전략연구 2014-15

빅데이터를 이용한 충남도 정책 키워드 분석

임화진

CDI 충남발전연구원
Chungnam Development Institute

발 간 사

최근 중앙정부나 지방정부가 직면하고 있는 정책환경에 있어 가장 큰 변화는 아마도 빅 데이터라는 개념의 등장과 이를 정책에 어떻게 활용할 것인가라는 고민일 것이다. 빅 데이터를 활용한 민간부문의 마케팅이나 수익창출 모델의 개발은 이미 수년전부터 국내외를 막론하고 활발하게 추진되고 있다. 한편, 행정부문에 있어서도 이를 활용한 혁신에 관한 연구도 활발하게 진행되고 있으며, 몇몇 부분에서는 실질적으로 이를 적용하고 있기도 하다. 우리 정부도 정부 3.0을 내세워 데이터 과학을 공공서비스 분야에 적용함으로써 효율적인 정책추진을 시도하고 있다. 즉, 행정에도 SNS 및 빅 데이터를 활용하여 효율적인 정책 추진 및 보다 질 높은 서비스를 주민에게 제공하려는 노력들이 이루어지고 있다.

이러한 측면에서 본 연구는 충남도와 관련된 언론 데이터와 SNS 데이터를 분석하여 지역현안 및 정책에 대한 도민의 생각을 추출하여 이를 도정에 적극적으로 활용할 수 있는 방법을 모색하고 있다. 본 연구는 구체적 정책내지 방향을 제시하고 있지는 않지만, 충남도와 관련된 언론 데이터의 활용을 통하여 정책 모니터링 및 도민의 의견을 효율적으로 수렴하여 도민의 삶의 질을 높일 수 있는 맞춤형정책수립을 위한 기초연구로서의 의미가 있다. 후속 연구를 통하여 빅 데이터를 활용하여 도민이 체감할 수 있는 좋은 정책을 만들어 낼 수 있기를 기대해본다.

끝으로 본 연구를 수행한 임화진 박사와 본 연구를 수행하는데 도움과 조언을 아끼지 않은 원내외 자문위원 및 심의위원들에게 감사의 말씀을 드립니다.

2014년 8월 31일

충남발전연구원장 **강 현 수**

연 구 요 약

본 연구는 충남도와 관련된 언론 데이터와 SNS 데이터를 분석하여 충남도에 지역 현안을 보도한 언론기사 분석을 바탕으로 한 도민 의견을 구조적으로 평가하고, 여론 분석 결과 도출해 낸 키워드를 정책과 연결하여 파급도와 수요 등을 가시적으로 표현하였다.

분석결과는 다음과 같다.

언론기사 추출결과로는 상반기에는 주로 정치, 경제적 이슈가, 하반기에는 문화관련 이슈 등 조금 더 폭넓은 이슈를 발견할 수 있었다.

트위터 분석을 통한 충남도 정책 관련 키워드 구조에서는 천안시가 전체 정보네트워크 안에서 중요한 HUB로서 추출되었으며 이는 충남 내의 다양한 화제들이 천안시와 밀접한 관계를 이룬다고 볼 수 있다.

한편 사회적경제와 3농혁신에 관한 분석에서는 각 키워드의 특성과 충남도와의 관계를 조망할 수 있었다. 사회적경제는 전국적인 화두로 인식되고 있으며 구체적으로 서울, 수원, 성남 등과 같은 지역명이 대두되고 있고 이 중 하나로 충남이 언급되고 있다는 것을 알 수 있다. 이와는 대조적으로 3농혁신은 충남고유의 정책으로 거의 대부분의 언급이 충남도와 직접적으로 연관이 있는 키워드지만 전국적인 파급효과가 있다고는 보기 힘들며 아직 추상적인 단계의 사업들이 대부분이다.

즉 사회적경제는 전국적인 화두로 인식되고 있으며 구체적으로 서울, 수원, 성남 등과 같은 지역명이 대두되고 있고 이 중 하나로 충남이 언급되고 있었다. 또한 여러가지 주체가 얽혀있는 열린 네트워크를 발견할 수 있었으나 충청남도 관련 키워드가 전체 네트워크에서 중심적인 역할을 하고 있지는 않았다. 따라서 향후 사회적경제에 관한 정책은 전국적인 네트워크 형성에 주력하고 그 무대를 확장시켜 나가는 것이 필요하다.

한편 3농혁신은 충남고유의 정책으로 거의 대부분의 언급이 충남도와 직접적으로 연관이 있는 키워드지만 아직 추상적인 연관어가 대부분이었고 구체적인 사례나 사업에 관한 연관어가 많지 않았다. 따라서 향후 3농혁신 정책은 도내외 여론이 더욱 관심을 가질 수 있는 구체적인 사업과 대중적이고 체감할 수 있는 언어로 전달하는 것이 필요하다는 점을 도출할 수 있었다.

본 연구의 분석 절차 및 활용방안을 참고하여 향후 충청남도 도정에 관련된 키워드를 적절히 모니터링 하여 빅데이터를 적극 활용하는 도정방안을 마련하는 것이 중요하다고 생각한다.

목 차

표 목 차

그 림 목 차

제1장 서 론

1. 연구의 배경 및 목적

1) 연구의 배경

제 3의 물결인 정보화 사회를 지나 제 4의 물결이라 칭할 수 있는 데이터 사회로 들어서게 되었다. 앨빈토플러가 언급한 제 4의 물결의 세 가지 중요요소인 시간, 공간, 지식을 다루는 중요한 수단으로 부상하고 있는 것이 바로 빅데이터다. 빅데이터는 사전적 의미로 본다면 큰 용량의 데이터지만 최근에 통용되는 정의로는 기존의 저장기술 및 분석기술로 대응할 수 없는 대용량 데이터를 지칭한다. 빅데이터는 3V(크기, 속도, 다양성)+1V라는 개념으로 대용량의 여러 가지 형태의 데이터를 빠른 속도로 처리하여 새로운 가치를 창출하는 패러다임으로 주목받고 있다.

한편 2000년대 이후 스마트폰 사용자가 급증하면서 모바일을 중심으로 한 시대가 오게 되었다. 즉 데이터를 이용 및 생산하는 스마트폰이라는 플랫폼이 보급되면서 모바일 시대에는 언제 어디서나 개인기반으로 데이터를 구축할 수 있는 시대가 오게 되었다. 이러한 스마트폰과 같은 플랫폼은 SNS 보급에 큰 역할을 하게 되어 SNS를 통한 실시간의 개인의 의견을 활용하고자 하는 수요가 급격히 증가하게 되었다. 이러한 변화는 기존의 정형화된 데이터를 일정 시간 이상 공들여 생산하는 것이 아닌 실시간으로 각 개인이 대량의 데이터를 생산하게 되면서 이러한 데이터가 빅데이터로 자리잡게 되었다.

최근에는 행정과 빅데이터 및 오픈데이터가 융합한 정부 3.0 그리고 지방 3.0이 큰 반향을

일으키고 있다. 데이터 과학을 공공서비스 분야에 적용함으로서 효율적인 정책 추진 및 평가를 시도한 것이라고 볼 수 있는데 최근 정부 3.0이라는 구상을 발표하였고 이에 발맞추어 지방자치단체에서도 지방 3.0이라는 정책 아젠더를 내걸고 정책 구상 단계에 돌입하여 행정과 관련된 제반 데이터를 오픈하는 방안을 추진하고 있다.

이러한 배경 하에 충남도민의 의견을 충실히 반영하기 위한 취합체계 구축이 필요하게 되면서 설문조사 및 여론 조사에 선행하여 정보를 효율적으로 수집하기 위한 수단으로서 유용한 SNS데이터가 각광을 받고 있다. 또한 정책 키워드와 여론 분석 결과를 비교 검토하여 정책평가를 도출하고 새로운 정책과제를 발견하는 모니터링이 시급하다. 이러한 모니터링을 통하여 맞춤형 정책수립이 가능하며 행정비용을 대폭 감소시킬 수 있다는 것이 큰 이점이라고 할 수 있다.

빅데이터를 활용한 민간부문의 마케팅이나 수익창출 모델은 이미 수년전부터 정책추진 및 평가를 시도하는 연구가 국내외를 막론하고 활발하게 추진되고 있으며 행정혁신부문에서도 많은 연구가 실시되어 왔다. 이 때 체계화된 데이터 이외에도 비정형데이터를 효율적으로 분석하는 방안이 필요하다.

정부3.0의 공공분야의 데이터 공개사례를 살펴보면 먼저 정형데이터 공개사례로서 국내 사업 중에서는 국토교통부와 LH공사와 지적공사 등이 함께 구축한 온나라 부동산 시스템을 통한 부동산 공시지가 및 관련 정보 공개 사례가 있다. 한편 비정형데이터를 활용한 이동통신 통화 데이터를 이용한 서울시 심야버스 노선설정, 다음 지도와 재해 정보 연계 및 범죄데이터 결합 등의 사례가 있다. 이처럼 다양한 분야에서 많은 정보를 공개하고 이용하는 움직임이 가속화 되고 있으며 이러한 흐름에서 빅데이터를 적극적으로 활용하여 도정에 유용한 시사점을 얻는 것이 무엇보다 중요해지고 있다.

2) 연구의 목적

본 연구는 이러한 배경을 바탕으로 이하와 같이 연구의 문제의식을 설정하였다.

- ◆ 행정혁신의 관점으로 볼 때 정책투입 일변도가 아닌 수요응답형 모니터링이 중요해지고 있지 않는가?
- ◆ 빅데이터의 흐름을 이용한 수요자 중심의 맞춤형 정책과제 도출이 가능하지 않을까?
- ◆ 충남도가 적극적으로 추진하고 있는 정책에 대한 여론의 반응과 수요자(도민)의 반응은 어떠한가?

이러한 연구의 문제의식을 바탕으로 본 연구는 다음과 같은 목적을 설정하였다.

먼저 기존 데이터와 빅데이터의 융합 활용을 통한 언론통계 및 비정형 데이터 통계 시스템을 구축하고 비정형데이터의 정책평가 활용방안을 제안하는 것이다.

다음으로 빅데이터를 이용한 충남 정책 모니터링을 실시하고자 하는데 이에 지역별, 시기별 맞춤형 정책수립을 위한 기초분석 결과를 바탕으로 향후 정책 모니터링의 여론 및 나아가서 민원 등의 반영 체계에 대한 실험적 분석을 실시한다. 또한 이를 통해 방향성을 도출하고자 한다. 나아가 비정형데이터를 효율적으로 정책평가에 활용할 수 있는 제안을 검토한다.

◆ 지역 현안을 보도한 언론기사 분석을 바탕으로 한 도민 의견을 구조적으로 평가하고,
◆ 여론 분석 결과 도출해 낸 키워드를 정책과 연결하여 파급도와 수요 등을 가시적으로 표현하도록 한다.

<표 1> 연구질문과 연구목적

연구질문	연구목적
빅데이터를 이용한 수요자 중심 정책과제를 발굴할 수 있지 않을까?	기존 데이터와 빅데이터의 융합 활용: 언론통계 및 비정형 데이터 통계 시스템 구축
행정혁신의 관점으로 볼 때 정책투입 일변도가 아닌 수요응답형 모니터링이 중요해지고 있지 않는가?	빅데이터를 이용한 충남 정책 모니터링의 실험적 분석
충남도가 적극적으로 추진하고 있는 정책에 대한 여론의 반응과 수요자(도민)의 반응은 어떠한가?	도민 여론 분석을 통한 사회적경제 추진정책 평가

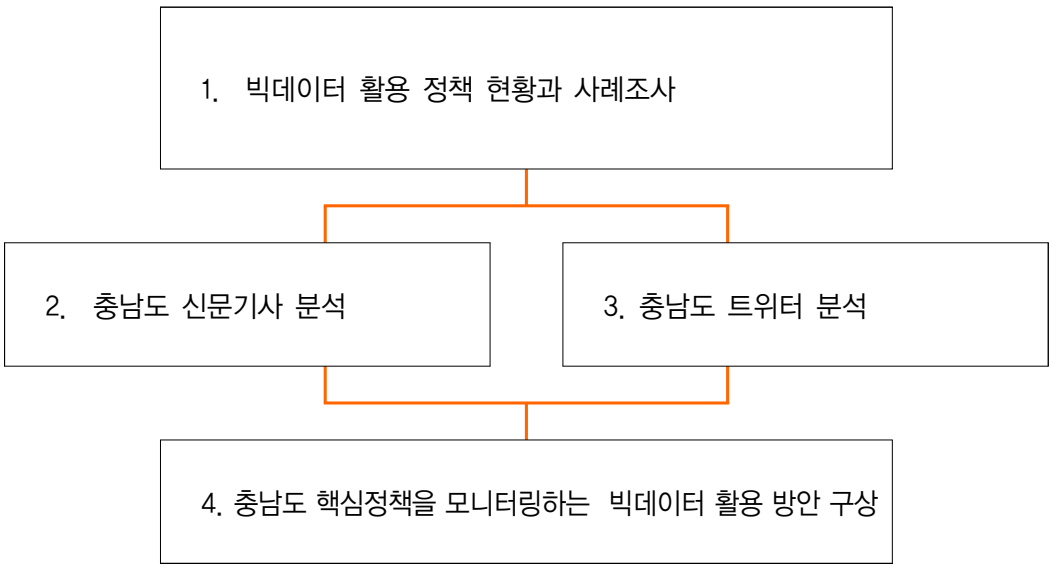
3) 사용 개념의 정의

본 연구에서는 정책 파급도란 개념을 이용하여 충청남도의 정책에 관련된 키워드를 분석하도록 한다. 또한 연관어를 통한 정책 이미지를 평가하도록 한다.

