토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구*

A Study on Opinion Mining of Newspaper Texts based on Topic Modeling

강 범 일 (Beomil Kang)**
송 민 (Min Song)***
조 화 순 (Whasun Jho)****

목 차

초 록

이 연구에서는 토픽 모델링 기법을 이용하여 신문 기사를 대상으로 주제 기반의 오피니언 마이닝을 수행하였다. 언론 매체가 가지는 정파성을 일종의 오피니언으로 간주하여 대선이라는 거대 이슈에 대한 각 매체들의 입장을 분석하였다. 먼저 대량의 신문 기사에서 주제를 추출한 후 매체별 주제 구성의 차이를 살펴보았다. 그리고 주제별 네트워크 분석을 통해 주제의 구조와 내용을 분석하였다. 마지막으로 시계열 분석을 통해서 시기별 주제 분포의 차이를 매체별로 살펴보았다. 그 결과 모든 분석에서 진보매체와 보수매체 모두 자신들의 이데올로기를 따라 기사를 보도하는 경향성이 확인되었다. 이를 통해 주제 기반 오피니언 마이닝이 타당성 있는 의견 분석의 기능을 수행할 수 있음을 확인할 수 있었다.

ABSTRACT

This study performs opinion mining of newspaper articles, based on topics extracted by topic modeling. We analyze the attitudes of the news media towards a major issue of 'presidential election', assuming that newspaper partisanship is a kind of opinion. We first extract topics from a large collection of newspaper texts, and examine how the topics are distributed over the entire dataset. The structure and content of each topic are then investigated by means of network analysis. Finally we track down the chronological distribution of the topics in each of the newspapers through time serial analysis. The result reveals that both the liberal newspapers and the conservative newspapers exhibit their own tendency to report in line with their adopted ideology. This confirms that we can count on opinion mining technique based on topics in order to analyze opinion in a reliable fashion.

키워드: 토픽 모델링, 오피니언 마이닝, 네트워크 분석, 언론의 정파성
Topic Modeling, Opinion Mining, Network Analysis, Newspaper Partisanship

---

* 이 논문은 2010년도 정부재원(교육부 인문사회연구역량강화사업비)으로 한국연구재단의 지원을 받아 연구되었음(NRF-2010-330-B00028).
** 연세대학교 언어정보연구원(kangbeomil@gmail.com) (제1저자)
*** 연세대학교 문헌정보학과 부교수(min.song@yonsei.ac.kr) (교신저자)
**** 연세대학교 정치외교학과 부교수(wsjho@yonsei.ac.kr)
논문접수일자: 2013년 10월 17일 최초심사일자: 2013년 11월 8일 게재확정일자: 2013년 11월 13일
한국문헌정보학회지, 47(4): 315-334, 2013. [http://dx.doi.org/10.4275/KSLIS.2013.47.4.315]

1. 서 론

각종 스마트 기기의 보급으로 인한 소셜 미디어 콘텐츠의 증가로 대중과 소비자의 의견을 수집할 수 있는 기회가 생겨났다. 이에 따라 소셜 네트워크 서비스나 블로그, 댓글 등을 통해 오가는 텍스트 데이터들은 대중의 생각을 쉽게 들여다 볼 수 있게 해 주는 새로운 원천으로 다루어지게 되었다. 이러한 데이터의 증가와 텍스트 마이닝 기술의 발전으로 텍스트에 담겨 있는 의견을 자동으로 추출하는 오피니언 마이닝의 중요성 또한 부각되기 시작하였다.

오피니언 마이닝은 주로 기업이나 각종 기관이 사회적 사건이나 정치적 이슈, 기업 전략이나 마케팅, 제품 선호에 대한 대중들의 의견을 수집하여 의사 결정에 활용하기 위한 목적으로 활발히 사용되어 왔다. 따라서 기존의 오피니언 마이닝에 관한 연구들은 주로 상품평이나 영화평과 같이 의견이 감정 언어로 표현되는 텍스트를 바탕으로 감정의 극성을 판단해 내기 위한 기법이나 그것의 기반이 되는 감정어 사전을 구축하는 방법에 초점을 맞추어 왔다.

이와는 다르게 본 연구에서는 단어나 문장 단위가 아닌 텍스트에서 자동으로 추출된 주제들을 기반으로 오피니언 마이닝을 수행하고자 하였다. 특정 이슈를 다루는 텍스트에서 발견되는 주제들의 성격과 그 주제들의 시기별 분포 변화를 분석함으로써 해당 이슈를 바라보는 텍스트 생산자의 관점이나 의견이 파악될 수 있을 것으로 보았기 때문이다. 시계열 분석이 가능한 대량의 자료로는 시기별로 비교적 일정한 분량의 텍스트로 구성돼 있는 신문 자료가 적당할 것으로 보았기 때문에 이를 연구 자료로 선택하였다. 그리고 기사를 통해 드러나는 언론 매체의 정파성을 일종의 의견으로 간주하여 이를 오피니언 마이닝의 대상으로 삼았다.

이를 위해, 18대 대선 후보들에 대한 기사를 진보매체의 기사와 보수매체의 기사로 나누어 수집한 후, 각 매체별로 주제를 자동으로 추출하기 위해 토픽 모델링(Topic Modeling) 기법을 사용하였다. 토픽 모델링 알고리즘은 방대한 양의 텍스트에서 사용된 단어들을 분석하여 주제들을 발견하고 그 주제들이 서로 어떻게 연결되어 있는지, 시간에 따라 어떻게 변화되어 가는지를 분석해 주는 통계적 방법이다(Blei 2013).

이렇게 추출된 주제들은 매체별 정파성 분석을 위해 세 단계로 분석되었다. 먼저, 각 매체가 따르는 정파적 입장에 따라 대선 후보 관련 기사에서 형성되는 주제가 다르게 나타날 것으로 보고, 매체별로 고정적으로 나타나는 주제들을 비교하여 그 차이를 분석하였다. 그리고, 양 매체에서 공통적으로 나타난 주제일지라도 주제의 세부 내용은 매체별로 차이를 보일 것으로 보고, 주제를 구성하는 단어들을 네트워크 분석 기법을 통해 비교하였다. 마지막으로, 공통 주제에 대한 시기별 분포 또한 매체별로 다르게 나타날 것으로 보고 시계열 분석을 통해 주제 분포의 차이를 매체별로 분석하였다. 이를 통해 주제 기반 오피니언 마이닝이 타당성 있는 의견 분석의 기능을 수행하는지 규명하고자 하였다.

2. 관련 연구

2.1 오피니언 마이닝

오피니언 마이닝은 주로 다양한 소셜 미디어 콘텐츠로부터 상품 및 서비스의 선호도, 사회적 사건이나 정치 이슈 등에 대한 대중들의 의견을 분석하는 데 적용되어 왔고, 관련 연구들 또한 이러한 의견들의 추출, 분류, 이해, 평가를 위한 방법론에 관한 논의를 주로 다루고 있다(Chen and Zimbra 2010; Liu 2010).

문헌정보학 분야의 오피니언 마이닝 관련 연구로는 의견이나 감정을 담고 있는 의견 문서들의 자동 분류의 성능을 향상시키기 위해 잠재의미색인(LSI) 기법을 이용해 분류 실험을 한 연구(이지혜, 정영미 2009)와, 의견 검색을 위한 사용자 정보 요구를 표현하는 방법과 의견 자질들의 결합 방법에 대한 실험을 진행한 연구(한경수 2010)가 있었다. 이들은 모두 의견의 수집과 분석이라는 오피니언 마이닝의 주된 목적을 수행하기 위한 연구라기보다는 정보검색의 관점에서 오피니언 마이닝을 다룬 연구라고 할 수 있겠다.

오피니언 마이닝은 텍스트 마이닝의 하위 분야이기 때문에 의견의 수집 및 분석을 위해 자연언어처리 기법, 전산언어학적 방법들과 더불어 텍스트 마이닝 기법들이 적용된다. 대부분의 오피니언 마이닝 연구는 감정 어휘나 문장의 통사 구조를 분석하여 텍스트로부터 의견을 추출하고 텍스트 마이닝기법을 이용하여 의견의 극성을 분류하는 방식으로 이루어진다. 이때 주로 사용된 기법은 기계학습이었으며 의견 분석의 대상이 되는 단위는 개별 어휘나 구, 문장 등이었고 본 연구에서와 같이 토픽 모델링을 통해 추출된 주제들을 바탕으로 의견을 분석한 연구는 찾아볼 수 없었다.

