

빅데이터를 이용한 지역미래전략 수립에 관한 시론적 연구 - SNS 및 언론기사 분석을 통하여 -

A Tentative Approach for Regional Futures Strategy with Big Data - Through the Analysis using the Data of SNS and Newspaper

임 화 진* · 박 성 현**

Lim, Hwa Jin · Park, Sung Hyun

요 지

본 연구는 빅데이터를 활용하여 지역미래전략을 수립하기 위해 충청남도를 사례로 선정하여 실험적인 분석을 실시하였다. 구체적으로 지역 현안을 보도한 언론기사와 SNS분석을 바탕으로 한 여론을 구조적으로 분석하고, 이러한 여론 분석으로 도출해 낸 키워드를 정책과 연결하여 정책과급양상을 공간적·가시적으로 표현하였다. 언론기사 분석 결과, 2013년 상반기에는 주로 정치, 경제적 이슈가, 하반기에는 문화 관련 이슈 등이 중점적으로 표출되는 것을 발견하였다. 트위터 분석을 통한 충청남도의 정책 관련 키워드 구조에서는 천안시가 전체 정보네트워크 안에서 중요한 허브(hub)로 추출되었다. 또한 충청남도의 대표 정책인 사회적 경제와 3농혁신에 관한 각 키워드의 특징을 분석하여 각 키워드의 특성과 충청남도와의 관계를 조망할 수 있었다. 이러한 분석은 빅데이터 분석을 이용한 광역지방 정부의 정책현안과 특징을 반영한 전략수립의 기초자료로 활용될 것으로 기대한다.

핵심용어 : 빅데이터, 미래전략, 텍스트마이닝, 충청남도, SNS

Abstract

This paper experimentally analysed Big Data for futures strategy in local government, the case of Chungcheongnam-do province. The analysis consists the structural analysis of the evaluating for policy, analysing public opinion using the newspaper which reported on regional issues and SNS data as public opinion, and the visualization of the result. The result of analysis using newspaper mainly indicates political and economic issues in the first half, and issues related to culture in the second half. From the result of Twitter analysis, the policy keyword structure are relevant of Cheonan City as network hub, which means all of the information concentrated. In addition, we analysed the characteristics, the relationship and structure in SNS big data of each representative policy keyword of Chungcheongnam-do, for example 'social economy' policy and '3 nonghyeoksin'(3 innovation for agriculture). These analyse could be expected to be a material for the implementation of futures strategy of wide-area local government.

Keywords : Big Data, Futures Strategy, Text-mining, Chungchonnang-do, SNA

투고 : 2015년 2월 25일 · 심사(수정) : 2015년 3월 19일 (4월 14일) · 게재확정 : 2015년 4월 25일

본 논문은 충남발전연구원 전략과제 '빅데이터를 이용한 충남도 정책 키워드 분석'을 토대로 재분석·구성한 것임

* 주저자, KAIST 미래전략연구센터 선임연구원 (E-mail : verite2484@kaist.ac.kr)

** 정회원 · 교신저자, 목포대학교 임해지역개발연구소 특별연구원 (E-mail : sunghyun7908@gmail.com)

1. 서 론

최근 행정과 빅데이터(Big Data) 및 오픈데이터를 융합한 정부 3.0과 지방 3.0이 큰 반향을 일으키고 있다. 정부 3.0과 지방 3.0은 각각 중앙정부와 지방정부의 정책 추진과정에서 데이터 과학(data science)을 적용하여 효율적인 정책을 추진하고자 하는 움직임이다.

빅데이터를 활용한 민간부문의 마케팅이나 수익창출 모델은 이미 수년전부터 정책추진 및 평가를 시도하는 연구에서 국내·외를 막론하고 활발하게 추진되고 있으며 행정혁신부문에서도 많은 사례가 존재한다. 그 중 정부 3.0의 공공분야의 데이터 활용사례로 국토교통부와 LH공사와 LX공사 등이 함께 구축한 온나라 부동산 시스템, 비정형데이터를 활용한 이동통신 통화 데이터를 이용한 서울시 심야버스 노선 설정, 다음(Daum) 지도와 재해 정보 연계 및 범죄데이터 결합 등을 들 수 있다. 이처럼 다양한 분야에서 많은 정보를 공개하고 이용하는 움직임이 가속화 되고 있다.

최근, 지방자치의 성숙과 초국가적 도시권의 발달로 지역의 독자적인 전략수립이 시급해지면서 미래전략에 대한 광역정부의 관심이 높아지고 있다. 즉, 지역의 비전을 수립하고 그에 걸맞은 전략을 수립하기 위해 광역정부의 이슈를 효율적으로 모니터링하고 현황을 진단하는 것이 시급한 과제로 여겨지고 있다. 이 때 빅데이터를 통해 새롭고 입체적인 대안의 도출이 가능해짐에 따라 불확실한 미래를 다양하고 더욱 정확하게 예측하

고 대비할 수 있다.

이상의 시대적 흐름을 포착한 LX공사는 최근에 빅데이터를 활용, 지방행정을 지원하는 사업에 본격적으로 착수하고 있다.¹⁾ 전국 최초로, 공사 본사가 소재한 전주에 전라북도와 공동으로 '빅데이터 활용센터'를 개설하였다. 이 센터는 전라북도와 관련된 공공데이터를 확보하고 가공·분석하여 정책수립과 신사업 창출, 의사결정 등을 지원하는 역할을 하게 된다. 한편, 국가차원에서 미래예측 및 전략 수립과 관련하여 한국정보화진흥원과 미래창조과학부가 함께 추진하고 있는 데이터 기반 미래전략 수립이 격화되고 있다. 이 사업을 통해 빅데이터를 기반으로 미래예측 및 연구의 허브를 조성하여 활발한 연구를 전개해 나갈 예정이다.²⁾ 이처럼 빅데이터를 응용한 미래전략 수립연구는 국가차원에서는 상대적으로 활발히 진행되고 있으나, 이를 광역정부에 걸맞게 응용하는 단계에는 전라북도의 사례와 같이 아직 맹아기(萌芽期)에 머무르고 있다.

이러한 배경 하에 본 연구에서는 빅데이터를 활용한 지역미래전략을 도출하고자 한다. 연구 사례 지역은 광역정부 중 충청남도로 선정하고,³⁾ 해당 광역정부에서 적극적으로 추진하고 있는 정책에 대한 여론의 반응과 수요자의 반응이 어떠한지를 측정하여 정책의 파급효과에 대해 고찰해 보고자 한다. 즉, 빅데이터를 이용한 충청남도의 정책에 관한 모니터링을 실시하여 지역별, 시기별 맞춤형 정책수립을 위한 기초분석 결과를 바탕으로 향후 정책 모니터링의 여론 및 나아가서 민원 등의 반

1) 최근, LX공사에서는 지적측량 사업의 한계를 인식하고 새로운 먹거리 창출을 위하여 공간 빅데이터에 대한 관심을 표명하고 있다. Franklin & Hane(1992)가 언급한 "정보의 80%는 지리공간과 관련 있다"는 말처럼 모든 빅데이터의 80%는 지리공간과 관련 있기 때문에 빅데이터를 공간정보 기반으로 융합하여 구축하게 된다면, 빅데이터를 융합한 공간정보시장에서 공사가 주도적인 역할을 할 것으로 기대하고 있다. LX공사는 어느 조직보다 국토공간에 대한 정보를 다량으로 보유하고 있기 때문에 보다 유리한 입장에 있다.

2) 빅데이터 기반 미래 전략 수립 본격화...미래부-NIA, 미래전략센터 개소식, 2014.9.19. 전자신문 <http://www.etnews.com/20140918000301>

3) 충청남도를 사례지역으로 선정한 이유는 본 연구자가 소속되었던 기관을 통해 공식적으로 비정형데이터의 수집이 가능하였기 때문이다.