언론과 트위터로 대표되는 어떠한 현상에 대한 글을 통하여 정책자체의 만족도나 선호도를 구할 수는 없다. 다만 파급도라고 명명한 일정 이상의 관심도는 그 빈도수나 구조로서 파악할 수 있고 함께 언급된 관련 키워드를 통해 단어에 관한 이미지를 파악할 수 있다.

따라서 본 연구에서는 정책에 관련된 핵심 키워드를 도출하고 키워드간의 연관관계를 통해 충청남도 관련 여론의 구조를 파악하도록 한다.

2. 연구의 흐름


1. 빅데이터 활용 정책 현황과 사례조사
2. 충남도 신문기사 분석
3. 충남도 트위터 분석
4. 충남도 핵심정책을 모니터링하는 빅데이터 활용 방안 구상

제2장 빅데이터에 관한 선행연구 및 활용사례

1. 빅데이터의 개념[1)]

1) 빅데이터의 정의 및 특징

○ 빅데이터의 개념 및 범위

빅데이터란 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 확대되는 경향도 보이고 있다. 빅데이터의 가장 큰 특징은 지금까지 잘 다루어지지 않았던 비정형데이터를 데이터분석에 연계하여 활용한다는 점이다. 특히 SNS데이터와 같은 방대한 텍스트데이터를 손쉽게 처리 및 분석하는 부분이 빅데이터 관련 분야 중에서도 가장 주목받고 있는 부분이다.

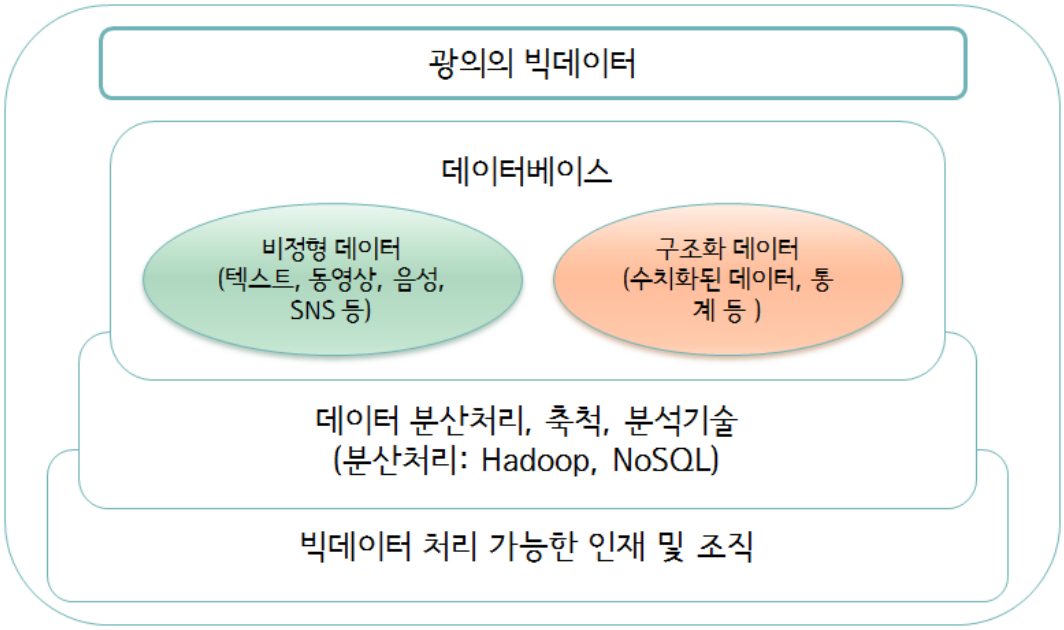
1) 정책FOCUS를 바탕으로 작성

○ 빅데이터의 특징

빅데이터의 특징은 크게 3V로 설명된다. 여기서 3V는 데이터의 규모 (Volume)와 데이터의 종류(variety), 데이터의 속도(Velocity)를 일컫는다. 최근에는 3V에 가치(Value)를 추가하여 3V+V라고도 한다.

빅데이터의 효용가치의 가장 대표적인 것으로 다양한 정보를 파악할 수 있다는 것을 들 수 있다. 빅데이터의 가장 대표격인 SNS데이터는 개개인의 취향을 직접 반영한 시간, 공간이 파악 가능한 데이터이다. 이를 통해 기업 및 고객, 행정 및 정책수요자간의 쌍방향 소통에 있어서 유용하게 활용될 수 있다.

빅데이터의 저장 기술과 데이터 구축 기술, 데이터 분석기술이 함께 융합된다면 기존에 시도하지 못했던 데이터 사이에 상호융합이 가능해지며 새롭고 입체적인 대안의 도출이 가능해진다.



[그림 1] 빅데이터의 개념과 범위
노무라연구소 빅데이터 시대 도래(2012)를 참고 및 수정

2) 빅데이터에 관한 선행연구

○ 빅데이터 분석 방법

빅데이터 분석방법은 크게 텍스트마이닝, 웹 마이닝, 소셜 마이닝, 현실 마이닝이라는 단계로 구분할 수 있는데 텍스트마이닝과 웹마이닝, 소셜마이닝까지는 현재 빅데이터 분야에서 주된 방법론으로 사용되고 있다.

즉 마이닝이라는 개념은 분석방법을 일컫는 것이며 텍스트, 웹, 소셜이라는 것이 분석 데이터를 지칭한다고 볼 수 있다. 여기서 현실마이닝은 이러한 복합적인 데이터를 융합하여 현실에 더욱 근접한 모델을 구성하는 것이라고 할 수 있다. 각각의 마이닝에 대해서는 수집, 저장, 처리, 해석의 측면에서 다분야에 걸친 연구성과가 존재한다. 이 중 자연어 처리분야인 텍스트마이닝 분야가 발달하면서 대용량의 텍스트를 기계적으로 분석할 수 있는 기반이 갖추어지게 되어 빅데이터 분석 분야에 큰 지평을 열었다고 할 수 있다

한편 빅데이터 흐름과는 달리 기존의 언론정보학에서 다루던 내용분석이나 대응 분석 등은 사회학 분야를 포함하여 오랜 시간 연구성과가 축적되어 왔다.

본 연구에서는 수집, 저장의 측면은 컴퓨터 과학과 같은 기술적인 분야기 때문에 다루지 않고 텍스트 처리 및 해석방법에 초점을 두어 선행연구를 검토하도록 한다.

○ 국외연구 동향

빅데이터와 SNS를 결합한 최초의 사례라고 평가되는 연구는 08년 인디애나 주립대 Johan Bollen에 의해 행해졌다. 이 연구는 트위터데이터가 개인의 의사를 반영한 정보가 포함된 것을 발견하고 이에 착안했던 연구라고 할 수 있다. 구체적인 분석 과정은 반년간의 트위터의 데이터를 이용하여 오피니언 분석을 통해 기분변화를 측정하고 이와 관련된 연간 이벤트의 연관성을 도출하였다. 이러한 연구를 바탕으로 MIT연구그룹은 트위터데이터를 활용하여 행복도를 측정하는 연구를 시도하기도 하였다.

○ 국내연구 동향

국내 연구에서는 특히 2012년 이후 한국에서는 SNS데이터, 특히 트위터에 대한 분석이 활

발하게 실시되고 있다. 주로 기술적인 측면은 정보학분야에서 실시된 연구가 많으나 기술을 적용하여 새로운 시사점을 뽑아내기 위한 시도도 사회학이나 정치학 등 다양한 분야에서 이루어지고 있다.

한 예로 배정환 외 2 (2013)는 국내 트위터를 분석하여 선거 결과를 조망하여 선거전의 과정과 결과에 있어서 트위터의 구조 등을 본 연구이다.

또한 박재희(2013)의 연구는 트위터 데이터를 이용하여 도시정책지표를 구성하고 주거환경만족에 대한 공간적 특성을 도출하는 연구를 수행하였다. 여기서는 트위터의 공간정보와 트위터 텍스트를 주거만족도로 해석하는 텍스트 마이닝 기법이 결합된 연구라고 할 수 있다.

○ 데이터별로 본 텍스트마이닝 기존연구

빅데이터의 성질을 규정하는 가장 큰 부분인 비정형데이터는 음성, 사진, 동영상 등을 일컬으나 주로 텍스트데이터로 환원되어 분석에 사용된다. 즉 텍스트데이터는 빅데이터에서 큰 비중을 차지하고 있으며 이를 분석하는 텍스트마이닝 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 연구 방법이 주로 행해져 왔다. 최근들어 데이터 마이닝 툴이 발달함에 따라 대용량의 신문기사 및 여론 자료를 분석하는 논문들이 발표되기 시작하였다. 그 예로 감미아 외 1 (2012)를 들 수 있다. 이 연구에서는 주요 신문사의 논조 비교를 대용량의 데이터 분석을 통해 구현했으며 구체적으로 어떤 단어에 대한 논조의 차이가 존재했는지 검토를 시도하였다.

구글검색어를 이용한 텍스트마이닝도 새로운 시도로 조망받고 있다. KISTEP에서 수행한 미래 트렌드 분석 연구는 텍스트마이닝과 네트워크 분석을 활용한 연구로서 구글검색어를 데이터베이스로 정하고 기존의 텍스트마이닝 지표를 개량한 새로운 빈도수 지표를 활용하는 등 다양하고 새로운 시도를 행한 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.

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 TOP100을 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.

○ 빅데이터 연구의 유의점

빅데이터 연구에서 가장 유의해야 할 점은 빅데이터라는 단어에 지나치게 고착되어 있어서는 안된다는 것이다. 즉 단순히 데이터베이스만을 구축하는 것을 목표로 해서는 안 되고 데이터를 활용해야 할 목표가 있어야 하며 목적에 맞는 방법을 구체화해야하는 것이 가장 필요하다. 빅데이터에 관한 전문적인 시각을 가지고 있는 가트너의 부사장은 아래와 같이 빅데이터에 도입 시에 주의해야 할 부분에 대해 언급하고 있다.

빅데이터를 도입할 때 확인할 부분은 투자 대비 원하는 만큼의 효과를 낼 수 있는지 확인하는 것이다. (중략) 빅데이터는 새로운 것이 아니라 이전부터 존재했던 데이터를 모은 것에 불과하다. 기업 활동에 의미가 없는 다크 데이터와 가치있는 데이터를 구분할 수 있어야 경영에 도움이 되지만 대부분은 분위기에 휩쓸려 빅데이터를 무조건 받아들이고 있다. (중략) 보유한 데이터 중 가치있는 부분을 발견하고 분석하는 것에 집중하는 것이 중요하다.

– 도널드 페인버그, 가트너 부사장, 2013.10.21.

이 글에서도 알 수 있듯이 빅데이터를 데이터구축에 한정시키는 것이 아니라 새로운 방법을 동원하여 가치있는 사실을 도출해 내는 것이 가장 중요한 목적이며 이것은 지금까지 행한 데이터분석과 크게 다르지 않다는 것을 의미한다. 즉 빅데이터는 조금 더 새롭고 다양한 데이터를 추출과 정제를 통해 분석할 수 있는 형태로 구비하고 해법을 찾아내는 과도기적인 개념이라고 할 수 있다.

2. 국내, 외 빅데이터 활용사례

1) 민간부문의 활용사례

○ [국내] 포스코의 원료가격 효율적 구매 관리

포스코는 빅데이터를 이용하여 효율적이고 빠른 속도로 원료가격을 예측하여 구매 관리를 실시하고 있다. 가격 변동이 큰 철광석 등 자원을 적시에 조달하기 위하여 데이터 분석을 통하여 최적 구매 시기와 가격대를 결정하는 것이다. 이를 통하여 고객의 수요 데이터, 남미·호주 광산의 상황, 런던 금속 거래소의 광물 가격 데이터를 분석하여 미래의 철광석 가격을 예측하고 있으며 이를 통해 생산 공정별 온도, 습도, 압력, 성분 등의 데이터와 불량률을 결합하여, 생산 효율성이 높아지도록 실시간으로 공정 제어도 실시하고 있다.

○ [국내] SK텔레콤의 Tmap

공간정보와 통신정보를 결합한 SK텔레콤은 지도와 연결된 유동인구, 업종별·월별 매출 정보 등으로 상권분석서비스를 제공하고 있다. 자영업 창업 희망자가 업종별 매출 현황, 경쟁 매장, 잠재 수요고객, 유동인구 등의 정보를 지도에서 직접 분석 가능한 시스템을 제공하고 있다.