2.2 토픽 모델링

토픽 모델링은 구조화되지 않은 방대한 문헌집단에서 주제를 찾아내기 위한 알고리즘으로, 맥락과 관련된 단서들을 이용하여 유사한 의미를 가진 단어들을 클러스터링하는 방식으로 주제를 추론하는 모델이다(Steyvers and Griffiths 2007; Blei 2012). 이러한 특징 때문에 토픽 모델링은 문헌들을 연구 자료로 사용하는 다양한 분야의 연구에서 분석 도구로 사용되어 왔다. Griffiths와 Steyvers(2004)는 토픽 모델링을 통해, PNAS에 실린 1991년부터 2001년 사이의 논문들의 초록에서 주제를 추출한 후 시기별로 각광받는 주제와 소멸되는 주제를 파악하였다. Newman과 Block(2006)은 초기 미국 사회와 출판 문화의 이해를 위해, 토픽 모델링을 이용하여 18세기 신문 텍스트에서 주제를 추출하고, 각각의 주제들이 시간의 흐름에 따라 어떻게 변하는지를 분석하였다. Gerrish와 Blei(2010)는 동적 토픽 모델(Dynamic Topic Model)을 이용하여 논문 코퍼스에서 시간의 흐름에 따른 주제들의 내용 변화를 파악하여 이를 개별 문헌의 영향력을 측정하는 데 적용하였다. Grimmer(2010)는 미 상원의원들의 보도 자료에서 의원들이 강조하는 어젠다를 추출하는 데 토픽 모델링을 이용하여, 입법자들이 유권자들에게 어떤 방식으로 그들의 업무를 홍보하는지를 분석하였다. Song과 Kim(2012)은 생물정보학 분야의 논문 자료에서 주제를 추출하여

해당 학문 분야의 지적 구조를 분석하고 하위 분야에 대한 연구 경향을 살펴보았다. 토픽 모델링을 적용한 문헌정보학 분야의 국내 연구로는 문헌정보학의 연구 동향을 분석한 박자현, 송민(2013)과 한국의 경제 연구 동향을 분석한 송혜지 외(2013)가 있었다. 이처럼 토픽 모델링은 특정 분야에 대한 동향 분석에 주로 사용되어 왔고 오피니언 마이닝에 적용된 사례는 없는 것으로 확인되었다.

2.3 정파성 분석

본 연구에서 다루는 언론의 정파성은 대부분 언론학 분야에서 활발히 연구되어 왔다. 주로 연구자가 갈등적 요소가 강한 몇 개의 주제를 임의로 선정하고, 그 주제와 관련된 몇 개의 키워드를 정하여 해당 키워드로 검색된 기사들의 내용을 직접 분석하는 방식으로 정파성이 분석되었다. 이때 측정되는 요소들은 크게 취재원, 프레임, 논조로 분류해 볼 수 있다.

취재원이란 기사의 출처를 의미하는데 이는 다시 개인 취재원과 기관 취재원(청와대, 정당 등)으로 나누어진다. 취재원을 분석한 연구들에서는 이들 개인 또는 기관 취재원들의 유형이나 성격, 특정 이슈에 대한 입장 등을 코딩하여 비교하는 방식으로 기사의 정파성을 판단하였다(윤영철 2000).

프레임 분석은 언론이 뉴스를 제공할 때 일정한 틀을 도입함으로써 수용자들이 어떤 메시지를 유목화(categorization)해서 해석하고 평가하도록 작용하는 역할을 한다는 프레이밍 이론을 바탕으로 한 것이다. 즉, 프레임은 언론이 취한 입장과 그에 따른 보도 양식을 나타내는 것으로서 뉴스 수용자의 의견 형성에 중요한 영향력을 행사하게 된다(이준웅 2001). 이러한 측면에서, 특정 이슈를 다룬 기사들에 대한 프레임의 비교·분석은 정파성을 판단하기 위한 방법으로 활용되어 왔다(김정아, 채백 2008).

논조는 논설이나 평론 등의 경향에 관련된 것으로 이러한 경향이 잘 드러나 있는 사설이나 칼럼 자료는 정파성 분석을 위한 주요 자료로 사용되어 왔다. 이를 이용한 연구들은 특정 사안이나 인물에 대한 언론의 관점을 파악하기 위해 사설이나 칼럼에서 연구 대상에 대한 논조를 긍정, 부정, 중립의 의견으로 구분하거나 비판의 빈도 등을 산출하는 방식으로 정파성을 판단하였다(최진호, 한동섭 2012).

내용 중심의 기사 분석에 계량적인 방법을 도입하여 정파성을 분석한 연구도 있었다. 박재영(2009)에서는 기존의 연구에서 기사 내용의 불균형을 파악하기 위해 만들어진 수식을 정파성 분석에 적합하도록 수정하여 정파적 정도와 정파성의 방향성을 살펴보기 위해 사용하였다. 최현주(2010)에서는 지지와 반대 입장의 불균형 정도를 나타내는 '관점 균형지수'와 취재원의 불균형 정도를 나타내는 '취재원 균형지수'를 도입하기도 하였다.

이상에서 살펴본 언론학 분야의 연구들은 대부분 질적 분석 방법을 사용하였기 때문에 분석의 대상이 되는 자료의 수가 제한적일 수밖에 없었다. 정보학 분야에서 기계학습 방법을 활용하여 신문사별 논조의 차이를 살펴본 연구(감미아, 송민 2012)에서도, 엄밀한 의미에서의 논조 분석을 위해 사용된 기사는 4대강 관련 기사 229건에 불과해, 본 연구에서와 같이 대량의 자료를 대상으로 정파성 분석을 시도한 연구는 없

었던 것으로 파악되었다.

3. 연구 방법

3.1 데이터 수집

최민재, 김재영(2008)에 따르면, 불공정 편파 보도는 특히 선거 때 현저해지고 방송보다는 신문에서 두드러지게 나타나는 경향이 있다. 이를 고려하여 18대 대선 국면에서 보도된, 주요 대선 후보와 관련된 기사들을 정파성 분석의 대상으로 정했다. 자료의 수집을 위해 웹 크롤링 기법을 사용하여, 선거 직전 2년 동안(2011년 1월 1일~2012년 12월 18일) 신문 지면에 게재된 기사 중 박근혜, 문재인, 안철수가 등장하는 기사들을 수집하였다.[1)]

기사 수집의 대상이 된 일간지는 국내 신문 시장에서 가장 영향력이 크다고 판단되는 보수 성향의 일간지 2종과 진보 성향의 일간지 2종이다. 보수 성향의 일간지로는 조선일보와 동아일보, 진보 성향의 매체로는 한겨레신문과 경향신문을 택했다. 〈표 1〉은 기사 수집 결과를 나타낸 것이다.

3.2 데이터 전처리

모든 자료는 인터넷 사이트를 통해 수집된 웹 페이지 형식의 기사이기 때문에 마크업 기호, 연관 기사 제목, 신문사별 고유 문구[2)] 등의

〈표 1〉 대선 후보 기사의 수집 결과

후보	성향	매체	기사수	단어 수
박근혜	보수	조선	2,506	652,147
		동아	2,364	562,933
	진보	한겨레	2,376	704,787
		경향	2,510	710,502
문재인	보수	조선	1,300	317,397
		동아	1,335	317,723
	진보	한겨레	1,193	364,216
		경향	1,306	390,083
안철수	보수	조선	1,594	413,631
		동아	1,401	338,415
	진보	한겨레	1,336	412,771
		경향	1,475	460,255

1) 언론사가 인터넷을 통해 배포하는 기사 중에는 지면에 게재되지 않는 기사나, 일부 내용이 수정된 동일한 제목의 기사, 내용은 동일하지만 제목만 다른 기사 등 정제되지 않은 기사들이 다수 존재하기 때문에 보다 신뢰성 있는 결과를 얻기 위해 지면에 게재된 고유한 기사들만을 수집하였다.

2) '경향닷컴은 한국온라인신문협회(www.kona.or.kr)의 디지털뉴스이용규칙에 따른 저작권을 행사합니다.' 등의 고정된 문구.

