커뮤니케이션 이론
2018년 14권 1호 50~98쪽
https://doi.org/10.20879/ct.2018.14.1.050

변화하는 시대에 커뮤니케이션학 연구하기

: 컴퓨터연산 사회과학방법을 중심으로 한 동향과 전망*

윤호영 연세대학교 커뮤니케이션 연구소 전문연구원

본 글은 새롭게 부상하고 있는 컴퓨터연산 사회과학 방법(CSS)이 커뮤니케이션 연구와 어떻게 서로 만날 수 있는 지에 관해 논의한다. CSS를 행위자기반 모형, 연결망 분석, 텍스트 마이닝 및 분석의 3가지로 요약하여 정리하고 해당 방법론이 가진 특징이나 지금까지의 경과를 논의하면서 이러한 방법론의 동향과 앞으로의 전망을 생각해본다. 그리고 교육 및 연구, 데이터와 관련된 사항 역시 몇 가지 고찰해 보고, CSS 연구 방법론이 가져올 수 있는 함의를 한 가지 지적하였다.

KEYWORDS 컴퓨터연산 사회과학, 데이터 사이언스, 행위자기반 모형, 연결망 분석, 텍스트 마이닝

* 본 논문의 초고는 〈정치커뮤니케이션 연구회〉와 방송학회 〈방송과 정치 연구회〉의 2017년 합동 연말 세미나에서 발표되었다. 자리를 마련해 주시고, 사회를 보아주시고, 참여하고, 토론해주신 김기태, 김소형, 김윤환, 노성종, 민영, 양정애, 이재국, 정일권, 최수진, 최지향 (이상 가나다순) 선생님들께 감사드린다.

1. 하나의 시선

컴퓨터 연산 사회과학(Computational Social Science, 이하 CSS) 연구 방법은 2000년대 이후 눈부신 발전을 거듭해오고 있다. 2004년 "새로운 과학(New Science)"을 선언하였던 연결망 이론 및 분석(Watts, 2004)은 2000년대 중반 연결망 확률분포 분석(예. Exponential Random Graph Model)을 발전시켰고, 이후 연결망 시계열 분석 모형을 개발하였다. 텍스트 마이닝을 비롯한 텍스트 분석은 과거에는 텍스트 파일을 프로그래밍 언어로 사전처리하여 저장된 텍스트로 불러온 후, 형태소 분석기로 품사를 태깅하고, 태깅된 말뭉치 파일을 색인 사전을 동원하여 단어로 추출하는 등 - 각 단계마다 해당 작업에 맞는 프로그램을 동원해야 하는 - 여러 개의 프로그램을 다루는 작업이었다. 그러나 현재는 대규모 자료의 토픽모델링을 비롯하여 기계학습과 인공지능으로 텍스트 및 단어 분류를 하는 등의 작업이 하나의 프로그래밍 혹은 프로그램으로 충분히 소화될 수 있는 방식으로 진화하였다. 흔히 쓰는 토픽모델링 방식인 LDA(Latent Dirichlet Allocation)의 정립이 2002년이었고, R프로그램의 토픽모델(Topicmodel) 패키지가 '통계 프로그램 학술지(Journal of Statistical Software)'에 소개된 것이 2011년, Word2vec 알고리즘이 발표된 것이 2013년이었으니, 텍스트 마이닝 역시 지난 15년간 엄청난 발전을 거듭한 것이다. 빅데이터가 활성화되기 이전 시기, 분석 사회학을 이끌던 사람들이 이론화 작업을 위해 많이 사용하였던 행위자 기반 모형(Agent-Based Model)은 사회과학에서 병리/역학 등의 연구로 이어지며 전파와 전염에 특화되어 발전해 왔는데(예. Epstein, Cummings, Chakravarty, Sinha, & Burke, 2004), 커뮤니케이션학에서는 소셜미디어 시대를 맞아 소통행위의 이론화 작업에 다시 활용되고 있다.

이들 연구방법들은 빅데이터 시대를 맞아 커뮤니케이션학에 특히 활

발히 응용되기 시작했는데, 이미 상당부분 대중화되고 오래전부터 논의되어 왔던 연결망 분석방법을 제외하고 아직까지 국내에서는 이와 관련된 논의가 부족한 상황이다.[1] 물론 이들 방법론들은 갑자기 등장한 것이 아니라 각각 독립된 영역에서 급속도로 발전해오고 있었다. 그러나, 최근 그 활용성에 대한 기대가 높아지면서 데이터 시대의 방법론으로 소개되고 있다(예. 황용석 외, 2017). 이러한 기대감에는 여러 가지가 있겠으나, 무엇보다 연구의 환경 변화가 새로운 질문을 던질 수 있게 하고, 새로운 질문이 기존의 방법론이 아닌 새로운 방법론을 필요로 하게 된 배경이 있는 것으로 사료된다. 따라서 이들 연구 방법의 유명세는 독립된 영역에서만 주로 사용되던 해당 방법론의 범용성이 여러 학문 영역으로 확장되는 과정에 기반한다고 보면 될 것이다.

이와 같은 배경하에, 이 글을 다음과 같이 구성하여 보았다. 우선, 새로운 방법론의 – 그 중에서도 CSS를 중심으로 하여 – 핵심적인 내용과 그 발전 과정을 살펴본다. CSS를 활용한 커뮤니케이션 연구에 대해 다양한 연구 방법을 다루며 논의된 바 있으나(황용석 외, 2017) CSS로 묶어서 설명하고 이들 연구방법론이 발전한 맥락을 통시적으로 들여다보면서, 현재 시점의 연구 지형의 변화로 살펴보지 않은 아쉬움이 있다. 그런 의미에서 CSS를 먼저 간략하게 정리하고 앞으로 변화될 것으로 예상되는 연구 지형에서 과연 새로운 연구 방법이 필요한 이유가 무엇일지 생각해 본다. 여기서는 새로운 연구 방법이 기존 연구 방법과 다른 점도 언급될 필요가 있으나, 그보다는 동향과 전망으로 대신하기로 한다. 이러한 관점의 채택은 비록 우리가 보는 특정한 현상은 일시적인 현상일 뿐이지만, 사실은 과거로부터 존재하던 흐름이고, 현재는 해당 흐름이 구조화되어가는 과정으로 여길 필요가 있다는 점을 강조하기 위한 것이다. 페르낭

1) 연결망 분석과 관련된 논의는 이미 〈커뮤니케이션 이론〉에서도 2번 다루어진 적이 있다. 김경모(2005), 송현주(2011)를 참조.

브로델의 표현을 빌면 우리가 보는 건 새로운 연구 방법이라는 '사건'이지만, 이 사건은 하나의 장기적인 '국면'을 이루고 있는 상태가 현재의 상태일 터, 이러한 국면이 어떠한 구조로 변화하여 끊임없이 우리에게 영향을 미치는지 살펴 보아야 할 필요가 있다는 것이다. 마지막으로는 이러한 논의들을 기반으로 연구와 교육에서 어떤 점을 고려할 수 있을 것인지 몇 가지 짚어보았다.

2. 새로운 연구 방법? : 컴퓨터연산 사회과학(Computational Social Science)

최근 회자되고 있는 새로운 연구 방법은 여러 가지 방식으로 불려지고 있다. 빅데이터 분석방법, 데이터 사이언스 등이 그 예이다. 각각의 명명이 의미하는 바가 있겠으나, 본 글에서는 새로운 연구 방법을 CSS로 한정하여 정의하기로 한다. 최근 각광받고 있는 분류와 연관 등을 중심으로 한 Big Data 분석방법, 그 중에서도 데이터 마이닝 또는 자연어 처리와 관련된 기계학습(Machine Learning) 또는 인공지능을 활용한 방법이 무엇인가를 논의하는 것도 필요한 작업일 것이다. 그러나 이들 분석방법은 연구방법론(Methodology)라기 보다는 여러 맥락에 사용될 수 있는 기술(technique)에 가깝다는 점에서 각각의 빅데이터 분석방법들이 커뮤니케이션학 전반과 어떻게 만날 수 있을 것인가 논의하는 것은 따로 이야기되어야 할 것으로 보인다. 비유로 설명하면, 통계 모수 추정 방법에 있어서 최대우도비나 베이지안 추정 모형 등 여러 가지 방법을 활용할 수 있지만, 연구 지형의 변화가 베이지안 추정 모형을 사용하도록 만들었다고 진술하기는 어렵다는 것이다. 모형선택에 있어서 베이지안 추정 모형이 최대우도비에 대비해서 더 나은 잇점을 설명하는 것이 올바른 설명방법일 것이다. 그런 측면에서 보면, 별개의 지면을 통해 빅데이터 분석 방법과 관련된 이야기가 다루어져야 할 것으로 보인다. 그럼 CSS의 첫 번

째 이야기로 행위자 기반 모형(Agent-Based Model, 이하 ABM)부터 시작하자.

1) CSS의 출발: 행위자 기반 모형(Agent-Based Model)

2000년대까지만 하더라도 분석 사회학(Analytical Sociology)의 한 기법으로, CSS라 하면 대개 시뮬레이션에 기반한 행위자 기반 모형을 의미하는 것이었다. ABM은 특히 사회학에서 각광을 받아왔다. 사회학 이론의 가장 큰 난제 중 하나는 미시(Micro)와 거시(Macro)의 연결이 이어지는 중간 단위(Meso)의 동학을 설명하기 어렵다는 점에 있었다. 예를 들어, 자본주의의 발전을 전체 사회의 생산력의 증가와 생산 양식의 변화에 의해 이행되었다고 보고, 거시 구조의 환경 변화로 설명할 수 있지만 이러한 설명에는 행위자가 보이지 않게 되며, 역사는 개인들의 의지와 상관없이 흘러가는 구조주의적 결정론에 근거한 행위가 된다. 그렇다고 행위자의 개인 동기로 설명하는 것이 쉬운 것은 아니다. 개개 행위자들의 행위가 어떻게 집단화하여 거대한 역사 변화로 이어지는지, 미시적 동기가 어떻게 거시적 구조 변화로 이어지는지에 대한 설명은 현실자료로 증명하기 매우 어렵다. 예를 들어, 베버가 방법론적 개인주의로 설명하듯 세속에서의 성공이 신에게 봉사하는 것이라고 믿었던 개신교 자본가들의 믿음이 자본주의 발전을 이끌었다고 설명할 수는 있지만, 어떻게 개개 자본가들의 자본 축적이 전체 사회의 생산력의 확장으로 이어졌는지, 자본가들의 다른 행위자들을 어떻게 이끌었는지 등을 설명하기 위해서는 과학기술의 발전과 경쟁, 국가의 역할, 자원의 개발 등 더 많은 변수들을 필요로 하기 마련이다. 따라서, 사회운동론의 발전에서 보듯 이러한 미시와 거시의 연결은 모든 현상을 설명할 수 있는 일반 이론으로서의 과학 법칙(Covering Law)이 아닌, 특정한 구조적 조건에 놓인 행위자들의 상호작용이 특정한 사건을 촉발시키면, 이에 따라 규칙적인 결과가 나타나는 메커니즘을 중시하는 설명방식을 추구하게 된다.[2)]

ABM은 이러한 설명방식에 기반하여, 미시적 행위, 조건들의 상호작용이 어떻게 거시적인 현상, 구조를 만들어내는지 반복적인 컴퓨터 시뮬레이션을 통해 검증하는 방법이다(Epstein, 1999). 미시적 동기들이 모였을 때 나타나는 의도하지 않은 효과 역시 ABM이 자주 검증하는 내용이다. 예를 들어, 내 주변에 나와 비슷한 취향을 가진 사람과 함께 살고 싶다는 개인적 동기가 의도하지 않게 취향에 따른 거시적 집단 거주지 분리 현상이 되는 식이다. 행위자들의 상호작용을 모형화하는 ABM에서 행위자들은 자율적, 독립적, 적응력을 가지고 성찰할 수 있는 존재로 정해진 규칙에 따라 행위한다고 가정된다(Cederman, 2005; Macy & Willer, 2002). ABM은 그 자체만으로도 연구방법으로서의 독특한 의미가 있지만, 연결망과 만나면서 연구 방법으로서의 힘이 배가되어 왔다. 특히 분석 사회학과 연결되는 주요 인물들이 현재 펜실베니아 대학 애넌버그 커뮤니케이션학과에 있는 것이 흥미롭다. 코넬대학(Cornell University)의 마이클 매이시(Michael Macy)와 함께 이 분야를 이끌었던 사람은 옥스퍼드(Oxford) 사회학과에 있던 피터 헤드스톰(Peter Hedström)이었는데 이 사람의 제자로 연결망 기반 시뮬레이션을 연구하는 산드라 곤잘레스(Sandra Gonzàlez-Bailó)는 옥스퍼드 인터넷 연구소(Oxford Internet Institute)를 거쳐 현재 펜실베니아 대학 애넌버그 커뮤니케이션학과에 있다. 복잡한 전염(Complex Contagion) 연구를 통해 연결망 시뮬레이션으로 유명해진 마이클 매이시의 제자 데몬 센톨라(Daemon Centola) 역시 MIT를 거쳐 같은 학과에 정착해 있다.

