 파악하는 인식 성능을 개선할 필요가 있다. 이는 다시 세 측면에서 가능하다. 관련 있는데 관련 없다고 판정하지 않는 재현율의 개선, 관련 없는데 관련 있다고 판정하지 않는 정확도의 개선, 그리고 하나의 차원에서는 관련 없지만 다른 차원에서는 관련 있을 수 있는 다차원 분석의 심화 측면이다. 또한 이 연구에서는 동일 정보원의 발언

날짜나 문장 간 유사도 등 연결을 정의하는 계수를 임의로 정했다. 문장 수준의 분석은 아직 충분하지 않고 이 연구에서 제시한 IBM 프로젝트 디베이터 역시 정확한 성능을 공개하지 않아서 이 연구와 유사 연구 간의 성능 비교를 제시하지 못한 점도 한계다. 다만 이 연구는 딥러닝과 고유 기술을 사용한 IBM에 비해 단순하고 투명한 알고리즘을 사용했다는 장점이 있다. 덕분에 이를 의미경로나 핵심, 요약, 상술 등을 수학적으로 정의했음에도 불구하고 사회과학자도 비교적 분석 과정을 쉽게 해석할 수 있다. 특히 매우 기초적인 유사도 계산과 규칙 기반의 알고리즘을 사용하고 각종 계수도 임의로 선정했음에도 불구하고 자동화된 담론분석에 활용할 수 있는 유용한 프로토타입 프로그램을 만들 수 있었다.

셋째, 방향성을 고려할 수 있다. 〈쿼트넷〉은 연결망의 방향을 고려하지 않고 분석한다. 이 연구에서는 다루지 않았지만 연결정도 중앙성을 통해 무방향 문장 연결망의 분포를 분석하면 각 결점의 연결정도 중앙성의 값이 뉴스 정보원 연결망이나 뉴스 주제 연결망 등 일반적인 경우보다 대략 2배 큰 것을 알 수 있었다(박대민, 2014b). 만일 연결에 방향을 부여하고 인링크만 고려한다면 연결정도 중앙성의 값은 절반으로 떨어질 가능성이 있다. 이는 뉴스 문장 연결망이 뉴스 단어 연결망과 유사한 분포를 갖는다는 것을 의미한다. 직관적으로도 문장은 순서가 있기 때문에 앞서 등장한 문장이 이후 등장한 문장으로 방향성을 갖는 연결망을 그릴 수가 있다.

넷째, 문장 연결망의 분포를 좀 더 다양하게 파악해 분석의 타당성을 심화시킬 필요가 있다. 문장 연결망의 전형적인 분포를 알면 수집한 데이터가 분석에 충분하고 타당한지를 간접적으로 파악할 수 있다. 실제로 시행연구에서 분석한 인공지능 관련 기사의 문장 연결망은 충분히 성장한 의미연결망이 아닌 성긴 연결망처럼 보인다. 이는 인공지능이라는 주제 자체가 언론에서 충분히 논의되지 않아서, 언론 보도 이외의 문서에 담긴 문장들과 그 문장들 간의 의미론적 관계가 잠재적인 결점과 연결로만 남

아있고 이 논문에서 다룬 기사와 연결망에는 반영되지 않았기 때문일 수 있다. 문장 연결망의 보편적 분포를 파악하기 위해서는 다른 다양한 주제에 대해 더 많은 데이터를 바탕으로 분석해볼 필요가 있다. 충분히 성장한 문장 연결망을 분석해, 사회 연결망과 유사하게 문장 연결망도 평균 경로거리가 짧은 좁은 세계인지, 척도 없는 연결망 내지 두터운 꼬리 분포를 갖는지, 선호적 연결이나 스며들기 군집, 일정 의미거리 내에는 관련도가 높은지 등을 파악하고 그러한 분포의 의미를 사회 연결망이 아닌 의미연결망의 관점에서 해석해볼 필요가 있다.

다섯째, 자연어처리 성능 개선도 필요하다. 〈빅카인즈〉의 알고리즘은 개략적으로만 알려져 있으며, 현재 성능이나 수집 데이터 현황이 적시에 업데이트되고 있지는 않다. 자연어처리 성능이 개선되고 명시된다면 〈쿼트넷〉의 성능도 높아질 수 있다.