불필요한 요소들을 제거한 후 기사의 본문만을 추출하였다. 기사의 본문은 UTagger[3]를 이용하여 형태소 분석을 하였는데, 그 결과 중에는 합성명사가 과분석되거나 고유명사 또는 전문용어가 오분석되어 나타난 경우가 있었다. 이러한 오류를 개선하기 위해 '명사+명사' 형태의 바이그램(bigram)에 대해 상호정보량(Mutual Information: MI로 약칭)을 산출하여 적정 수준 이상의 MI값을 가지는 결합을 합성명사로 보고 이를 결합된 형태로 치환하는 작업을 거쳤다. '경제+민주화'(MI값:8.01)나 '여론+조사'(MI값:8.87) 등이 그 대상이 되었다. 또한 수집한 기사 내에서 '이 대통령'과 '나 후보' 등은 각각 '이명박', '나경원'을 지칭하는 이형태 표현으로 볼 수 있는데 이와 같은 경우 하나의 대표 형태로 통합하는 과정을 거쳤다. '한나라당', '새누리당'과 같이 수집 대상 기간 중 명칭이 변경된 용어들은 모두 바뀐 용어로 통일하였다.

본 연구에서 정파성 판단의 기준이 되는 것은 기사에서 추출되는 주제이기 때문에 이러한 주제는 명사 개념어를 통해 드러난다고 보고 일반명사, 고유명사를 제외한 다른 문법 범주에 속하는 용어들은 모두 불용어(stopwords)로 처리하였다. 여러 문서에 걸쳐 나타나는 명사들 중 '위원회', '대표', '관계자' 등과 같이 주제를 파악하는 데 도움이 되지 않는 다고 판단되는 어휘들도 함께 제거하였다. 같은 이유에서 각 후보들의 이름도 분석에 대상에서 제외시켰다.

3.3 데이터 분석

3.3.1 LDA 모델을 이용한 주제 추출

본 연구에서는 대선 후보 관련 기사에서 주제가 되는 키워드를 추출하기 위해, MALLET의 LDA(Latent Dirichlet Allocation) 기반 토픽 모델링 알고리즘을 사용하였다.

토픽 모델링 기법 중 하나인 LDA는 단순하다는 특징과 함께 데이터의 차원을 축소하는데 유용하며, 의미적으로 일관성이 있는 주제들을 생산한다는 장점을 가지기 때문에 텍스트 분석에서 인기 있는 모델로서 사용되어 왔다(Mimno and McCallum 2008). LDA는 모델링을 수행하기 위한 샘플링의 반복 횟수와 추출할 주제의 수를 지정해야 하는데 본 연구에서는 샘플링을 3,000회 반복하여 총 15개의 주제를 추출했다. 이러한 설정값들은 여러 번의 시행착오를 거쳐 얻어진 것으로, 연구에 사용된 데이터 집합으로부터 가장 해석이 가능한 주제 집합을 추출해 주는 것으로 파악된 값들이다. 추출된 주제들에 대해 적절한 주제명을 레이블링하고 진보매체와 보수매체에서 나타난 주제들의 구성을 서로 비교하여 주제적 차이를 통해 정파성을 분석하였다.

3.3.2 네트워크 분석

두 종류의 매체에서 공통으로 추출된 주제라 하더라도 각각의 주제를 구성하는 단어들의 내용과 단어들 간의 관계는 매체별로 차이를 보일 것으로 보았다. 이러한 차이를 살펴보기 위해 동시 출현 단어 네트워크를 구성해 분석하였다.

3) 울산대학교 한국어처리연구실(http://nlplab.ulsan.ac.kr)에서 만든 한국어 문장에 대한 품사 및 동형이의어 분별 시스템이다.

각각의 주제를 구성하는 20개의 단어들 중 동일한 기사에 함께 출현한 단어들과 그들 간의 동시 출현(co-occurrence) 빈도로 행렬을 구성하여 벡터 간 코사인유사도(Cosine Similarity)를 산출한 후 이 값을 이용하여 동시 출현 단어들 간의 패스파인더 네트워크(Pathfinder Network: PFNet으로 약칭)를 구성하였다(Schvaneveldt 1990). PFNet은 삼각부등식(Triangle Inequality)을 위반하는 경로를 제거하여 네트워크를 생성하는 알고리즘이기 때문에 이를 이용할 경우 주요한 링크만 남겨진, 요약된 네트워크를 얻을 수 있게 되는 장점이 있다(이재윤 2006a). 또한 이재윤(2006b)에서 제안한 삼각매개중심성(Triangle Betweenness Centrality: TBC로 약칭)을 이용하여 각각의 주제 네트워크를 이루는 노드들의 입지와 영향력을 살펴보았다. 마지막으로 노드들의 군집화를 위해 PNNC(Parallel Nearest Neighbor Clustering) 알고리즘을 적용하였다. 이는 PFNet상의 노드를 최근접 이웃끼리 연결하여 군집을 형성해 나가는 일종의 계층적 클러스터링 기법이다(이재윤 2006c). PFNet 형성과 TBC 측정 및 PNNC 수행을 위해 WNET0.4(이재윤 2012)를 사용하였고 NodeXL을 이용하여 이를 시각화하였다.

3.3.3 DMR 모델을 이용한 시계열 분석

18대 대선 후보들의 기사로부터 추출된 각 주제들의 분포가 시간의 흐름에 따라 어떻게 변화하는지 살펴보기 위해 DMR(Dirichlet-multinomial Regression) 기반의 토픽 모델링을 수행하였다. 이는 기본 LDA와는 다르게 문헌의 텍스트 자료 외에도 메타데이터를 추가로 입력 받아 해당 메타데이터에 따라 주제의 분포가 변화되는 양상을 추적하게 해 주는 모델이다(Mimno and McCallum 2008). 여기서는 먼저 LDA를 이용해 각 매체의 후보별 기사에서 주제를 추출한 후, MALLET의 DMR 기반 토픽 모델 알고리즘을 사용해 각 주제 분포의 시기별 변화 양상을 파악하여 주요 정치적 이슈가 일어났던 시기를 중심으로 매체별 주제 분포의 차이를 분석하였다.

4. 분석 결과

4.1 매체별 주제 구성 분석

언론은 특정 사안을 보도하지 않는 방법을 통해 정파성을 드러내기도 하는데 이를 '매체의 정파성에 의한 의도적 배제'라고 표현한다(김영욱 2011). 따라서 어떠한 주제가 특정 매체에서만 등장한다면 이는 매체의 정파적 성향과 밀접한 관련이 있는 현상으로 해석할 수 있을 것으로 보았다. 이를 확인하기 위해, 진보매체와 보수매체의 기사를 대상으로 각각 모델링을 수행하여 각각 20개의 단어로 이루어진 총 15개의 주제를 각각 추출하였는데 그 결과는 〈표 2〉와 같다. 표의 우측에 위치한 주제 구성 단어들은 토픽 모델링의 수행 결과로 출력된 것이고, 좌측의 주제명은 단어들의 의미적 연관성을 고려하여 연구자가 직접 부여한 것이다.

MALLET의 LDA 기반 토픽 모델링은 동일한 자료와 동일한 설정값을 적용하더라도 모델링을 실행할 때마다 조금씩 다른 결과를 생성하는 특징이 있다. 따라서 매체별 모델링 결과의 신뢰성을 확보하기 위해 모델링을 반복적으로 수행하여 각 매체별로 비교적 꾸준히 나타난