영 체계에 대한 실험적 분석을 실시하고자 한다.⁴⁾

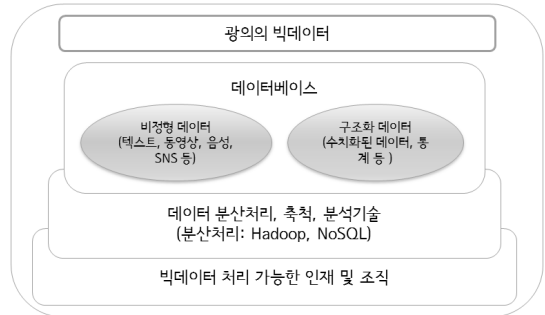
이상의 방법론과 연구결과의 의의는 ‘전라북도·LX 빅데이터 활용센터’에서 참조 가능한 모델로 활용가능하며, 맞춤형 정책수립이 가능하며 행정비용을 대폭 감소시킬 수 있다. 또한 정책투입 일변도가 아닌 수요응답형 모니터링(Demand Response Monitoring)으로 행정혁신의 접근이 가능하게 된다.

2. 이론적 배경

2.1 빅데이터와 미래전략

빅데이터(Big Data)란 대용량 데이터를 활용 및 분석하여 가치 있는 정보를 추출하고, 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술과 그 데이터 자체를 일컫는다. 초기에는 기술 측면에서 접근하여 데이터 자체만을 지칭하였으나, 현재는 수집, 저장, 검색, 공유, 분석, 시각화 등 관련 제반 기술을 폭넓게 포함하고 있으며 분석도구 및 인재, 조직으로 그 의미가 확대되고 있다.

빅데이터의 가장 큰 특징은 지금까지 잘 다루어지지 않았던 비정형데이터를 활용한다는 점이다. 특히, SNS데이터와 같은 방대한 텍스트데이터를 손쉽게 처리 및 분석하는 부분이 빅데이터 관련 분야 중에서도 가장 주목받고 있는 부분이다. 또한 SNS데이터는 개개인의 취향을 직접 반영한 시간, 공간이 파악 가능한 데이터로서, 이를 통해 기업 및 고객, 행정 및 정책수요자간의 쌍방향 소통에 있어서 유용하게 활용될 수 있다.



자료 : 시로타 마코토, 2012, “일본의 빅데이터의 현황과 과제 -사업전략을 발굴하기 위한 조직체제와 인재가 키-”, *IT프론티어*, 노무라종합연구소, pp.6-9.

그림 1. 빅데이터의 개념과 범위

빅데이터를 정책에 활용하는 방안으로 여러 가지 방안이 존재하나, 본 논문에서는 미래학 분야 중 행정기관의 미래전략수립을 응용한 빅데이터 분석 방안에 착안하고자 한다.

미래전략은 미래예측(Foresight)을 통한 미래 먹거리 발굴과 정책의제 발굴 등 미래발전전략의 줄임말로 사용되고 있으며 미래예측을 통한 시나리오 및 대응방안을 구상하는 것을 일컫는다. 나아가 데이터기반의 미래전략이란 데이터를 정책 설계 과정에서 필요한 정책수단 또는 도구로써 활용하여 전략을 수립한다는 의미를 지닌다.⁵⁾ 다음 표 1은 데이터기반의 국가미래전략을 전개하고 있는 국가와 국제기구의 현황을 보여주고 있다. 이와 같이 다양한 국가미래전략이 존재하나, 싱가포르의 RAHS와 영국의 호라이즌스캐닝 센터, EU의 iKNOW프로젝트 등은 빅데이터 분석을 결합한 정보의 광범위한 탐색을 실시하여 전략을 도출하는 프로세스로 진행하였다. 본 연구에서 시도하고자 하는 분석 또한 이들과 맥락을 같이한다고 볼 수 있다.

4) 구체적으로 지역 현안을 보도한 언론기사와 SNS(Social Networking Service)분석을 바탕으로 한 여론을 구조적으로 평가하고, 여론 분석 결과 도출해 낸 키워드를 정책과 연결하여 정책파급양상을 공간적·가시적으로 표현하고자 한다.

5) 송영조, 2013, “성공적 공공정책 수립을 위한 미래전략연구방법론”, *IT & Future Strategy*, 한국정보화진흥원.

표 1. 데이터기반의 국가미래전략 현황

국가(기구)	미래전략 현황
싱가포르 RAHS 프로그램	• 싱가포르 정부의 국가안전보장사무국(NSCS)는 국가안전을 위협하는 요소에 대한 평가와 주변 환경변화를 탐지하여 새로운 기회를 발굴하는 RAHS 프로그램을 2004년부터 운영 - RAHS는 싱가포르 환경탐색(Horizon Scanning)을 통해 획득한 데이터를 분석하여 국가적 영향을 미칠 수 있는 잠재적 위협요소와 불확실성 요소를 탐색하여 이머징 이슈를 분석 - 데이터분석 실험센터(REC, RAHS Experimentation Centre)를 2007년 창설하여 새로운 미래 연구 방법론 및 기술들을 개발·탐구하고, 해외 연구소와 긴밀한 협력관계를 유지
미 국	• 미국 과학정책자문위원회는 ‘Designing a Digital Future’를 발표(2010.12)(하고 대규모 데이터(large-scale data)의 가치에 주목하고 데이터 취합, 관리 및 분석의 중요성을 제시 - 미국 정부기관들은 데이터를 지식으로 지식을 행동으로 변환하는 전략에 집중해야 함을 주장 • 그리고 2010년 “모든 연방정부는 빅데이터 전략 수립이 필요”하다고 제시하고, 2012년 3월 빅데이터 연구개발에 2억 달러 이상을 투입하는 빅데이터 연구개발 이니셔티브를 발표 - 빅데이터 연구개발 이니셔티브 참가기관은 유전자 연구 및 의료, 교육, 지구과학 및 국방분야 등 빅데이터 활용 효과가 뛰어난 분야의 기관들이 우선적으로 참여 - 국립과학재단(NSF), 국립보건원(NIH), 국방부(DoD), 고등방위연구계획국(DARPA), 에너지부(DoE), 지질조사원(USGS) 등이 참여
영국의 호라이즌 스캐닝센터 (The Foresight HSC)	• 최근 과학이론과 데이터 등 증거기반의 정책분석 서비스를 제공하고 정부의 혁신적 전략 및 중장기 미래 정책개발을 지원하기 위해 2005년 수립 • 주요기능 - 미래신성장동력 발굴(Future Projects) : HSC는 다양한 주제와 이슈들의 연관분석을 통해 10~15년 이후 미래를 전망 - 미래연구역량강화 및 기법개발(Training and tools) : HSC는 정부의 미래전략 및 정책개발을 위해 필요한 미래 분석 가이드 및 방법론 제시 - 시그마 스캔(Sigma Scan) : STEEP(정치, 경제, 사회, 환경, 기술)별로 50년 이후의 정책 이슈를 포괄하는 연구 및 정책 보고서 검색 서비스 - 미래연구 분석가 네트워크 구축 : 호라이즌 스캐닝과 미래연구자들간의 인적 네트워크
한 국	• 국가정보화전략위원회는 ‘빅데이터를 활용한 스마트 정부 구현(안)’을 제시(2011.11)하고 빅데이터 시대에 대비하여 융합 지식과 분석 행정을 통한 스마트 정부 실현을 위한 세부과제를 도출
OECD	• 빅데이터를 비즈니스 효율성을 제공하는 새로운 자산으로 인식 • 제15차 WPIS회의에서 빅데이터의 경제학 측정을 의제로 채택 : 무형 자산인 빅데이터의 경제적 가치를 실증하고 정책적 함의 제공이 목표
EU	• EU는 금융위기 극복과 사회의 복잡성을 이해하기 위한 FuturICT와 불확실한 미래탐구를 위한 iKnow 프로젝트를 추진 - FuturICT는 지구 신경망시스템, 전지구 시뮬레이터와 글로벌 참여 플랫폼을 구현하여 세계 변화의 방향과 새로운 지식을 탐구 - iKnow프로젝트는 전세계의 약신호(weak signal)와 와일드 카드(wild cards)를 포착함으로써, 미래를 형성하는 지식과 전략적 이슈를 발굴

자료 : 한국정보화진흥원, 2013, “데이터 기반 국가미래전략을 위한 정책과제”, IT & Future Strategy 제3호, pp.17-18.