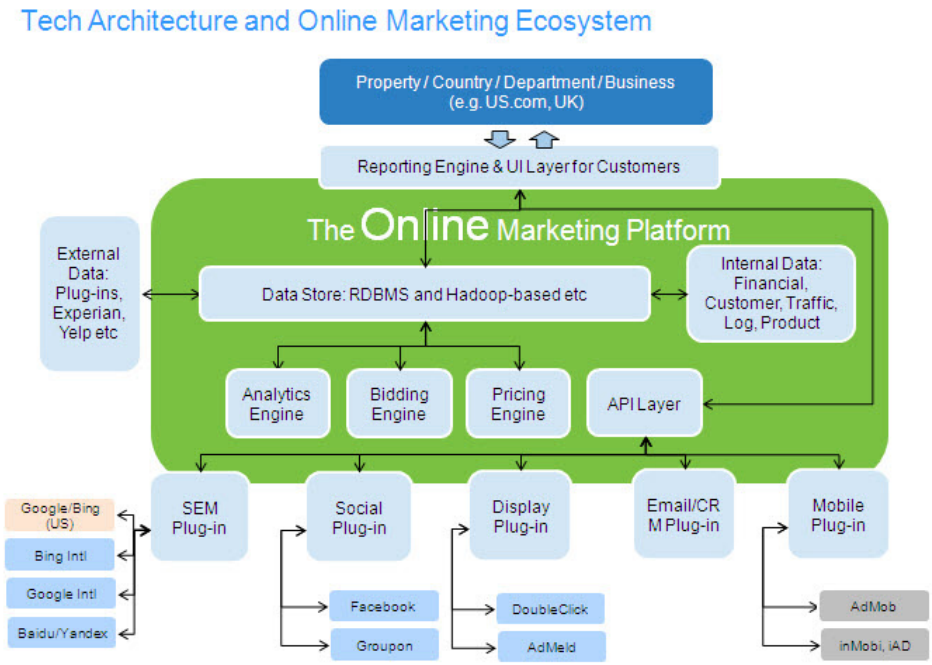
총 연결정보는 10종으로 2,650만 SK텔레콤 가입자 동선(유동인구), 3,000만 OK캐쉬백 회원 소비패턴, 현대카드 가맹점 결제, 부동산114의 상권 시세 등을 제공하고 있다.

○ [국내] 다음소프트 소셜 인사이트

다음소프트 소셜 인사이트라는 분석 플랫폼을 통해 소셜 네트워크의 이슈와 관심 키워드의 실시간 모니터링으로 상황에 맞는 대응전략 및 마케팅 전략의 수립을 지원하고 있다. 자연어 처리, 분석 기법 등을 이용하여 블로그와 트위터에서 형성되는 트렌드와 여론을 일반과 공공으로 구분*하여 제공하고 있다. * 일반 : insight.some.co.kr, 공공 : pub.some.co.kr

○ [국외] 월마트(Walmart)의 social genome을 통한 고객관리

월마트는 각 지점의 모바일과 소셜 쇼핑의 특징을 이용한 월마트랩을 운영하고 있으며 웹사이트에서 발생하는 거래 데이터를 재고 예측에 이용하여 적절하고 효율적인 재고 관리를 도입하였다. 또한 소셜 미디어 회사인 코스믹스(Kosmix) 인수를 통해 소셜 네트워크와 콘텐츠를 관리하여 고객정보에 맞게 각 지점을 운영할 방침을 가지고 있다.



[그림 2] 월마트 소셜 게놈 시스템(출처: 월마트랩 HP)

2) 공공부문의 활용사례

정부3.0과 밀접하게 연결되는 공공부문의 데이터베이스 공개에 관한 선행사례는 국내, 외에 다양하게 존재한다.

○ 국외 공공분야 빅데이터 활용 사례

먼저 외국의 데이터개방제도에 관하여 살펴보도록 한다. 유럽연합은 빅데이터 물결이 닥치기 이전에도 '공공정보 재활용 지침'(DIRECTIVE on The Re-Use of Public Sector Information, 2003)을 제정하고 매년 회원국의 이행 여부를 심사하는 제도를 마련한 바 있다. 이후 2012년에 27개 회원국 공공기관의 모든 공공데이터의 온라인 개방을 의무화한 '오픈 데이터 전략(Open Data Strategy, ODS)'을 발표하고 적극적으로 공공 데이터 개방에 나서고 있는 모습을 보이고 있다.

영국에서도 이미 2005년에 정보화 물결에 대응하면서 민간이 더욱 활발하게 공공데이터를 이용할 수 있도록 '공공정보 재활용 규칙'(The Re-use of Public Sector Information Regulations, 2005)을 제정하였고 ODI(Open Data Institute, 국가전반의 공공데이터 활용 정책 총괄)를 설립하여 공공정보 재활용 정책을 총괄하도록 하였다. 이를 통해 공공데이터의 접근성이 높아지고 정보공유인프라가 구축되어 왔다. 최근에는 data.gov.uk을 통해 인구, 범죄, 건강 등 공공정보를 개방하고 있다. 이와 같은 데이터 원스톱 서비스는 정부의 투명성을 높이고 국민의 알 권리를 향상시켜 경제 및 사회적 가치를 증대시키도록 도모한다. 나아가 제 4의 물결에서 주도권을 획득하려는 전략이라고도 할 수 있다.

한편 미국 또한 오래전부터 공공 데이터 개방에 적극적인 자세를 보여 왔다. 일례로 1996년에 민간이 공공데이터를 자유자재로 활용할 수 있는 권리를 '정보자유법'(Electronic Freedom of Information Act, 1996)에서 규정하고 웹2.0기반의 정부2.0정책으로 172개 공공기관의 38만개가 넘는 정보를 개방해 온 실적이 있다. 오바마 정부 이후에는 대대적인 공공데이터 개방에 더욱 박차를 가해, 국방부 등 6개 연방기관의 주체가 되어 빅데이터 선진기술개발에 2억 달러 규모의 '빅데이터 연구개발 이니셔티브'가 추진되었다. 구체적으로 의료부문에서는 국립보건원의 필박스(Pillbox) 서비스로 검색통계를 활용하여 약 검색 등을 제공하고 주요 질

병의 분포, 연도별 증가 등을 분석하고 있다. 이 사업은 연간 5천만 달러의 비용절감 효과를 거둘 수 있다고 기대하고 있다.

다음으로 호주에서는 '범정부 차원의 정보공개 체계 정립에 관한 지침'(Whole of Government Information Publication Scheme, 2009)이 마련되어 data.australia.gov.au을 통해 연방, 주, 지방정부에서 생성되는 1,100개의 공공정보를 개방하고 있다.

[그림 3] 영국, 미국, 호주의 공공데이터 개방 홈페이지

이처럼 빅데이터 흐름이 본격적으로 전개되기 이전부터 미국과 영국등지에서는 공공데이터를 적극적으로 개방하는 등 체계적인 준비를 해왔으며 이를 통한 새로운 사업 창출을 독려해왔다.

○ 국내 공공분야 빅데이터 활용 사례

다음으로 국내사례에 대해 살펴보도록 한다. 우리나라에서는 2013년 이후 정부3.0이라는 기조를 내걸고 본격적으로 공공 데이터 개방을 추진하고 있다. 이는 다른 선진국들의 흐름에 힘입어 세계화 3.0과 자본주의 4.0등 새로운 물결 하에 새로운 성장 동력을 찾기 위한 돌파구로서 행정혁신을 이루고자 함이다. 정부는 '공공데이터의 제공 및 이용 활성화에 관한 법률'을 제정하는 등 제도적인 기반을 마련하고 사업을 추진하고 있다.

정부3.0의 대략적인 효과로서 한국정보화진흥원(2011)의 추산에 따르면 영국의 행정혁신 정책을 한국의 공공시스템에 적용하면 약 10.7조원의 비용 절감을 기대할 수 있다고 한다. 또한 비용절감뿐 아니라 다양한 의견을 수렴하고 정책을 설계 및 평가 할 수 있다는 점에서 시너지 효과를 낼 수 있다.

<표 2> 정부 3.0의 추진 방향 및 전략

실현 목표	실천 방안	세부 전략
소통하는 투명한 정부	공공정보의 적극공개	적극적이고 능동적인 공개를 통해 정책사업에 대한 사전 공표를 확대. 원문공개, 전과정 공개, 국민중심공개
	공공데이터 개방	민간수요가 많은 공공데이터의 대폭 개방. 개방 로드맵 수립
	민관협치 강화	인터넷투표, 전자공청회, 토론회 등 의견수렴의 장 마련.
일 잘하는 유능한 정부	칸막이를 없애는 국정운영 시스템 혁신	국정, 협업과제의 보다 근원적이고 본질적인 해결 도모. 국민 체감할 수 있는 가시적 성과 창출
	데이터 기반의 과학적 행정	클라우드 컴퓨팅 환경 구축으로 지식공유 기반 마련 및 빅데이터를 활용한 과학적 행정 구현
국민 중심의 서비스 정부	민원서비스 혁신	생애주기별 맞춤형 민원서비스 제공
	원스톱 기업민원 서비스 제공	중소기업 지원사업 통합관리시스템 구축

* 공공기관 정부3.0 책임관 워크숍 내용 저자 편집

○ 민원분야 국내 활용사례

국민권익위원회 민원정보분석시스템은 민원정보분석시스템 구축사업('10년~'12년)을 통해 국민신문고를 운영하고 있다.[2] 정부민원에 대한 민원통계DB 구축 및 분석기반 마련하기 위해 구축되었으며 그동안 사회 이슈 등 단기 분석을 수행하며 축적된 노하우와 분석기법을 활용하여 장기적·거시적 관점에서 민원분석을 이하와 같이 시도하였다.

1차년도 사업에 대한 이용활성화 및 지속적인 확대를 위한 2차 사업 추진(2011년)

- *안정적 서비스를 위한 인프라 증설 : 110콜센터 연결 및 하드웨어 증설*
- *민원분석 업무지원 강화 : 의미기반의 민원분석을 위한 클러스터링 및 의견분석기능 개발*
- *보다 다양한 국민의 소리 분석 : 194개 교육청, 16개 광역시도, 29개 중앙행정기관 민원게시판 수집, 뉴스, 아고라 등 외부 정보 수집*
- *부처별 공동활용 서비스 제공 : 고용노동부, 국토해양부, 보건복지부, 경찰청 분류체계*

○ 충남도의 활용사례

충남도에서도 빅데이터와 행정혁신에 대응하기 위해 다방면으로 전략을 수립하고 있다. 『zero 100 프로젝트』를 통해 빅데이터를 연계 활용하여 지역현안을 해결하는 사업을 추진하고 있다. 또한 이 밖에도 공공기관 보유정보의 활용과 다양한 사업화 지원을 위하여 『충남도민 발명 아이디어 공모전』을 개최하였다.

최근에는 다수 부처에 분산되어 있는 안전과 관련된 데이터를 활용하여 분석 및 공유를 통한 재난예측체계 도입하고 재난 대응기능의 보강을 넘어선, 재난의 사전예측·대비 기능을 도입하는 것을 목표로 하고 있다. 즉 이를 통해 결과적으로는 재난정보 빅데이터를 활용한 미래 위기대응 및 대비전략 수립하는 것을 목표로 하고 있다. [3]

2) 민원분석 서비스를 제공하였으나 현재는 잠정적으로 운영이 되지 않고 있다.(2014년 2월 기준)
3) 출처: 2013년 충청남도 시책토론회 자료

○ 공공분야 활용 가능성

이상으로 현재 공공분야의 빅데이터 활용의 국내외 사례를 살펴보았다. 향후 공공분야의 빅데이터 활용은 다양한 측면으로 정리될 수 있다. OECD가 지적한 빅데이터 중요 5대 분야로 첫째, 온라인 마케팅 (맞춤형 서비스), 둘째, 보건의료(smart health-care), 셋째, 지능형 교통, 넷째, 스마트 에너지, 그리고 마지막으로 행정의 효율화를 들 수 있다. 여기서 공공데이터 활용 또한 전분야에 걸쳐서 유용하게 사용될 수 있다. 각 분야별로 살펴보면 온라인 마케팅은 민간 부문에서 주력을 기울이고 있는 분야이나 공공분야에서도 관광 및 지역마케팅으로 이용될 수 있다. 지능형 교통 분야는 교통 데이터베이스가 대부분 공공에서 관리하고 있는 만큼 잠재력이 높은 분야이다. 실제 사례로서 서울시 심야버스 노선 설정 과정에서 통신사의 통화량을 이용하여 통행 인구를 예측하여 노선을 합리적으로 설정한 것을 들 수 있다.

이처럼 공공분야의 데이터 개방은 다양한 활용사례들이 등장하고 있으며 향후 활용 가능성도 무궁무진하다. 활용 분야에 대해 분류해 보면 표 4와 같이 크게 사회조사 및 민원, 의료보건복지, 환경 및 도시, 방범보안의 분야로 나눌 수 있다.