〈표 2〉 토픽 모델링 수행 결과 중 일부

주제명	주제 구성 단어
경제 민주화	대기업, 경제민주화, 일자리, 중소기업, 등록금, 양극화, 재벌개혁, 순환출자, 비정규직, 반값등록금, 김종인, 자본주의, 서비스, 이명박정부, 소득세, 성장률, 중산층, 계열사, 공무원, 경제위기
여론조사	여론조사, 지지율, 포인트, 이명박, 유권자, 단일화, 지지층, 지지자, 수도권, 단일후보, 투표율, 야권후보, 양자대결, 노무현, 미디어리서치, 응답자, 정몽준, 오차범위, 이회창, 무소속
남북관계	김정일, NLL, 정상회담, 한반도, 천안함, 연평도, 탈북자, 남북관계, 미사일, 노무현, 대화록, 북방한계선, 통일부, 김정은, 이명박정부, 러시아, 김일성, 대사관, 아시아, 국정원
정수 장학회	정수장학회, 이사장, 문화방송, 부산일보, 김지태, 최필립, 김재철, 방문진, 언론사, 지분매각, 박정희, 이사회, 한국방송, 청와대, 한겨레, 장학금, 과거사, 청문회, 이진숙, 인터뷰
과거사	박정희, 아버지, 인혁당, 과거사, 쿠데타, 민주화, 피해자, 전두환, 대법원, 장준하, 역사인식, 대한민국, 어머니, 민주주의, 지도자, 유가족, 육영수, 의문사, 독재자, 이승만
새누리당	최고위원, 지도부, 비대위, 비상대책, 이재오, 대변인, 비대위원, 지역구, 전당대회, 김문수, 쇄신파, 홍준표, 당선자, 이명박, 여의도, 황우여, 청와대, 정몽준, 이한구, 수도권
민주 통합당	최고위원, 손학규, 김두관, 문고문, 이해찬, 노무현, 지도부, 박지원, 정세균, 전당대회, 선거인단, 도지사, 한명숙, 모바일투표, 정동영, 김대중, 유시민, 문성근, 강원도
지도자, 리더십	이명박, 리더십, 지도자, 박정희, 노무현, 김대중, 쿠데타, 기득권, 이명박정부, 이미지, 오바마, 아버지, 부정적, 대한민국, 산업화, 긍정적, 공동체, 스타일, 민주주의, 민주화
복지문제	일자리, 비정규직, 복지국가, 노동자, 반값등록금, 등록금, 소득세, 정규직, 대학생, 법인세, 테마주, 중산층, 무상보육, 사교육, 가계부채, 건강보험, 이명박정부, 자유무역협정, 서비스, 무상급식
SNS, 미디어	트위터, SNS, 인터넷, 프로그램, 소프트웨어, 소셜네트워크서비스, 콘서트, 사이트, 컴퓨터, 서비스, 미디어, 스마트폰, 사용자, 페이스북, 시스템, 메시지, 온라인, 진행자, 리트위터, 사이버
서울시장 보궐선거	서울시장, 박원순, 보궐선거, 무상급식, 나경원, 주민투표, 오세훈, 서울시, 최고위원, 단일후보, 투표율, 시민단체, 제작소, 네거티브, 홍준표, 교육감, 무소속, 서울대융합과학기술대학원, 서울시민, 포퓰리즘
선거운동	선거운동, 선관위, 후보자, 무소속, 지역구, 사무실, 아파트, 토론회, 여의도, 선대위, 강원도, 선거관리, 선거법, 네거티브, 후원금, 보좌관, 지지자, 공천헌금, 국정원, 공직선거법
통합 진보당	노동자, 통합진보당, 비정규직, 쌍용차, 해군기지, 이정희, 전태일, 민주노동당, 정규직, 심상정, 진보적, 건강보험, 국가관, 정리해고, 이석기, 당권파, 노회찬, 노동계, 김재연, 지부장
안철수 연구소	안철수연구소, 테마주, 소프트웨어, 투자자, 상한가, 글로벌, 코스닥, 서비스, 시스템, 프로그램, 인터넷, 컴퓨터, 대기업, 거래소, 스마트폰, 수수료, 모바일, 테크노밸리, 컴퍼니, 대주주
단일화	단일화, 무소속, 선대위, 대변인, 정치쇄신, 정권교체, 서울대, 정치개혁, 야권후보, 지지자, 브리핑, 공보단장, 단일후보, 선거운동, 토론회, 메시지, 정치혁신, 출마선언, 박선숙, 간담회

다고 판단되는 주제들을 해당 매체에서 두드러지는, 신뢰성 있는 주제로 파악하기로 하였다. 〈표 3〉은 모델링을 10회 수행한 결과, 각 주제들이 출현한 횟수를 매체별로 나타낸 것이다.

〈표 3〉 모델링 10회 수행 결과

토픽	출현 횟수	
	진보매체	보수매체
MB정부실정	8	0
과거사	10	3
복지문제	7	0
정수장학회	9	1
남북관계	3	9
SNS, 미디어	7	9
경제민주화	10	10
단일화	10	10
민주통합당	8	8
새누리당	9	8
서울시장보궐선거	10	9
여론조사	9	10
지도자, 리더십	6	9
선거운동	2	2
안철수연구소	0	4
통합진보당	3	0

음영으로 표시된 부분은 과반수 넘게 출현한 경우로, 이들만을 신뢰성 있는 주제로 판단하여 비교 대상으로 삼았다. 진보매체에서 등장한 신뢰성 있는 주제는 'MB정부실정', '과거사', '복지문제', '정수장학회'이다. 이 중 이명박 정부에 대한 비판적 내용을 담고 있는 'MB정부실정'이나, 선거 국면에서 박근혜 후보의 아킬레스건으로 작용했던 '과거사'와 '정수장학회'는 이들과 입장을 같이하는 보수매체에서 불편하게 느낄 수 있는 주제들이기 때문에 진보매체에서 주로 나타났던 것으로 보인다. '복지문제'는 보수매체에서는 나타나지 않고 진보매체에서만 활발히 거론되고 있는데 이는 복지가 전통적인 진보의 어젠다라는 측면에서 이해될 수 있을 것이다.

언론사의 정파적 차이를 가장 극명하게 드러내는 영역이 남북문제라는 사실은 여러 연구(윤영철 2000; 김재홍 2003 등)를 통해 지적된 바 있다. 위 결과에서 '남북관계'의 매체별 출현 빈도는 이러한 지적을 뒷받침해 준다. 이 주제는 보수매체에서만 과반수 넘게 출현했는데 이는 안보 현안을 중요하게 다루는 매체의 보수적 성향이 그대로 드러난 결과라 하겠다. 또한 선거 때마다 큰 변수로 작용하는 소위 '북풍'이라고 불리는 북한 변수가 보수매체에서 상대적으로 부각돼 나타난 결과로 풀이될 수도 있을 것이다.

'SNS,미디어', '경제민주화', '단일화' 등 8개의 주제들은 양 매체에서 모두 과반수 넘게 출현한 것들로, 정파적 특성과는 관계없이 대선 국면에서 중요하게 다뤄진 주제들이라고 볼 수 있을 것이다. 한편 '선거운동', '안철수연구소', '통합진보당'의 출현 빈도는 모두 과반수에 미치지 못했다.

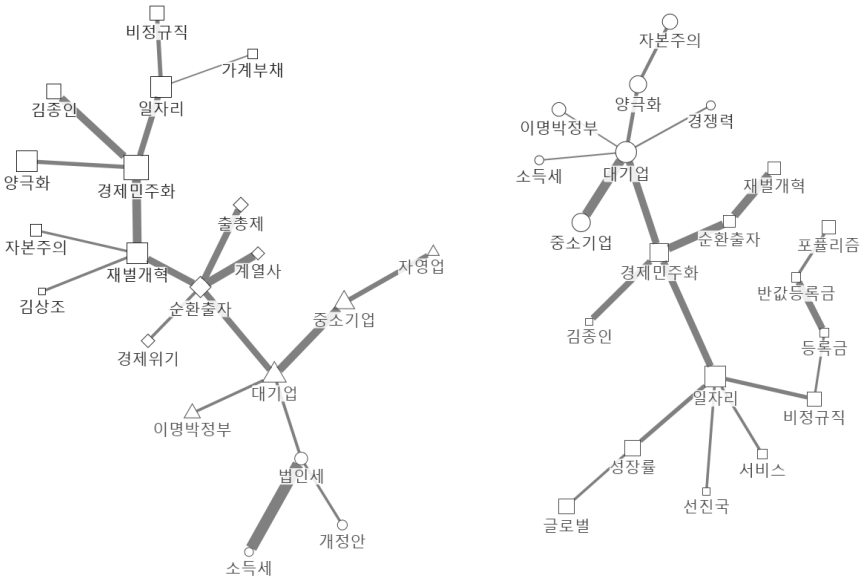
4.2 주제별 네트워크 분석

진보매체와 보수매체에서 공통으로 나타난 주제일지라도 각 주제를 구성하는 단어군의 세부 내용은 차이를 보일 것이라고 판단하였다. 이를 확인해 보기 위해 4.1에서 수행한 토픽 모델링 10회 결과 중 한 회의 결과를 택하여 두 매체에서 공통적으로 출현한 몇 가지 주제에 대한 네트워크 분석을 수행하였다. 각각의 주제를 구성하는 20개의 단어들이 한 기사 내에서 동시 출연한 빈도를 파악하여 코사인 유사도를 구한 후 이를 PFNet으로 나타냈다.

〈그림 1〉은 '경제민주화'에 대한 진보매체와 보수매체의 네트워크를 나타낸 것이다. 노드의 크기는 TBC의 정도를 나타내고 링크의 굵기는 두 노드 간의 연결 가중치를 반영한 것이다. 또한 PNNC 기법으로 네트워크상에 단어들의 군집을 표시하였는데, 동일한 군집에 속한 단어들의 노드는 같은 모양으로 표시하였다.