2.2 선행연구 검토

빅데이터의 성질을 규정하는 가장 큰 부분인 비

정형데이터는 음성, 사진, 동영상 등을 일컫는다. 따라서 이들을 구성하는 텍스트데이터는 빅데이터에서 큰 비중을 차지하고 있으며 이를 분석하는

텍스트마이닝(Text Mining) 또한 빅데이터 분석의 중요요소라고 할 수 있다. 최근까지 주로 행해져 온 텍스트마이닝의 주요 데이터소스는 신문기사, 검색어, 트위터, 그 외 데이터를 들 수 있다.

신문데이터는 언론정보학 분야에서 연구가 풍부하게 진행되었다. 그러나 이전까지는 기계적인 텍스트마이닝보다는 주로 일대일로 읽고 내용을 분석하는 연구방법이 주로 행해져 왔다. 최근 들어 데이터 마이닝 툴이 발달함에 따라 대용량의 신문기사 및 여론 자료를 분석하는 논문들이 발표되기 시작하였다. 그 예로 감미아·송민(2012)⁶⁾를 들 수 있는데, 이 연구에서는 주요 신문사의 논조 비교를 대용량의 데이터 분석을 통해 구현했으며 구체적으로 어떤 단어에 대한 논조의 차이가 존재했는지 검토를 시도하였다.

최근에는 구글(Google) 검색어를 이용한 텍스트마이닝도 새로운 시도로 조망 받고 있다. 한국과학기술기획평가원(KISTEP)에서 수행한 미래 트렌드 분석 연구⁷⁾는 텍스트마이닝과 네트워크 분석을 활용한 연구로서 구글검색어를 데이터베이스로 정하고 기존의 텍스트마이닝 지표를 개량한 새로운 빈도수 지표를 활용하는 등 다양하고 새로운 시도를 통해 이슈를 도출한 연구라고 할 수 있다. 이 연구에서는 검색엔진 및 논문 등을 통한 새로운 지식에 대해 추이를 분석하고 관계도를 그려보는 것이 미래예측의 하나의 방법론이 될 수 있다고 지적하고 있다.

이 외에 행정기관에서 다룰 수 있는 대표적인 데이터로 민원데이터를 들 수 있다. 중앙정부의 사례를 보면, 국민신문고 출범 이후 국민신문고

처리 민원을 분석한 연구로 민원 텍스트마이닝을 실시한 연구사례가 존재한다.⁸⁾ 그 분석 내용을 살펴보면 연도별 민원현황 및 총량적인 민원 추이 분석, 연령대별, 성별, 지역별 민원 추이 및 특성 분석, 민원 주제별, 민원 키워드별 추이 및 특성 분석 등이며 연도별·성별 민원키워드 TOP 100을 추출·분석하여 남·여 주요 민원 주제 5개를 각각 선정하여 키워드 분석 결과 공통 주제 4개와 성별 특성 주제 각각 1개씩 총 6개 민원 주제에 대해 분석하였으며 연령대와 성별로 각 부문별 민원 키워드 수를 집계하였다.

빅데이터와 SNS를 결합한 최초의 사례라고 평가되는 연구는 2008년 인디애나 주립대 Johan Bollen에 의해 행해졌다. 이 연구는 트위터⁹⁾데이터가 개인의 의사를 반영한 정보가 포함된 것을 발견하고 이에 착안했던 연구라고 할 수 있다. 구체적인 분석 과정은 반년간의 트위터데이터를 이용하여 오피니언 분석을 통해 기본변화를 측정하고 이와 관련된 연간 이벤트의 연관성을 도출하였다. 이러한 연구를 바탕으로 MIT연구그룹은 트위터데이터를 활용하여 행복도를 측정하는 연구를 시도하기도 하였다.

한편, 국내 연구에서는 특히 2012년 이후 한국에서는 SNS데이터, 특히 트위터에 대한 분석이 활발하게 실시되고 있다. 주로 기술적인 측면은 정보학분야에서 실시된 연구가 많으나 기술을 적용하여 새로운 시사점을 뽑아내기 위한 시도도 사회학이나 정치학 등 다양한 분야에서 이루어지고 있다. 예를 들어, 배정환 외 2(2013)¹⁰⁾는 국내 트위터를 분석하여 선거 결과를 조망하여 선거전

6) 감미아, 송민, 2012, “텍스트 마이닝을 활용한 신문에 따른 내용 및 논조 차이점 분석”, *지능정보연구*, 18(3), 한국지능정보시스템학회, pp.53-77.

7) 정근하, 2011, “텍스트마이닝과 네트워크 분석을 활용한 미래예측 방법 연구”, 한국과학기술기획평가원.

8) 국민권익위원회, 2013, “민원 키워드 중심으로 살펴본 국민신문고 출범 이후 민원 분석”

9) 트위터란 소셜 네트워크 또는 마이크로 블로그 서비스로서 1회 투고가능한 글자가 140자로 제한되어 있다. 전송하는 메시지를 트윗(tweet)이라고 하며, 상대방을 팔로우(follow)함으로써 상대방의 메시지를 구독할 수 있다. 다른 사람이 작성한 메시지를 재전송하는 것을 리트윗(RT)이라고 한다.

의 과정과 결과에 있어서 트위터의 구조 등을 본 연구이다. 또한 박재희·강영옥(2013)¹¹⁾의 연구는 트위터 데이터를 이용하여 도시정책지표를 구성하고 주거환경만족에 대한 공간적 특성을 도출하였다. 여기서는 트위터의 공간정보와 트위터 텍스트를 주거만족도로 해석하는 텍스트 마이닝 기법이 결합된 연구라고 할 수 있다.

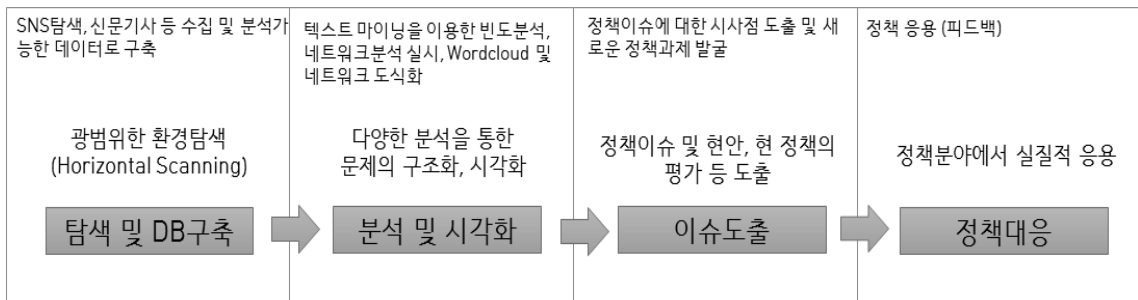
이상 기존 연구들을 살펴보면, 이미 빅데이터의 분석기술의 발전은 꾸준히 이뤄지고 있으며, SNA와 신문기사를 응용한 연구도 점차 늘어나고 있다. 그러나 정책적 관점에서 실제 지역을 대상으로 분석한 사례는 부족한 실정이다. 본 연구의 의의는 기존의 연구들을 지역미래 전략수립의 관점에서 응용한 점이다.