<표 3> 빅데이터의 공공분야 활용가능성

(알기쉬운 공공부문 빅데이터 분석, 활용 가이드 2013를 참고로 연구자 재구성)

	구분	활용방안	이용가능 데이터
사회조사·민원	시민 목소리 이해	특정주제에 대한 시민의 목소리를 이해하고 추이분석하여 민원센터와 소셜데이터에 기반을 둔 정책의제발굴과 전략확보방안	민원, 소셜데이터
	사회이슈분석	이슈의 발굴과 연관검색어 등을 통한 주제 분석을 통해 정책수요 발굴 및 지역별 이슈도출. 또한 맞춤형 대국민 서비스 전략수립 가능하도록 분석	일간지, 소셜데이터, 민원센터 로그
	기관, 인물 평판분석	지정된 기관의 시민 인식 및 평판에 대한 소셜미디어 분석.	SNS, 일간지, 포털 게시판 등
	맞춤형민원 서비스	지역별, 기관별로 주민의 민원을 분석하여 개인맞춤형 시스템을 구축	서비스 사용 로그, 게시판 및 민원센터 로그, 포털 게시판 등
의료·보건·복지	의료 및 복지 서비스	의료보험비용분석, 부당청구방지, 복지정책의 수요 및 만족도 분석, 불균형 해소	의료보험데이터, 민원센터 로그, 소셜데이터, 서비스 기관 홈페이지, 주요 일간지 통합분석
	전염병, 질병 관리	유행전염병, 질병예측, 대응 및 지역적 전파, 연도별 거시분석 등	검색어, 보고데이터 및 GIS등
	교육정책 및 현안분석	교육환경개선 및 민원처리를 포함한 합리적 교육예산의 집행 및 절감	예산집행데이터, 소셜데이터, 민원센터 등

환경·도시	재난대응, 도시관제	사고다발지역 대책 및 재난예측을 통한 사고 및 재난 방지. 응급시 시민의 목소리를 적확하게 반영	CCTV, 도로센서, SNS, 전화량 등
	교통상황관리 및 최적화	교통흐름 모델링을 통한 예측, 최적화 시스템, 교통신호 체계 및 교통유지보수 활용가능	도로센서, 사건사고 기록, 날씨, 명절, SNS
	환경감시 및 대응	환경데이터와 다양한 연구결과를 메타분석하여 환경오염과 변화상황을 모니터링. 또한 중장기적 수행 전략 수립을 위한 기초자료 수집	리모트 센싱, 측량, 각종 연구결과 등
방범·보안	범죄예방과 대응	지역별, 시간별, 이벤트별, 유형별, 범죄패턴문석 및 지역별, 시기별 예방전략 수립	뉴스, 언론사, 소셜데이터
	금융감독 및 세금, 내부 감사	조세회피 및 탈세의 패턴감지 및 조기대응. 지역별 기간별 동향 파악 내부 담합 및 보안 등 감시기능	조세, 금융거래 데이터와 소셜데이터의 통합분석
	국방 및 국가안보	주요 이슈 모니터링 및 정책근거자료 수집 및 분석	보고서, 뉴스, 정보사이트 등

제3장 충청남도 정책키워드 분석 방법

1. 충남도 정책키워드 분석 개요

어떠한 정책에 관련된 키워드를 추출하고 이를 통한 파급양상을 분석함으로써 정책키워드에 대한 대중의 관심도 및 인지양상을 간접적으로 파악할 수 있다. 이러한 관심도 및 인지도를 바탕으로 정책 파급도를 정의해 볼 수 있다.

즉 정책이 파급된다는 것은 넓은 의미로 해석한다면 정책에 대해 사회적으로 인지도가 상승한다는 의미와 정책 자체의 전파를 통한 실천주체의 증가 및 정책 효과의 극대화 및 만족도 증대로 표현할 수 있다.

한편 기존의 대표적인 매스미디어의 형태로서 신문기사를 들 수 있으나 2000년대 후반부터 빠른 속도로 전파된 SNS의 미디어로서의 기능이 새롭게 부각되고 있다. 여기서 신문기사가 주로 언론사를 중심으로 전개되는 공급자 위주의 미디어라면 SNS는 쌍방향 소통을 통한 공급자와 수요자가 동시다발적으로 정보를 생산해 내는 새로운 미디어 채널로서 주목받고 있다.

따라서 본 연구에서는 보도자료 및 언론사를 중심으로 한 공급자의 시선을 대표하는 신문기사와 공급자와 수요자가 상호 연관적으로 실시간으로 소통하는 미디어인 SNS에서 충청남도가 적극적으로 추진해 온 주요 정책에 관한 키워드를 파악하여 도정에 관한 여론을 간접적으로 측정하도록 한다.

2. 분석자료의 범위 및 자료구축 방법

1) 분석자료의 범위

본 연구의 분석 자료는 크게 신문기사와 SNS데이터로 나누어진다. 흔히 이러한 문자(텍스트)가 기반인 데이터를 비정형데이터라고 하는데 이는 여러 통계표로 대표되는 정형화된 수치데이터와 대비되는 표현으로 사용하는 용어이다. 이러한 비정형데이터를 수집하고 분석하기 위해서는 기존의 정형데이터와는 다른 분석방법론이 필요하다.

먼저 데이터의 수집에 관해 살펴보면 신문기사에 대해서는 자체적인 스크랩이나 검색엔진에서 제공되는 데이터를 사용하여 독자적으로 전산화를 거쳐 텍스트데이터화 시키는 것이 필요하다. 그러나 전자화되어있지 않은 데이터를 처리하기 위해서는 OCR(Optical Character Recognition)이라는 광학문자인식기술이 필요한데 전반적으로 한국어에 관해서는 인식도가 많이 부족한 것이 사실이다.

한편 SNS데이터에 관해서는 유명한 데이터가 Facebook, 트위터(Twitter), 카카오톡 등이 존재하는데 이 중 가장 분석이 용이한 매체가 트위터 분석이다. 그러나 이 경우에도 개인정보관리 등과 관련하여 이전보다 일반인이 데이터를 제공받기 힘든 상황이 전개되고 있다.

따라서 늘어나는 분석수요를 조달하고 비정형데이터에 관한 자체적인 아카이브를 구축하는 것이 힘들 경우에는 데이터를 보유하고 있는 기업과 적극적인 연계를 통해 해법을 찾는 것이 중요하다고 할 수 있다.

한편 최근에는 서울시가 심야버스 노선 선정에 KT의 통신데이터를 이용한 것으로 유명하다. 이러한 통신사 데이터 이외에는 행정에서 습득할 수 있는 빅데이터로서 민원데이터와 센싱(측정)데이터 등 이 있을 수 있다. 그러나 이와 같은 데이터베이스는 축척에 의미를 두기 때문에 활용도가 낮다는 것이 문제점으로 지적되고 있다.

2) 분석자료의 수집

본 연구에서는 다음 표5 와 같은 자료를 구축하여 충청남도 정책에 관한 키워드 분석을 실시하였다.

구축한 자료는 크게 언론기사와 트위터 데이터로 나눠진다. 언론데이터는 주로 웹페이지에서 모은 기사를 활용하였고 기간은 2013년 1년간을 수집하였다. 또한 트위터는 공개트위터만을 기준으로 분석하였다.

트위터 데이터를 추출한 방법은 먼저 공개 트위터 데이터안에서 각각 주제1과 주제2를 포함한 트위터데이터를 추출한 뒤 관계가 적은 데이터는 삭제하는 방법을 거쳤다. 이 때 주제1은 지역에 관련된 주제어이며, 주제2는 충남의 핵심정책과 전반적으로 연결되어 있는 주제어와 지명 등 고유명사를 포함한 주제어로 나누어진다.

<표 4> 수집 데이터 개요

데이터 종류	언론 데이터	트위터 데이터
수집방법	- 포털사이트(네이버) 웹 페이지의 뉴스 리스트	공개 트위터 분석 - 수집 및 전처리는 ㈜사이람에 의뢰
기간	2013.1.1.~2013.12.31	2013.1.1.~2013.12.31
건수	12,959건(충청남도 검색) 언론사: 384개 (상세 목록은 부록참조)	주제1: 약 30만 건 주제2: 약 125만 건

<표 5> 트위터 추출 주제어

주제1	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&계룡 공주, 부여, 천안, 예산, 아산 등의 지역명은 충청남도와 전혀 상관없는 트윗를 추출하는데 영향을 미쳐 추출 키워드를 수정 ex) 백설공주, 동기부여, 천안함, 정부 예산, 아산병원 등
주제2	사회적경제, 사회적기업, 공유경제, 착한기업, 마을기업, 내포신도시, 3농혁신, 서해안, 농업직불금, 농촌마을, 지속가능발전, 행복지표, 협동조합, 6차산업, 미더유, 경제선순환, 에너지, 균형발전, 송전선로, 전통시장

3. 분석방법의 개요

1) 텍스트마이닝(Text Mining)

빅데이터 분석에 특화된 분석기법으로 가장 대표적인 것이 데이터마이닝(Data Mining)이다. 데이터마이닝은 가설을 정확하게 수립하고 모델을 검증하는 기존의 통계분석과는 달리 데이터의 홍수 속에서 해답을 찾아내는 기법으로 귀납적이고 경험적인 방법이다. 또한 그러한 분석을 실시하기 전까지 수많은 과정의 전처리가 필요한데 이러한 과정도 데이터마이닝에 속한다.

데이터마이닝이 수치정보에 근거한 정형화된 데이터 (Structured Data)를 처리 및 분석하는 방법이라면 텍스트마이닝은 비정형데이터(Unstructured Data, or unstructured information)을 처리 및 분석하는 가장 대표적인 방법이라고 할 수 있다. 왜냐하면 현재 비정형데이터라고 불리는 대다수의 정보들은 사람의 말이나 글이기 때문이다. 대표적인 비정형데이터로

e-mail, 논문, 책, 사진, 오디오, 비디오 등을 들 수 있으며 이러한 데이터는 언어데이터를 디지털화 시켜야 하는 과정을 반드시 수반한다. 이러한 처리 및 분석과정에서 쓰이는 기법을 텍스트마이닝이라고 한다.

텍스트마이닝의 일반적인 과정은 크게 4단계로 이루어져 있다. 데이터 수집, 가공, 정보추출, 분석의 절차이다.

2) 텍스트마이닝 기법 및 지표

텍스트마이닝의 기법은 크게 정보추출, 문서 클러스터링, 토픽 추출, 웹마이닝, 질의응답시스템 등을 들 수 있다. 이러한 여러 기술 중에서 본 연구에서는 문장의 분절을 통한 단어별 통계작성기법을 적용한 뒤 단어를 분석하는 빈도수나 여러 지표를 사용하였다.

○ R Package: tm, KoNLP

분석에 사용한 R 패키지는 tm 과 KoNLP이다. 여기서 tm은 텍스트마이닝을 위한 패키지이며 KoNLP는 한국어 처리를 위한 패키지이다. [4]또한 한국어 사전은 sejong.dic에 충남관련 정책 키워드를 포함한 user dic을 구축하여 수행하였다.

○ TF-IDF

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. 즉 여러 문서에 동시에 출현하는 단어는 출현 확률이 높다는 전제하에 역문헌 빈도수를 계산하여 DF가 커질수록 중요도가 감소하는 효과를 볼 수 있다.

---

4) 자세한 사항은 R홈페이지(http://www.r-project.org/) 를 참조

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

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

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

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

TF: 문서내 특정 단어의 빈도수

DF: 여러 문서내의 특정 단어 빈도수

IDF: DF의 역수

그러나 TF-IDF모델은 일반적인 단어를 분석하는 측면에서는 매우 유용하나 고유명사나 새로운 개념을 분석할 때는 중요도가 과소 및 과대평가 될 수 있어 주의를 요구한다.

따라서 본 연구에서는 이러한 한계점을 보완하기 위하여 키워드 빈도수를 중심으로 TF-IDF를 보완적으로 사용하였다.

○ 연관어 분석

연관어 분석에 앞서 연관관계의 정의가 불가피하다. 연관관계(association relationship)란 어떠한 단어에 대한 다른 단어들이 가진 패턴의 유사성을 의미한다. 연관규칙의 기본개념은 도로 파악할 수 있다.

○ 트위터 노출도 분석

트위터 분석에서 어떠한 트위터 또는 URL이 얼마만큼 전파되었는지를 나타내는 정도를 노출도라고 명명하였다.

즉 노출도는 트윗의 총 노출 범위를 말하는 것으로, 트윗을 보게 되는 전체 유저수를 일컫는

다. 계산방법은 해당 트윗을 작성한 사람의 팔로워와 RT한 사람들의 팔로워를 모두를 합하여 중복을 제거한 유저수로 환산하였다.

또한 '자주인용된 미디어'의 노출도는 그 미디어(URL)를 포함한 트윗들의 노출도를 모두 합한 값이다.

3) 분석의 구성

본 연구에서는 충청남도 정책에 관한 키워드를 분석하는 방법으로 신문기사와 SNS데이터를 이용한 텍스트마이닝을 실시하였다. 구체적으로는 텍스트마이닝을 통한 키워드의 TF-IDF분석과 관련어 분석을 실시하고 관련어를 바탕으로 한 키워드 네트워크 맵을 작성하여 단어 간의 연관성에 대해 가시적으로 검토하는 과정을 거칠 것이다. 나아가 네트워크상의 중심적인 단어에 대해 중심지 지표의 간략한 분석을 통해 중심 키워드를 도출하도록 한다.

