빅데이터 분석론에 대한 비판적 검토

윤석진*, 신상기**
*연세대학교 경영대학 산학협력중점교수
**배재대학교 미디어콘텐츠학과 교수
e-mail: ynskjn@yonsei.ac.kr

A Critical Review of Big Data Analytics

Seokjin Yun*, Shang Ki Shin**
*School of Business, Yonsei University, Seoul
**Dept. of Media & Contents, Paichai University, Daejeon

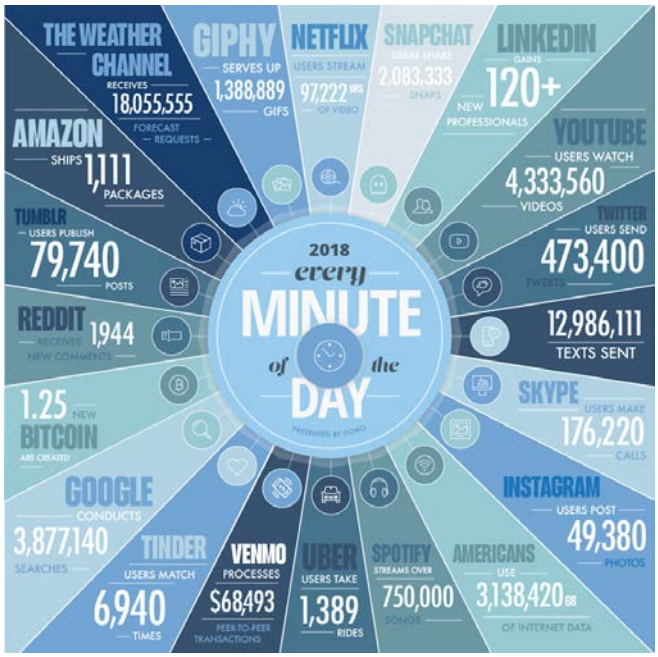
요 약

빅데이터의 출현으로 인한 데이터 기반의 변화와 더불어 빅데이터의 분석 가치를 옹호하는 빅데이터 분석론의 입장을 소개하고, 이에 대한 비판적 검토를 통해, 특히 사회과학 연구의 방법론 입장에서 빅데이터 분석론의 한계와 방향성에 대해 다양한 측면을 시론적으로 고찰한다.

1. 빅데이터 분석론의 배경

최근의 데이터 통신, 네트워크 및 플랫폼 기술과 서비스의 비약적인 발달에 따른 데이터량의 급증 추세는 단순히 증가라고 부르기에는 그 규모나 속도면에서 기존과 너무나 달라 데이터의 '폭주(deluge)'라고 부를 만큼 기존의 도구로는 감당하기 힘든 정도가 되었다. 이처럼 '빅데이터(big data)'라는 별도의 새로운 용어가 필요해질 정도로 데이터 기반의 물적 토대가 바뀌는데 원동력을 제공한 것은 디지털 혁명이다. 저장매체의 고용량화, 저비용화가 이루어지고, 스마트 기기를 포함한 데이터 수집 및 저장 기기가 소형화, 저렴화, 보편화되고, 네트워크의 보급·확산과 고속화로 데이터의 이동과 수집이 활성화되었으며, 연산능력이 향상되고, 인공지능, 기계학습 등 데이터 처리 기술이 발달하는 등 빅데이터의 등장 배경에는 하드웨어와 소프트웨어 부문 모두에서 일어난 기술환경의 혁신이 자리 잡고 있다.

[표. 2018 년 기준 1 분당 발생하는 데이터의 양상]



하지만 규모는 빅데이터를 특징짓는 요소 중 하나일 뿐이다. 물론 상대적인 것이기는 하지만 크기 자체로만 본다면 빅데이터 시대 이전에도 큰 규모의 데이터는 존재했다. 문제는 규모의 변화와 더불어 달라진 데이터의 성격이다. 즉, 예전과 같은 형태와 내용의 데이터가 단순히 양적으로만 늘어난 것이 아니라, 그렇게 늘어나는 데이터들이 그 원천과 생산방식, 구성과 사용방식에 있어서 지금까지와는 근본적으로 다른 새로운 형태와 내용의 것이라는 질적 변화의 측면에도 주목해야 한다.