보다 구체적으로 ABM에 대해 이야기 해보자면, ABM은 3가지 핵심 요소, 즉 1) 행위자 설정 2) 환경 설정 3) 규칙을 통해 구현된다. 행위자 설정은 개인 행위자가 가진 속성을 정의하는 것을 말한다. 환경 설

---

2) 이러한 설명방식을 채택한 대표적인 사례가 Contentious Politics (예. Tilly & Tarrow, 2015)이다.

정은 행위자가 위치하는 공간의 물리적 설정과 행위자 행위에 영향을 미치는 행위 공간 또는 물체에 대해 컴퓨터 시뮬레이션상에서 정의하는 것이다. 규칙은 행위자간 상호작용의 규칙이거나 혹은 환경과 행위자간의 상호작용 규칙이다. 예시를 들어 설명해 보자면, 앞선 거주지 분리 현상에서 개인 행위자의 속성은 자신의 주변에 자신과 비슷한 취향을 가진 사람들과 살고 싶어한다고 정의할 수 있다. 환경은 이렇게 생각하는 사람들이 살고 있는 지역 즉 거주 공간이자, 행위 공간으로서 사람들이 '내 주변'이라고 정의하는 공간의 의미를 갖는다. 규칙은 개인이 생각하는 비슷한 취향을 가진 사람들이 어느 정도가 내 주변에 있다고 판단하는 임계점을 넘지 않는 한 이동하거나 이주하지 않고 그 자리에 머무는 것이며 임계점을 넘으면 이동한다고 정하면 된다.

이러한 내용을 침묵의 나선이론으로 적용하면 다음과 같이 된다.[3] 행위자 속성은 개인이 정치적 견해로 주변에서 분란을 일으키기 싫다고 정의한다. 환경은 정치적 견해를 가진 사람들이 무작위로 공간에 흩어져 있으며, 개인이 내 주변이라고 정의하는 공간이다. 규칙은 내 주변에서 다수인 정치적 의견이 나와 다를 경우 침묵하거나 혹은 내 주변이 다수가 아니라 내 주변 모두가 나와 다른 의견이면 다른 곳으로 이동한다고 정의한다. 이렇게 정의될 경우, ABM은 침묵의 나선이론이 어떤 식으로 진행되는지 볼 수 있게 한다. '내 주변'에 대한 정의가 변화할 때 침묵하는 사람의 비율이 어떻게 변화하는 지, 침묵을 견디지 못하고 이동하는 사람들의 비율은 어떻게 되는지, 과연 이러한 현상들이 정치적 양극화를 어느 정도 가져오는 지 등을 연구할 수 있는 훌륭한 이론적 방법이 된다. 모형에서 필요로 하는 여러가지 계수와 규칙 – 예를 들어 공간에 퍼져있는 사람들의 밀도나 '내 주변'에 대한 정의 등 – 을 현실에서 가져온다면 현실의 내용을 적용하여 현실과 얼마만큼 똑같아 지는지 검증하면서 계수와

3) 실제 ABM이 커뮤니케이션학에서 가장 먼저 적용된 것도 침묵의 나선이론이다.

규칙의 변화에 따라 모형이 어떻게 변화하는지도 볼 수 있다. 즉, ABM의 장점은 미시적 행위자 수준에서 출발한 거시적 현상에 대한 이론화에 있으며, 동시에 사회과학에서 통제하기 어려운 복잡한 요소들을 가정을 통해 통제해보는 실험적 설계방식에 있다.

2) CSS의 대중화: 연결망 분석(Social Network Analysis)

이제는 대중화되었다고도 볼 수 있는 연결망 분석에 대한 배경설명이나 그 내용은 이미 정리된 바가 있으므로 특별히 필요치 않을 것으로 보인다. 그럼에도 불구하고, 연결망 분석의 핵심 3요소만 간단히 복기하고 커뮤니케이션학의 연결망 분석이 다른 학과(예. 사회학)과 다른 차이점 한 가지만 부각시키도록 한다.

연결망 분석에서 중요한 3가지 개념은 1) 구조 2) 위치 3) 역할이다. 여기서 구조란 연결망 전체의 구조(global) 또는 내 주변(local)의 구조이다. 연결망 전체 구조는 물리학자들이 법칙을 발견하는 주요 무대였다. '작은 세상(small world)', '지수함수 분포의 법칙(power-law distribution)'. '척도 없는 네트워크(Scale-Free Network)' 등이 물리학자들이 관심을 가진 연결망 전체 특징의 예이다. 물리학자들은 또한 사회연결망 구조가 다른 여타 연결망과 어떤 점에서 다른 구조인가 연구하여 왔다. 잠시 설명하면, 사회연결망은 연결을 많이 가진 사람들 주변에 비슷한 정도로 연결을 많이 가진 사람들이 모여서 군집화하여 중복된 연결을 가진 상태로 집단을 이루는 특징을 보인다(Newman & Park, 2003). 이러한 연결망 구조는 신호 전달을 최적화하기 위해 말단에 많은 신경이 퍼져있고, 중간은 통로 그리고 신호를 만들어내는 신경에 또 다른 복잡한 연결망을 지님과 동시에 연결망 중복을 최대한 배제하고 신호전달의 효율성을 극대화하는 생물학적 연결망과는 매우 다른 형태이다. 또한 여러가지 복잡한 신호를 체계적으로 분류하고 정보를 취합하는 뇌신경 신호처리 방식과도 다르다. 이러한 연결망 구조에 대한 관심은 연결망

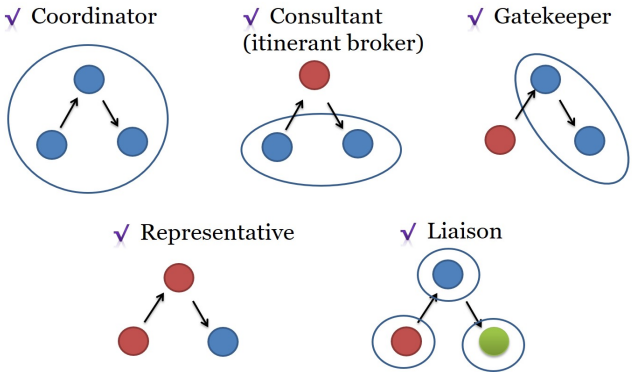
내 행위자에 대한 관심에 반비례한다.

구조에 대한 관심과 달리, 연결망내 위치와 역할에 대한 관심은 주로 행위자에 대한 직접적인 관심이다. 위치와 역할은 서로 엮여있는 경우가 많은데, 연결망의 노드(node)가 어떤 위치에 있는가가 그 노드의 역할을 결정하거나 혹은 특정한 역할은 반드시 특정한 위치에 있어야 하는 경우를 생각하면 될 것이다. 예를 들어, 3대로 이루어진 가족관계 연결망 그래프가 있다면, 할아버지와 자식 세대를 잇는 '아버지'의 위치는 일정하게 나타날 것이고, 그 위치가 담보해야 하는 역할을 생각해 볼 수 있다. 집단과 집단을 연결해주는 역할은 연결망 내 특정한 위치가 아니면 불가능하며, 연결망 내 위치는 특정한 역할을 할 수 있도록 만들 수 있는 기반이 된다. 생각해보면 연결망 분석의 많은 지표들이 개인 노드에 대한 수치정보이고, 이러한 수치 정보는 각각의 노드들이 가진 위치를 알기 위한 것이다 (그리고 위치로 역할을 설명한다). 위치와 역할을 통합하여 사고한다면 '지위(Status)'가 되는데, 이러한 지위는 특정한 행위에 대한 신뢰나 판단의 기제가 되는 신호로서 역할을 할 수 있다. 예를 들어 회사 조직의 연결망에서 최상위 위치와 그 위치에서 하는 역할을 생각해보면 왜 위치와 역할이 만나면 지위가 되는지 이해하기 쉬울 것이다. 이러한 특성을 요약하면 연결망은 프리즘이자 파이프로 볼 수 있다(c.f. Podolny, 2001). 정보가 흐르는 파이프이자, 신호를 전달하여 특정한 효과를 굴절시키는 프리즘 두 가지 기능을 하는 셈이다.

연결망 분석 방법은 분석 방법이 가진 많은 분석 지표와 방법에도 불구하고, 그동안 시각화(Visualization)와 중앙성(Centrality) 지표만이 남용이라 불릴 정도로 집중적으로 활용되어 왔으며, 각 중앙성 지표의 의미도 맥락을 고려치 않고 기계적으로 해석되는 등 다양하게 활용될 수 있는 가능성이 상당히 위축되어 왔다. 연결망은 중앙성 지표외에도 다양한 많은 지표가 있고, 연결망 분석을 통한 모든 연구 가설이 중앙성 지표로만 해결되는 것도 아니다. 예를 들어, 연결망내 중개(Brokerage) 역

할만 하더라도 다양한 형태가 존재하는데 이를 활용한 연구를 찾기 어렵다. 〈그림 1〉은 연결망의 중개 역할을 5가지로 구분한 것으로 집단간 소통 형태에 따라 중개자의 역할이 다를 수 있다는 점을 보여주고 있다. 같은 색깔은 같은 집단에 속한 것을 의미한다. 연결망으로 정의되는 문지기(Gatekeeper)의 경우는 정보를 받아서 전달하는 사람과 받는 사람이 같은 집단이고, 대변인(Representative)은 정보 제공자와 정보전달자가 같은 집단이다. 즉, 중개자가 정보를 연결해 줄 때 어떤 집단에 속해 있는가에 따라 그의 역할이 달라진다.

그림 1. 중개(Bridging)의 5 가지 역할 구분



* 진행자(coordinator), 상담역(consultant), 문지기(Gatekeeper), 대변인(representative), 연락책(Liaison)

여기서는 한 가지 예만 들었지만, 연결망 연구의 심화모형인 연결망에 대한 통계적 추론 및 시간 모형 등을 활용한 연구 역시 찾아보기 어렵다. DBpia 데이터베이스를 통해 지수분포기반 무작위 그래프 모형(Exponential Random Graph Model, ERGM)을 신문방송학 분야로 한정시켜 찾아보면, 해당 방법을 쓴 논문이 거의 보이지 않는다. 해외 저널에서는 정적인 형태의 연결망 통계 모형인 ERGM이 이미 보편화되어

가는 추세이긴 하나, 해외 역시 기초적인 연결망 지표에 집중된 것은 마찬가지이며, 연결망내 커뮤니티 분석이나 구조적 등위성 지표를 통계 분석 방법과 결합한, 이른바 그래프 없는 사회연결망 분석은 사회학 학술지나 전문 사회연결망 분야에만 집중적으로 활용되고 있을 뿐이다.