'경제민주화'는 보수와 진보를 막론하고 모든 후보 진영에서 공약으로 내걸었던 이슈였기 때문에 두 매체에서 모두 두드러지게 나타난 것으로 보인다. 진보매체의 네트워크에서는 3개의 군집이 나타났고 보수매체의 네트워크에서는 2개의 군집이 나타나 있다. '경제민주화'와 직접적으로 연결되는 노드를 비교해 보면 진보매체의 경우는 '양극화', '일자리', '재벌개혁'과 관계를 맺고 있고, 보수매체에서는 '대기업', '순환출자', '일자리'와 연결되어 있다.

진보매체에서 '경제민주화'와 직접 관계를 맺는 단어들은 모두 경제민주화를 통하여 해결되

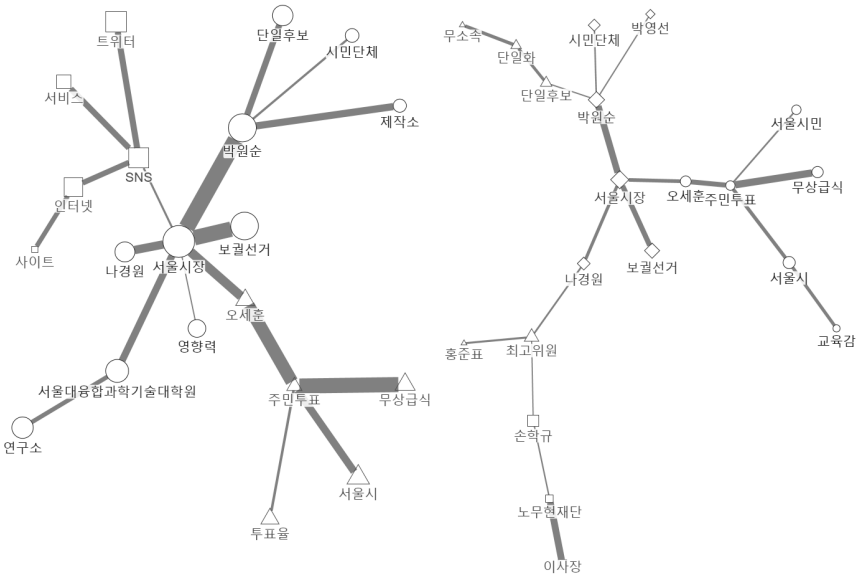


〈그림 1〉 '경제민주화' 네트워크(좌: 진보매체, 우: 보수매체)

거나 추진되어야 하는 문제들로 볼 수 있다. 따라서 진보매체에서는 '경제민주화'에 대한 기사에서 경제민주화를 통해 얻을 수 있는 효과를 주로 관련지어 보도하면서 이슈에 대한 긍정적인 시각을 드러낸 것이라고 볼 수 있다. 특히 '재벌개혁'과의 링크는 매우 굵게 나타나 있는데 이는 진보매체가 가지는 반대기업·반재벌 정서가 반영된 것이라고도 볼 수 있다.

보수매체의 네트워크에서는 진보매체에서와는 다르게 '경제민주화' 노드와 직접 연결되는 단어들이 동일한 군집에 속해 있지 않다. '대기업'이 바로 그러한 경우인데, 경제민주화 정책이 대기업에 대한 규제를 바탕으로 한다는 측면에서 대기업은 이 정책에 대한 부정적인 입장을 견지하는 집단이다. 따라서 보수매체는 '경제민주화'를 이러한 '대기업'과 자주 관련지으며 입장을 같이하려 했던 것으로 해석해 볼 수 있다. '대기업'은 '경쟁력'과도 연결되어 있는데 이 둘을 자주 함께 언급했다는 것은 보수매체가 '경제민주화'에 대한 기사에서, '경제민주화'를 통해 약화될 수 있는 '대기업'의 '경쟁력'에 대한 우려의 시각을 자주 드러냈다는 것을 의미한다고 볼 수 있겠다.

〈그림 2〉는 '서울시장 보궐선거'에 대한 매체별 네트워크를 나타낸 것이다. 2011년에 있었던 '서울시장 보궐선거' 또한 무상급식 문제를 핵심 어젠다로 하여 진보 대 보수의 구도로 펼쳐진 선거였기 때문에 두 매체에서 모두 두드러지게 나타난 것으로 보인다. 진보매체에서는 3개의 군집이 발견되었고 보수매체에서는 5개의 군집이 나타났다. '서울시장'과 직접 연결된 노드를 살펴보면 진보매체에서는 '나경원', '박원순', '오세훈', '서울대융합과학기술대학원' 등의 단어들이 발견되고 보수매체에서는 '나경원', '박원순', '오세훈'과 연결되어 있는 것을 확인할 수 있다.



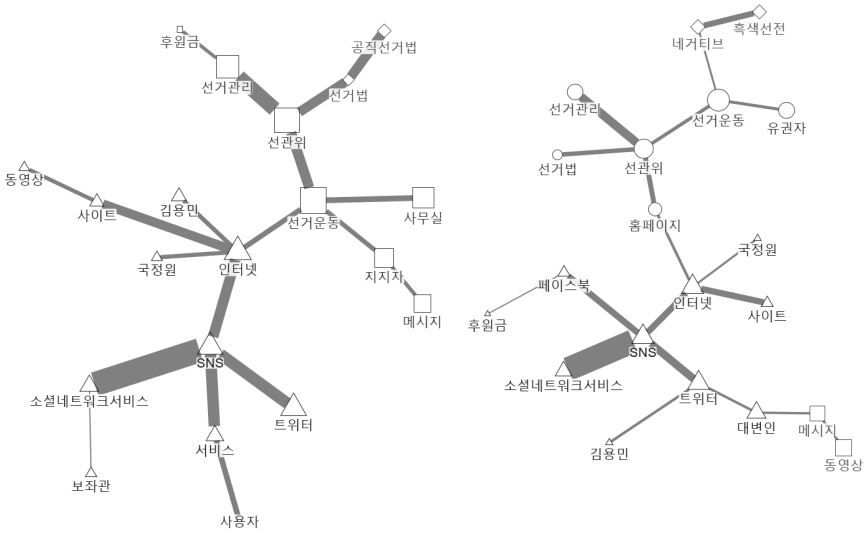
〈그림 2〉 '서울시장 보궐선거' 네트워크(좌: 진보매체, 우: 보수매체)

진보매체의 네트워크에서 '서울시장'과 연결된 단어들 중 특징적인 노드는 'SNS'이다. 이는 보수매체의 네트워크에서는 나타나지 않은 단어이다. '서울시장'과 'SNS'가 관계를 형성하는 것은 당시의 서울시장 보궐선거가 국내에서 치러진 최초의 본격적인 소셜네트워크서비스(SNS) 선거였다는 세간의 평가와 맞물려 있는 것으로 보인다. 당시 선거에서 진보 진영의 유력 인사들과 젊은 유권자들이 SNS를 이용해 선거 여론을 주도한 바 있는데 진보매체에서는 이러한 SNS 여론을 적극적으로 보도했던 것으로 파악할 수 있다.

'서울대융합과학기술대학원'과의 직접적인 연결 또한 보수매체의 네트워크와 차이를 보이는 점이다. '서울대융합과학기술대학원'은 당시 안철수가 소속된 기관이었으므로 사실상 '안철수'를 의미한다고 볼 수 있다. 앞서 데이터 전처리 단계에서 세 명의 대선 후보자의 이름을 제거했다고 밝힌 바 있는데, 만약 이러한 작업을 거치지 않았다면 '서울대융합과학기술대학원'이 위치한 노드에 안철수가 자리했을 가능성이 크다. 안철수는 당시 유력한 야권 후보로 거론되다가 박원순과 단일화를 이루었기 때문에 '서울시장'과 연결되어 있는 것으로 보인다. 이 '안철수' 노드는 다른 두 후보(나경원, 박원순)와 같은 군집에 속해 있기도 하는데, 이를 통해 안철수를 당시 선거에서 큰 영향력을 가지는 인물로 다루었던 진보매체의 시각을 엿볼 수 있다.

보수매체의 네트워크에서는 '서울시장'이 전임 시장과 두 후보와만 관련을 맺고 있어 별다른 특징을 찾아보기 어렵다.