뉴스는 100년 이상 매일마다 수많은 주제에 대한 사회적 논쟁을 간결하고 형식화해 정리해 놓은 자료이다. 특히 인용문은 기사가 다루는 핵심 의제에 관련되면서도, 서로 구분되는 견해를 대표하는 내용을 압축해 정제됐지만 구어체로 제시한다. 따라서 인용문 연결망을 통해 대화 중 특히 논쟁을 모사하기 위한 인공지능, 일종의 토론기계를 만들기 위한 말뭉치로 활용될 수 있을 것으로 기대한다. 기존의 말뭉치가 형태소분석과 구문분석을 활용한 형식적인 말뭉치였다. 반면 뉴스 문장 연결망 분석 방법으로 구축된 말뭉치는 형태소분석을 유사도 분석에만 활용하기 때문에, 일정 수준의 성능만 보인다면 형태소분석기에 크게 구애 받지 않고 만들어 질 수 있다. 구문분석 역시 형태소분석에서 필요한 정도면 충분하기 때문에 그 성능도 크게 중요하지는 않다. 다만 문장을 형식이 아닌 내용 중심으로 다중 분류하는 작업은 보다 나은 문장 연결망 말뭉치 구축을 위해 필요할 수는 있다. 예컨대 어떤 문장이 어떤 지면, 어떤 인물, 기관, 장소, 수치와 연결됐는지 파악하는 것이 더 중요할 수 있다.

더 나아가 IBM의 디베이터와 같은 본격적인 대화형 토론기계를 설

계하려면 음성인식이나 자연어생성, 감성표현, 인터페이스 등 다양한 측면의 보완이 요청된다. 장기적으로는 뉴스 문장연결망 정보를 포함한 메타데이터가 부착된 학습 데이터를 활용해 인공지능을 학습시키고, 새로운 주제에 대해서도 관련된 전문가나 기관을 제시하고 고려할 논쟁적 요소에 대해 조언하는 진일보한 능동형 토론기계 제작을 기대해본다.

참고문헌

곽재현 · 홍지숙 (2018). 빅데이터를 활용한 욜로 현상 분석. 〈관광연구저널〉, 32권 2호, 21 34.

권혜진 · 김송이 · 신혜원 · 이완정 (2017). 영유아 행복권 보장을 위한 부모지원 보육정책의 고찰 및 제언. 〈한국아동학회 학술발표논문집〉, 15 47.

김남규 · 이동훈 · 최호창 (2017). 텍스트 분석 기술 및 활용 동향. 〈한국통신학회논문지〉. 42권 2호. 471 492.

김대진 · 박다인 · 박종석 (2018). 데이터 마이닝 기법을 통한 마케팅 전략 변화에 대한 연구. 〈Korea Business Review〉, 22권 2호, 177 194.

김동환 · 이준환 (2015). 로봇 저널리즘 : 알고리즘을 통한 스포츠 기사 자동 생성에 관한 연구. 〈한국언론학보〉, 59권 5호, 64-95.

김민준 · 윤종묵 · 맹욱재 · 이중식 (2017). 링크드 데이터의 시각화: 뉴스 도메인을 중심으로. 〈한국HCI학회 학술대회 논문집〉, 292 295.

김봉제 (2018). Big Data 분석을 통한 한국사회의 도덕 · 윤리 용어 사용 특성 연구. 〈도덕윤리과교육연구〉, 58호, 27 58.

김선호 · 박대민 · 오세욱 (2015.12.). 스트럭처 저널리즘, 데이터 저널리즘을 넘어서. 〈2015 해외 미디어 동향〉. 서울: 한국언론진흥재단. 213 255.

김재욱 · 김한수 (2018). 빅데이터 분석을 통한 건설 메가트렌드 관심도에 관한 연구. 〈대한건축학회 학술발표대회 논문집〉, 38권 1호, 709 710.