연결망과 관련하여 또 다른 흐름으로 커뮤니케이션학에서는 정보의 전파나 사회운동의 확산 등에 관한 연구를 하는 연구자들(앞서 언급한 곤잘레스 또는 센톨라)이 연결망과 관련된 '발견'을 주도하여 나가고 있는데, 이들이 발표하는 저널이 주로 커뮤니케이션학이 아닌 것에서 알 수 있듯이(예. Science, Scientific Report, Social Networks), 연결망과 관련되어 커뮤니케이션학의 고유한 이론이나 흐름이 있다고 보기 어려운 측면이 있다. 이러한 현상은 국내외를 막론하고 관찰되는 것은 물론이요, 통시적으로는 애초에 연결망 이론이 커뮤니케이션학에 적용되기 시작하면서부터 나타난 흐름이라 볼 수 있다. 피터 몽그(Peter Monge)와 노쉬 컨트랙터(Noshir Contractor)가 2003년에 "커뮤니케이션 연결망 이론들(Theories of Communication Networks)"라는 저서를 출판하기도 하였지만 핵심적 개념과 내용에 있어서 커뮤니케이션학만의 고유한 특징이 있다고 보기 어렵다. 서문에서 인용하고 있는 학자들도 임마뉴엘 카스텔과 그라노베타 등 일련의 사회학자들의 연구이며, 연결망의 창발성(emergent property)에 대한 언급 역시 커뮤니케이션학이 발전시켜왔다고 보기 어려운 사회학의 영역이다.

3) 컴퓨터 과학(Computer Science)의 침입: 텍스트 마이닝(Text Mining)

문헌정보학과(원래는 도서관학)에서 텍스트 마이닝은 정보 검색(Information Retrieval), 추출(Extraction), 요약(Summarization)을 통해 책을 비롯한 기존의 문서 정보를 정리하고 분류하는 목적이었다. 이후 1980년대에서 90년대사이 촘스키의 언어론이 자연어 처리의 지배적인 철학으로 자리 잡으면서, 글을 토큰화(tokenization)하고 이를 분석하는 알고리즘

의 기술적 혁신이 이루어졌다. 2000년대 들어서면서 텍스트 마이닝은 비약적으로 발전하는데 전통적인 정보 검색, 추출, 요약, 분류 뿐만 아니라 기존에 구분해놓은 주제 범주에 문서를 배치하는 주제 분류툴의 발전이 이루어졌다(Minfer, G., Delen, D., Elder, J., Fast, A., Nisbet, R. A.. 2012).

무엇보다, 현대적 텍스트 마이닝의 기술발전을 두고 기계학습을 비롯한 컴퓨터 알고리즘의 발전을 빼 놓을 수 없다. 알고리즘을 통한 문서 분류법의 혁신, 특히 단어뭉치(Bag-of-Word, BOW) 개념이 고차원 벡터 공간(High-dimensional space)으로 확장되고, 신경망을 이용한 분류법이 개발되었다. BOW만 가장 단순하게 설명해 보자면, 특정한 주제를 다루는 문서가 있을 경우 해당 주제를 다루는 단어들이 빈번하게 나오기 마련인데 그렇다면 해당 문서에 나온 단어들의 빈도나 조합으로 문서가 특정 주제에 관한 내용이라고 분류할 수 있게 된다. 쉽게 생각해서 대출 스팸 메일을 구분하기 위해서 메일에서 단어들이 쓰이는 순서나 문장 구조등의 내용 혹은 단어들이 쓰이는 문맥을 읽어볼 필요 없이, 해당 메일에 "오늘", "대출", "승인" 등의 단어가 빈번하게 나올 경우 쉽게 스팸메일로 구분이 가능하다는 모형이다. 가장 단순한 알고리즘 하나만을 이야기했지만, 텍스트 마이닝은 그 목적에 따라 매우 다양한 컴퓨터 알고리즘을 적용할 수 있다(예. Grimmer & Stewart, 2013). 기계학습, 인공지능 등의 알고리즘의 활용은 기본적으로 데이터 마이닝 기법에서도 자주 활용되는 분류 또는 집단화(클러스터링), 연관관계 분석도 활용 가능하다.

컴퓨터를 이용한 텍스트 분석과 관련해 디마지오(DiMaggio, 2015)는 흥미로운 설명을 제시한 바 있다. 그는 토픽모델링 분석방법을 컴퓨터과학자들과 협업하면서, 사회과학자와 컴퓨터과학자간에는 적어도 세 가지의 차이가 있음을 발견했다. 첫 번째 차이는 선호하는 모형의 차이이다. 사회과학자들은 비지도 기계학습(unsupervised machine learning)을

선호하는데 반해 컴퓨터과학자들은 지도기계학습(supervised machine learning)을 선호한다는 것이다. 단순하게 말해서 지도 기계학습은 컴퓨터에게 사람이 선험적으로 자료를 투입하며 해당 자료가 어떤 내용인지 미리 컴퓨터를 학습시키는 과정을 거쳐서 알고리즘의 정확성을 최적화하는 방법이다. 비지도학습은 컴퓨터에게 사전 자료를 주지 않고 분석되어야할 데이터만을 제공하여 문서 분류 등의 작업을 수행하도록 만드는 방법을 말한다. 사회과학에서 흔히 사용되는 토픽모델의 기법인 LDA 같은 방법은 텍스트 말뭉치의 잠재적 구조를 파악하기 위해 알고리즘을 특별히 사전에 훈련시키지 않는다. 이와 연관되어 (DiMaggio는 세번째 지점으로 말하고 있지만) 두 번째 차이는 컴퓨터과학자들이 역설적으로 사람을 사회과학자보다 더 신뢰한다는 것이다. 예를 들어, 텍스트를 분석하여 감성 분석을 잘 해내었는가에 대한 평가에 있어서 최고의 기준은 "사람의 판단"에 부합하는가라는 것이다. 이는 IBM의 인공지능인 왓슨의 암 진단에 대한 평가가 인간 의사의 판단과 얼마나 일치하는가에 의해 판정되는 것과 같은 원리이다. 앞서 언급한 사회과학자들의 비지도학습 선호 문제는 분석에 있어서 최대한 인간을 배제하고 객관성을 확보하기 위한 것인데, 컴퓨터 과학자들에게 중요한 것은 사람의 개입 여부가 아니라 자신들이 만든 알고리즘이 정확하게 목적을 달성하는가의 문제가 중요하고 따라서 인간의 판단에 기반한 지도학습에 대한 선호가 있는 것도 이해할 수 있다는 것이다. 그런 측면에서, 컴퓨터 과학자들은 계산 중간에 컴퓨터가 어떠한 과정을 거쳐서 나름의 결론을 내었는지, 그 과정에 인간이 이해하기 어려운 어떠한 과정이 숨어있는지에 대한 관심이 결여되어 있다는 것이다. 세 번째 차이가 이러한 태도와 연결되는데 사회과학자들은 모델을 테스트하는데 있어서 한번 익힌 방법으로부터 최대한 많은 것을 얻어내려고 노력하는 반면, 컴퓨터과학자들은 한번 만든 모형에 대한 관심이 있는 것이 아니라 새로운 모형을 만들어서 성능을 향상시키는데 관심이 있다는 것이다. 극도로 절제된 외교적 표현인 이 말은 결국 사회과

학자들과 컴퓨터과학자들 사이에는 채워지지 않는 간극이 있다는 설명이다. 사회과학자들은 모형이 분석한 과정에도 관심이 있어서 해당 모형 분석 결과가 나오게 된 과정에 대한 설명을 요구하는데 반해, 컴퓨터 과학자들은 이를 가볍게 넘긴다는 것이다. 디마지오는 토픽 모델링 초보자로서, 모형이 정확한 해답을 내놓고 있는지 평가할 수 있는 가장 단순한 형태의 모형적합도(goodness-of-fit)조차 없는 것을 믿을 수 없었다고 언급하면서, 컴퓨터 과학자들은 인과관계보다도 모형이 연구자가 원하는 바를 제대로 해내었나, 새로운 모델이 이전 모델보다 더 향상되었나를 본다고 말하고 있다.

그렇다면, 최근 상당히 주목받고 있거나 혹은 그 활용도가 증가하고 있다고 여겨지는 이들 CSS 연구 방법은 과연 어떤 문제 해결에 도움이 되는 것인가? 과연 연구 패러다임이 변화하고 있어서 이들 연구방법의 활용은 피할 수 없는 것인가? 만약 변화하고 있다면 실제로 해당 연구방법을 사용하는데 특정한 어려움은 없는가? 등등 관련 질문과 의문이 밀려올 수 밖에 없다.

3. 새로운 연구 방법의 필요성? : 변화는 어디에서 오고 있는가?

우선 CSS 발전과 관련하여 앞으로의 미래에 대해 언급한 경우로 카펠라(Cappella, 2017)의 4 벡터(Vector)론이 있다. 카펠라는 매스 및 대인 커뮤니케이션과 관련하여 4가지 방향성을 제시하면서 미래를 전망하고 있는데, 1) 연결망 분석 2) 요인이 아닌 행위자로 옮겨가는 미디어 효과 3) 추천시스템 4) 내로우캐스트(narrowcast)와 브로드캐스트(broadcast)의 결합으로 인한 (소셜미디어와 같은) 플랫폼 영향력의 증대가 바로 그것이다. 흥미롭게도 추천 시스템은 알고리즘의 영향력으로 볼 수 있고, 다양한 미디어 플랫폼에 따른 영향력에 관한 사항은 글에서도 제시하고 있

듯이 직접적으로는 소셜 미디어에 관한 사항이다. 요인에서 행위자로의 이행과 관련하여서는 "요인에서 행위자로(From Factors to Actors)"라는 정확히 말해 매이시와 센톨라의 ARS(Annual Review of Sociology) 학술지 2002년 논문의 제목을 직접적으로 언급하고 있고, 또한 센톨라의 Complex Contagion 논문인 2010년 사이언스(Science) 논문을 언급하고 있는데, 이는 행위자 기반 모형과 관련된 것이며, 특별히 빅 데이터와 상관없이 이론을 정립하는 연구로 코넬 사회학과에서 꾸준히 해 오고 있었던 것들이다. 즉, 앞서 설명하였듯이 이러한 벡터들은 사실상 앞서 설명한 텍스트 마이닝을 제외한 2가지의 CSS 방법이 CSS로 묶이기 이전부터 전통적으로 해오고 있던 것들인데, 이러한 방법이 구체적으로 데이터의 축적과 분석방법의 발전과 연관되는지는 의문이다. 결국, 카펠라의 이야기는 소셜 미디어와 같은 개인화되어 있으나 서로 연결되어 있는 연결망 플랫폼에 대한 연구가 매스 및 대인 커뮤니케이션 연구의 미래라는 것으로 요약될 수 있는데, 일정 부분 공감이 되기는 하나 – 에넌버그 학교가 성장시키고자 한 영역을 앞으로의 미래로 투사한 것은 아닌가 하는 생각과 더불어 – 미래를 이야기하기보다는 현재의 모습을 일반화 시킨 것이라고 여겨진다.

그렇다면, 다시 질문을 던져보자. 연구의 패러다임이 변화하고 있어서 새로운 연구 방법들에 대한 수요는 필연적인가? 패러다임의 기본적 정의로 살펴보면, 새로운 연구 방법들이 연구 패러다임의 변화를 이끌고 있다고 보기는 어렵다. 토마스 쿤(1970)이 이야기 했듯이, 패러다임의 변화라는 것은 정상과학의 비정상성이 정상과학의 기반을 흔드는 증거들이 지속적으로 관찰되어 비정상과학에 있던 비정상과학이 정상과학의 지위를 차지하는 것이다. 이는 연속적인 진행의 개념이 아니라 분절과 단절을 기반으로 한다. CSS에 의해 기존 연구와 분절, 단절이 이루어지고 있는가? 그렇다고 보기는 어렵다. 그러면 패러다임의 변화까지는 아니더라도, 새로운 연구 방향이 제시되고 있고 그에 따라 새로운 연구 방법이 각

광받고 있는 것일까?