〈그림 3〉은 'SNS/미디어'라고 이름 붙인 주제에 대한 진보매체와 보수매체의 네트워크를 나타낸 것이다. 앞서도 언급했듯이 서울시장 보궐선거를 계기로 선거에서 SNS가 가지는 영향력이 증대되기 시작하였는데 'SNS/미디어'



〈그림 3〉 'SNS/미디어' 네트워크(좌: 진보매체, 우: 보수매체)

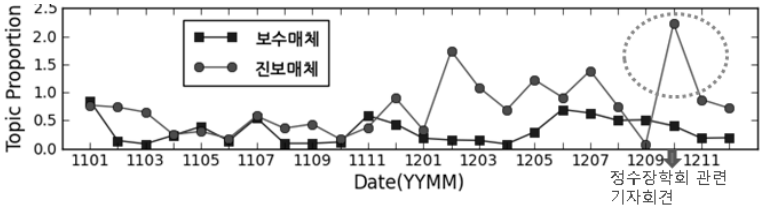
가 두 매체에서 공통된 주제로 추출된 것을 보면 SNS의 영향력은 18대 대선에서도 어김없이 발휘되었던 것으로 볼 수 있다. 클러스터링 결과를 보면 진보매체에서는 3개의 군집이 나타났고 보수매체에서는 4개의 군집이 나타났다. 두 매체에서 공통적으로 나타난 주요 군집은 'SNS'와 '선거운동'을 중심으로 한 것들이다. 따라서 이 주제는 SNS를 이용한 선거 운동이라는 관점에서 해석되어야 할 것이다.

매체 간의 네트워크에서 큰 차이를 찾아볼 수는 없지만 보수매체의 네트워크에서만 '네거티브', '흑색선전'으로 구성된 군집이 '선거운동'과 연결되어 나타나는 것은 특징적이라고 할 수 있다. 이는 SNS 이용자 중에 진보적 이념 성향을 가진 사람이 많다는 윤성이(2012)의 조사 결과를 고려할 때, 보수매체의 네트워크에서는 SNS를 이용한 선거운동을 네거티브, 흑색선전과 연관시키며 그 가치를 폄하하는 시각이 드러난 것이라고 볼 수 있을 것이다.

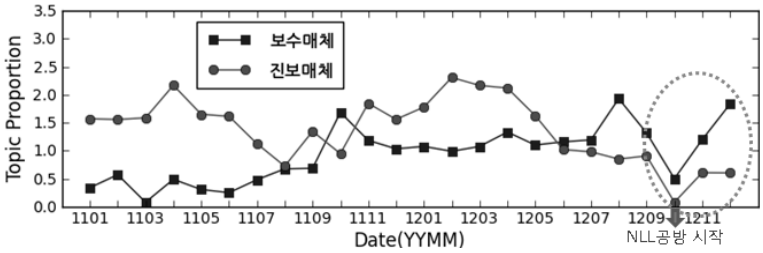
4.3 주제 분포의 시계열 분석

이민웅(2003)에 의하면 언론에서 인물 중심의 정파성은 '특정후보 밀어주기'의 형태로 나타난다. 언론이 일종의 킹 메이커(king maker) 역할을 하게 된다는 것이다. 이러한 문제의식을 바탕으로 특정 인물 보도에 대한 정파성을 살펴보기 위해, 진보매체와 보수매체의 각 후보 관련 기사에서 매체별로 주제를 추출한 후, 두 매체에서 공통적으로 출현한 몇 개의 주제에 대하여 시간의 흐름에 따른 주제 분포의 변화 추이를 비교하였다.

〈그림 4〉는 박근혜 후보의 기사에서 추출된 '정수장학회' 이슈의 시기별 분포를 나타낸 것이다. 정수장학회 관련 비리에 대한 기자회견이 있었던 2012년 10월에 진보매체의 추세선은 최고점에 위치한 반면 보수매체의 추세선은 앞선 시기와 비교했을 때 오히려 서서히 하강하는 모습을 보이고 있다. 이 주제는 이슈화 될 수



〈그림 4〉 박근혜 기사에서 추출된 주제 '정수장확회'의 시기별 분포

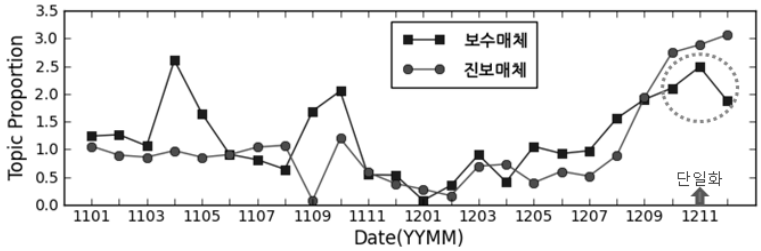


〈그림 5〉 박근혜 기사에서 추출된 주제 '남북관계'의 시기별 분포

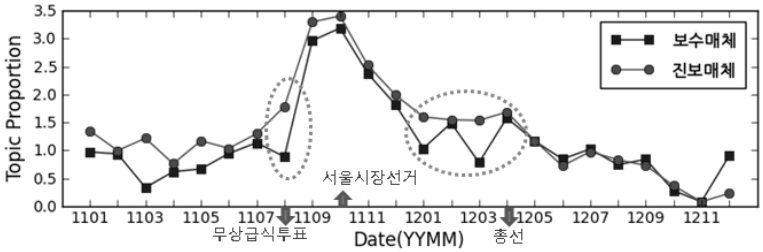
록 박근혜에게 불리하게 작용되는 사안이므로 보수매체는 박근혜의 편에서 이를 소극적으로 다루었던 것으로 풀이할 수 있다. 진보매체는 10월 이후에 정수장학회 이슈를 더 이상 쟁점화하지 않는 양상을 보이는데, 10월 21일에 있었던 박근혜의 정수장학회 관련 기자회견이 이러한 논란을 잠재우는 역할을 했던 것으로 보인다.

〈그림 5〉는 박근혜 후보의 기사에서 추출된 '남북관계'의 시기별 분포를 나타낸 것이다. 대선을 몇 달 앞둔 시점인 2012년 5월부터, 보수매체는 진보매체보다 관련 주제를 활발히 거론하고 있는 것으로 나타났다. 이는 선거 때마다 북한과 관련한 안보 문제를 이슈화시키는 보수 진영 고유의 성향이 그대로 드러난 결과로 볼 수 있다. 두 매체 모두 2012년 10월을 기점으로 상승하는 추세를 보이고 있는데 이 시기는 노무현 전 대통령의 NLL 포기 발언에 대한 진실 공방이 시작된 시점이다. 그러나 보수매체는 이후 계속해서 대선이 끝날 때까지 가파르게 상승하는 모습을 보이는 반면 진보매체에서의 분포는 더 이상 상승세를 이어가고 있지 않다. 이는 NLL 관련 사안은 공방이 오갈수록 야권에게 불리하게 작용되는 사안이므로 보수매체에서는 활발히 보도하고 진보매체에서는 상대적으로 침묵했던 것으로 해석될 수 있겠다.

〈그림 6〉은 안철수 후보 기사에서 추출된 주제인 '서울시장 보궐 선거'에 대한 시기별 분포이다. 앞서 언급했듯이 안철수 후보는 서울시장 보궐 선거를 계기로 정치인으로서의 역할을 요구받기 시작했다. 이를 반영하듯 두 매체의 추세선 모두 서울시장 선거 시점에서 가파른 상승곡선을 그리고 있다. 매체별로 차이를 보이는 부분을 살펴보면 시장 선거에 앞서 무상급식 투표 이슈가 부각되었던 2011년 7~8월에 보수



〈그림 6〉 문재인 기사에서 추출된 주제 '단일화 · 정권교체'의 시기별 분포



〈그림 7〉 안철수 기사에서 추출된 주제 '서울시장 보궐선거'의 시기별 분포

매체의 추세선은 하강하고 있는 반면 진보 매체는 상승세를 보이고 있다. 이는 진보매체가 당시 선거 국면에서 야권의 전폭적인 지지를 받으며 나타났던 안철수에 대해 보다 먼저 적극적으로 보도한 반면, 보수매체는 보도에 소극적이었다는 것을 보여 준다. 또 다른 차이는 총선을 앞둔 시점에 나타난다. 서울시장 선거 이후부터 두 매체의 추세선이 대체로 하강하는 양상을 보인다는 점은 같지만, 보수매체에서는 총선 무렵까지 변동을 거듭하는 반면 진보매체에서는 서서히 하강하다가 총선을 몇 달 앞둔 시점부터는 일정 수준을 유지하고 있는 것으로 나타났다. 이는 진보매체가 서울시장 보궐선거와 안철수를 꾸준히 관련지어 보도하면서 서울시장 보궐선거에서 나타났던 안철수의 긍정적인 영향력을 총선까지 이어가려 했던 것으로 풀이된다.