빅데이터 시대에는 생활의 모든 측면에서 데이터가 생산되고, 측정되고, 기록되고, 저장된다. 이런 데이터의 대부분은 일상생활에서의 부산물로 나온다. 이메일과 메시지를 주고받고, 상품을 주문하고, 파일을 공유하는 과정에서 남기게 되는 전자흔적들(digital footprints)이 자동적으로 그리고 실시간으로 축적되고 구성된다. 그 결과 데이터의 형태도 지금까지처럼 정형화된 것뿐만 아니라 자연언어 텍스트, 사진이나 음악, 동영상, 위치 데이터 등 다양한 형태의 비정형적인 데이터까지 포함하게 된다. 또한 스마트폰을 통해 생산되는 다양한 데이터들의 경우처럼 이 데이터들을 서로 연결시키는 것이 가능해진다. 지금까지는 수집할 수 없었거나, 수집 대상으로 삼지 않았거나, 수집은 하더라도 분석할 수단이 없어 버려지던 데이터들이 모두 포함된다는 점에서 데이터의 성격이 달라진다.

이처럼 데이터의 양과 질이 달라지자 이를 어떻게 다루어야 할 것인가 하는 문제에 대한 시각도 달라지

기 시작했다. 이 변화를 어쩔 수 없이 떠안아야 하는 부담으로만 보는 것이 아니라 오히려 그것을 이용할 수 있고, 또 이용해야 하는 자원으로 보는 적극적 시각이 나타나고 기술적인 해결책을 찾기 시작한 것이다. 이렇게 빅데이터가 가지는 잠재력과 가능성에 주목하는 사람들은 데이터의 성격 자체, 또는 정의 자체가 변했다고 본다. 복합적 형태의 빅데이터는 미시 수준에서나 거시 수준에서나 종래에는 가능하지 않았던, 즉 전통적인 사회과학의 틀에서는 할 수 없었던 새로운 형태와 내용의 분석을 가능하게 할 것이라는 점을, 또 기술환경의 변화추세가 앞으로도 더 큰 규모로, 더 빠른 속도로 지속될 것이라는 점을 감안할 때 이렇게 생산되는 데이터의 비중은 계속 커질 것이라는 점을 강조한다. 그리고 데이터가 이렇게 양적으로, 질적으로 계속 변화한다면, 그것에 맞는 새로운 사회과학 분석 방법이 나와야 한다고 주장한다.

19 세기 프랑스의 수학자이자 물리학자였던 앙리 프앙카레(Henri Poincaré)는 '돌을 쌓아 집을 만드는 것처럼 과학은 사실(데이터)로 만들어진다. 그렇지만 돌무더기가 집이 아닌 것처럼 사실(데이터)의 더미가 과학은 아니다' 라고 말했다. 물론 돌이 많으면 그만큼 큰 집을 지을 수 있는 여지가 생기지만, 그 자체가 좋은 집을 짓는 충분조건은 아니다. 오히려 돌이 많아질수록 그것들을 어떻게 쌓느냐가 더 중요해진다. 즉, 데이터를 연결하는 틀과 조직하는 시각, 다시 말하면, 넓은 의미의 이론과 방법론이 여전히 핵심이라는 말이다.

2. 빅데이터 분석론의 입장

빅데이터 분석론에서 바라보는 데이터 규모(more), 데이터 구성(messy), 데이터 분석 준거(good enough) 이상 세 가지에 대한 입장은 서로 지원하고 증폭시키는 고리로 연결되어 있다. 첫 번째는 데이터의 양적인 측면이다. 지금까지 상상하지도 못했던 규모와 정확도로 수십억 인구의 일상생활을 관찰할 수 있게 되고, 특히 이를 통해 기존의 표집에서는 접근할 수 없었던 소수 하위집단들에 대해서도 세분화된 관찰이 가능해진다는 점이 강조된다. 두 번째는 데이터의 구성에 대한 입장이다. 빅데이터는 이전의 데이터처럼 일관되게, 체계적으로 축약되고 정리된 데이터가 아니다. 이는 빅데이터의 생성과정이 가져오는 불가피한 측면이기도 하고, 빅데이터가 실제 세상의 복잡다기함을 직접적으로, 즉각적으로 반영하기 때문에 생기는 측면이기도 하다. 따라서 예전의 작은 규모의 데이터에서만 가능했던 엄밀하고 통제된 정확성을 더는 요구할 수도 충족시킬 수도 없게 된다. 하지만 이런 한계점에만 집착하지 않고 이런 양상을 효용 측면에서 보다 주목한다면, 득이 더 많다는 것이다.