1) 전망과 동향 1: 데이터에 따른 이원화 경향

CSS를 적극적으로 활용하는 연구자들은 1) 대규모 데이터의 생산 및 축적과 2) 대규모 데이터를 분석할 수 있는 도구의 개발이라는 두 가지 환경적 요인이 CSS 연구 활성화에 기여했다고 본다(Lazer et al, 2009; King, 2017; Watts, 2014). 일부 학자들은 CSS분야 자체를 아예 빅데이터 분석의 하위 범주로 구분하면서 대규모의 데이터를 컴퓨터 연산 또는 알고리즘을 통해 분석하여 사회적 질문(social inquires)에 대답하는 접근 방법이라고 정의하는 학자들도 있다(Shah, Cappella, & Neuman, 2015). 하지만, CSS를 두고 빅데이터 분석의 하위 범주로 구분한 것이나 데이터의 용량만을 강조하는 것은 CSS를 '새롭게' 접했다고 하는 고백에 다름이 아니다. 비록 CSS가 대규모 데이터가 축적되면서 그 방법론들이 크게 발전하였고 개별적으로 발달하고 있던 방법론들을 한데 묶어서 일컫는 말이기는 하나, CSS를 활용하는 분석이 반드시 빅데이터일 필요는 없다. 개리 킹(Gary King)의 표현을 빌면, 비록 데이터의 종류와 다양성이 이전에는 상상도 할 수 없는 규모로 증가했으나, 빅데이터 혁명이라는 것은 데이터에만 있는 것이 아니며, 데이터가 말할 수 있도록 - 원문에서는 "actionable" - 데이터 분석 방법이 지난 수십년간 발전한 것에도 있다는 것이다.

따라서, CSS방법을 활용한 연구 경향은 (때에 따라 대규모인) 해당 분석 방법에 적합한 데이터를 구할 수 있는가의 여부에 의해 결정되기 마련이다. 그렇다면 CSS와 관련하여 구할 수 있는 데이터란 무엇인가? 시뮬레이션에 기반한 ABM을 제외하고 보면, 이 글에서 정의한 CSS와 관련된 데이터란 텍스트로 된 자료, 사람들간 상호작용의 기록 그리고 이 모든 것들의 연결망 세 가지로 볼 수 있다. 그런데, 이 세 가지가 모든 경우에 다 구할 수 있는 것은 아니다.

(1) 텍스트 분석: 소통기록과 정보텍스트 사이에서, 현실은 언론 분석

텍스트 분석은 당연하게도 분석할 수 있는 텍스트가 생산되는 영역에서 활성화될 것인데, 정보는 대개 제도적 행위자들이 생산하는 문서 및 발언이 텍스트 분석의 대상으로, 소통은 시민들의 대화 기록이 분석의 대상으로 나타나고 있다. 현재 분석대상이 된 텍스트로 볼 때에는 소통 기록은 발언의 내용이 소셜 미디어에 집중되어 있으며 정보에 대한 텍스트로는 언론 기사가 대부분을 차지한다. 흥미로운 것은 정부가 많은 정보를 공개하고 있는 흐름에도 불구하고, 커뮤니케이션학에서는 언론 텍스트에 대한 분석에 집중되어 있다는 점이다. 〈표 1〉은 2017년 ICA CSS세션에서 발표된 텍스트 분석 관련 논문들의 제목을 취합하여 분류한 것인데, 방법론을 제외하고 보면 대부분 언론이 생산해낸 텍스트인 뉴스를 분석하고 있으며, 정부가 발행하는 텍스트나 시민들의 소통에 대한 분석은 상대적으로 적은 것을 알 수 있다.

표 1. 2017년 CSS 세션에 발표된 텍스트 마이닝 관련 논문제목

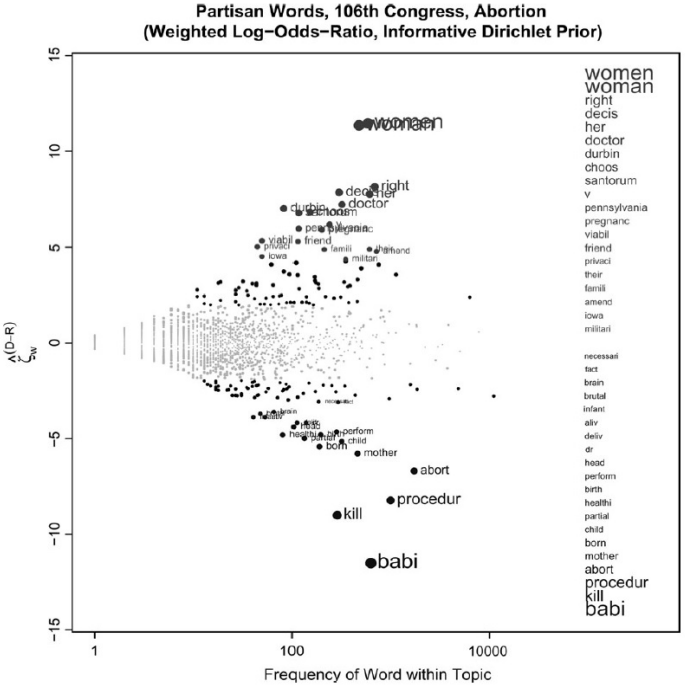
	정부	언론	시민	방법론
Mapping Out China's Online Public Diplomacy: A Computational Textual Analysis Approach	○			
Applying LDA Topic Modeling in Communication Research: Towards a Valid and Reliable Methodology				○
Combining Expert Surveys and Topic Modeling to Identify Unknown Issues in Multinational Media Discourses				○
What a Difference a Day Makes: Topics in Time-Aggregated User Comments on an Anti-Immigration Facebook Page			○	
Homophily Versus Prestige: Adding Topics to the Explanation of Link Formation on the Web				○
Best Practices in Using Topic Models for Communication Research				○
Facebook News Captions as Outsourced Emotions: A Computational Analysis of Subjective Language of Newspapers on Facebook			○	

	정부	언론	시민	방법론
A Case Study Using Syntax Dependencies to Find Differences Between News and Nonnews Tweets		○		
Evaluating the 'Liquidity' of Online News: An Assessment of the Homepages of U.S. News Organizations		○		
Finding Topics Associations in Political News Using Deep Learning		○		
Investigating the Gatekeeping Role of Social Media Editors of Newspapers Using Computational Text Analysis		○		
Same News, Different Platforms: A Computational Approach for Characterizing User Comments Across Platforms and Journalistic Genres			○	
Simulating Who Sees What: Dismantling the Babooshka Doll of Layers of Diversity in News Recommender Systems		○		
Visualization, Technologies, or the Public? A Text Mining Analysis of Tweets on Data-Driven Journalism		○		
Software Demonstration: Collecting and Analyzing Social Media Data Using SocialMediaLab				○
Software Demonstration: Virtual Observatory for the Study of Online Networks (VOSON)				○
Introduction to rtweet: Collecting Twitter Data				○
Same, Same? Ensuring Comparative Equivalence in the Semantic Analysis of Heterogeneous, Multilingual Corpora				○
Methodological Challenges in Estimating Tone: Application to News Coverage of the U.S. Economy		○		
Supervised Sentiment Analysis of Parliamentary Speeches and News Reports	○	○		
Best Practices in Computer Automation: Dictionary-Based Sentiment Analysis				○
Using Crowdsourcing for Developing an Attributed Sentiment Analysis Tool				○
Distributed Sentiment Analysis of Real-Time Political Tweets				○
Sentiment Analysis of Twitter Data of a Crisis: Supervised Machine Learning Method				○
Technical News Logics Amid the Proliferation of Computational and Data-Driven Journalism		○		

	정부	언론	시민	방법론
A Development and Evaluation of Keywords to Automatically Categorize Research Methods for Research Synthesis				○
Extracting Moral Foundations From Text Narratives: Relevance, Challenges, and Solutions				○
Finding Voices in the Margins: Computer-Assisted Discovery of Naturally Belonging Names				○
Scaling Up Content Analysis				○
Using Text Mining to Measure Diffusion of Innovation				○

정치학에서는 정부나 의회의 기록과 관련된 분석이 그나마 잘 보이는 편이다. 특히 몬로, 콜라레시, 퀴인(Monroe, Colaresi, & Quinn, 2008)의 분석이 시각화에 대한 고민에 있어서는 매우 훌륭한데, 이들은 동일한 한 개 주제에 대해 미 공화당과 민주당 의원들이 발언이 어떻게 다른지를 비교했다. 텍스트 분석도 분석이지만 분석의 내용을 한 눈에 알아볼 수 있는 시각화(Visualization)가 매우 인상적이다. 〈그림 2〉를 보면 토픽내 단순 단어 빈도를 보여주는 것이지만, 공화당과 민주당 의원들의 단어 사용 차이가 어디서 나는지 베이지안 추정을 통해 매우 효과적으로 계산화하여 시각화함으로써, 각 당별 주요 단어 사용의 차이가 어떤 관점의 차이로 나타나는지 그리고 그러한 관점이 무엇인지 확연히 드러나도록 한 점이 돋보인다.

그림 2. 낙태 문제를 둘러싼 공화당과 민주당 상원의원들의 발언 단어 (Monroe, Colaresi, & Quinn, 2008)



흥미롭게도 시민들의 소통 기록에 대한 텍스트 분석은 많지 않은데 그 이유는 무엇보다 분석할 대상으로서 시민의 범위를 한정짓기 어렵기 때문이기도 하거니와, (공개 데이터만 분석할 수 있는 등) 불완전한 데이터가 가진 한계 때문으로 생각된다. 예를 들어, 어떤 현상에 대한 시민들의 발언을 수집하려면 트위터와 같은 경우 손쉽게 해쉬태그를 이용할 수 있지만, 페이스북과 같은 경우는 API가 제공하는 내용이 오로지 공개된 팬페이지의 포스팅만 이용할 수 있기 때문에, 사용자들이 공개적으로 올린 내용을 특정한 검색어를 기반으로 수집하기 어렵다. 언어상의 차이 역시 흥미로운 지점이다. 트위터에서 해쉬태그를 사용한다 하더라도 140자면 (띄어쓰기없이) 성명서 수준의 발언을 할 수 있는 한글과 달리, 다른

언어 문자의 경우는 트위터가 140자라는 제한을 풀기 이전까지 발언할 수 있는 내용이 한정적일 수 밖에 없었다. 반면, 한글은 많은 내용을 담을 수 있었어도 R 프로그램과 같이 텍스트 마이닝과 분석을 동시에 수행할 수 있는 분석 프로그램의 한글 사전이나 형태소분석기 모듈의 효용성이 영문보다 뛰어나지 않은 난점과 또한 기술적 장벽을 가지고 있다. 따라서 해쉬태그를 사용하여 자료를 수집하더라도 대개 텍스트 분석이 아닌 오히려 연결망 분석이 주가 된 경우가 많다.