3. 연구방법 및 지표

3.1 자료수집

본 연구에서는 신문기사와 SNS를 이용해 충청남도가 적극적으로 추진해 온 주요 정책에 관한 키워드를 파악하여 도정(道政)에 관한 여론을 간접적으로 측정하도록 한다. 이처럼 두 가지 데이터를 병용하는 이유는 특징이 다른 두 매체를 통합적으로 이용하여 풍부한 해석을 얻고자 함이다. 신문기사는 주로 언론사를 중심으로 전개되는 공급자 위주의 미디어라고 한다면, SNS는 쌍방향 소통을 통한 공급자와 수요자가 동시다발적으로 정보를 생산해 내는 새로운 미디어 채널이라고 할 수 있어 두 데이터의 통합적 분석을 통한 정책 공급양상을 분석해 볼 수 있다.

본 연구는 빅데이터 기반의 과학적 정책수립의 방법론으로 그제 4단계의 과정을 거쳐 빅데이터 분석 프로세스에 따라 수행하였다(그림 2).



자료 : 박정은, 2014, “미래전략센터의 의미, 성과와 발전 방향”, *Future Horizon*, 제22호 수정·보완

그림 2. 빅데이터 기반 과학적 정책수립 방법론

언론데이터는 주로 웹페이지에서 모은 기사를 활용하였고 기간은 2013년 1월 1일부터 12월 31일

까지로 1년간을 수집하였다. 또한 트위터(twitter)는 공개트위터만을 기준으로 분석하였다(표 2).

10) 배정환, 손지은, 송민, 2013, “텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석”, *지능정보연구*, 19(3), 한국지능정보시스템학회, pp.141-156.

11) 박재희, 강영옥, 2013, “트위터 데이터를 통해 본 생활환경 만족도의 공간적 특성”, *대한지리학회지 학술대회논문집 6*, 대한지리학회, pp.383-386.

표 2. 수집 데이터 개요

데이터 종류	언론 데이터	트위터 데이터
수집 방법	포털사이트(네이버) 웹페이지의 뉴스 리스트	공개 트위터 분석 (수집 및 전처리하는 (주)사이람에 의뢰)
기간	2013.1.1. ~ 2013.12.31	
건수	12,959건 (충청남도 검색) 언론사 : 384개	주제1 : 약 30만 건 주제2 : 약 125만 건

언론데이터는 네이버 뉴스 라이브러리에서 2013년 1월부터 12월까지 1년 동안의 신문 기사를 수집하였고 이를 OCR¹²⁾ 등으로 정리하였다. 또한 100퍼센트 일치하는 기사는 보도자료 등의 복사로 간주하여 중복 수집하지 않고 1회로 제한하였다. 그 결과 총 12,959건의 기사가 추출되었다.

트위터 데이터를 추출한 방법은 먼저 공개 트위터 데이터 안에서 각각 주제1과 주제2를 포함한 트위터 데이터를 추출한 뒤 관계가 적은 데이터는 삭제하는 방법을 거쳤다. 이 때 주제¹³⁾은 지역에 관련된 주제어이며, 주제²¹⁴⁾는 충청남도의 핵심정책과 전반적으로 연결되어 있는 주제어와 지명 등 고유명사를 포함한 주제어로 나누어진다.

3.2 분석방법 : Text Mining

분석에 사용한 프로그램은 R이며, 구체적으로

사용한 R 패키지는 tm과 KoNLP이다.¹⁵⁾ 여기서 tm은 텍스트마이닝을 위한 패키지이며 KoNLP는 한국어 처리를 위한 패키지이다. 또한 한국어 사전은 sejong.dic에 충남관련 정책 키워드를 포함한 user dic을 구축하여 수행하였다.

텍스트마이닝의 여러 지표들은 수학적 알고리즘을 바탕으로 구축되어 있다. 가장 보편적인 지표 중 하나로 TF-IDF(Term Frequency-Inverse Document Frequency)를 들 수 있다. 이 TF-IDF는 TF와 DF의 역수를 곱한 지수로서 어떠한 단어의 중요도를 추출하는데 사용될 수 있다. 즉, 단순한 빈도처리가 아닌 단어의 출현 확률을 기준으로 출현빈도를 한 번 더 가공처리 하여 단어의 빈도수를 나타내고 있다. 여러 문서에 동시에 출현하는 단어는 출현 확률이 높다는 전제하에 역문헌 빈도수를 계산하여 DF가 커질수록 중요도가 감소하는 효과를 볼 수 있다.

$$TF_{i,j} = \frac{f_{i,j}}{\max f_{i,j}} \quad (1)$$

$$DF_i = \frac{n_i}{N} \quad (2)$$

$$IDF_i = \log \frac{N}{n_i} \quad (3)$$

$$w_{i,j} = TF-IDF = TF \times \frac{1}{DF} \quad (4)$$

TF : 문서내 특정 단어의 빈도수

DF : 여러 문서내의 특정 단어 빈도수

IDF : DF의 역수

12) OCR(optical character reader)이란 광학 문자 판독 장치의 약칭으로, 이것은 용지상에 타이프라이터나 라인프린터로 인쇄된 문자 또는 사람이 손으로 쓴 문자를 읽는 장치이다. 본 연구에서는 ReTIA의 ABBYY OCR 프로그램을 사용하였다.

13) 충청남도, 충남, 안희정, 충청&논산, 충청&서산, 충청&공주, 충청&부여, 충청&천안, 충청&예산, 충청&아산, 충청&서천, 충청&당진, 충청&홍성, 충청&보령, 충청&청양, 충청&금산, 충청&태안, 충청&계룡공주, 부여, 천안, 예산, 아산 등의 지역명은 충청남도과 전혀 상관없는 트윗을 추출하는데 영향을 미쳐 추출 키워드를 수정하였다.

ex) 백설공주, 동기부여, 천안함, 정부 예산, 아산병원 등이다.

14) 사회적경제, 사회적기업, 공유경제, 착한기업, 마을기업, 내포신도시, 3농혁신, 서해안, 농업직불금, 농촌마을, 지속가능발전, 행복지표, 협동조합, 6차산업, 미더유, 경제선순환, 에너지, 균형발전, 송전선로, 전통시장 등이다.

15) 통계 계산과 그래프를 위한 오픈소스 프로그램 언어이자 소프트웨어 환경이다. 자세한 사항은 R홈페이지(<http://www.r-project.org/>)를 참조하기 바란다.

그러나 TF-IDF모델은 일반적인 단어를 분석하는 측면에서는 매우 유용하나 고유명사나 새로운 개념을 분석할 때는 중요도가 과소 및 과대평가가 될 수 있어 주의를 요구한다. 따라서 본 연구에서는 이러한 한계점을 보완하기 위하여 키워드 빈도수를 중심으로 TF-IDF를 보완적으로 사용하였다.

다음은 연관어 분석을 사용하였다. 이 때, 연관어 분석에 앞서 연관관계의 정의가 불가피하다. 연관관계(association relationship)란 어떠한 단어에 대한 다른 단어들이 가진 패턴의 유사성을 의미한다. 즉, 함께 등장한 비율이 높은 경우 유사한 패턴이라고 정의한다. 연관어 분석 결과를 바탕으로 단어간의 연관 네트워크를 분석하였다. 이 때 연관등장 빈도를 네트워크의 가중치로 정하고 매개중심성¹⁶⁾ 지수와 Newman방식의 클러스터링을 통한 서브그룹 분석을 실시하였다.

4. 분석 결과

4.1 충청남도 여론 데이터의 공간분포

여기서는 ‘충청남도’ 또는 ‘충남’ 검색어가 지역별로 어떻게 분포되어 있는지 살펴보도록 한다. 데이터는 구글 트렌드를 이용하여 집계하였고 기간은 2010년 1월부터 2013년 12월까지이다. 전국 현황 중 가장 높게 나타난 아산시의 검색회수를 100으로 놓고 다른 지역을 상대적으로 표시한 수치를 나타냈다. 이 때 아산시를 100으로 봤을 때 천안시가 80, 그 다음으로 대전광역시가 28, 서울시가 10으로 나타났다. 즉 충청남도 내에서는 아산시, 천안시지역의 사람들이 충청남도라는 검색어로 가장 많이 검색했으며 그 외에는 충청남도

내 지역이 아닌 대전광역시와 서울시에 있는 사람들이 검색을 실시한 것으로 보인다.