김종성 (2017). 스포츠관광 활성화와 빅데이터 활용에 관한 연구. 〈한국엔터테인먼트산업학회논문지〉, 11권 3호, 99 109.

김혜원 · 이정욱 (2018). 신촌 연세로 대중교통전용지구는 어떻게 성공적으로 도입되었는가?. 〈지방정부연구〉, 21권 4호, 209 237.

맹미선 (2017). 〈알파고 쇼크와′4 차 산업혁명′담론의 확산: 과학기술 유행어(Buzzword)의 수사적 기능 분석을 중심으로〉. 서울대학교 박사학위논문.

박대민 (2017). 중국 보도 10년: 뉴스 빅데이터 분석으로 본 10대 사건. 〈미디어이슈〉, 3권 5호. 한국언론진흥재단.

박대민 (2016). 장기 시계열 내용 분석을 위한 뉴스 빅데이터 분석의 활용 가능성: 100만 건 기사의 정보원과 주제로 본 신문 26년. 〈한국언론학보〉, 60권 5호, 353 407.

박대민 (2015). 사실기사의 직접인용에 대한 이중의 타당성 문제의 검토. 한국언론학보, 59권 5호, 121 151.

박대민 (2014a). 하버마스, 루만, 들뢰즈, 가타리의 이론을 통한 일반 대중매체 체계이론의 제안. 〈한국언론정보학보〉, 67호, 119 151.

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

박대민 (2013). 뉴스 기사의 빅데이터 분석 방법으로서 뉴스정보원연결망분석. 〈한국언론학보〉, 57권 6호, 233 261. .

박대민 · 서봉원 · 이중식 (2016). 〈사용자 참여 뉴스 빅데이터 서비스 연구〉. 서울: 한국언론진흥재단.

박이수 (2017). 〈조선족 밀집지에 대한 내 · 외부 인식 비교 서울시 구로구 가리봉동을 중심으로〉. 서울대학교 대학원.

박지영 · 김태호 · 박한우 (2013). 의미연결망 분석을 통한 셀러브리티의 SNS 메시지 탐구. 〈방송통신연구〉, 82호. 36 74.

박창섭 (2010). 〈한국 언론의 야마 관행과 뉴스의 현실 구성〉. 서울대학교 석사 학위논문.

박현정 · 김한나 · 홍유정 (2017). 토픽모델링을 활용한 학생인권조례의 사회적 이슈 분석. 아시아교육연구, 18권4호, 683 711.

박희봉 · 이민화 (2016). 3 차원 미래예측 기법을 활용한 OTT 서비스 시장 전망 연구. 〈한국경영학회 통합학술발표논문집〉, 2319 2343.

배진수 (2016). 한일 간 독도 이슈의 추이와 일본의 도발 패턴. 〈독도연구〉, 21호, 309 349.

성미애 (2017). 건강가정기본법 시행 및 호주제 헌법불일치 판결 이후 일간지에 나타난 가족주의 인식. 〈한국가정관리학회지〉, 35권 4호, 113 139.

손기준 · 김찬우 · 김정대 · 김진택 (2015) 빅데이터 분석기법을 이용한 농업가뭄 평가 연구. 〈한국농공학회 학술대회 논문집〉.

솔트룩스 (2015.11.). 〈분석엔진 품질 평가서 버전 1.2〉.

신사임 · 이해솔 · 이종설 (2017). 시공간 기반 미디어 서비스 지원용 멀티미디어 콘텐츠 지식베이스 자동 구축 기술 연구. 〈한국정보과학회 학술발표논문집〉, 370 372.

신유현 · 안연찬 · 이상구 (2013). 신문 기사의 사건 탐지를 위한 문서 클러스터링. 〈한국정보과학회 학술발표논문집〉, 575 577.

윤호영 (2018). 변화하는 시대에 커뮤니케이션학 연구하기. 〈커뮤니케이션이론〉, 14권 1호. 50-98.

이은별 · 전진오 · 백지선 (2017). 서울의 다문화 공간 연구. 〈미디어 경제와 문화〉, 15권 2호, 7 43.

이정석 (2017). 부산시 외국인주민 사회통합 정책연구. 〈부산발전포럼〉, 163호, 120-125.