〈그림 7〉은 문재인 후보의 기사로부터 추출된 '단일화 · 정권교체' 주제의 시기별 분포이다. 두 매체의 분포 모두 야권 후보 단일화가 이루어진 시점인 2012년 11월 무렵까지 상승하는 추세를 보이는데 그 이후에는 보수매체에서는 하강하는 반면 진보매체에서는 조금 더 상승하는 모습을 보이고 있다. 이는 당시 새누리당에서 단일화의 의미를 축소시키고 그것의 파급 효과를 차단하려는 전략을 구사했던 점, 민주통합당에서 안철수의 지지층을 흡수하여 지지율 상승을 이끌어 내기 위해 단일화의 의미를 부각시키려 했던 점과 결코 무관하지 않은 것으로 보인다. 따라서 이러한 추이는 모두 각 정당들의 정파적 입장이 반영되어 나타난 것으로 해석할 수 있겠다.

5. 결 론

본 연구에서는 토픽 모델링 알고리즘을 사용하여 신문기사에서 추출한 주제를 바탕으로 오피니언 마이닝을 수행하였다. 먼저 18대 대선 후보들의 기사에서 형성되는 주제들을 추출하고, 이러한 주제들이 매체별로 차이를 보이는지, 주제를 구성하는 단어들의 내용에 차이가 있는지, 시기별로 주제 분포의 차이가 드러나는지를 각각 살펴보았다.

매체별 주제 구성의 분석을 통해서는 'MB정부실정', '과거사', '정수장학회'와 같이 보수 세력에게 불리한 주제들이나 '복지문제'와 같이 진보 세력에게 유리한 주제는 진보매체에서 주로 나타난 것으로 파악되었다. 반면 안보와 관련 주제인 '남북관계'는 보수매체에서만 두드러지게 나타난 것을 확인할 수 있었다.

주제별 네트워크 분석을 통해서는 두 매체에서 동일하게 나타난 주제들이라도 정파적 입장에 따라 주제를 구성하는 단어들 또는 단어들의 관계가 차이를 보이는 것을 확인할 수 있었다. '경제민주화'의 경우 진보매체에서 보다 긍정적으로 보도한 측면이 있었고 '서울시장 보궐선거'의 경우에는 진보매체에서만 당시 진보 진영에 유리하게 작용했던 'SNS'와 '안철수'를 선거와 연관시키고 있는 것으로 나타났다. 'SNS/미디어'에 대한 네트워크에서는 보수매체가 이를 '네거티브', '흑색선전'과 관련지어 바라보는 부정적인 시각이 드러났다.

시기별 주제 분포의 분석을 통해서는 특정 시기의 주제 분포가 각 진영의 유불리에 따라 다르게 분포하는 특징을 확인할 수 있었다. 박근혜 기사에서의 '정수장학회'는 의혹이 크게 불거진 시점에 진보매체에서는 높은 분포를 보이는 반면 보수매체에서는 변동이 없었다. 박근혜에게 유리했던 주제인 '남북관계'는 NLL 관련 공방이 시작된 이후 보수매체에서만 분포가 지속적으로 상승했던 것으로 확인되었다. 안철수 기사에서 추출된 '서울시장선거'의 경우는 진보매체에서 이를 먼저 이슈화한 후 총선 무렵까지 보도를 이어감으로써 안철수의 긍정적인 영향력을 유지하려고 했던 것으로 나타났다. 문재인의 기사에서 나타난 '단일화'의 경우, 진보매체에서는 단일화 시점 이후에도 분포가 지속적으로 증가한 반면 보수매체에서는 하락하는 양상이 나타났다.

이 세 가지 분석을 통하여 진보매체와 보수매체 모두 자신들과 유사한 이데올로기 지형에 위치해 있는 후보 및 진영의 입장에서, 본인들에게 유리한 이슈는 적극적으로 보도하고 불리한 이슈는 소극적으로 보도하는 경향성이 파악되었다. 이를 통해 대선이라는 거대 이슈와 그와 관련된 주제들에 대해 각 매체가 가지는 입장을 엿볼 수 있었다.

본 연구는 일종의 탐색적 연구로서, 토픽 모델링 기법을 이용한 주제 기반의 오피니언 마이닝이 타당성 있는 의견 분석의 기능을 수행하는지를 밝혀 보고자 했다는 점에서 기존의 오피니언 마이닝 관련 연구들과 차별화된다. 또한 토픽 모델링 알고리즘이 단순히 방대한 문헌에서 주제를 추출하기 위해 사용되어 왔다는 점을 감안할 때, 네트워크 분석을 통해 개별 주제의 구조와 내용을 살펴본 것은, 토픽 모델링이 가지고 있지 않은 개별 주제 분석의 기능을 보완하는 방법이라는 측면에서 의미를 부여할 수 있을 것이다.

그러나 각 매체별 분석 결과의 차이가 정파성의 차이로 인한 것임을 증명하는 데 보다 과학적인 방법들을 적용하지 못한 점은 본 연구가 가지는 한계이다. 또한 각 주제에 대한 레이블링 작업이 세 명의 연구자에게만 의존해 진행되었고, 각 주제를 구성하는 단어들이 정확히 일치하지 않음에도 레이블링 결과가 같을 경우 동일한 주제로 취급하여 비교 대상으로 삼은 점도 한계로 지적될 수 있겠다.

향후 이러한 점들이 보완된다면 본 연구에서 시도한 주제 기반의 오피니언 마이닝이 방대한 양의 자료를 대상으로 언론의 정파성을 분석하는 데 효과적으로 사용될 수 있을 것으로 보인다.

참 고 문 헌

[1] 감미아, 송민. 2012. 텍스트 마이닝을 활용한 신문사에 따른 내용 및 논조 차이점 분석.『지능정보연구』, 18(3): 53-77.

[2] 강명구. 2004. 한국 언론의 구조변동과 언론전쟁,『한국언론학보』, 48(5): 319-421.

[3] 김영욱. 2011. 한국 언론의 정파성과 사회적 소통의 위기.『한국언론학회 심포지움 및 세미나』, 107-136.

[4] 김재홍. 2003. 김대중 정부의 대북 포용정책에 대한 언론노조와 국민여론의 비교분석.『한국정치학회보』, 37(2): 197-218.

[5] 김정아, 채백. 2008. 언론의 정치 성향과 프레임: '이해찬 골프'와 '최연희 성추행' 사건의 보도를 중심으로.『한국언론정보학보』, 41: 232-267.

[6] 박자현, 송민. 2013. 토픽 모델링을 활용한 국내 문헌정보학 연구동향 분석.『정보관리학회지』, 30(1): 7-32.

[7] 박재영. 2009. 한국 언론사들의 정파성 지형.『한국언론재단 세미나 종합 보고서』, 17-65.

[8] 신태범, 권상희. 2013. 국내 청소년의 포털뉴스 이용특성과 뉴스신뢰, 공공성인식에 관한 연구.『사이버 커뮤니케이션 학보』, 30(1): 241-294.

[9] 송혜지, 박경수, 정혜은, 송민. 2013. 텍스트 마이닝 기법을 활용한 한국의 경제연구 동향 분석.『한국정보관리학회 학술대회논문집』, 20: 47-50.

[10] 윤성이. 2012. 소셜 네트워크의 확산과 민주주의 의식의 변화.『한국정치연구』, 21(2): 145-168.

[11] 윤영철. 2000. 권력 이동과 신문의 대북정책 보도: 신문과 정당의 병행관계를 중심으로.『언론과 사회』, 27: 48-81.

[12] 이민웅. 2003.『저널리즘: 위기 변화 지속』. 서울: 나남.

[13] 이재경. 2004. 저널리즘의 위기와 언론의 미래.『신문과 방송 40주년 세미나』. 2004년 3월 18일.

[서울: 프레스센터].

[14] 이재윤. 2006a. 지적 구조의 규명을 위한 네트워크 형성 방식에 관한 연구.『한국문헌정보학회지』, 40(2): 333-355.

[15] 이재윤. 2006b. 계량서지적 네트워크 분석을 위한 중심성 척도에 관한 연구.『한국문헌정보학회지』, 40(3): 191-214.