제4장 충청남도 정책키워드 분석 결과

1. 충청남도 언론기사 분석

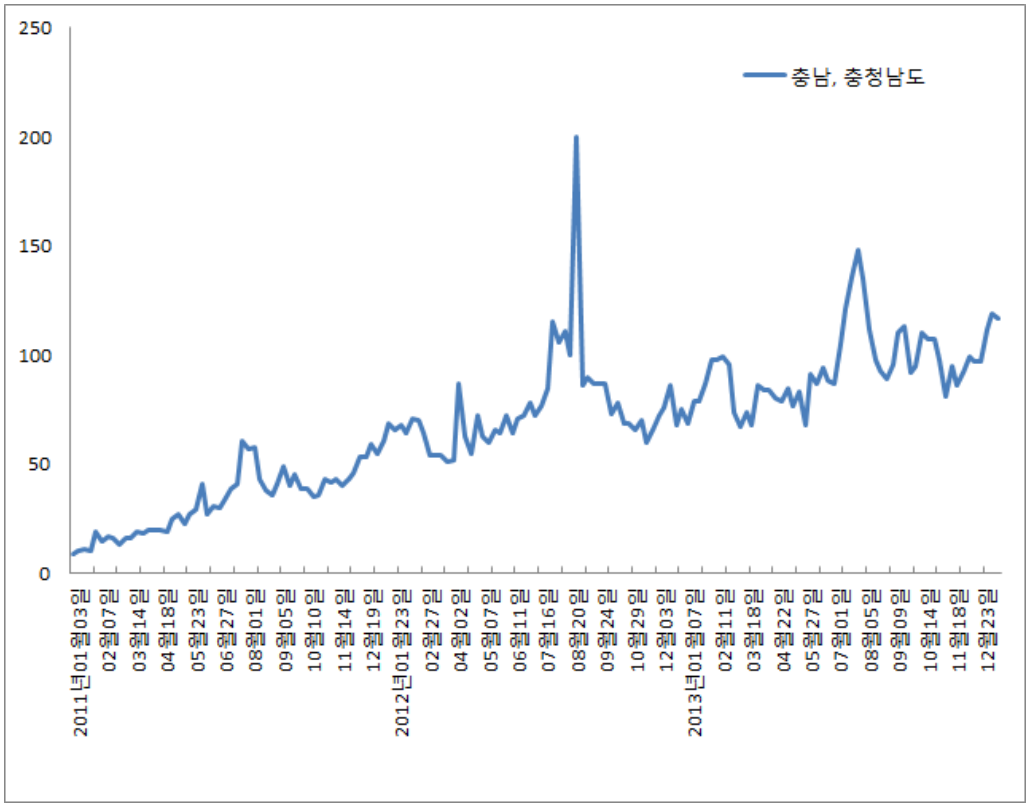
여기서는 웹에서 수집한 충청남도 언론기사를 대상으로 텍스트마이닝을 이용하여 분석한 결과를 살펴보도록 한다.

1) 기본 통계 - 언론기사 전체 월별현황 및 검색어

먼저 충남에 관련된 언론기사의 전체적인 경향을 파악하기 위해 네이버 트렌드(http://trend.naver.com/)를 통해 [충청남도] 와 [충남] 이라는 단어로 검색어 추이를 알아보았다.

그림5는 각각 충청남도와 충남의 검색어 횟수를 살펴보면 2010년도이후 건수가 지속적으로 증가하는 경향을 띠고 있으나 간헐적으로 검색어가 급속히 증가하는 시점이 등장한다.

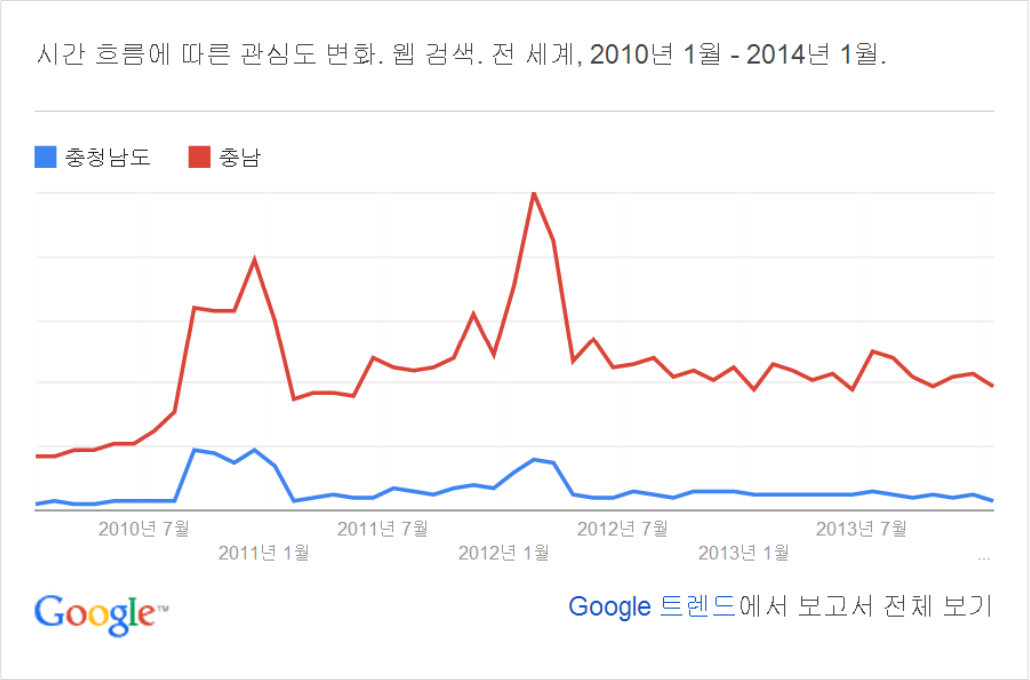
이때 주로 충청남도에 대한 관심도가 상승했다고 할 수 있다.



[그림 4] 충남, 충청남도 키워드 네이버트렌드 검색결과

다음으로 구글트렌드(http://www.google.com/trends/)도 네이버와 같이 검색어에 대한 정보를 제공하고 있는데 네이버트렌드에 비해서 시, 군 단위의 지역도 파악할 수 있어 더욱 자세한 경향파악이 가능하다.

구글트렌드 결과를 살펴보면 2010년 1월부터 2014년 1월에 걸친 기간 중, 2010년 하반기부터 2011년 연초, 2012년 6월 즈음의 검색어 결과가 급상승했다는 것을 알 수 있다.



[그림 5] 충남, 충청남도의 구글 검색수

다음으로 지역별로 집계된 [충청남도]라는 검색어의 검색 현황을 살펴보도록 한다. 기간은 2010년부터 2013년 12월까지이고 아산시의 검색회수를 100으로 놓고 아산시를 기준으로 다른 지역을 상대적으로 표시한 수치를 나타내었다. 그 결과 전국 현황 중 아산시에서 검색한 횟수가 가장 높게 나타났다. 이 때 아산시를 100으로 봤을 때 천안시가 80, 그 다음으로 대전광역시가 28, 서울시가 10 으로 나타났다. 즉 충청남도 내에서는 아산시, 천안시지역의 사람들이 충청남도라는 검색어로 가장 많이 검색했으며 그 외에는 충청남도 내 지역이 아는 대전광역시와 서울시에 있는 사람들이 검색을 실시하였다. 한편 아산시와 천안을 제외한 충청남도 내 다른 지역들은 충청남도에 대한 검색어 빈도수가 많지 않다는 것을 알 수 있다.

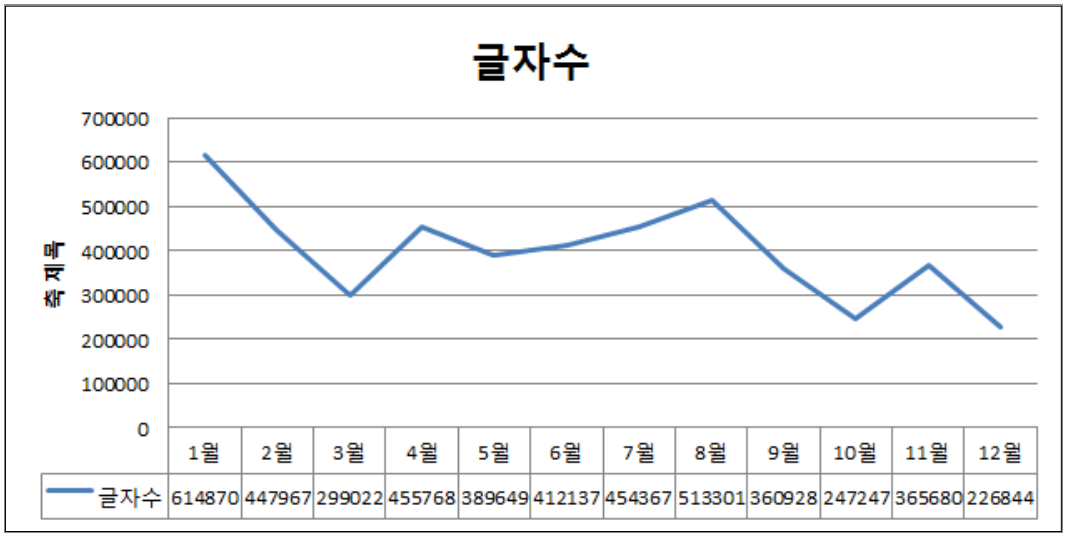
<표 6> 충청남도에 대한 관심이 높은 도시

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

다음으로 웹에서 [충청남도]라는 단어가 들어간 신문기사를 추출한 데이터의 단어수를 살펴보았다.

데이터베이스의 출처는 네이버 뉴스 라이브러리에서 2013년1월부터 12월까지 1년 동안의 신문기사를 수집하였고 이를 OCR등으로 정리하였다. 또한 100퍼센트 일치하는 기사는 보도자료 등의 복사로 간주하여 중복 수집하지 않고 1회로 제한하였다. 그 결과 총 12950건의 기사가 추출되었다.

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



<그림 6> 구축한 신문기사 데이터 월별 글자 수 및 단어 수 그래프

2) 신문기사 키워드 분석 결과

○ 단어(명사) 추출 및 주요 키워드 분석

여기서는 2013년도 충청남도라는 단어를 포함한 신문기사를 수집한 데이터베이스를 이용하여 형태소 분석을 거친 후, 명사를 중심으로 단어를 추출한 결과를 살펴보도록 한다. 이 때 출현빈도가 12회[5]가 넘는 단어를 추출하고 동일한 의미를 가지는 단어를 통합한 결과를 표 10에 나타내었다.

표 10을 통하여 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.

5) 최소 출현빈도 설정에는 이견이 있을 수 있으나 12달치라고 생각할 경우 한달에 한번 이상 언급되었다고 가정할 경우 총 빈도수가 12회일 때라고 할 수 있음.

<표 7> 신문기사 주요 키워드 리스트

단어	빈도수
충청남도	4961
사업	3483
안희정(도지사, 충남도지사)	1312
내포신도시 (내포신청사)	851
활성화	653
디스플레이	397
중소기업	368
농업기술	299
서해안	296
신도시	294
에너지	294
백제문화	285
친환경	229
글로벌	199
균형발전	167
네트워크	162
문화예술	160
지역발전	158
주민자치	156
고속도로	152
지역주민	148
사회복지	145
워크숍	142
협동조합	136
지속가능 (지속가능 발전)	136

단어	빈도수
사회적기업	114
황해경제자유구역	112
기후변화	99
온실가스	97
농공단지	76
출연금	68
환황해권	67
보령머드축제	65
사회복지시설	61
노동조합	58
경부고속도로	56
송전선로	53
공유재산	50
온양온천	47
거버넌스	46
국제통상	43
롯데아울렛	40
서해안고속도로	40
농촌마을	38
도시계획	34
에너지사업	33
환경오염	33
고마나루	29
상생산단	26
환경정화	26

단어	빈도수
농업정책	23
롯데백화점	23
발광다이오드	23
벤처기업	23
허베이스피리트호	23
지역균형발전	21
국립부여박물관	18
김수근문화재단	18
재정자립도	17
리솜스파캐슬	16
송산일반산업단지	16
스토리텔링	16
에너지관리	16
외국인투자기업	16
투자설명회	16
환경영향평가	16
충남테크노파크	15
부곡산업단지	14
탄천산업단지	14
친환경농업	13
자유무역협정	12
중소기업지원	12
천안국제웰빙식품엑스포	12
총 73개	

○ TF-IDF를 이용한 월별 이슈 추출

다음으로 신문기사텍스트를 이용하여 키워드를 분석한 결과를 살펴보도록 하겠다. 여기서는 TF-IDF를 이용하여 3글자 이상인 키워드 중에서 TF-IDF가 0.5 이상인 단어를 월별 이슈로 선정하였다.