세 번째에 대한 입장은 그동안 사회과학 연구의 틀을 정립하는데 있어서 이상적 준거가 되었던 인과관계로부터 상관관계로의 방향전환이다. 한편으로는 더 커지고, 더 다양해지고, 더 직접적이 된 데이터와 이를 다룰 수 있는 향상된 연산능력에 힘입어 상관관계 자체를 다양한 측면에서 분석할 수 있는 가능성이 더 커졌다는 것이고, 그럼으로써 이렇게 많고, 복잡하고, 빠르게 생산·축적되는 빅데이터를 효율적으로 분석하는 현실적인 대안으로서의 상관관계의 입지가 인과관계의 그것보다 더 강화된다는 것이다.

지난 수 세기에 걸쳐 지속되어온 '과학적 방법' 의 작동구조는 관찰된 데이터로부터 파생된 변수들 간의 관계(상관관계)를 원인과 결과의 관계(인과관계)라는 결론으로 엄밀한 사고를 통해 도출해내는 과정으로서의 '인과관계 추론(causal inference)' 을 중심에 두고 있다. 이 과정에서 상관관계는 인과관계를 규정하는 데 있어서의 쉽게 말해 세 가지 필요조건 중 하나일 뿐이다(나머지 두 가지는 누락 변수의 통제, 시간우선성이다). 가장 중요한 것은 어떻게 원인이 결과를 만들어내는지를 설명하고, 그 기제를 제시하는 이론적 논의가 있어야 한다는 것이다. 따라서 이 기존의 틀에서는 인과관계가 데이터에서 경험적으로 찾는 상관관계보다 한 단계 위에 놓인다.

하지만 이런 사고의 틀에서 벗어날 수 있게 되었다는, 따라서 벗어나야 한다는 급진적 빅데이터 주창자들의 입장을 가장 단적으로 표현한 것이 아래에 인용된 미국의 와이어드(Wired)지 편집장이었던 크리스 앤더슨(Chris Anderson)의 주장이다.

'사람들이 왜 어떤 행동을 하는지를 누가 다 알 수 있단 말인가? 중요한 점은 사람들이 그런 행동을 한다는 것이며, 이제 그 행동들을 지금까지 상상도 하지 못 했던 정도로 정교하게 추적하고 측정할 수 있게 되었다는 것이다. 데이터가 충분히 주어진다면, 숫자가 스스로 자기 얘기를 할 수 있게 된다. 상관관계가 인과관계를 대체하게 되고, 일관성을 갖춘 모형이나 통합된 이론 등의 체계적 설명이 없이도 과학은 발전해 나갈 수 있다.'

양적으로, 또 질적으로 근본적으로 달라진 데이터가 이런 극단적 형태의 귀납적 경험주의를 가능하게 하며 이제 상관관계만으로도 충분한 시대가 왔다는 것이다.

이런 문제들의 대부분은 빅데이터 논의에 때로는 명시적으로, 때로는 암시적으로 깔려 있는 '데이터의 양이 많아지면 그에 따라서 질도 좋아진다' 는 양에서 질로의 전이(轉移)라는 가정에서 비롯된다. 객관성이나 정확성 등 사회과학의 데이터로서 갖춰야 할 속성들이 다다익선(多多益善)의 논리에 의해 확보될 것이라는 가정이다. 이 가정이 간과하는 것은 양과 질 사이의 충돌 지점들이다. 이런 맥락에서 사회과학 방법론의 활용 측면에서 아직은 간헐적 연구들을 통해 탐색적 수준에 머무르고 있는 빅데이터를 효과적으로 수용하기 위해 비판적으로 검토해야 할 사항들에 대해 기존 연구들을 참고하여 다음에서 간략하게 정리해 본다.

3. 빅데이터 분석론의 비판적 검토

첫째, 모든 데이터는, 수량적 크기와 상관없이, 신호(signal)와 잡음(noise)의 두 요소로 구성된다. 즉,

데이터에서 실제로 얻을 수 있는 유용한 데이터의 양은 데이터의 총수량만으로 결정되는 것이 아니라 신호 대 잡음의 비율(signal-to-noise ratio, SNR 혹은 S/N)에 의해서도 영향을 받는다. 그래서 이 둘을 분리해내고 잡음 요소(불완전 데이터 때로는 무관한 데이터)을 제거하거나 처리하는 것이 과학적 데이터 수집의 핵심적 단계 중 하나이다. 빅데이터 시대에 들어서도 이 단계를 생략할 수는 없다. 예전과 같이 데이터 수집 단계에서의 기획과 조직, 통제가 이루어지지 않은 채 네트워크에 의해 자동적으로 축적된 데이터이기 때문에 더욱 그렇다. 데이터의 규모가 커지면 이 문제는 오히려 더욱 증폭될 위험성이 있다. 그래서 역설적으로 데이터가 많아질수록 쓸 만한 데이터를 골라 내는 일이 더욱 어려워질 수 있다.