(2) 연결망: 참여, 사회운동 - 대규모 소셜 미디어 연결망 또는 시뮬레이션과 결합

행위자간 관계를 설정할 수 있다면 연결망으로 분석할 수 있다. 사람간의 관계 뿐만 아니라 텍스트 단어간 관계도 연결망 분석이 가능하다. 그러나 사람간 연결망으로 한정지어 본다면, 실시간 자료 수집을 통한 사회운동의 전파 또는 시뮬레이션과 결합한 전파 원리 실험이 현재로서는 커뮤니케이션 및 컴퓨터연산 사회과학과 관련된 연결망 분석의 큰 줄기를 이루고 있다. 사람간 연결망 분석은 지금까지 대부분 트위터를 사용하여 왔는데, 이는 무엇보다 연결관계를 수집하는 자료 접근성의 문제로 볼 수 있다. 과거에는 트위터에서 화이트 리스트라 불리는 API 개방이 이루어 졌고, 현재는 트위터가 대규모 자료를 판매하고 있기 때문이다. 자료를 구매하지 않는 이상 데이터 수집을 위해 API를 우회하는 다양한 방식을 시도해야 하며, 아마존 AWS와 같은 클라우드 서비스와 같은 플랫폼을 사용하는 방식 등이 필요하다. 그러나 이러한 장벽을 넘어 병렬처리 클라우드 서비스를 활용할려면 하둡(Hadoop)과 같은 데이터베이스 관리시스템을 쓸 수 있어야 하거나 자료 수집이후 대용량 데이터에 대한 전처리(Wrangling, Parsing) 과정을 거쳐야 하는 등 여러 가지 기술이 필요하다. 다나 보이드(Dana Boyd)와 케이트 크라포드(Kate Crawford) 등의 연구자들이 페이스북이나 구글과 같은 기업과 관련을 맺어야만 데이터에 접근할 수 있다는 데이터 접근의 불평등성과 동시에 컴퓨터 작업 능력의 격차로 인해 나

타나는 불평등을 언급한 이유가 바로 여기에 있다(2012).

다시 돌아와서, 시뮬레이션과 연결망 분석이 결합된 CSS는 매우 정밀한 형태의 모형화를 통해 이론화 작업이 이루어지고 있다. 투표 참여에 대한 개인 동기의 복잡성을 ABM으로 모형화한 필드하우스, 레사드, 에드먼드(Fieldhouse, Lessard-Phillips, & Edmonds, 2016) 연구를 비롯하여, 사회운동에 참여하는 개인의 시점을 개인 주변에 있는 주변인들이 개인에게 미치는 영향이 각각 달라질 때뿐만 아니라 확산에 있어서도 연결망 형태(Network Topology)에 따라 어떻게 달라지는지 매우 정밀하게 사고하여 이론화 작업을 시도하는 등의 작업(Piddrahita, Borge-Holthoefer, Moreno, & Gonzàlez-Bailó, 2017)이 나타나고 있는데 이들 연구들이 출판된 저널에서도 알 수 있듯이 커뮤니케이션학에서는 상대적으로 이와 관련된 발전이 더딘 편이다.

2) 전망과 동향 2: 정보에서 소통으로, 태도에서 행동으로

각 방법론의 구체적인 측면과는 별개로, 전체적인 연구 방향이 어떻게 변화하여 왔는지를 이야기하여 보면 1) 연구의 무게중심이 정보에서 소통으로 이동하고 있으며, 또한 2) 개인들의 태도나 의견보다는 행동이나 행위에 더욱 관심이 증대할 것이라는 두 가지 전망으로 요약할 수 있다. 이 전망은 해당 연구에 관심을 가진 사람들이 공통적으로 체감하는 사항이기는 하나, 무작정 이야기를 던지는 것만으로는 불충분할 것이기에, 이러한 주장에 대한 간단한 증거로 저널에 실린 논문의 제목들이 어떻게 변화하여 왔는지만 가볍게 살펴보기로 본다. 이는 과거의 사건이 있고 현재의 사건이 있으면, 과거의 사건으로부터 현재의 사건으로 선을 이을 경우 하나의 흐름이 된다는 점에 착안한 것이다.

〈표 2〉의 왼편은 New Media & Society 학술지를 특별호(Speical Issue)를 제외하고 2004년, 2011년, 2017년 각 년도의 논문 제목들을 모아 어떤 단어들이 가장 많이 사용되었나를 단어 어근으로 축

약 수집하여 표기한 것이다. 논문이 많지 않고, 초록과 같은 보다 검증된 형태의 내용을 사용한 것도 아니며 매우 간단하게 보기 위한 것임에도 불구하고, 2017년을 보면 소셜 미디어(social media)의 상위권 점유, data라는 단어의 등장, Facebook, Twitter 그리고 social network site로 쓰인 SNS 등을 볼 수 있다. 반면에 사라진 단어들은 Divide, Interact 등의 단어인 것을 볼 수 있다. 아래에 있는 정치커뮤니케이션 학술지의 변화는 더 극명하다. New Media & Society에 비해 연간 발행 논문의 수가 적어 2개 년도를 합산했는데, 짧은 기간 동안 몇 안되는 논문에서도 Information 이라는 단어의 상대적 순위가 줄어들었고, Coverage는 2016-2017년에서는 한번도 언급되지 않았다. 없어진 것으로는 Knowledge, Television 등이고 새로 생긴 단어로는 Social Media가 출현하였다. Participation과 Partisan의 단어들이 2016-17년을 차지하고 있다면 Civic, Citizenship, Community 등은 2003-2004년의 단어들이다. 소셜미디어가 하나의 유행이라고 볼 수도 있지만, 커뮤니케이션의 다양한 효과들이 나타나는 연구 대상 플랫폼이고 이 플랫폼에서 다양한 미디어 효과가 관심사라는 점, 결국 정보 그 자체보다 정보의 '흐름'이라는 점에 주목해야 한다고 본다.

변화에 대한 가벼운 접근이지만, Journal of Information Technology and Politics와 Int'l Journal of Press/Politics에 대한 2016-2017년 논문 제목들에 대한 단어 빈도와 Journal of Communication 및 Information, Communication, & Society 저널에 대한 논문 제목 단어 빈도를 부가적으로 더 만들어 보았다(〈표 3〉). 한 가지 분명한 경향은 모두가 Social Media의 강조가 부각되어 나타난다는 점이다. 사실 그보다 주목하는 것은 중요도가 떨어지고 있는 단어인데, 역시 Information의 상대적으로 낮은 위치 그리고 selective 보다 더 낮은 message 등이 눈에 띈다. 표 3와 표 4에서 보듯, 정치 커뮤니케이션과 같은 세부 주제로 옮겨지면 변화가 더 잘 드러나는 경향이 있다.

표 2. 해외 일부 학술지들의 논문 제목 단어 빈도 (1): 2003-4, 2010-11, 2016-17년

New Media & Society					
2004	빈도	2011	빈도	2017	빈도
media	9	internet	12	digit	14
internet	7	digit	10	onlin	13
studi	7	media	9	media	11
onlin	6	onlin	8	socialmedia	8
polit	6	interact	7	public	7
communic	5	mobil	7	communic	6
life	4	social	7	ident	6
social	4	cultur	6	internet	6
web	4	game	6	polit	6
age	3	communiti	5	approach	5
develop	3	case	4	cultur	5
digit	3	model	4	data	5
interact	3	network	4	experi	5
tie	3	product	4	facebook	5
consumpt	2	twitter	4	network	5
divid	2	video	4	privaci	5
implic	2	young	4	site	5
influenc	2	campaign	3	socialnetwork	5
journal	2	capit	3	studi	5
mobil	2	children	3	youth	5
Political Communication					
2003-2004	빈도	2010-2011	빈도	2016-2017	빈도
polit	14	news	12	polit	26
news	11	polit	9	media	18
media	8	effect	8	effect	11
coverag	6	media	8	parti	10
elect	5	elect	6	news	8
inform	5	campaign	5	elect	6
citizen	2	campaign	5	vote	6
citizenship	2	coverag	4	digit	5
civic	2	candid	3	socialmedia	5
communiti	2	inform	3	campaign	4
cultur	2	network	3	compar	4
debat	2	opinion	3	experi	4
effect	2	strategi	3	particip	4
evalu	2	attitud	2	partisan	4
evid	2	choic	2	presidenti	4
foreign	2	communic	2	public	4
knowledg	2	congress	2	social	4
lang	2	experiment	2	voter	4
letter	2	interest	2	action	3
televis	2	internet	2	inform	3

표 3. 해외 일부 학술지 논문 제목 단어 빈도 (2) : 2016-17년

JITP & IJPP	빈도	JOC & ICS	빈도
polit	30	communic	29
media	21	news	20
elect	15	onlin	19
campaign	14	polit	19
news	11	socialmedia	19
effect	9	media	18
socialmedia	9	social	17
twitter	9	analysi	15
digit	8	digit	15
onlin	8	public	14
agenda	6	facebook	13
facebook	6	twitter	13
analysi	5	network	12
candid	5	studi	12
communic	5	populist	10
democraci	5	role	10
impact	5	select	10
interact	5	inform	9
press	5	internet	9
public	5	support	9
studi	5	campaign	8
chang	4	chang	8
data	4	content	8
network	4	mobil	8
particip	4	particip	8
inform	4	privaci	8

소통으로의 무게중심이동 및 태도보다는 행동으로의 이행에 대한 주장을 펼치는 다른 논의를 제시하자면, 무엇보다 정치 정보를 선택할 수 있는 다양성의 증가로 인해, 정보 효과를 특정하기 어려운 복잡성의 증가를 들 수 있다(예. Aelst et. al, 2017). 국내로 예를 들면, 연령에 따라 다를 것으로 예상되나 포털과 TV가 여전히 정치 정보 또는 뉴스를 보는

채널인 상황에서, 포털에서 무엇을 보는가 TV도 어떤 채널을 보는가에 따라 정보효과가 매우 다르다(예. 윤호영・이종혁・길우영, 2017), 또한 1차적인 뉴스 정보습득 경로뿐만 아니라 해당 뉴스에 대한 개인적인 종합이 여러 가지 다양한 경로를 통해 이루어지고 있는 시대이다. 즉, 다변화된 미디어 환경은 뉴스 노출에 대한 특정한 효과, 예를 들어 언론에 의한 틀짓기(Framing)효과가 있다면 그 효과의 원천을 단 하나의 정보 소스로 한정지을 수 없게 한다. 그런 의미에서 베넷과 아이엔거(Bennett & Iyengar)가 2008년 A New Era of Minimal Effect 논문에서 선택적 노출(Selective Exposure)의 영향력을 비판한 것을 이해할 수 있다. 이와 관련하여 쉐하타와 스톰(Shehata & Strömbäck, 2013)는 여전히 전통적인 미디어의 아젠다 셋팅 효과는 사라지지 않았으나, 다양한 온라인 뉴스의 활용은 이러한 효과를 감소시킨다고 보고하면서 아직은 미니멀 효과의 새 시대가 아니다("Not (Yet) a New Era of Minimal Effects") 라고 한 바 있다.

정보 홍수와 관련 하여 설득의 맥락을 잡아내기가 더 어려워지고 있는 가운데, CSS방법이 앞으로 활성화될 것이라는 의미는 정치 정보 습득, 지식 및 태도보다는 데이터와 증거가 남는 개인으로부터 집단, 집단간의 결합에 의한 결과 및 대규모 행동 등에 주목한다는 것을 의미한다. 기존 미디어 뉴스의 내용과 시민들의 소통 기록간의 비교, 뉴스가 퍼지는 흐름 추적과 같은 연구 내용들은 개개인의 태도 변화와 같은 미시적 효과와는 거리가 있다. 오히려 개인의 태도의 다양성을 인정한 상태에서 시작하여, 주변과의 상호작용을 통해 이러한 태도가 어떻게 변화하는가에 대한 것이 더욱 활성화될 것이라고 본다. 그런 의미에서 프레이밍과 같은 실험을 통한 전통적인 연구는 CSS와 길을 달리하게 된다. 논란을 불러일으켰던 페이스북 사용자 8백만명을 대상으로 했던 대규모 실험은, 본인 연결망을 타고 오는 정보가 미치는 영향이 '개인 행동에 의한 결과'에 의해 측정되었지, 태도에 대한 정보 데이터를 – 그 자체도 얻기 어렵지만

- 얻는 방식으로 이루어지지 않았다. 물론, 엔트만(Entman)과 같은 학자는 2016년 ICA 패널세미나에서 정보홍수 시대에 틀짓기라는 단어를 버리고, 아예 점화(priming)만 남았다고 선언하기도 했다.