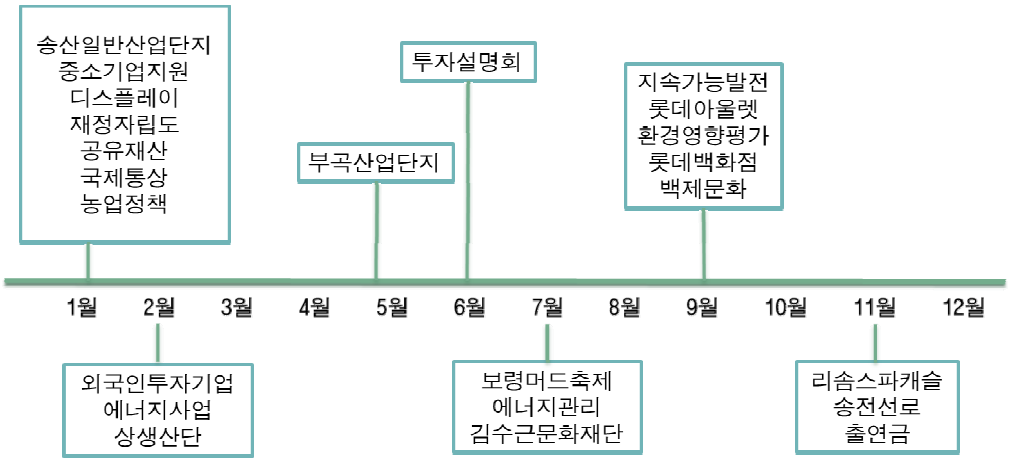
결과를 살펴보면 3월, 4월, 8월, 9월, 12월에는 특별한 이슈는 추출되지 않았다. 한편 1월은 중소기업지원, 산업단지, 디스플레이, 국제통상 등의 지역경제와 관련된 이슈가 추출되었는데 이는 연초의 도정 계획 및 언론의 관심이 주로 경제분야에 주목되어 있다는 것을 알 수 있다.

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

9월에는 다른 달에 비해 부여 롯데아울렛 및 롯데백화점에 관한 화제가 주된 기사였다. 이와 관련해서 지속가능발전도 함께 언급되었던 점이 특징적이라고 할 수 있다. 11월에는 리솜스파캐슬에 관한 기사가 많았다. 한편 충남내 송전선로에 관한 이슈도 많이 언급되었다.



[그림 7] 월별 키워드 추출(TF-IDF 이용)

2. 충남도 관련 트위터 분석

여기서는 트위터에서 수집한 충청남도 언론기사를 대상으로 텍스트마이닝을 실시한 결과를 살펴보도록 한다.

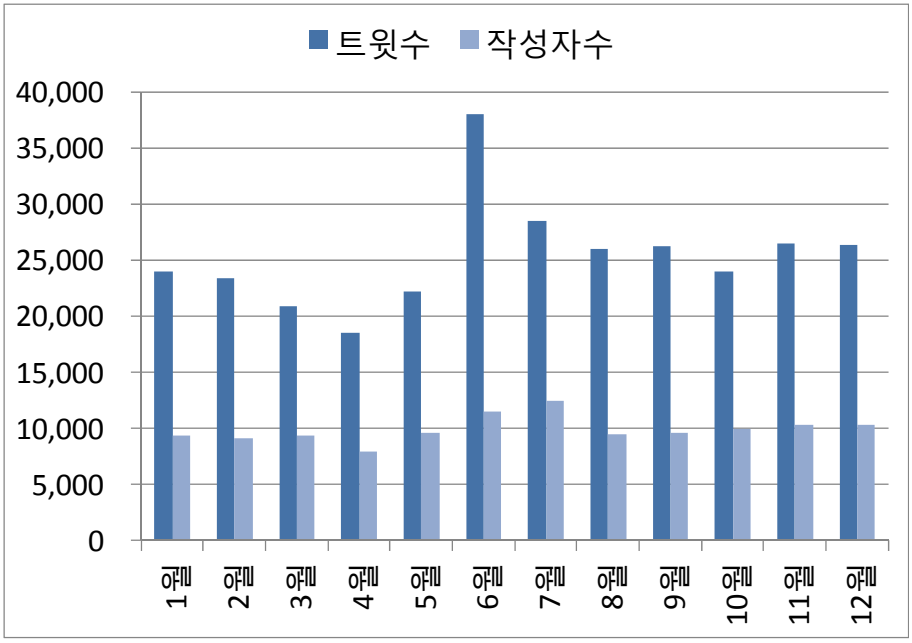
1) 기초 현황 분석

먼저 기초 데이터를 살펴보면 충청남도 관련 키워드를 언급한 트윗수와 작성자수를 나타낸 것이 그림9이며 표6에 나타낸 주제2의 정책별 키워드를 언급한 트윗수와 작성자수를 나타낸 것이 그림10이다.

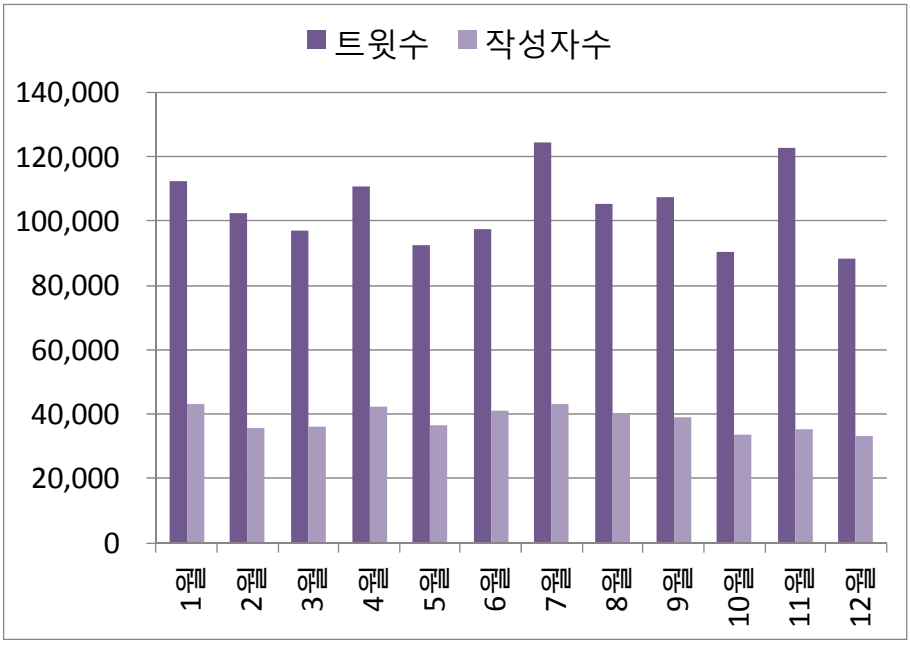
먼저 그림9를 보면 2013년 한 해 동안 6월이 가장 많은 트윗수를 나타냈고 6월을 기준으로 상반기보다 하반기가 트윗수가 많아진 것으로 보아 충남에 대한 트윗노출도가 하반기로 갈수록 증가했다는 점을 알 수 있다. 이를 바탕으로 하반기에 들어서서 충청남도에 대한 관심도가 상승했다고 유추할 수 있다. 주요 트윗 내용을 살펴보면 6월에는 충남도민체전에 아이돌스타 출연으로 인한 팬들의 관심, 안희정 충남지사 관련뉴스, 국정원관련 충남지역 시국선언이 있었고 7월에는 해병대사설캠프 사건, 호두과자 업체 노무현 비하 사건, SKT 홍보글 등이 화제를 일으켜 트위터수를 증가시킨 요인으로 작용했다.

한편 충남도 정책키워드 관련 트윗수는 이와 다르게 월별 경향이 크게 다르지 않은 것을 발견할 수 있다. 전체 트윗수는 충남을 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 않는 것으로 보아 고정된 인원이 정책키워드에 관해 관심을 가지고 언급하였다는 점을 알 수 있다.

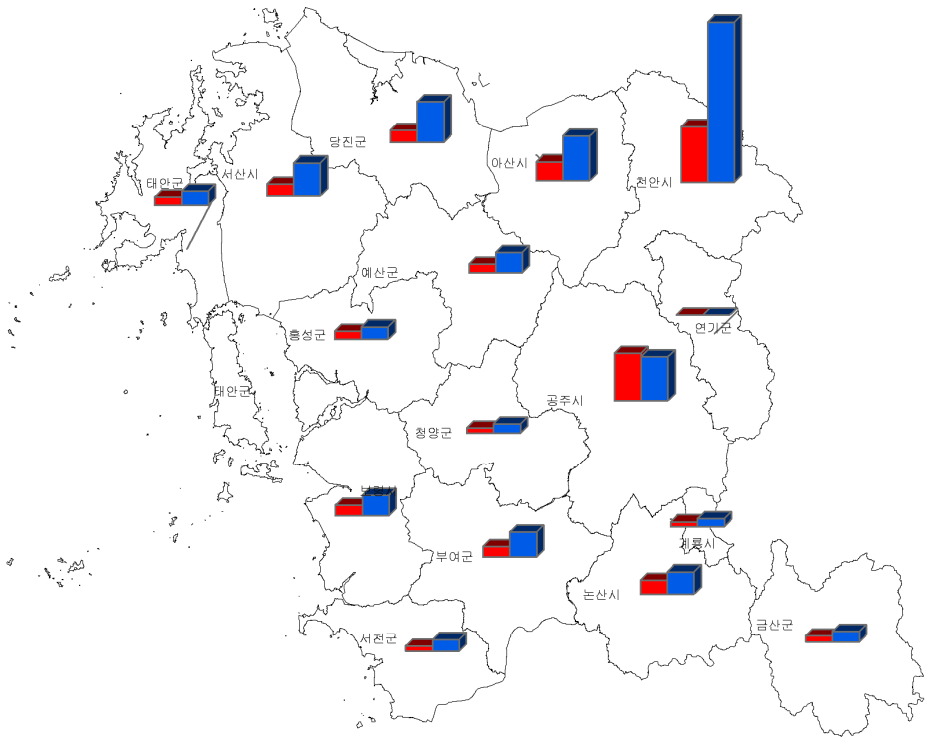


[그림 8] 충청남도 관련 트윗수 및 작성자수



[그림 9] 정책키워드 관련 트윗수 및 작성자수

다음으로 각 시, 군에 관한 트위터 현황을 살펴보도록 한다. 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 높았다.



[그림 10] 각 시군별 관련 트위터 현황(붉은색: 1월~6월, 파란색: 7월~12월)

2) 연관어 분석

충남 또는 충청남도에 관한 연관어분석 결과로 다음과 같은 결과를 얻을 수 있다. 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. 한편 충남도의 중요한 정책 이슈이자 많은 사람들이 주목하고 있는 [행복]이라는 키워드 또한 1635회를 차지하는 높은 수준의 빈도를 나타내고 있다.

다음으로 지역명에 대한 결과를 살펴보면 천안, 아산, 대전 순으로 빈도가 높았으나 천안이 다른 지역에 비해 압도적인 빈도수를 나타냈다. 한편 다른 충남지역의 노출도는 대전이나 충북, 경기보다 낮은 수준이었다.

<표 8> 빈출 키워드 1: 고유명사, 인물

충남 고유명사		인물	
키워드	빈도수	키워드	빈도수
충남도청	8623	안희정	9982
충남대(학교)	15540	노홍철	2628
내포신도시(청사)	11771	김호연충남	2017
충남교육청	3278	김종성	1661
		이회창충남	1657
		충남도지사	1574
		김형오부산	1564

<표 9> 빈출 키워드 2: 일반명사, 지역명

일반명사		지역명	
키워드	빈도수	키워드	빈도수
경품	8287	천안	15407
추천	5732	아산	6887
이벤트	5605	대전	6539
확인	5167	공주	5664
희망	4999	충북	5085
개청	4879	강원	4829
상품권	4755	경기	4677
지역	4732	서산	4610
퀴즈	4680	전북	4085
농축수산물	4490	세종	4083
추첨	4191	논산	4066
도전	4151	경북	3988
참여자중	4059	전남	3928
스마트폰	4008	보령	3791
개최	3999	당진	3782
세계최초	3898	예산	3175
속도	3597	전국	2965
학생	3492	홍성	2929
명단	2796	서북구	2813
찬성의원	2768	부여	2733
지사	2631	동남구	2580
시작	2399	태안	2086

학교	2083	부산	1918
실시	2037	수원시	1606
사랑	2035	영도구	1596
대통령	1999	안성시	1594
이전기념	1912	연제구	1573
친구	1888	팔달구	1564
현재	1862		
사람	1849		
장학사	1833		
경찰	1824		
SNS공유이벤트	1741		
등록	1644		
행복	1635		
의원	1617		
지원	1566		

다음으로는 각 정책별로 나눈 키워드를 살펴보도록 하겠다. 여기서는 전체 정책에 대해 살펴보는 것은 불가능하기 때문에 그 중 3농혁신에 관한 키워드를 추출해서 살펴보았다.

3농혁신에 관한 키워드 중 가장 높은 빈도를 차지하는 것은 충청남도라는 지역이었으며 그 다음이 안희정지사의 이름이었다. 그 다음으로 많았던 것이 개최, 농업, 의회 등이었다. 이로 미루어보아 3농혁신에 관련된 기사는 충청남도가 주도적인 역할을 한 기사가 대부분이었다는 점을 알 수 있다. 이것은 충남에서 보도자료나 사업홍보를 위한 트윗이 대부분이었고 다른 이미지나 일상적인 언어에서는 많이 언급되지 않는다는 것을 알 수 있다. 한편 3농혁신에 관해서는 다른 지역에 대한 언급은 거의 없는 것으로 보아 충남도 고유의 정책으로 봐도 무관하다고 할 수 있다.