둘째, 많을수록 좋다는 빅데이터의 논리에 내포된 또 하나의 가정은 데이터의 수량이 커지면 그만큼 포괄적이 되고, 따라서 데이터의 대표성도 확보된다는 가정이다. 데이터 소스가 수백이나 수천에 그쳤던 지금까지의 제한된 사회과학 표집자료(sampling data)와는 달리 빅데이터의 소스는 전수 수준에 이른다는 점이 자주 강조되고 있고 있는데, 이로써 기존 표집의 한계, 특히 표집오차로부터 자유로워진다는 것이다. 하지만 빅데이터 분석과 전수조사(complete enumeration)는 개념적으로 별개의 것이고, 연구대상의 완전한, 또는 체계적인 대표성도 보장되지는 않는다. 가령, 트위터(Twitter)에서 생성된 트윗 데이터를 사용한 기존 연구에서 볼 수 있는 것은, 그 수량이 얼마나 큰가에 상관없이, '트위터 사용자'라는 특정한 범주에 관한 것일 뿐, 전체 인구를 대상으로 한 것이 아니고, 따라서 그 데이터가 누구를 대표하는지의 문제에 주의해야만 한다. 또한 기존 데이터와 다르게 실시간으로 지속적으로 생성된 데이터라는 특성 측면에서 볼 때도 어느 시점에 획득되었느냐의 문제도 중요하게 된다.

셋째, 관찰 또는 분석의 단위(unit of observation, unit of analysis)가 무엇인가 하는 문제이다. 실제 트위터 사용자와 트위터 계정은 반드시 일인일계정으로 짝지어지지 않는다. 데이터의 구조가 관찰하고자 하는 실제 사회와는 다른 방식으로 짜여 있기 때문이다. 이 상황에서 어느 쪽을 기준으로 'N'을 집계할지는 쉽게 결정할 수 없다. 스마트폰에서 생성되는 데이터나 과거 빅데이터의 관심에 불을 지핀 구글 독감추세 예측에 사용된 검색의 경우뿐만 아니라 최근 우리나라에서도 심심치 않게 문제가 되는 댓글 조작 이슈도 마찬가지이다.

넷째, 빅데이터의 분석 준거로 앞서 얘기한 상관관계에 내포된 문제점들도 고려해야 한다. 데이터의 규모가 커질수록 통계적으로 유의미한 상관관계를 찾을 가능성은 기하급수적으로 커진다. 하지만 이렇게 찾아지는 상관관계들의 대부분은 실제 상황을 이해하는 데 도움이 되지 않는 소위, 의사(擬似) 상관관계들이라 할 수 있으며 이를 오판할 경우 오히려 분석을 저해할 수 있다. 이 문제를 보여주기 위해 자주 드는 예 중 하나는 빅데이터 분석을 통해 주가지수(S&P 500)와 방글라데시의 버터 생산량 사이에서 높은 상관관계를 찾은 경우이다. 빅데이터 분석이 오히려 더 쉽게 '아포페니아(apophenia, 서로 연관성이 없는 현상이나 데이터에서 규칙성이나 연관성을 추출하려는 인식 작용)'에 빠질 수 있다는 점을 경고하는 예이다. 즉, 빅데이터에서는 상관관계를 너무나 쉽게, 너무나 많이 찾을 수 있기 때문에 오히려 분석과 해석은 더 어려워질 수 있으며, 따라서 데이터의 속성, 방법의 선택, 분석틀의 기본 가정 등에 대해 더 많은 질문과 고민을 할 필요가 있는 것이다.