3) 전망과 동향 3: 시너지와 갈라섬 사이

새로운 자료와 미디어 환경이 새로운 연구 방법을 동원한 흐름을 만들고 있다면, 기존 연구 방법은 기존에 해오던 것외에 다른 방식으로 CSS와 결합하여 성숙할 수 있을 것인가 역시 중요하게 다루어야 할 문제라 볼 수 있다.

우선, 분석 기술과 관련해서 통계분석과 CSS 관련 분석 방법은 서로 결합하여 시너지를 낼 수 있을 것으로 기대된다. 특히 기계학습을 포함한 데이터 마이닝 기법은 전통적인 통계 분석이전에 빠르게 모형을 추론하거나 통계 모델링을 통한 인과 관계 추론을 위한 사전 작업에 도움이 될 것이다. 콘과 라미레즈(Conn & Ramirez, 2017)는 변수들간 상관관계가 서로 중첩되어 있는 경우 그리고 고차원의 상호작용이 작용하는 경우에 대한 모델링 기법으로 랜덤 포레스트 분류 분석 방법을 개량한 퍼지 포레스트(Fuzzy Forest) 방법을 보여주고 있다. 랜덤 포레스트는 기본적으로 의사 결정 나무를 활용한 방식인데, 한 개의 노드가 가지치기 하듯이 그 다음 나무로 갈라지고, 또다시 여러 속성에 따라 갈라지기를 반복하면 가지치기가 반복되는 방식을 통해 복잡한 문제를 계층 구조로 하나씩 풀어가며 예측치를 만들어내고 이에 따라 데이터를 분류하는 방법을 말한다 (참고로, 랜덤 포레스트는 트리를 활용한 앙상블 학습 방법 - 단일 모형이 아니라 여러가지 다양한 모델을 통해 자료를 학습 - 을 활용하고, 과적합(overfitting)문제를 해결하기 위해서 제시된 방법이다).

콘과 라미레즈(2017)는 전통적인 통계 분석 방법을 동원한 SAS 통계 패키지로는 9.3시간 동안 걸린 작업을 퍼지 포레스트 방법을 통해 11.1초만에 해내었다고 보고하였다. 185개의 잠재적인 예측 변수를 30

개의 중요 변수로 줄이고 각 변수들의 중요성을 매우 빠르게 계산하였다. 이들의 글은 이러한 작업 효율성에 관해 이야기하고 사회과학에 이러한 분석 방법이 어떻게 쓰일까에 대해 이야기하는 예로써, 이들 분석 방법의 효율성에 맞추어져 있지만 더욱 중요한 것은 이러한 방식이 다양한 변수들이 섞여 있고 변수간 상호작용이 매우 복잡하게 얽혀 있어 고차원 방정식이 필요할 때를 대체하여 이를 사전에 풀어내는 작업으로서 더욱 기대가 된다고 할 것이다.

눈치챈 독자도 있겠으나, 지금껏 이야기 해 온 변화의 방향은 매우 사회학적이며, 심리학 기반의 이론이나 연구와는 거리가 있다. 그 이유는 기존 연구방법과 관련하여 심리학 기반의 이론이나 연구와 CSS의 적합성은 사회학기반 연구에 비해 상대적으로 제한적이라고 볼 수 있기 때문이다. 물론, 심리학적인 내용을 통계적으로 분석하는 데 있어서는 앞서 랜덤 포레스트와 같은 기술이 사용될 수 있으나, ABM을 제외한 데이터 기반 CSS가 심리학 기반 연구들과 어떻게 만날 수 있는 지에 대해서는 더욱 고민해 볼 필요가 있다. 최근 출판된 컴퓨터연산 사회 심리학(Computational Social Psychology)라는 책을 보면 수많은 수식과 ABM만이 언급되고 있을 뿐, 그 외의 구체적인 연구 결과나 사례를 찾기 어렵다(Vallacher, Read, & Nowak, 2017). 정치 커뮤니케이션으로 연결 짓는다면, 태도 변화와 설득 관련 연구와 관련하여 어떤 데이터 생성을 이야기할 수 있는지, 그리고 그러한 데이터 생성이 변화하는 커뮤니케이션 지형을 잘 담아낼 수 있는지 등 여러가지 문제가 제기될 수 있다. 예를 들어, 텍스트 분석을 통해 특정한 커뮤니티내 메시지를 분석할 수는 있지만, 이러한 메시지가 개인의 어떠한 심리적 동기 혹은 유인에 의해서 그 효과를 발휘하는지 CSS 방법으로 연구하기는 어렵다는 것이다. 오히려 메시지를 주고 받는 사람들간의 연결망을 그려서 메시지에 의해 특정한 행위(예를 들어, 투표)가 나타나지는 관찰하기가 상대적으로 쉽다. 따라서, 변화하는 미디어 지형 – 특히나 선택지가 높아지는 상황 - 의 영

향으로 미디어에 의한 설득 효과를 약화시킨다는 지적(Iyengar, 2017)과 더불어, 비단 CSS 뿐만 아니더라도 앞으로 심리학기반 연구들이 마주치게 될 연구는 무엇인지 생각해 보아야 할 필요가 있다. 그 과정에서 주의할 점은 심리적 영향과 심리적 요인을 구분할 필요가 있다는 점이다. 예시를 들자면, 우울한 메시지를 페이스북 타임라인에 더욱 자주 노출시켜 통제 집단과 비교하기 위해 텍스트 마이닝을 동원한 연구가 있었는데(Kramer, Guillory, & Hancock, 2014), 그 연구 결과인 일부 사용자들이 그러한 메시지 효과를 받아 상대적으로 우울한 감정에 대응하는 단어들을 많이 사용하였다는 내용은 연결망 효과를 보는 연구로 심리학적 연구라기보다는 오히려 이 글에서 말하는 오히려 사회학적 연구이다.

그런 점에서 기존의 연구 방법을 통해 개인 차원 흔히 미시적 차원이라 불리는 효과를 찾는 심리학적 연구들은 그러한 효과의 지평을 사회적 차원으로 확대하여, 개별 연구 질문에 대한 복합적인 차원을 고려하는 방식으로 진화될 가능성이 높아졌다고 보지만, 구체적인 그림이 잘 그려지지 않는 점이 있다. 실제로 〈표 4〉에 나타난 ICA에서 발표된 CSS 세션의 연결망 분석을 보아도 이 지점은 분명해진다. 연결망과 관련되어 심리학적 연구로 사료될 수 있는 연구를 찾기 보기 어렵다.

표 4. 2017년 CSS 세션에 발표된 연결망 분석 관련 논문제목

연결망 분석	방법론	실증연구
New Bottle of Old Wine vs. Old Bottle of New Wine: What Contributions Have Computational Methods Made to Social Media Research?	△	
Computational Approaches and Diffusion of Health Information		△
Studying the Political Communication of Opaque Regimes Through Computational Methods		△
Computational Approaches and Public Relations	△	
The Implications of Computational Methods for Comparative Studies	△	

연결망 분석	방법론	실증연구
Knowledge-Structuring Practices: Using Semantic Network Analysis to Assess the Influence of Design Team Structure on Knowledge Emergence		△
Are We Who We Follow? Analyzing Human Personality and Brand Following on Twitter		△
Correlation Between Linguistic Content and Social Links in an Online Network		△
Exploring the Optimized Network Structure for Discourse Quality: Evidence From Analyzing Reddit Politics		△
The Hidden Costs of Requiring Accounts Online: Quasiexperimental Evidence From Peer Production		△
The Cost of Communication Failure: Validating an Empirical Case With Individual-Based Modeling on MERS outbreak		△
Intervening "Just-in-Time": Supervised Machine Learning to Detect Recovery Problems in an Online Support Forum		△
It's Not Just About Mobility: How Can Mobility Datasets Advance our Understanding of Information Diffusion	△	
Recommendation With Spatio-Temporal Information	△	
A Multimethod Approach to Collecting and Understanding Mobile Log Data	△	
Analyzing Mobile Phone Data With Network Science		△
Stability in the World Wide Web: A Longitudinal Two-Mode Network Analysis		△
Can Big and Broad Data Motivate New Network Theories &Methods	△	
Network Methods for the Analysis of Online Data		△
Ontologically Aware Social Media Network Research: Concepts and Tools	△	
Trends, Tools, and Techniques in Network Visualization for the Social Sciences		△
Connecting Issues and Actors in Online Activist Fields	△	
Who Sets the Political Agenda? Applying the Network Agenda-Setting Model to Twitter		△
Hybrid Contagion: Theoretical and Practical Implications for Communication Campaigns	△	△
The Contagion Effects of Recurrent Activation in Social Networks		△
The Predictability of Communication Patterns in Spontaneous Teamwork	△	
Social Hypergraph Analysis: Towards an Operationalisation of Actor-Network Theory Using Hypergraphs	△	

마지막으로 생각해보는 한 가지 전망은 기존의 통계 분석 방법이나 CSS든 데이터 축적으로 인해 장기간의 관찰을 통한 분석은 각광을 받을 것이라 조심스럽게 예상해 본다. 세월호 사건이후 1년간 언론의 정치적 편향성을 연구한 박종희(2016)는 텍스트 분석과 베이지안 모델 추청을 통해서 언론의 보도가 어떠한 편향성을 가졌는지 장기간 추적하였다. 이러한 장기간 추이 변화에 관한 연구들은 지금까지 강조되어왔지만 앞으로도 더욱 강조될 것이다

4) 국내 연구 동향의 한 단면: 탐색과 모색

최근 데이터 시대를 맞이하는 커뮤니케이션 연구에 관한 책이 출간되고, 학회차원에서의 방법론에 대한 워크샵이 개최되고 있는 것과는 별개로, 실질적인 연구 동향은 국제적 흐름과 별개로 국내 상황에 맞추어져 있으며 아직은 탐색과 모색기간이라고 볼 수 있다.

〈표 5〉는 한국언론학회보에 실린 2011년과 2017년 논문 제목들을 앞서 사용했던 방법으로 다시 한번 정리한 것인데, 사용된 단어들이 거의 일치하는 가운데 2011년에 '스마트' 단어가 상위권에서 사라지고 2017년에 '이슈'라는 단어가 부상한 점과 '미디어'가 2011년에는 상위권으로 빈번히 쓰인 단어였지만, 2017년에는 밑으로 처진 점을 제외하고는 거의 비슷한 단어들이 반복적으로 2011년과 2017년을 채우고 있는 것을 알 수 있다.

표 5. 한국언론학보 논문 제목 단어 빈도: 2011년, 2017년

2011년		201년	
연구	40	중심	25
중심	35	연구	15
영향	25	영향	15
뉴스	23	뉴스	10
대한	18	대한	10

2011년		201년	
미디어	17	분석	10
이용	16	정치	9
보도	12	효과	9
분석	12	이슈	8
신문	12	보도	7
커뮤니케이션	12	언론	7
인식	11	이론	7
효과	11	인식	7
사회	10	변화	6
언론	9	커뮤니케이션	6
유형	9	방송	5
관련	8	사회	5
기사	8	유형	5
비교	8	정책	5
스마트	8	낙인	4
탐색	8	매개	4
관계	7	미디어	4
문화	7	요인	4
방송	7	이용	4
정치	7	정부	4

실제로 국내 CSS와 관련한 연구에서도 커뮤니케이션학의 국내 연구는 다른 학문 분야에 비해 많지 않다. 최근 부상하고 있는 텍스트 마이닝을 통한 토픽 모델링만 놓고 보더라도, 토픽 모델링을 사회과학 분야로 한정하여 DBpia에서 검색하여 보면, 48개의 연구 중에 3개만이 신문방송학 연구로, 본래 텍스트에 대한 분류가 주 관심사여서 그 양이 많을 것으로 예상되는 문헌정보학의 17개의 연구는 차치하고서라도, 경제/경영학의 14개에 비해 매우 적은 편이다. 3개의 연구들은 정부와 언론의 커뮤니케이션 전략 비교(이태준・이승배, ・오창동, 2017), 인터넷 게시판 분석(최선영・고은지, 2017), 미디어 언론 분석(김영욱・함승경・김영지・최지명, 2017) 등으로 정부, 언론, 시민 3개의 영역에 대해 각각 다

루고 있는 점이 흥미로운데, 모두 2017년에 출판되어 이제 탐색과 모색이 이루어지고 있음을 보여주고 있다.