한편, 아산시와 천안을 제외한 충청남도 내 다른 지역들은 충청남도에 대한 검색어 빈도수가 많지 않다는 것을 알 수 있다.

표 3. ‘충청남도’ 검색이력이 많은 지자체

해당 지자체	비중
아산시 (대한민국)	100
천안시 (대한민국)	80
대전광역시 (대한민국)	28
서울특별시 (대한민국)	10
인천광역시 (대한민국)	7
부산광역시 (대한민국)	6

4.2 언론기사 분석결과

먼저, 언론기사 전체 월별현황 및 검색어에 대해 살펴보면 다음과 같다. 네이버(www.naver.com)를 기준으로 ‘충청남도’라는 단어가 들어간 신문 기사를 추출한 데이터의 단어수를 살펴보았다.

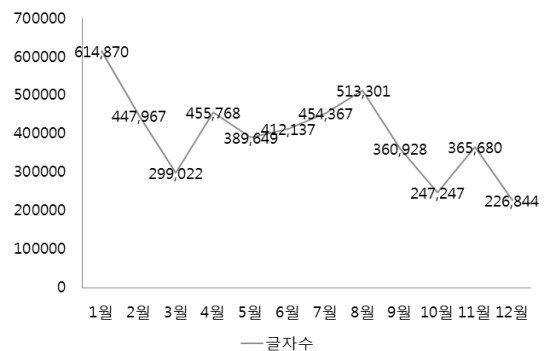


그림 3. 월별 기사 현황

16) 여기서의 매개중심성은 연관어 네트워크 상에서 어떠한 단어를 제외한 다른 모든 단어들의 최단경로 개수와 해당 최단경로에서 특정 단어가 존재하는 개수의 비율을 나타낸다. 즉, 어떠한 단어가 문장 구성의 네트워크 안에서 HUB역할을 하고 있는지를 판단할 수 있는 지표이다.

월별 기사량을 살펴보면, 1월이 가장 많았고 10월이 가장 적었다. 또한 전체적인 경향을 보았을 때 하반기보다는 전반기가 훨씬 많았다. 2013년 전체 기사의 단어 수는 4,787,780개이며 매월 평균은 398,981개로 분석되었다.

표 4. 신문기사 추출 키워드(빈도수 높은 순)

단어	빈도수	단어	빈도수
충청남도	4,961	글로벌	199
사업	3,483	균형발전	167
안희정(도지사, 충남도지사)	1,312	네트워크	162
내포신도시 (내포신청사)	851	문화예술	160
활성화	653	지역발전	158
디스플레이	397	주민자치	156
중소기업	368	고속도로	152
농업기술	299	지역주민	148
서해안	296	사회복지	145
신도시	294	워크숍	142
에너지	294	협동조합	136
백제문화	285	지속가능(발전)	136
친환경	229		

2013년도 충청남도라는 단어를 포함한 신문 기사를 수집한 데이터베이스를 이용하여 형태소 분석을 거친 후, 명사를 중심으로 단어(키워드)를 추출한 결과를 살펴보았다. 이 때 출현빈도가 120회¹⁷⁾가 넘는 단어를 추출하고 동일한 의미를 가지는 단어를 통합한 결과를 표 4에 나타내었다.

전체 키워드 빈도수를 살펴보면, 지역명인 ‘충청남도’가 가장 높은 빈도수를 나타냈으며, 다음으로 ‘사업’이라는 단어의 빈도가 높은 것을 알 수 있다. 다음으로 내포신청사를 포함한 ‘내포신도시’

가 높게 나타났다.

다음은 신문기사텍스트를 이용하여 키워드를 분석한 결과를 살펴보았다. 여기서는 TF-IDF를 이용하여 세 글자 이상인 키워드 중에서 TF-IDF가 0.5 이상인 단어를 월별 이슈로 선정하였다. 분석결과, 3월, 4월, 8월, 9월, 12월에는 특별한 이슈는 추출되지 않았다. 한편 1월은 중소기업지원, 산업단지, 디스플레이, 국제통상 등의 지역경제와 관련된 이슈가 추출되었는데 이는 연초의 도정 계획 및 언론의 관심이 주로 경제분야에 주목되어 있다는 것을 알 수 있다.

2월에는 외국인 투자기업, 에너지사업, 상생산업단지에 관한 기사가 주된 키워드로 부상하였다. 이는 1월에서 이어진 경향으로 산업분야의 화제를 보여주고 있으며, 에너지 등에 관한 언급도 있었다는 것을 알 수 있다. 또한 상생산업단에 대한 관심이 급증했던 때라고도 할 수 있다. 이후 5월, 6월에는 부곡산업단지, 투자설명회로 이어지는 지역경제와 관련된 이슈가 추출되었다. 7월부터는 경제적인 화제가 아닌 다른 측면에서도 많은 키워드들이 부상하였다. 보령머드축제와 같은 문화관광에서의 이슈도 주목을 받았다.¹⁸⁾ 9월에는 다른 달에 비해 부여 롯데아울렛 및 롯데백화점에 관한 화제가 주된 기사였다. 이와 관련해서 지속가능발전도 함께 언급되었던 점이 특징적이라고 할 수 있다. 11월은 충청남도내 송전선로에 관한 이슈도 많이 언급되었다.

이를 종합하면, 크게 경제정책(산업, 통상, 농업, 외국인 투자 등)과 개발사업(아울렛, 리조트, 송전선), 이벤트(축제 등)가 주요 키워드로 추출되었다. 그러나 충청남도의 정책적 화두인 3농혁신, 사회적 경제, 지속가능발전 등과 같은 키워드는 크게 관심을 끌지 못했다고 할 수 있다.

17) 최소 출현빈도 설정은 한달에 10번 이상 빈도로 언급되었다고 설정하였으며, 이 때 총빈도수는 120회가 된다.

18) 김수근문화재단의 경우에는 7월에 건축가 김수근을 기리는 기사가 작성되고 반향을 부름에 따라 국립부여박물관과 함께 언급된 것이 영향을 미쳤다고 할 수 있다.

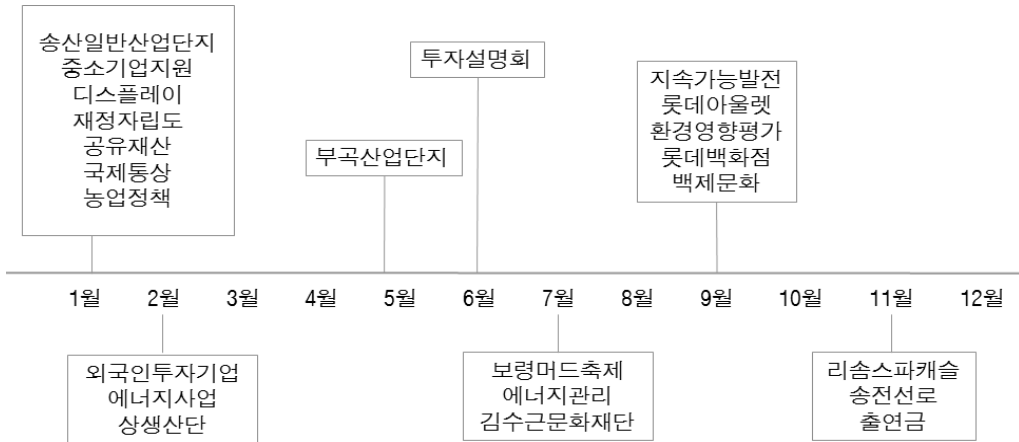


그림 4. 월별 키워드 추출(TF-IDF 이용)

4.3 트위터 분석결과

먼저, 그림 5의 기초 데이터를 살펴보면, 2013년 한 해 동안 6월이 가장 많은 트윗수를 나타냈고 6월을 기준으로 상반기보다 하반기가 트윗수가 많아진 것으로 보아 충청남도에 대한 트윗노출도가 하반기로 갈수록 증가했다는 점을 알 수 있다. 이를 바탕으로 하반기에 들어서서 충청남도

에 대한 관심도가 상승했다고 유추할 수 있다. 주요 트윗 내용을 살펴보면, 6월에는 충남도민체전에 아이돌스타 출연으로 인한 팬들의 관심, 안희정 충남지사 관련뉴스, 국정원 관련 충남지역 시국선언이 있었고, 7월에는 해병대 사설캠프 사건, 호두과자 업체 노무현 비하 사건, SKT 홍보글 등이 화제를 일으켜 트윗터수를 증가시킨 요인으로 작용했다.