다섯째, 온라인 환경에서 만들어지는 디지털 기록들이 무엇을 보여주는 것인지, 만약 그것이 사람들의 의견, 태도, 행위 등을 보여주는 것이라면 그것을 얼마나 어떻게 보여주는지 등의 문제도 다시 생각해봐야 한다. 사회 조사 데이터를 다루는 연구자들에게 이미 잘 알려져 있는 것 중 하나가 사람들이 뭘 하는지, 뭘 생각하는지, 어떻게 얘기하는지는 각각 서로 다르다는 것이다. 여기에 무엇을 어떻게 찾는지(search), 무엇을 얼마나 보는지(view), 어디에서 어디로 움직여 가는지(click) 와 같은 온라인 환경에서 생산되는 행위 기록까지 더해지면 데이터는 그만큼 더 복잡해질 수밖에 없다. 서치, 뷰, 클릭과 같은 행위들이 무엇을, 얼마나, 그리고 어떻게 보여주는지에 대한 심층적인 연구가 더 필요해진다. 이와 관련해 '보니니의 역설(Bonini's Paradox)'은 컴퓨터 모델링과 시뮬레이션에서 데이터 사이즈가 크면 클수록 좋을 것이라고 생각하지만, 실제로는 그렇지 않다는 것을 얘기한다. 현실을 이해하기 위한 모델링에서 모델이 또 하나의 현실이라고 할 만큼 거대하고 복잡해지면, 이해할 수 없는 또 다른 현실이 등장한 상황과 다름없어진다는 역설은 타당하게 보인다.

여섯째, 디지털 혁명을 통해 그동안 볼 수 없었던 새로운 영역들을 포함하는 다양한 생활의 영역들이 데이터화되고 있기는 하지만, 빅데이터가 보여주는 내용 또는 대상 영역은 사실상 매우 제한되고 편향되어 있다는 점이다. 그 예로 티라노사우루스라고 하는 공룡이 현재 많이 알려진 이유가 그 공룡이 중요하거나 특이해서가 아니라, 단지 그 공룡의 화석이 많이 남아 있어서 쉽게 발견되고 채집되기 때문이라는 점을 환기할 필요가 있다. 마찬가지로 빅데이터가 만들어지는 영역과 그렇지 않은 영역을 구분하는 기준에는 기술적 가능성, 상업적 필요성 등이 개입한다. 사회과학에서 다루는 영역 중 일부만이 전자에 속해 있고, 나머지 영역 중 어떤 것들은 앞으로도 당분간 디지털화되지 않을 것이다. 따라서 선택된 영역에서 자세한데이터들은 항시적으로 생산되기는 하지만 그런 영역들은 생활의 제한된 일부만을 보여주는 '깊기는 하지만 넓지는 않은' 상황이 생겨나는 것이다.

일곱째, 그냥 있는 그대로의 것이라는 뜻에서의 '원(原)자료(raw data)'는 논리적으로도, 실제적으로도 있을 수 없다는 점이다. 데이터는 그것이 만들어지는 역사 현실적 맥락과 과학기술적, 사회문화적 환경이 그 데이터의 성격을 규정한다. '자동적으로' 생산된다는 빅데이터의 경우에도 그 생산과정을 '자

동’이게 하는 ‘알고리즘(algorithm)’이 있게 마련이고, 그것을 통해 온라인상에 남겨지는 흔적들과 그것을 필요하게 하고, 가능하게 하고, 해석하게 하는 사회 문화적 논리가 연결된다. 그리고 이런 배경이 위에서 논의한 어떤 영역이 분석의 대상으로 선택되는지의 기준이 되고, 때로는 제약으로 작용한다. 아마존(Amazon)이나 넷플릭스(Netflix) 등의 회사가 만들어내는 데이터, 트위터나 페이스북(Facebook) 등의 소셜 미디어를 통해서 만들어지는 데이터 등의 빅데이터도 그것이 생산된 맥락과의 연결 속에서, 환경과의 관계 속에서 분석하고 이해하여야 하는 이유이다.

4. 결론

빅데이터의 출현으로 인한 데이터 기반의 변화가 사회과학의 대상과 방법에 큰 변화를 가져오고 있다는 것은 돌이킬 수 없는 사실이고, 그렇다면 그에 맞춰 사회 현상의 분석 전략도 바뀌어야만 할 것이다. 그런 의미에서 앞으로 사회과학은 방법론의 전환기를 맞게 될 것이다. 이 전환의 시기는 기존의 틀이 흔들리고 있다는 점에서 위기이기도 하고, 기존 방식의 한계를 넘어 새로운 가능성을 찾아본다는 점에서 기회이기도 하다.

이 위기와 기회를 통해 어떻게 사회과학을 할 것인가라는 새로운 방법론적 구상의 입장에서 본다면, 사회과학 연구는 사회현상에 대해 질문을 던지고 그에 대한 답을 찾는 연속적 과정이며, 그 과정에서 끊임없이 이론과 데이터, 데이터와 이론 사이를 오간다는 점이다. 데이터를 독립적인 별개의 것으로 보기보다는, 사회과학이라는 구조의 일부로, 또 그 안에서 이론을 보완하기 위한 상대역을 하는 역할로 보는 것이다. 이런 측면에서 빅데이터 시대에 사회과학이 성공적으로 적응하는 데 관건이 되는 것은 이렇게 변화된 데이터 환경 속에서 무엇을 어떻게 수확하고 소화해내는가 하는 것이다.