지금까지 논의한 전반적인 국내외 동향과 전망을 정리하면, 변화하는 연구 지형속에 새로운 연구 방법을 결합시키는 방식으로의 새로운 시도들이 다각도로 이루어지고 있다고 볼 수 있다. 새로운 연구 방향으로 이동하는 실험들이 이어지고 있으며 새로운 연구 방법들이 이러한 시도에 적극적으로 결합되고 있는 형태이다. 따라서, 새로운 연구 방법들이 패러다임의 변화를 가져온 것은 아니더라도 새로운 연구 방향을 만들어 내고 있다고 보기에는 어려움이 없을 것이다.

4. 어떻게 할까?: 현실적인 문제, 생각해 볼 문제.

변화가 나타나고 있다면 무엇이 이러한 변화에 잘 대응하는 방법인가에 대한 고민 역시 필수적이다. 그리고, 변화하는 지형속에 CSS과 관련된 현실적 문제는 무엇보다 CSS를 수행하기 위해 필요한 세 가지 영역과 관련될 수 밖에 없다. 통계 방법론, 컴퓨터 프로그래밍 기술, 그리고 도메인에 관련된 실질적 지식이다. 이 세 가지는 데이터 사이언티스트가 가져야 할 3가지 능력과 정확히 일치한다. 도메인에 관련된 지식과 통계 방법론은 양적 방법론을 활용하고 있는 연구자라면 대개 갖추고 있다고 전제할 수 있기 때문에, 관심은 프로그래밍 능력에 쏠릴 수 밖에 없다. 컴퓨터 과학자들과 협업도 사실은 프로그래밍 능력을 활용하기 위함이고, 컴퓨터 과학자들은 사회과학자들로부터 혜안(insight)를 얻고자 하는 협업하는 것은 주지의 사실이다. 그렇다면, 당연한 질문이 따라온다. 앞으로 연구와 교육에 있어서 어떤 것이 필요할까? 예를 들어, 데이터기반 분석이라면 데이터는 어디서 구하나? 와 같은 질문이다.

1) 연구와 교육

우선 통계 분석 및 CSS와 관련된 컴퓨터 프로그램부터 살펴보자. 기존에는 분석 모형에 따라 통계 분석 프로그램의 특징을 잘 활용하는 연구자들이 각광을 받아 왔다. 예를 들어, 패널 데이터 분석이나 시계열은 STATA가 좋지만, 패널 데이터의 자료 변형이나 방대한 데이터에서 필요 데이터만 변형하여 추출하는데에는 SQL 명령어를 지원하는 등 SAS가 월등하기에 SAS를 쓰는 식이다. 하지만 R은 이 모든 걸 하나로 통합시켰다. 그리고 R은 통계 분석뿐만 아니라, 웹크롤링, 네트워크 분석, 토픽 모델링을 비롯한 텍스트 분석 등도 할 수 있다. 그런 의미에서 분석과 관련하여서는 R 혹은 더 나아가 파이썬과 같은 하나의 프로그래밍 언어로 모든 것을 다 할 수 있는 컴퓨터 프로그램이 각광을 받을 것이다. 물론, 기존 통계 패키지 회사들도 느리긴 하나 혁신을 거듭하여 여러 가지 새로운 분석 패키지 라이브러리를 늘리고 있기는 하다. STATA를 예로 들면, 발빠르게 랜덤 포레스트를 활용할 수 있도록 하고 있다.

그러나 이러한 컴퓨터 프로그램과 관련된 보다 근본적인 문제는 대규모 데이터의 수집과 같은 활동에서 프로그래밍 언어 지식에 의해 연구가 제약당하거나 혹은 R에서 나타나는 오류들을 찾아서 해결하는 능력, 특별한 책이나 강의가 없어도 메뉴얼만으로 원하는 것을 구현해내는 능력 등 깊이가 더해질 때 나타나는 것들이다. 이러한 것들은 축적의 시간을 거쳐 경험으로 만들어지는 것이고, 앞으로 코딩 교육을 통해 익숙해진 학생들이 나타나면 괜찮아지겠지만 그 과도기에 속한 학생들 지금부터 약 4-5년 정도의 학생들에게는 쉽지 않은 일이다. 그런데, 문제는 이 과정에서 학습 진도(Learning Curve)가 연속적인 함수가 아니라 (비단 통계나 프로그래밍만의 일은 아니겠으나) 계단식으로, 일정정도 시간을 투자하여 앞이 안 보이다 어느 순간 도약하고 그러다가 어느순간 또 도약하는 순간이 온다는 것이다. 여기까지 이끌어 줄 수 있는 교육 인력의 확보가 CSS와 관련된 커뮤니케이션학의 과제가 될 수 있다.

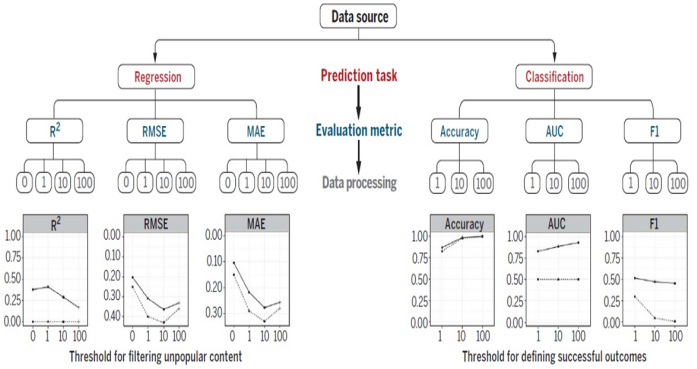
연구와 관련하여서 시선을 돌리면, 컴퓨터 과학자들이나 자료를 수집하여 던져줄 수 있는 연구자들과 협업을 하면 사실 특별한 어려움이 없다고 볼 수는 있다. 프로그래밍의 원리나 논리를 알고 있지 못하면 협업하면서 원하는 자료를 조율하는데 상당한 노력을 쏟아야 하기는 하지만, 자료를 얻어서 분석하면 되니 연구를 하는 것 자체에는 크게 문제가 되지 않을 것이라 볼 수 있다. 그럼에도 불구하고 본인이 자율적으로 조율할 수 없다는 측면에서 제약당하는 어려움, 프로그래밍 지식에 의해 무엇이 가능하고 불가능하고 또는 시간이 얼마나 걸리는 지에 대한 가늠 등 깊이의 문제가 다시 제기될 것이다.

프로그래밍 능력과 더불어 분석 방법 사용에 있어서의 편차가 CSS 분야에서는 앞으로 나타날 것이라고 본다. 빅 데이터는 단순히 사이즈가 커져서 빅데이터가 아니다. 예를 들어, 1500명의 설문으로 전체 인구 1인당 알콜 소비량을 알아내는 것과 3000만명을 설문조사해서 1인당 소비량을 알아내는 것에서 큰 차이가 없다. 이미 통계적 방법으로 1500명 정도면 충분하다는 것이 밝혀졌기 때문이다. 빅데이터가 빅인 이유는 데이터의 복잡성이 커지기 때문이다. 즉, 단순히 사이즈가 커져서 빅 데이터가 아니라 데이터의 복잡성이 커져서 여러가지 복합적인 것들을 동시적으로 고려하여 판단할 수 있기 때문이다. 그런 의미에서 여기에서도 분석의 깊이를 어디까지 가져갈 것인가가 연구의 핵심적 요소가 될 것이다.

분석 방법에서의 편차는 분석 검증의 문제와도 맞닿아 있다. 통계는 연구자가 연구 문제를 해결하기 위해 여러가지 분석방법 중에서 가장 적절한 방법을 선택하지만 각 모형에 대한 검증이 알려진 수치에 의해 이루어질 수 있다. 그러나 앞서 디마지오가 놀랐듯이, 흔한 적합도 검사 하나가 없는 상황에서 어떤 모형이 더 적합할 것인가에 대한 선택이나 검증이 오히려 연구자에게 맡겨지는 상황이 될 수 있다. 또한 컴퓨터를 통한 객관화의 시도가 오히려 더 주관적이 될 수 있다. 아래 그래프는 연구자가 어떤 분석방법을 선택하냐에 따라 설명력을 주장하는 정도의 차이를 보

여주고 있는데, 맨 왼쪽의 R 제곱에 기반하여 설명력을 주장하면 약 23%의 설명력을 갖지만 동일한 문제에 대해 AUC를 활용하면 거의 완벽에 가깝게 예측했다고 이야기할 수도 있는 상황을 보여주고 있다.

그림 3. Figure 1 in 호프만, 사라마, 와츠(Hofman, Sharama, & Watts, 2017)



2) 데이터 문제

데이터와 관련되어 공개되어 있거나 웹에서 수집할 수 있는 데이터는 크게 문제가 되지 않을 것으로 보인다. 페이스북과 트위터에 모든 연구 역량을 집중할 것이 아니라면 수집할 수 있는 데이터는 생각외로 많고 할 수 있는 것도 무궁무진할 것으로 예상된다. 다만, 우리나라에서 데이터 문제는 기업의 데이터나 공개된 데이터가 아니라 비공개되어 소수가 들고 있거나 수집해 온 데이터의 문제에 있다고 생각한다. 우리나라의 경우, 기업이 여러가지 이유로 내부적 사용외에 외부에 데이터를 공개하지 않으니 접근성의 불평등이 아니라 데이터 자체의 부재가 문제다. 예를 들어, 필자는 예전에 네이버와 다음에 문의하여 년간 인터넷 카페의 증가수치를 알려달라고 요청한 적이 있다. 연간 인터넷 카페의 증가치는 영업비밀이라며 비공개하는 상황에서 연구자가 해당 데이터를 다른 경로로 얻기는 어려웠고, 해당 수치는 연구에서 제외되었다.

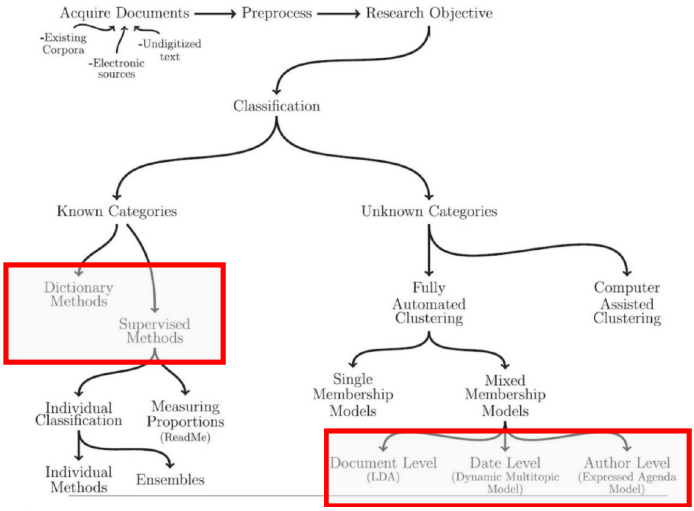
이와 더불어서, 데이터가 기자들에게는 유통되는 가운데 연구자들에게는 유통되지 않는 상황도 생각해 볼 거리다. 필자는 국정원 댓글 분석을 위해서 '원장님 말씀' 원본을 얻고자 이러저리 뛰어다닌 적이 있다. 당시 모 국회의원에게 자료를 요청하려고 전화도 하고, 페북으로 연락하기도 했으나 답변은 못 들었다. 또한 법원과 검찰청으로 가서 공판 기록을 통해 알아내고자 방문을 했으나 파기환송으로 인해 구할 수 없었다. 하지만, 정작 해당 자료들은 기자들의 경우 이미 들고 있었으며, 필자가 문의하였을 때는 기자가 분석해서 기사로 활용하고, 해당 자료를 더 이상 활용하지 않음에도 기사를 위해 국회의원에게 전달받은 것이기 때문에 전달할 수 없다는 답변을 들었다 (이러한 자료들이 전문 연구자가 아닌 기자들에 의해서만 활용되는 것도 생각해 볼 문제이다).