<표 10> 3농혁신 키워드

3농혁신 키워드	빈도수
충남도(충남, 충청남도)	231
안희정(도지사, 지사)	124
충남타임뉴스홍대	42
개최	37
농업	31
충남도의회	31
가속	24
다짐	23
전진대회	21
추진	20
사업	20
하반기	19
가시	18
내포	17
농업기술원	17
선정	16

다음으로 실제 트위터에서 노출된 URL을 추적하여 어떠한 정보가 가장 많이 노출되었는지를 알아보았다. 많이 노출 되었다는 것은 많은 사람들이 이 정보를 접했다는 것을 의미한다.

여기서는 상위 10위까지의 노출도를 보이는 URL을 정리해 보았다.

<표 11> 노출도 상위 10위

분류	미디어명	URL	컨텐츠명	인용한 트윗 수	인용한 작성자 수	노출도
뉴스미디어	중도일보	http://www.	오늘의 대전 충남 충북-중도일보(3월 19일 화요일자)	1	1	271,370,256
뉴스미디어	뉴스1	http://news	충남대 학생들, 도내 초중고생 멘토로 활동	1	1	94,113,479
뉴스미디어	동아일보	http://news.	[대전/충남]대전 문화체육계 '트리플 펀치'	1	1	90,764,618
커뮤니티	메디톡	http://medi	메디톡	38	1	59,597,575
뉴스미디어	위키트리	http://www.	Social Network News Service	1	1	42,958,992
뉴스미디어	한국일보	http://medi	[여행] 시간이 멈춘 그 곳..충남 강경	1	1	27,941,323
동영상	유스트림	http://www.	인터넷종합편성방송 팩트TV 입니다.	5	2	24,158,349
뉴스미디어	노컷뉴스	http://www.	노컷뉴스	5	5	14,362,910
뉴스미디어	SBS	http://news.	김종성 충남교육감 음독 시도...병원 후송	10	10	14,302,519

먼저 가장 높은 노출도를 보이는 기사 중 충남과 관계가 깊은 기사를 보면 충남대 학생들에 관한 기사와 여행정보에 관한 기사를 볼 수 있다. 한편 부정적인 기사인 경우 높은 노출도를 보이는 것도 알 수 있다.

다음으로는 노출도만이 아니라 인용한 트윗수 및 작성자 수를 살펴보도록 하겠다. 인용 트윗수 및 작성자수가 많을수록 많은 사람들이 이 정보를 알리고자 했다는 것을 나타낸다. 전체적으로 이웃돕기, 롯데아울렛 오픈, 정치적인 화제 등이 주로 언급되었다. 공주시 수돗물에 관련된 글이 상위로 올라온 점도 특징적이다.

<표 12> 인용 트윗수 10 이상 미디어

분류	미디어명	URL	컨텐츠명	인용한 트윗 수	인용한 작성자 수	노출도
블로그	네이버블로그	http://blog.naver.co	신천지 자원봉사단, 충남 논산에서 농촌봉사 활동 펼쳐 : 네이버 블로그	192	145	309,037
블로그	네이버블로그	http://blog.naver.co	충남 논산 농촌 일손돕기 : 네이버 블로그	185	149	218,419
블로그	다음블로그	http://blog.daum.n	병풍 김대업 ,안희정 50억 배달사고 주장! [단독] '종북 in USA' 실체를 밝힌다	74	3	502,313
블로그	롯데그룹블로그	http://blog.lotte.co	충청권 최초의 관광쇼핑 테마파크, 부여 롯데 아울렛	70	18	369,952
블로그	롯데그룹블로그	http://blog.lotte.co	충청권 최초의 관광쇼핑 테마파크, 부여 롯데 아울렛	52	12	369,952
블로그	네이버블로그	http://blog.naver.co	"헌 집 줄게 새집 다오!" - 충청남도인터넷방송 2013 이벤트 : 네이버 블로그	50	15	814,044
커뮤니티	메디톡	http://medi-talk.co.	메디톡	38	1	59,597,575
뉴스미디어	뉴시스	http://media.daum.	운전기사 대신 운전하는 안희정 지사	38	36	329,402
블로그	네이버블로그	http://blog.naver.co	[뉴스기사] 공주시와 K-water 충남중부권관리단, 안전한 수돗물 만들기 : 네이버 블로그	31	2	630,461
커뮤니티	네이버카페	http://cafe.naver.co	교육주체 우선하기 운동 후원요청의 글	24	1	1,153,569
뉴스미디어	연합뉴스	http://media.daum.	안희정 안철수 민주당 입당해야""	23	23	183,507
커뮤니티	네이버카페	http://cafe.naver.co	교육주체 우선하기 운동 후원요청의 글	22	1	1,153,569
블로그	네이버블로그	http://blog.naver.co	[뉴스기사] 공주시와 K-water 충남중부권관리단, 안전한 수돗물 만들기 : 네이버 블로그	20	2	630,461
커뮤니티	네이버카페	http://cafe.naver.co	교육주체 우선하기 운동 후원요청의 글	19	1	1,153,569
뉴스미디어	뷰스앤뉴스	http://www.viewsnr	각대학 교수들 잇단 시국선언, 민주주의 생사기로""	17	17	464,578
뉴스미디어	조선일보	http://news.chosun	안희정 실체없는 친노 이름으로 책임공방 옳지 않아""	15	15	4,490,645
커뮤니티	네이버카페	http://cafe.naver.co	교육주체 우선하기 운동 후원요청의 글	15	1	1,153,569
블로그	네이버블로그	http://blog.naver.co	충남도, 1,194곳 세무조사로 157억대 추가 세원 발굴 : 네이버 블로그	14	14	341,470
블로그	네이버블로그	http://blog.naver.co	충남도, 고액 체납자 공동관리TF 가동 : 네이버 블로그	14	14	204,882
뉴스미디어	한겨레	http://www.hani.co	충남 부여 지적장애인 찾기에 군민 1000여명 나서	13	13	4,440,935
뉴스미디어	경향신문	http://news.khan.co	홀로서기 성공한 '친노' 안희정 충남지사	13	13	351,082
커뮤니티	네이버카페	http://cafe.naver.co	교육주체 우선하기 운동 후원요청의 글	12	1	1,153,569
뉴스미디어	한겨레	http://www.hani.co	시험유출 충남 장학사 '대포폰' 사용	11	10	4,486,229
뉴스미디어	한겨레	http://www.hani.co	충남 전 지역에 '보호자 없는 병실'	11	10	4,476,901
뉴스미디어	블로터닷넷	http://www.bloter.n	훌륭한 대학 자료들, 왜 '검색'은 막나요	11	11	413,339
뉴스미디어	고발뉴스	http://www.gobaln	한양·가톨릭·충남대 교수 '국정원 규탄' 시국선언 가세	11	11	239,495
뉴스미디어	민중의소리	http://www.vop.co.	'150여명 해고 위기' 충남 학비노조, 혹한 속 노숙단식농성 5일째	11	11	199,609
뉴스미디어	경향신문	http://news.khan.co	[속보] 김종성 충남교육감 구속	11	11	145,075
뉴스미디어	SBS	http://news.sbs.co.k	김종성 충남교육감 음독 시도...병원 후송	10	10	14,302,519

3) 연관어 네트워크

여기서는 앞에서 추출된 연관어를 이용하여 각 중심 키워드를 중심으로 형성되어있는 연관어 네트워크를 살펴보도록 한다. 여기서 네트워크의 강도는 연관등장 빈도이며 사각형은 Newman 방식으로 클러스터링한 서브그룹을 나타낸다. 20개의 정책 키워드 중 충남지역의 정책 아젠다로서 큰 역할을 하는 2가지 키워드[6](3농혁신, 사회적경제)에 대해 네트워크를 그려보았다.

○ [사회적경제] 연관어 네트워크

사회적경제를 중심으로 한 연관어 네트워크에서는 안내, 부탁, 사회적경제센터, 개최 등이 중요한 키워드로 나타났다.

사회적경제 연관어 네트워크에서는 총 5개의 클러스터를 추출할 수 있었다.

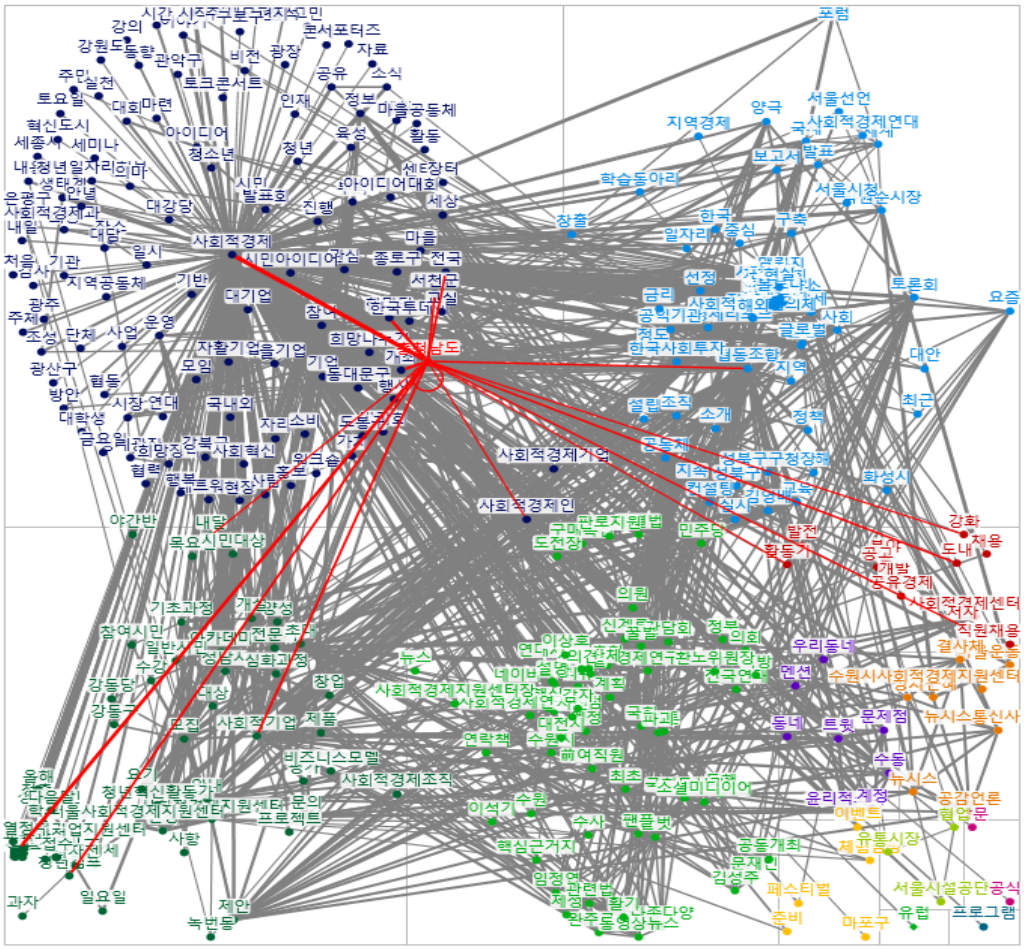
클러스터 1에서는 주로 이벤트 및 행사를 개최한다는 언급이 대부분이었다. 한편 각 충남, 수원시, 동대문구, 종로구, 도봉구 등 지역명도 거론되어 있었다. 한편 사회적경제와 비교적 가까운 거리를 유지하고 있는 클러스터 2에서는 충남도, 성남시, 화성시 등의 지자체와 교육, 판로지원, 구매촉진 등 실질적인 정책방향성을 제시하는 키워드도 발견할 수 있었다.

한편 클러스터 3 에서는 주로 사회적경제지원센터와 서울시에 관한 네트워크가 형성되어 있으며 협동조합 등의 언급도 볼 수 있다. 클러스터 4 수원 사회적경제센터가, 5에서는 강동구가 속해 있었다.

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

6) 여기서 정한 두가지 키워드는 충청남도 도정의 3대 혁신인 “3농혁신”, “행정혁신”, “자치분권”과, 충남발전연구원의 핵심 어젠다인 “행복”, “사회적경제”, “지역경제선순환”중에서 트위터 키워드 상위를 차지한 “3농혁신”과 “사회적경제”를 대상으로 분석하였다.