사회과학 분야에서 새로운 연구 문제를 찾는다는 관점에서 빅데이터의 속성 중 특히 주목할 만하고 효과가 클 것으로 보이는 것으로 두 가지를 꼽을 수 있다. 하나는 데이터가 커지면서 생겨나는 ‘깊이’를 잘 활용하는 것이다. 그동안 표집을 통해 얻은 데이터의 한계로 지적되어 온 소수집단과 희귀사건의 연구에 특히 활용의 가능성이 높아 보이는데, 산업과 경영 부문에서는 이미 상당한 진전을 보이고 있기도 하다.

또 하나는 쇤버거(Schönberger)와 쿠키어(Cukier)가 얘기한 ‘인과론의 포기, 상관성의 부활’이라는 빅데이터의 가치가 주는 의미 있는 메타포이다. 문제해결을 위한 기존 접근에 상호 보완적인 방법론으로서의 가치를 가져다줄 수 있다는 것이다. 그동안 논의만 되던 다양한 사회적 속성과 측면들간의 관련성과 복합성을 실질적으로 포착해낼 수 있는 물적 토대가 마련된 것이다. 빅데이터로 인해 가능하게 된 이 두가지 측면을 활용하려면 모집단의 대표값을 중심으로 하고 데이터의 복합적 다차원성을 축약해내도록 만들어진 현재의 분석틀은 다시 구상해야 할 필요가 생긴다.

그럼에도 불구하고 앞으로의 시도를 통해 빅데이터가 다양한 사회나 연구문제에 대해 기존 접근방법으로는 얻기 어려운 해답을 제공해 줄 수는 있겠지만, 유일한 방법도 아니고 또한 만병통치약은 더더욱 아닐 것이다.

현재 우리 사회에는 빅데이터가 인류를 더 나은 미래로 이끌어 줄 것이라는 전망이 우세하지만, 빅데이터를 통해 분명히 알 수 있는 것은 과거와 현재일 뿐, 그것이 곧 미래에 대한 정확한 예측을 의미하지는 않는다는 점이다. 빅데이터가 분명한 방향성을 제시할 수는 있지만, 그것이 인류의 건강한 미래로 직결된다고 단언하기도 어려울 것이다. 빅데이터가 인간과 삶, 인류의 미래를 위해 사용되는 것이라면, 다수(big)의 흐름을 ‘인간의 삶’이라는 기본 전제 위에서 의심하고 비판하며, 건강한 방향성을 설정하려는 인문학적 고민도 필요하다. 빅데이터가 창출한 가치에 대한 환호에 앞서, 그것이 과연 무엇을 위해 존재하는지에 대해 인문학이 던지는 근원적이고 큰 질문이 먼저 필요한 것이다.

참고문헌

[1]김기홍, “문화연구에서 빅데이터의 효용과 의미”, 인문콘텐츠 41, 2016.6.
[2]김동환, “빅데이터는 거품이다”, 페이퍼로드, 2016.
[3]김성태, “빅데이터 시대의 커뮤니케이션연구를 위한 방법적 경계확장과 논의”, 한국방송학회 학술대회 논문집”, 한국방송학회, 2014.11.
[4]김예란, “빅데이터의 문화론적 비판”, 커뮤니케이션이론 9(3), 한국언론학회, 2013.9.
[5]유강하, “빅데이터와 빅퀘스천”, 인문연구 82, 2018.3.
[6]이재현, “빅데이터와 사회과학-인식론적, 방법론적 문제들”, 커뮤니케이션 이론 9(3), 한국언론학회, 2013.9.
[7]임종수, “모나돌로지와 컴퓨터 연산 사회과학으로서의 미디어 연구”, 언론과사회 24(4), 2016.11.
[8]한신갑, “빅데이터와 사회과학하기-자료기반의 변화와 분석전략의 재구상”, 한국사회학 49(2), 한국사회학회, 2015.4.
[9]Anderson, Chris, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, Wired 16.07, 2008.6.23.
[10]Mayer-Schönberger, V., Cukier, K. N., 이지연 역, “빅데이터가 만드는 세상”, 21 세기북스, 2013.