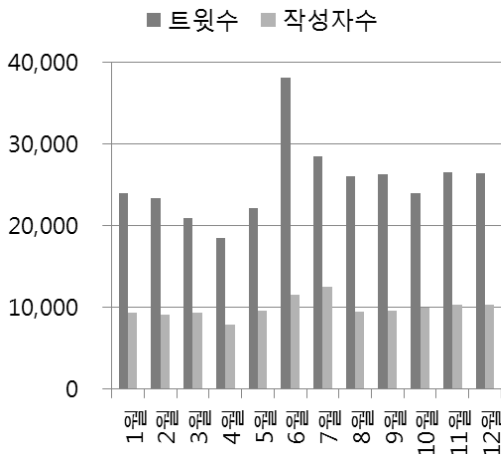


그림 5. 충청남도 관련 트윗수 및 작성자수

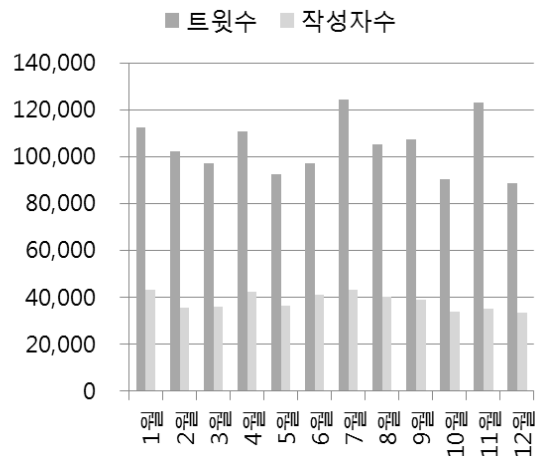


그림 6. 충청남도 정책키워드 관련 트윗수 및 작성자수

충청남도 정책키워드 관련 트윗수(그림 6)는 이와 다르게 월별 경향이 크게 다르지 않은 것을 발견할 수 있다. 전체 트윗수는 충청남도를 언급한 트윗보다 훨씬 많은데 이는 각 정책이 충청남도만이 아닌 다른 지역에서도 전개되고 있으며 보편적인 사항이기 때문이기도 하다. 세부적으로 살펴보면, 7월과 11월에 정책키워드를 언급한 트윗이 다른 달에 비해 많은 반면, 5월과 10월, 12월에는 다소 적었다. 그러나 작성자수는 크게 다르지 않은 것으로 보아 고정된 인원이 정책키워드에 관해 관심을 가지고 언급하였다는 점을 알 수 있다.



(왼쪽 : 1월~6월, 오른쪽 : 7월~12월)

그림 7. 충청남도 각 시군별 관련 트위터 수 현황

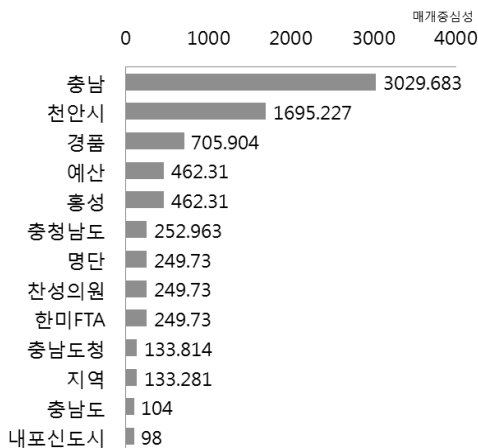


그림 8. 트위터 연관어 네트워크의 매개중심성 지수
(상위 13개 단어)

한편, 시군별 트윗수를 살펴보면 전체적으로 천안시, 공주시에 관한 언급이 주로 되고 있다는 점을 알 수 있다. 상반기와 하반기로 분류하면 주로 상반기보다 하반기에 들어서서 트윗수가 증가하는 것을 알 수 있다. 특히 하반기에 들어서는 천안시와 아산시에 관한 언급이 가장 많았다.

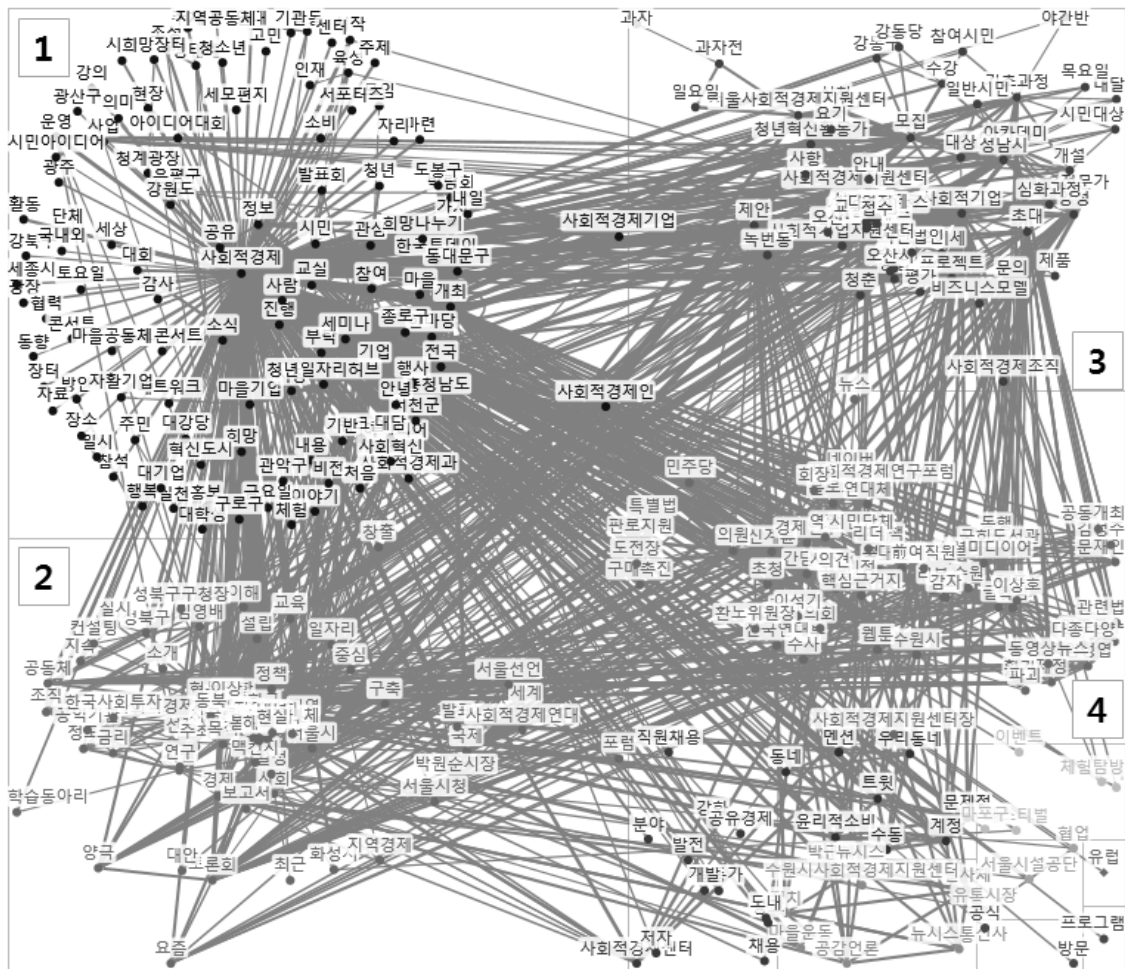
둘째, 충청남도(또는 충남)에 관한 연관어분석 결과로 다음 표 5와 같은 결과를 얻을 수 있었다. 충남에 관한 고유명사로는 충남도청, 충남대학교, 내포신도시, 교육청에 관한 단어가 다수 등장하였다. 다음으로 충남과 연관되어 등장하는 인물로서는 안희정 충남지사, 노홍철, 김호연, 김종성, 이회창 순이었다. 일반명사에 관해서는 행사에 관한 주목 및 관심도가 매우 높았던 관계로 경품, 추천, 이벤트 등과 같은 단어가 빈도수가 높았다. 행사관련 키워드를 제외한 다른 단어에서는 농축수산물에 관한 언급도 매우 높은 수준이었다. 지역명에 대한 결과를 살펴보면, 천안, 아산, 대전 순으로 빈도가 높았으며, 천안이 다른 지역에 비해 압도적인 빈도수를 나타냈다. 한편 다른 충남 지역의 노출도는 대전이나 충북, 경기보다 낮은 수준이었다.