같은색깔은 같은클러스터에 속해 있다는 것을 의미함



Created with NodeXL (http://nodexl.codeplex.com)

[그림 13] 사회적 경제 관련 트위터 네트워크 (충청남도)

○ [3농혁신] 연관어 네트워크

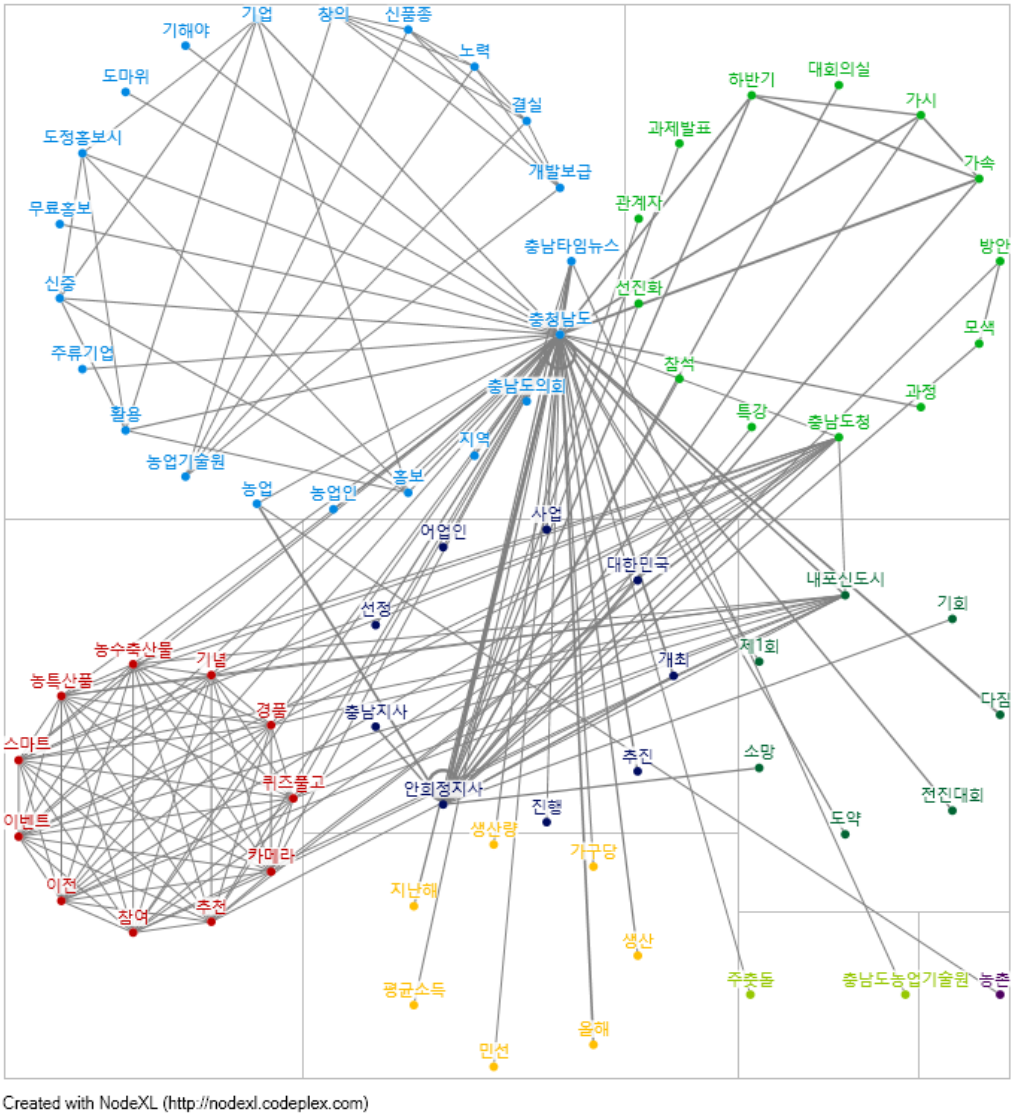
3농혁신을 중심으로 한 연관어 네트워크에서는 가장 중요한 키워드가 3농혁신이 아니라 충남, 충남도였다는 점이 특징적이다. 다만 충남에 관련된 명사가 대부분이며 주요 키워드들도 추진, 사업, 선정 등의 도정 보도자료나 정책기사를 벗어나지 못하고 있다는 점을 알 수 있다.

총 클러스터는 11개가 존재하나 의미 있는 클러스터는 3개정도이다.

클러스터 1에서는 충남도지사의 발언 및 관계자를 중심으로 한 농업 중요성 강조 및 행사 개최 및 참석에 대한 언급이 대부분이었다. 다음으로 클러스터 2에서는 민선5기, 충남도, 추진상황, 보고 등 민선5기의 정책성과로 연결짓는 언급이 연관되어있다.

마지막으로 클러스터 3 에서는 충남을 중심으로 몇 개의 촘촘한 네트워크로 연결되어 있는데 주로 행사 및 홍보, 농업기술 및 농업관련, 농어촌 등으로 나뉠 수 있다. 그러나 전체적으로 추상적인 명사가 많은 점도 특징적이다.

같은색깔은 같은클러스터에 속해 있다는 것을 의미함



Created with NodeXL (http://nodexl.codeplex.com)

[그림 14] 3농혁신 관련 트위터 네트워크

마지막으로 충남도 관련 트위터 연관어 네트워크 중 매개중심성이 높은 키워드를 도출해보면 표 15와 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.

<표 13> 중심성 지수

Vertex	Betweenness Centrality
충남	3029.683
천안시	1695.227
경품	705.904
예산	462.310
홍성	462.310
충청남도	252.963
명단	249.730
찬성의원	249.730
한미FTA	249.730
충남도청	133.814
지역	133.281
충남도	104.000
내포신도시	98.000

제5장 결론 및 제언

1. 주요 결론

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 충남도 정책 관련 키워드 구조에서는 천안시가 전체 정보네트워크 안에서 중요한 HUB로서 추출되었으며 이는 충남 내의 다양한 화제들이 천안시와 밀접한 관계를 이룬다고 볼 수 있다.

한편 사회적경제와 3농혁신에 관한 분석에서는 각 키워드의 특성과 충남도와의 관계를 조망할 수 있었다. 사회적경제는 전국적인 화두로 인식되고 있으며 구체적으로 서울, 수원, 성남 등과 같은 지역명이 대두되고 있고 이 중 하나로 충남이 언급되고 있다는 것을 알 수 있다. 이와는 대조적으로 3농혁신은 충남고유의 정책으로 거의 대부분의 언급이 충남도와 직접적으로 연관이 있는 키워드지만 전국적인 파급효과가 있다고는 보기 힘들며 아직 추상적인 단계의 사업들이 대부분이다.

2. 연구성과의 활용과 향후 과제

1) 충남도 빅데이터 활용 현황과 과제

충남도의 빅데이터 활용 및 발전방안으로 먼저 체계적인 민원조사 등을 통해 충남도민의

목소리를 적극적으로 반영하는 체계를 구축하여 기존 데이터의 체계적인 관리와 오픈데이터의 복합활용을 통해 맞춤형 정책과 효율적 의사결정을 실행하는 것이 필요하다.

한편 데이터 공유를 통한 효율적인 정보전달수단도 필요하다. 현재 충남도는 충남.NET이라는 포털사이트에 행정정보를 공유하여 충남도민의 정책에 대한 관심을 유도하고 행정데이터의 적극적인 활용을 도모하고 있다.

뿐만아니라 다수 부처에 분산되어 있는 안전과 관련된 데이터를 활용하여 분석 및 공유를 통한 재난예측체계를 도입하고 재난 대응기능의 보강을 넘어선, 재난의 사전예측·대비 기능을 도입하는 것을 목표로 하고 있다. 이를 통해 재난정보 빅데이터를 활용한 미래 위기대응 및 대비전략 수립하는 것을 목표로 하고 있다.[7)]

이와 같이 충남도에서는 다방면의 빅데이터를 이용한 행정혁신전략을 수립하고 있다. 이러한 기존의 대책들은 추후 지속적인 추진이 필요한 부분이며 추가적인 참여와 아이디어의 실사업화를 통한 가치창출에 초점이 맞춰져야 한다. 이를 위해서는 중앙정부에서 파악하지 못하는 지역밀착형 데이터 구축이 시급하다. 또한 기존의 데이터베이스를 더욱 효율적으로 결합하여 사용하는 방안을 끊임없이 고민해야한다. 또한 빅데이터를 통한 공공분야혁신의 가치사슬을 바탕으로 데이터 구축, 융복합, 의사결정, 공유 및 전파의 단계에서 신속한 의사결정이 이루어져야 한다.

빅데이터를 통한 충남도의 로드맵으로서 데이터를 통한 과학적인 정책평가 모니터링이 우선되어야 하고 행정 데이터를 체계적으로 공유함으로써 정책수요자인 도민이 융합을 통한 아이디어를 발굴 할 수 있는 체계를 마련해야한다. 나아가 아이디어를 통한 신규 사업들이 발굴되고 신속하게 사업화되는 것이 필요할 것이며 도정 전반과 새로운 아이디어에 대한 홍보 및 공유가 SNS와 같은 네트워크상에서 이루어져야 한다. 이를 위해서는 함축된 정보를 효과적으로 전달하는 인포그래픽과 같은 플랫폼이 중요할 것이다.

빅데이터 활용 시 주의해야 할 점도 많다. 빅데이터는 아직 과도기적인 개념으로 정확하게 그 의미가 정립되어 있지 않다는 점과 구체적인 사업화가 어렵다는 것이 큰 한계이다. 또한 실질적인 성과를 얻기까지는 오랜 시간과 노력이 투입되어야 한다는 것도 위험부담이 크다고 할 수 있다.

---

7) 출처: 2013년 충청남도 시책토론회 자료

또한 개인정보 및 보안상 문제에 대한 방안이 선행되어야 할 것이고 빅데이터의 규모에만 주목하거나 데이터를 구축하는 측면에만 치우쳐서는 안 된다.

빅데이터의 가장 핵심은 데이터간의 융복합을 통한 가치창출에 있으므로 현실의 문제점에 대한 질문을 가지고 이를 해결할 목적으로 빅데이터 분석을 진행하는 것이 바람직하다. 즉 데이터베이스가 풍부해 질수록 데이터에 관한 한계는 줄어들지만 얼마나 현명한 질문을 하느냐에 따라 결과가 좌우되게 된다.

이에 관해서는 충남도 정책에서도 빅데이터를 적극적으로 활용하여 현재 문제에 대한 대응방안과 끊임없는 질문, 그리고 데이터를 기반으로 한 정책판단이 이루어져야 할 것이다.

또한 향후 데이터과학과 기술이 발달함에 따라 많은 기술적 한계를 극복할 수 있을 것이고 이를 통해 행정혁신과 사회혁신을 도모할 수 있다는 가능성에 주목할 필요가 있다. 즉 공공분야의 빅데이터 정책을 효율적으로 추진 및 평가하는 툴(Tool)로서 적극적으로 활용할 것, 행정혁신을 리드하고 서포트 하는 체제를 구축한다면 충분히 성과를 거둘 수 있다.

현재 충남의 대표적인 문제인 한계마을문제, 환경문제, 지역경제선순환 등에 대해 빅데이터를 이용하여 해법을 찾아낼 수 있을 것이다.

2) 본 연구의 한계

본 연구의 한계로 먼저 자료의 한계를 들 수 있다. 본 연구에서는 트위터와 신문데이터만을 이용하였으나 더 많은 웹페이지나 다른 매체에서 얻을 수 있는 데이터를 이용한다면 조금 더 정확한 분석이 가능하리라고 생각한다.

또한 한국어 형태소 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 위치정보를 결합한 정보를 구득했다면 더욱 복합적인 결과를 얻을 수 있었을 것이다.

참고 문헌

김성웅. 2013. OECD의 빅데이터 관련 논의 동향, 제25권 10호 통권 555호

박주영. 2013. 공공혁신을 위한 떠오르는 키워드, 빅데이터, 예산춘추 NABO Budget & Policy, Vol.30

주 OECD 대표부. 2013.빅데이터를 활용한 창조경제 실현방안

충청남도. 2013. 2013년 충청남도 시책토론회 자료

박원준. 2012. '빅데이터(Big Data)' 활용에 대한 기대와 우려. Journal of Communications & Radio Spectrum

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

이유택, 홍영조. 2012. 알기쉬운 공공부문 빅데이터 분석 활용 가이드, 한국정보화진흥원

행정안전부. 2012. 스마트 행정 구현을 위한 빅데이터 마스터플랜 현황과 추진 계획

국가정보화전략위원회. 2011. 빅데이터를 활용한 스마트 정부구현

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

조수곤, 김성범,, 2011. 텍스트마이닝을 활용한 산업공학 학술지의 논문 주제어간 연관관계 연구, 2011년 대한산업공학회 추계학술대회

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

정근하. 2011. 텍스트마이닝과 네트워크 분석을 활용한 미래예측 방법 연구, 한국과학기술기획평가원

시로타 마코토 .2012. 일본의 빅데이터의 현황과 과제-사업전략을 발굴하기 위한 조직체제와 인재가 키-, IT프론티어, 노무라종합연구소, p6-9

■ 집 필 자 ■

연구책임 · 임화진 박사

전략연구 2014-15 · 빅데이터를 이용한 충남도 정책 키워드 분석

글쓴이 · 임화진
발행자·강현수 / 발행처·충남발전연구원
인쇄·2014년 8월 31일 / 발행·2014년 8월 31일
주소·충청남도 공주시 연수원길 73-26 (314-140)
전화·041-840-1123(기획조정연구실) 041-840-1114(대표) / 팩스·041-840-1129

ISBN · 978-89-6124-262-2 03350

http://www.cdi.re.kr