[16] 이재윤. 2006c. 지적 구조 분석을 위한 새로운 클러스터링 기법에 관한 연구.『정보관리학회지』, 23(4): 215-231.

[17] 이재윤. 2012. WNET. (version 0.4). (Software).

[18] 이준웅. 2001. 갈등적 이슈에 대한 뉴스 프레임 구성방식이 의견형성에 미치는 영향.『한국언론학보』, 46(1): 441-482.

[19] 이지혜, 정영미. 2009. 지도적 잠재의미색인(LSI)기법을 이용한 의견 문서 자동분류에 관한 실험적 연구.『정보관리학회지』, 26(3): 451-462.

[20] 진설아, 허고은, 정유경, 송민. 2013. 트위터 데이터를 이용한 네트워크 기반 토픽 변화 추적 연구.『정보관리학회지』, 30(1): 285-302.

[21] 차한필. 1989.『국내 신문 사설의 주제 분석과 각 신문 간 상관관계에 관한 연구』. 석사학위논문, 연세대학교 대학원, 도서관학과.

[22] 최민재, 김재영. 2008. 포털의 17대 대선 관련 뉴스서비스 공정성에 관한 탐색적 연구.『언론과학연구』, 8(4): 667-701.

[23] 최진호, 한동섭. 2012. 언론의 정파성과 권력 개입: 1987년 이후 13~17대 대선캠페인 기간의 주요일간지 사설 분석.『언론과학연구』, 12(2): 534-571.

[24] 최현주. 2010. 한국 신문 보도의 이념적 다양성에 대한 고찰: 6개 종합일간지의 3개 주요 이슈에 대한 보도 성향 분석을 중심으로.『한국언론학보』, 54(3): 399-426.

[25] 한경수. 2010. 효과적인 의견 자질 결합을 위한 실험적 연구.『정보관리학회지』, 27(3): 227-239.

[26] Blei, D., &Lafferty, J. 2006. “Dynamic topic models.” *The 23rd international conference on Machine learning*, 113-120.

[27] Blei, D. 2012. “Probabilistic topic models.” *Communications of the ACM*, 55(4): 77-84.

[28] Chen, H., & D. Zimbra. 2010. “AI and Opinion Mining.” *IEEE Intelligent Systems*, 25(3): 74-76.

[29] Gerrish, S., & Blei, D. 2010. “A language-based approach to measuring scholarly impact.” *The 27th International Conference on Machine Learning*, 375-382.

[30] Griffiths, T., & Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*.

[31] Grimmer, J. 2010. “A Bayesian hierarchical topic model for political texts: Measuring expressed

agendas in senate press releases." *Political Analysis*, 18(1): 1-35.

[32] Liu, Bing. 2010. "Sentiment Analysis: A Multifaceted Problem." *IEEE Intelligent Systems*, 25(3): 76-80.

[33] McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." 〈http://mallet.cs.umass.edu〉.

[34] Mimno, D., & McCallum, A. 2008. "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression." *The 24th Conference on Uncertainty in Artificial Intelligence*, 411-418.

[35] Newman, D., & Block, S. 2006. "Probabilistic Topic Decomposition of an Eighteenth-Century Newspaper." *Journal of the American Society for Information Science and Technology*, 57(5): 753-767.

[36] Schvaneveldt, Roger W. ed. 1990. *Pathfinder Associative Networks: Studies in Knowledge Organization*. US: Ablex Publishing.

[37] Song, Min., & Kim, Suyeon. 2013. "Detecting the knowledge structure of bioinformatics by mining full-text collections." *Scientometrics*, 96(1): 183-201.

[38] Steyvers, M., & Griffiths, T. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*. Edited by T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch. NJ: Erlbaum.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Kam, Miah, & Song, Min. 2012. "A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis." *Journal of Intelligence and Information System*, 18(3): 53-77.

[2] Kang, Myungkoo. 2004. "Media War and the Crisis of Journalism Practices." *Korean Journal of Journalism & Communication Studies*, 48(5): 319-421.

[3] Kim, Youngwook. 2011. "The Partisanship of Korean Media and The Crisis of Social Interaction." *Korean Society For Journalism And Communication Studies symposium seminar*, 2011: 107-136.

[4] Kim, Jaehong. 2003. "Editorial Tone of Major Korean Newspapers toward the Sunshine Policy during the Kim Dae Joong Government." *Korean Political Science Review*, 37(2): 197-218.

[5] Kim, Jungah, & Chae, Baek. 2008. "The Political Attitude of Newspapers and the Coverage of Political Scandal." *Journal of Communication & Information*, 41: 232-267.

[6] Park, Ja-Hyun, & Song, Min. 2013. "A Study on the Research Trends in Library &

Information Science in Korea using Topic Modeling." *Journal of the Korean Society for Information Management*, 30(1): 7-32.

[7] Park, Jaeyoung. 2009. "The Partisanship Topography of Korean Presses." *The Summary Report of The Seminar on Korea Press Foundation*, 17-65.

[8] Shin, TaeBeom, & Kweon, Sanghee. 2013. "A Study of The Relationship between Domestic Youth's Portal News Usage Characteristics and News Trust with Publicness Recognitions." *Journal of Cybercommunication*, 30(1): 241-294.

[9] Song, Hye-Ji, Park, Kyung-Soo, Jung, Hye-Eun, & Song, Min. 2013. "Trend Analysis of Korean Economy in the Economic Literature by text mining techniques." *Proceedings of the Korean Society for Information Management*, 20: 47-50.

[10] Yun, Seongyi. 2012. "Diffusion of Social Network Service and Its Challenge to Representative Democracy." *Journal of Korean Politics*, 21(2): 145-168.

[11] Yoon, Youngchul. 2000. "Power Shift and News Policy toward North Korea: An analysis of press-party parallelism." *Media and Society*, 27: 48-81.

[12] Lee, Minwoong. 2003. *Journalism: Crisis Change Endure*. Seoul: Nanam.

[13] Lee, Jaekyung. 2004. The Crisis of The Journalism and The Future of The Media. *The 40th Anniversary Seminar on Newspaper and Broadcasting*, Seoul: Korea Press Center

[14] Lee, Jaeyun. 2006a. "A Study on the Network Generation Methods for Examining the Intellectual Structure of Knowledge Domains." *Journal of the Korean Library and Information Science Society*, 40(2): 333-355.

[15] Lee, Jaeyun. 2006b. "Centrality Measures for Bibliometric Network Analysis." *Journal of the Korean Library and Information Science Society*, 40(3): 191-214.

[16] Lee, Jaeyun. 2006c. "A novel clustering method for examining and analyzing the intellectual structure of a scholarly field." *Journal of the Korean Society for Information Management*, 23(4): 215-231.

[17] Lee, Jaeyun. 2012. WNET. (version 0.4). (Software).

[18] Rhee, Junewoong. 2001. "Impacts of News Frames in the Coverage of Conflicting Issues on Individual Interpretation and Opinion." *Korean Journal of Journalism & Communication Studies*, 46(1): 441-482.

[19] Lee, Ji-Hye, & Chung, Young-Mee. 2009. "An Experimental Study on Opinion Classification Using Supervised Latent Semantic Indexing(LSI)." *Journal of the Korean Society for Information Management*, 26(3): 451-462.

[20] Jin, Seol-A, Heo, Coeun, Jeong, Yoo-Kyung, & Song, Min. 2013. "Topic-Network based Topic

Shift Detection on Twitter." *Journal of the Korean Society for Information Management*, 30(1): 285-302.

[21] Cha, Hanpil. 1989. *The Study on the Topic of Domestic Paper's Editorials and Correlation between Newspapers*. M.A. thesis, Yonsei University.

[22] Choi, Minjae, & Kim, Jaeyoung. 2008. "Fairness of Portal News Service in the 2007 Presidential Election." *Journal of Communication Science*, 8(4): 667-701.

[23] Choi, Jinho, & Han, Dongsub. 2012. "The Partisanship of Media and the Media Intervention in Political-power Creation in Korea: Focusing on the Analysis of the Major Newspapers' Editorial Articles during the 13-17th Presidential Election Campaigns." *Journal of Communication Science*, 12(2): 534-571.

[24] Choi, Hyunju. 2010. "A Study on the Diversity of Korean Newspapers: Analyzing the Tendencies of Covering Three Major Issues." *Korean Journal of Journalism & Communication Studies*, 54(3): 399-426.

[25] Han, Kyung-Soo. 2010. "Experimental Study for Effective Combination of Opinion Features." *Journal of the Korean Society for Information Management*, 27(3): 227-239.