셋째, 충청남도 관련 트위터 연관어 네트워크 중 매개중심성이 높은 키워드를 도출해 보면, 그림 8와 같다. 충남도 관련 트위터 중 천안시에 관한 매개중심성이 높은 것으로 보아 천안시가 충남도에 관련된 정보 중에서 다양한 상황에서 등장하고 있다는 점을 지적할 수 있다. 다음으로 충남도청 이전과 관련한 키워드에 관해서도 높은 매개중심성을 나타내고 있다고 볼 수 있다.

마지막으로 앞에서 추출된 연관어를 이용하여 각 중심 키워드를 중심으로 형성되어 있는 연관어 네트워크를 살펴보도록 한다(그림 10). 앞에서 언급한 바와 같이 여기서 네트워크의 강도는 연관등장 빈도이며 사각형은 Newman 방식으로 클러스터링한 서브그룹을 나타낸다. 20개의 정책

표 5. 트위터 추출 키워드(빈도수 높은 순)

충남 고유명사		인물		일반명사				지역명			
키워드	빈도수	키워드	빈도수	키워드	빈도수	키워드	빈도수	키워드	빈도수	키워드	빈도수
충남도청	8,623	안희정('충남도지사' 포함)	11,556	경품	8,287	개청	4,879	천안	15,407	강원	4,829
충남대(학교)	15,540	노홍철	2,628	추천	5,732	상품권	4,755	아산	6,887	경기	4,677
내포신도시(청사)	11,771	김호연	2,017	이벤트	5,605	지역	4,732	대전	6,539	서산	4,610
충남교육청	3,278	김종성	1,661	확인	5,167	퀴즈	4,680	공주	5,664	전북	4,085
		이회창	1,657	희망	4,999	농축수산물	4,490	충북	5,085	세종	4,083



주 : 사각형은 클러스터를 의미함

그림 9. ‘사회적 경제’ 관련 연관어 네트워크

있으며 협동조합 등의 언급도 볼 수 있다. 클러스터 4는 수원 사회적경제센터가, 클러스터 5에서는 강동구가 속해 있었다.

전체 네트워크에서 보았을 때 사회적경제와 굵은 선으로 연결되어 있는 단어가 서울시, 부락, 활성, 개최, 전문가, 전국, 사회적기업, 협동조합 등이다. 다만, 충남은 서울시보다는 약한 연결고리를 보였다.

한편, 3농혁신을 중심으로 한 연관어 네트워크에서는(그림 10) 가장 중요한 키워드가 3농혁신이 아니라 충남, 충청남도였다는 점이 특징적이다. 다만 충남에 관련된 명사가 대부분이며 주요 키워드들도 추진, 사업, 선정 등의 도정 보도자료나 정책기사를 벗어나지 못하고 있다는 점을 알 수 있다. 총 클러스터는 11개가 존재하나 의미 있는 클러스터는 3개 정도이다. 클러스터 1에서는 충남도지사의 발언 및 관계자를 중심으로 한 농업 중요성 강조 및 행사 개최 및 참석에 대한 언급이 대부분이었다. 다음으로 클러스터 2에서는 민선 5기, 충남도, 추진상황, 보고 등 민선 5기의 정책성과로 연결 짓는 언급이 연관되어 있다. 마지막으로 클러스터 3에서는 충남을 중심으로 몇 개의 촘촘한 네트워크로 연결되어 있는데 주로 행사 및 홍보, 농업기술 및 농업관련, 농어촌 등으로 나뉠 수 있다. 전체적으로 추상적인 명사가 많은 점이 특징적이다.

사회적 경제와 3농 혁신에 관한 분석을 종합해보면, 사회적 경제는 전국적인 화두로 인식되고 있으며, 구체적으로 서울, 수원, 성남 등과 같은 지역명이 대두되고 있고, 이 중 하나로 충청남도가 언급되고 있다는 것을 알 수 있다. 이와는 대조적으로 3농혁신은 충청남도 고유의 정책으로 거의 대부분의 언급이 충청남도와 직접적으로 연관이 있는 키워드지만 전국적인 파급효과가 있다고는 보기 힘들며 아직 추상적인 단계의 사업들이 대부분이라고 할 수 있다.

4. 결 론

본 연구에서는 언론기사와 트위터를 대상으로 충청남도와 관련된 기사 및 월별 이슈를 추출하였다. 언론기사 추출결과로는 상반기에는 주로 정치, 경제적 이슈가, 하반기에는 문화관련 이슈 등 조금 더 폭넓은 이슈를 발견할 수 있었다. 트위터 분석을 통한 충청남도 정책 관련 키워드 구조에서는 천안시가 전체 정보네트워크 안에서 중요한 HUB로서 추출되었으며 이는 충청남도 내의 다양한 화제들이 천안시와 밀접한 관계를 이룬다고 볼 수 있다.

SNS미디어인 트위터를 대상으로 충청남도의 대표적인 정책인 사회적 경제와 3농혁신에 관한 각 키워드의 특징을 분석하여 각 키워드의 특징과 충청남도와의 관계를 조망할 수 있었다. 현재, 사회적 경제는 전국적인 화두로 인식되고 있는 만큼 충청남도를 비롯하여 그 이외의 지역인 서울, 수원, 성남 등 다양한 지역과 연관관계가 있었다. 또한 여러 가지 주체들이 얹혀있는 열린 네트워크를 형성하고 있으나 충청남도 관련 키워드가 전체 네트워크 안에서 중심을 차지하고 있지 않았다. 이와는 대조적으로 3농혁신은 충남고유의 정책으로 여겨질 만큼 대부분의 언급이 충청남도와 직접적으로 연관되어 있는 키워드였다. 다만, 추상적인 연관어가 많았다. 또한 3농혁신은 사회적 경제의 분석결과보다는 일상생활에서 쓰이는 언어가 아닌 정책문서에 주로 쓰이는 단어들과 관계가 깊었다는 점에서 사회적 경제와는 대조적이었다. 이러한 원인으로 3농혁신이 충청남도의 대표적인 정책으로 인식되고 있으나, 구체적인 사업에 대한 언급이 부족하고 전국적인 홍보효과가 아직 미흡하다는 점을 지적할 수 있다.