이준웅 (2010). 한국 언론의 경향성과 이른바 사실과 의견의 분리 문제. 〈한국언론학보〉, 54권 2호, 187-209.

정태석 (2002). 〈사회이론의 구성: 구조/행위와 거시/미시 논쟁의 재검토〉. 서울: 한울아카데미.

조현채 · 박철용 (2018). 랜덤포레스트를 이용한 신문사들의 19대 대통령 선거 보도 특성 분석. 〈한국데이터정보과학회지〉, 29권 2호, 367 375.

차세대융합기술원 (2013.12.). 〈빅데이터 기술을 활용하여 스마트 뉴스를 제공하는 모바일 앱 개발〉.

채영길, 유용민(2017). 네이버 · 다음 모바일 포털 뉴스 플랫폼의 19대 대통령 선거기사 분석. 〈사이버커뮤니케이션학보〉, 34권 4호, 195-242.

최모나 · 김민지 · 이시욱 · 유지형 · 김준솔 · 염유식 (2016). 간호간병통합서비스 관련 온라인 기사 및 SNS 빅데이터의 의미연결망 분석. 〈한국간호과학회 학술대회〉, 352 352.

최영지 (2017). 〈소비사회와 청년세대의'여성혐오'〉. 서울대학교 박사학위논문.

최준희 · 김상현 · 복근성 · 김형철 · 김동성 (2017). 의료용 인공지능에 대한 동향 분석 및 방사선사 인식조사. 〈대한영상의학기술학회 학술대회 자료집〉, 137-145.

최충익 · 김철민 (2017) 빅데이터를 활용한 지진 위험정보의 사회적 확산 분석. 〈한국지역개발학회지〉.

Aiden, E., & Michel, J. B. (2014). *Uncharted: Big data as a lens on human culture*. Penguin.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Bharat, K., Curtiss, M., & Schmitt, M. (2009). *U.S. Patent No. 7,568,148*. Washington, DC: U.S. Patent and Trademark Office.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine learning research*, *3*(Jan), 993-1022.

Burt, R. S. (1992). Structural hole. *Harvard Business School Press, Cambridge, MA*.

Callaway, D. S., Newman, M. E., Strogatz, S. H., & Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical review letters*, *85*(25), 5468.

Callon, M., Rip, A., & Law, J. (Eds.). (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. Springer.

Carr, C. S. (2003). *Visualizing argumentation*. Springer, London.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. MA: Little, Brown and Company.

Curtiss, M., Bharat, K., & Schmitt, M. (2009). *U.S. Patent No. 7,577,655*. Washington, DC: U.S. Patent and Trademark Office.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., &

Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215-239.

Garfield, E. (1964). "Science Citation Index"-A New Dimension in Indexing. *Science*, *144*(3619), 649-654.

Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*(Vol. 8). New York: Wiley.

Granovetter, M. S. (1977). The strength of weak ties. In *Social networks* (pp. 347-367).

Habermas, J. (1984). *The theory of communicative action*(Vol. 1, 2). Beacon press.

Hofmann, T. (1999, July). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289-296). Morgan Kaufmann Publishers Inc..

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Springer, Berlin, Heidelberg.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, *28*(1), 11-21.

Kretschmer, H. (1994). Coauthorship networks of invisible colleges and institutionalized communities. *Scientometrics*, 30(1),

363-369.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International Conference on Machine Learning* (pp. 1188-1196).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788.

Lee, D. L., Chuang, H., & Seamons, K. (1997). Document ranking and the vector-space model. *IEEE software*, *14*(2), 67-75.

Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*(Vol. 33, pp. 81-93).

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B. & Sigelman, S. (2002, March). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 280-285). Morgan Kaufmann Publishers Inc..

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, *69*(2), 213-238.

Moreno, J. L. (1937). Sociometry in relation to other social sciences. *Sociometry*, *1*(1/2), 206-219.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006, May). Analyzing entities and topics in news articles using statistical topic models. In *International conference on intelligence and security informatics* (pp. 93-104). Springer, Berlin, Heidelberg.

Newman, M. E. (2001). The structure of scientific collaboration

networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the national academy of sciences, 101(suppl 1), 5200-5205.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1-2), 1-135.

Park, H. W., & Leydesdorff, L. (2004). Understanding the KrKwic: A computer program for the analysis of Korean text. *Journal of the Korean Data Analysis Society*, *6*(5), 1377-1387.

Park, S., Kang, S., Chung, S., & Song, J. (2009, April). NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443-452). ACM.

Protess, D. L. (Ed.). (1992). *The journalism of outrage: Investigative reporting and agenda building in America*. Guilford Press.

Roberts, M., & McCombs, M. (1994). Agenda setting and political advertising: Origins of the news agenda. Political communication, *11*(3), 249-262.

Rorty, R. (Ed.). (1992). *The linguistic turn: Essays in philosophical method*. University of Chicago Press.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Salton, G. (1971). The SMART retrieval system-experiments in automatic document processing. *Englewood Cliffs*.

Sowa, J. F. (2000). *Knowledge representation: logical, philosophical, and computational foundations* (Vol. 13).

Pacific Grove, CA: Brooks/Cole.

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM review*, *35*(4), 551-566.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Travers, J., & Milgram, S. (1967). The small world problem. *Phychology Today*, *1*(1), 61-67.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141-188.

Van Dijk, Teun A. (1988). *News as discourse*. NJ: Lawrence Erlbaum.

Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media.

Vater, H. (2001). *Einführung in die Textlinguistik: Struktur und Verstehen von Texten. 3*. überarbeitete Aufl.

Watts, D. J. (2013). Computational social science: Exciting progress and future directions. The Bridge on Frontiers of Engineering, 43(4), 5-10.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature*, *393*(6684), 440.

White, H. C. (1963). *An anatomy of kinship: mathematical models for structures of cumulated roles*. Prentice-Hall.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the*

American Society for information Science, *32*(3), 163-171.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American society for information science*, *49*(4), 327-355.

Zhao, D., & Strotmann, A. (2008). Information science during the first decade of the web: An enriched author cocitation analysis. *Journal of the American Society for Information Science and Technology*, *59*(6), 916-937.

투 고 일 자: 2018년 07월 09일

심 사 일 자: 2018년 08월 07일

게재확정일자: 2018년 08월 31일

Abstract

A news sentence network analysis algorithm based on similarity, cooccurrence, and news sources

Daemin PARK
Senior Researcher, Korea Press Foundation
Bongwon Suh
Associate Professor, Graduate School of Convergence Science and Technology, Seoul National University
Seonghyun Kim
Master, Graduate School of Convergence Science and Technology, Seoul National University
Jaeyoun You
Ph.D. Student, Graduate School of Convergence Science and Technology, Seoul National University
Jungwoo Song
Master's Student, Graduate School of Convergence Science and Technology, Seoul National University

In this paper, we point out the limitations of the research on word-based semantic network analysis and propose a news sentence network analysis method as a method of semantic network analysis at sentence level. In this study, we developed Quetnet as the prototype news quote analysis program, and conducted a pilot study. The news sentence network focusing on quotes is a semantic network that nodes are quotes, and edges are defined as the association of jacquard similarity, co-occurrence of articles, and whether the same sources uttered quotes within a short period of time.

The news sentence network has a semantic path and can be used to define the one main sentences, summarizing sentences, and sentences in details. In this study, we conducted a news sentence network analysis on 5,046 quotations of 2,337 articles quoted as ″artificial intelligence(AI)″ from January 1, 1990 to April 30, 2016. Articles were collected from news big data system 〈BigKinds〉. As a result of the analysis, it was found that 3,742 nodes and 6,708 edges were detected except for the isolated node when the similarity parameter was 0.333, and topics about AI technology and its soical shocks are classified and summarized better comparing with the analyis node when the similarity parameter was 0.450. Given that news is accumulating years of public debates of important social issues, it can also contribute to the design of computational argumentation such as the IBM Project Debater, espeically for the purpose of social sciences.

KEYWORDS news semantic analysis. QuoteNet, semantic network analysis, computational argumentation, news big data anlaysis