한편, 정책에 관한 홍보나 사회 인지도 등에서 사회적 경제는 다양한 지자체가 참여하고 있다는 점을 근거로 다양한 정책이 전개되고 있다고 볼

수 있다. 그러나 3농혁신은 충남만이 연관관계를 나타내고 있어 1인 독주 경향을 보이고 있다. 따라서 향후 사회적 경제에 관한 정책은 전국적인 네트워크 형성에 주력하고 무대를 확장시켜 나가는 것이 필요하며 서울 및 다른 지자체와 더욱 적극적으로 연계, 협력한다면 충청남도가 사회적경제의 허브로서 이미지를 구축해 나갈 수 있을 것이다. 향후, 3농혁신 정책은 도내·외 여론이 더욱 관심을 가질 수 있는 구체적인 사업과 대중적이고 체감할 수 있는 언어로 전달하는 것을 제안할 수 있다.

본 연구에서는 위와 같은 분석을 통해 광범위한 여론과 정책관련 키워드를 빅데이터를 통해 수집하고 분석하였다. 이러한 분석방법은 빅데이터를 이용한 광역정부의 정책현안과 특징을 반영한 미래전략 수립의 객관적인 기초자료로 활용 가능할 것으로 판단된다. 특히, 앞에서 언급한 바와 같이 전라북도·LX 빅데이터 활용센터에서의 활용 가능한 모델이 될 것으로 판단된다.

마지막으로 본 연구에서는 트위터와 신문데이터만을 이용하였으나 더 많은 웹페이지나 다른 매체에서 얻을 수 있는 데이터를 이용한다면 더욱 정확한 분석이 가능하리라고 생각한다. 또한 한국어 형태소(morpheme) 분석기법의 한계를 지적할 수 있다. 이는 긍정, 부정의 분석과 같은 오피니언 분석에서도 한계가 있다고 할 수 있다. 나아가 충청남도라는 구체적인 지역에 대한 정보를 수집할 때, 위치(공간)정보를 결합한 정보를 구독했다면 더욱 복합적인 결과를 얻을 수 있었을 것으로 판단되며, 이에 관한 추가적인 연구가 필요하다.

참고문헌

감미아, 송민, 2012, “텍스트 마이닝을 활용한 신

문사에 따른 내용 및 논조 차이점 분석”, *지능정보연구*, 18(3), 한국지능정보시스템학회, pp.53-77.

국가정보화전략위원회, 2011, “빅데이터를 활용한 스마트 정부구현”

국민권익위원회, 2013, “민원 키워드 중심으로 살펴본 국민신문고 출범 이후 민원 분석”

김성웅, 2013, “OECD의 빅데이터 관련 논의 동향”, *방송통신정책*, 제25권, 10호(통권 555호), 정보통신정책연구원, pp.85-92.

김수연, 2006, “마이닝 기법을 이용한 연관용어 집합 생성에 관한 연구”, 연세대학교 대학원 문헌정보학과 석사논문.

박원준, 2012, “빅데이터(Big Data) 활용에 대한 기대와 우려”, *방송통신전파저널* 통권 51호, 한국방송통신전파진흥원, pp.28-47.

박주영, 2013, “공공혁신을 위한 떠오르는 키워드, 빅데이터”, *예산준추 NABO Budget & Policy*, Vol.30.

박재희, 강영옥, 2013, “트위터 데이터를 통해 본 생활환경 만족도의 공간적 특성”, *대한지리학회지 학술회의논문집 6*, 대한지리학회, pp.383-386.

박정은, 2014, “미래전략센터의 의미, 성과와 발전 방향”, *Future Horizon*, 제22호.

배정환, 손지은, 송민, 2013, “텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석”, *지능정보연구*, 제19권, 3호, 한국지능정보시스템학회, pp.141-156.

시로타 마코토, 2012, “일본의 빅데이터의 현황과 과제 - 사업전략을 발굴하기 위한 조직체제와 인재가 키 -”, *IT프론티어*, 노무라종합연구소, pp.6-9.

송영조, 2013, “성공적 공공정책 수립을 위한 미래전략연구방법론”, *IT & Future Strategy*, 한국정보화진흥원.

윤미영, 권정은, 2012, “빅데이터로 진화하는 세상: Big Data 글로벌 선진사례”, 한국정보화진흥원.

이윤택, 홍영조, 2012, “알기쉬운 공공부문 빅데이터 분석 활용 가이드”, 한국정보화진흥원

이만재, 2011, “빅데이터와 공공 데이터 활용”, *Internet and Information Security*, 제2권, 제2호, pp.47-64.

조수근, 김성범, 2011, “텍스트마이닝을 활용한 산업공학 학술지의 논문 주제어간 연관관계 연구”, *2011년 대한산업공학회 추계학술대회*.

정근하, 2011, “텍스트마이닝과 네트워크 분석을 활용한 미래예측 방법 연구”, 한국과학기술

기획평가원.

충청남도, 2013, 2013년 충청남도 시책토론회 자료.

한국정보화진흥원, 2013, “데이터 기반 국가미래전략을 위한 정책과제”, IT & Future Strategy 제3호, pp.17-18.

행정안전부, 2012, “스마트 행정 구현을 위한 빅데이터 마스터플랜 현황과 추진 계획”

OECD 대표부, 2013, “빅데이터를 활용한 창조경제 실현방안”.

전자신문 <http://www.etnews.com/20140918000301>.

Franklin, Carl and Paula Hane, “An introduction to GIS: linking maps to databases,” Database. 15(2) April, 1992, 17-22.

Hwajin Lim, Sunghyun Park. : A Tentative Approach for Regional Futures Strategy with Big Data –Through the Analysis using the Data of SNS and Newspaper, *Journal of Korean cadaster information association*. 17-1. 75-90 (2015)

本研究は、ビックデータを活用し、地域の未来戦略を策定するための実験的な研究である。今まで地域に関する政策を SNS データを用いて分析した研究は初めてであり、先駆性が認められた。本研究で使われた方法は、科学的な政策策定プロセスのうち、パブリックコメントを効率よくまとめる方法としても活用できる。本研究は、応募者が Chungnam Development Institution で行った戦略研究「ビックデータを利用した忠清南道政策キーワード分析」の研究結果を修正及び発展させたものである。

研究の背景及び目的

近年、行政とビックデータを融合した政府 3.0 や地方 3.0 が大きな反響を呼んでいるなか、データ・サイエンスを活用して効率的かつ科学的な政策推進システム構築が注目されている。未来学の観点からも科学的政策策定方法論においてビックデータを利用した情報分析は重要とされている。このような状況のなかで、本研究はビックデータを利用し、広域地方公共団体である忠清南道に関する地域情報や世論を分析して、政策の波及度を間接的に測定することで、政策の方向性を見出すことを目的としている。

研究方法

ビックデータである SNS データと大量の新聞データを用い、テキストマイニング分析を行った。テキストマイニングは、テキストデータを定量的に分析する方法である。分析指標はテキストマイニングの分析指標の中で、TF-IDF と連関語分析を用いた。また、分析結果の可視化や関連する単語の関係を分析するため、ネットワーク分析の指標を用いた。事例地域として忠清南道を選定し、地域に関する新聞記事と SNS データであるツイッターの分析を通じ、世論を構造的に示した。

研究結果及び示唆

新聞記事の分析により、2013 年上半期には主に経済と政治に関する話題が、下半期には主に文化関連話題が出現していた。SNS 分析の結果、忠清南道では‘天安市’が全体のネットワーク内で重要なハブとして抽出され、代表的な政策である‘社会的経済’と‘3 農革新’に関するキーワードにおける忠清南道の位置づけを確認することができた。

本研究での分析手法は、今後ビックデータ分析を利用した広域地方政府の政策策定の方法論としての活用が期待できる。