텍스트 분석에 대한 다양한 기법들을 언급하고 있지만 표본에 대한 모수가 특정될 수있는 언론기사에 대한 분석으로 편향될 수 있는 가능성도 생각해보아야 할 문제이며, 전 국민이 뉴스를 본다고 여겨지는 포털의 영향력이 감퇴되면 모두가 보는 언론 기사라는 연구 자료의 대표성을 어떻게 확보할 것인가도 문제가 될 수 있다. 포털의 뉴스 집중 및 공정성 문제를 넘어서서 데이터 사이언스 입장에서는 이러한 분산되는 것이 오히려 재앙적인 상황이 될 수도 있다.

3) CSS 연구 방법의 한계

한편으로 이 글에서 제시하고 있는 CSS 연구 방법이 가진 한계 역시 충분히 인식해야 할 필요가 있다. 가장 먼저 지적할 수 있는 점이 컴퓨터연산을 한다고 해서 사람의 판단이 완전히 배제되는 것이 아니라는 점이다. 〈그림 4〉는 텍스트를 데이터로 처리하여 분석하는 방법을 그리머와 스테와트(Grimmer & Stewart, 2015)가 정리한 그림인데, 해당 그림에서 원문에 표기된 것은 아니나 네모로 상자를 그린 부분이 사람에 의한 판단이 필요한 부분이다.

그림 4. 데이터로서 텍스트 방법에 관한 개요(Figure 1, Grimmer & Stewart, 2013)



즉, 텍스트를 분류하기 위하여 기존에 만들어진 사전에 기반하는 방식(왼쪽 박스), 혹은 완전 자동화된 형태로 모델링(오른쪽 박스)을 한다고 하면, 사전을 사람이 사전에 만들어주거나 또는 해당 토픽이 어떤 내용을 담고 있는지는 사람이 보고 정리해야 하여 설명해야 한다.

두 번째 한계는 앞서 디마지오가 지적한 사회과학자와 컴퓨터 공학자의 차이에서도 언급된 바 있는데, 컴퓨터가 알고리즘에 의해 내리는 판단(예. 기계학습, 강화학습 등)을 사람이 알 수 없는 블랙박스의 문제도 있다. 예를 들어, 신경망 알고리즘을 활용하는 인공지능 알고리즘의 경우 데이터 입력과 결과 출력 사이에 숨겨진 레이어(Hidden Layers)를 사용하는데 이들 레이어내에서 가중치 산정이 어떠한 방식으로 최적화되어 결과가 도출되는지에 대해 인간이 이해하기 어려운 측면이 있다. 최근의 예로, 알파고와의 대결에서 신의 한수라 불린 이세돌과 알파고의 제 4대국 78수의 경우 구글 딥마인드팀은 해당 수가 인간이라면 십만명중에 7명이 선택할 확률이라는 것을 파악하며 말그대로 신의 한수라 놀랐지만

(Kohs, 2017) 실제로 왜 십 만명 중에 7명이 둘 확률인지, 그러한 확률이 어떻게 산출된 것인지 설명하기는 어려운 것과 마찬가지이다. 세 번째 한계는 기술적 한계이다. 특히 R과 같은 프로그램 패키지의 자연어처리 모듈의 기술적 문제는 그 자체로도 따로 논의되어야 할 것으로 보인다. 간단한 예를 〈표 6〉에서 들고 있는데. 〈표 6〉은 통상 R프로그램 텍스트분석에서 자주 사용되는 KoNLP 자연어처리 패키지를 활용하여 명사를 추출한 결과와 국립언어원에서 세종 프로젝트의 일환으로 2000년대 초반 개발한 형태소 분석기와 색인 사전의 결과를 나열한 것으로, 비교를 위해 차이나는 부분만 강조하였다. 기본으로 설정된 세종사전을 R 패키지에서 사용하였기 때문에, 두 분석 결과에 사전으로 인한 차이는 존재하지 않는다. 그럼에도 불구하고 R 분석 결과를 보면, 한 글자로 단어에 대한 잘못된 형태소 분석 및 복합어에 대한 태깅 오류 등으로 인해 같은 단어라 하더라도 나타나는 빈도의 차이를 볼 수 있다. 예를 들어, '정치'라는 단어에 붙은 괄호가 - 불용어처리를 활용했음에도 여전히 - 남아 있어서 정치라는 단어의 빈도에서 차이가 나는 것을 볼 수 있으며, '갈등' 단어에서도 갈등구조에 대한 복합어를 따로 구분하여 '갈등'이라는 단어와 '구조'라는 단어로 분류해 내지 못한 것 역시 나타나 있다. 물론 연구자들은 이러한 오류를 최소화하기 위하여 다양한 방식을 취하고 있지만, 실제로 이러한 오류를 제거하기 위해 어떤 처리를 하였는지 서술하고 있는 연구를 찾아보기 어렵다.

표 6. 한글 형태소 분석 비교: 세종 프로젝트와 R의 KoNLP 패키지

세종 프로젝트 형태소 분석기+색인		R 프로그램 KoNLP Package	
중심	25	중심	25
연구	15	**적**	19
영향	15	연구	16
뉴스	10	영향	15

세종 프로젝트 형태소 분석기+색인		R 프로그램 KoNLP Package	
분석	10	분석	10
정치	9	뉴스	9
효과	9	효과	8
이슈	8	이슈	7
		정치	5
		c(정치)	4
갈등	2	**갈등**	1
거리감	2	**갈등구조**	1
검열	2	**개**	1
검증	2	개발	1
경제	2	**만**	1
공영	2	만족감	1

기술적 난점은 이뿐만이 아니다. 텍스트 분석에서 자주 사용하는 R의 tm 패키지의 DocumentTermMatrix 함수는 한글 인코딩이 UTF-8이나 Unicode로 되어 있을 경우 한글표기의 인코딩을 풀어서 있어서 리눅스 운영체제가 아닌 윈도우 운영체제에서 이 함수만 거치면 한글이 깨진 결과가 나오기 일쑤이다. 즉, 한글이 영문이나 알파벳기반 영어의 인코딩으로 넘어가서, 분석 내용 한글이 깨져서 나와 개개 컴퓨터 특성에 맞게 인코딩 설정을 잡아줘야 하는 등의 연구 및 교육적 측면의 어려움이 존재한다.

5. 변화하는 시대: 우리의 연구는 숙성될까, 새 부대에 담길까?

지금까지 CSS 방법에 대해 개괄하고, 이와 관련된 동향과 전망 그리고 몇 가지 생각해 볼 지점들에 대해 상술하였다. 사실 모든 연구 방법은 연구자가 연구하는 방향과 내용에 맞추어 나갈 수 밖에 없다. 우문이었으나 통계 공부를 어떤 순서로 하면 좋은가 하는 질문을 던진 적이 있었는데,

당시 들은 답변은 어떤 연구를 하느냐에 따라 필요한 연구 방법이 있고, 그에 따라 특화되어 나갈 것이기 때문에 자연스럽게 알게 된다는 것이었다. 그리고 누구나가 그렇겠지만 실제로 어떤 방법이 필요하면 그것들을 하나씩 익혀가며 연구하게 된다.

그런 점에서 보자면, CSS가 대용량 데이터의 수집과 분석이 가능해짐으로 각광을 받고 있지만 CSS 방법론 자체는 대용량 데이터 그 자체보다는 양질의 데이터가 필요한 분석 방법이라는 점에 주의할 필요가 있다. '우리에게 필요한 것은 빅 데이터(Big Data)가 아니라 굳 데이터(Good Data)'라는 말이 가진 함의에 다시 주목해야 하며 컴퓨터 연산이 모든 문제를 해결하는 것도 아니라는 사실 역시 상기해야 한다.

하지만, CSS와 같은 연구 방법은 변화하는 연구 지형이 가져온 연구 관심사와 관련이 있기 때문에, CSS와 같은 연구방법을 통해 우리의 연구 문제, 풀어내고자 하는 연구 질문을 숙성시킬 필요가 존재한다. 만약 우리가 질문을 오래된 것과 새로운 것으로 나누고 방법을 오래된 것과 새 것으로 구분하는 2 x 2 표를 만들어 보았다면, 오래된 질문에 대한 오래된 방법의 사용은 아무도 관심을 갖지 않을 것이라 짐작할 수 있다. 그리고 오래된 질문에 대해 새로운 방법론을 적용할 경우 그 자체는 새로울 수 있으나 재현(Replication)의 문제가 아니거나 효과의 크기 문제가 매우 심각한 논의의 대상이 아닌 경우라면 새로운 방법론이 가져오는 혁신은 완전히 새로운 발견에 비해 크게 주목받지 못할 것이라는 것도 생각할 수 있다. 예를 들어, 불평등과 같은 문제는 언제나 불평등 존재의 문제가 아니라 정도의 문제이기 때문에 주목받기 어렵다(Lieberson, 1987). 그렇다면, 결국 우리가 추구하는 것은 새로운 문제에 대한 오래된 방법론의 적용 혹은 새로운 문제에 대한 새로운 방법론의 적용으로 요약된다.

CSS가 각광을 받은 이유는 무엇보다 대용량의 데이터를 기존의 방법으로 다루기기 힘들기 때문에, 새로운 변화에 걸맞는 새로운 이론이 필요하기 때문이다. 이는 CSS와 관련된 변화, 새로운 방법론의 적용 문제

가 아이디어 문제이기도 하지만 데이터 환경이 변화하면서 만들어내는 복잡성의 문제라는 점을 시사한다. 필자는 커뮤니케이션 연구로 넘어온 지 얼마 안되어서 미디어 사용이 정치 참여를 이끄나와 같은 연구 문제의 경우, 미디어가 케이블 TV, 브로드 밴드, 모바일, 소셜 미디어 등의 순서로 바뀌기만 하고 동일한 연구 프레임이 지속적으로 반복되는 것을 알게 되었다. 그리고, 이런 문제를 연구한 논문들이 대개 다중회귀분석만을 활용하며, 인과관계를 증명하기 위해서는 좀 더 나아간 구조방정식(Structural Equation Model)의 활용으로 귀결되는 패턴을 발견했다. 그리고 이러한 이유가 무엇일지 생각해 본적이 있는데, 아마도 그 이유는 연구 문제가 복잡한 연구 방법을 필요로 하지 않았기 때문이라고 본다. 미디어를 사용했나 안했나 여부가 중요한 것이지, 여러 미디어를 많이 사용했나, 많이 사용했어도 다양하게 사용했나, 다양하게 사용한 것과 다양하지 않은 것은 어떠한가 등의 문제들을 한 모형에서 연구하지 않았기 때문에, 복잡한 연구 방법이 필요하지 않았던 것이라 생각한다. 하지만 대용량의 실제 사용자들에 의한 데이터 접근이 가능해지면서 기존의 연구 질문이 비해 매우 복잡한 문제 설정이 가능해지고 있고 그런 점에서 CSS 방법이 주목을 받는 것이라 생각한다.

따라서, CSS와 같은 새로운 방법론을 어떻게 활용할 것인가의 문제는 무엇보다 연구 질문의 복잡성 문제라고 결론을 내릴 수 있을 것 같다. 복잡성이 필요한 질문 또는 그 복잡성이 대규모의 디지털화된 자료와 조우하지 않을 경우, 우리는 항상 해오던 것을 분석 대상만 바꾸며 반복적으로 하고 있게 될 지도 모르며, 항상 새로운 분석 대상을 찾아 같은 질문을 반복적으로 던지고 있을 수 밖에 없을 것이다. 새로운 연구방법을 맞아 기존의 연구 방법을 숙성시킬 것인가 아니면 새로운 연구 방법을 새 부대에 담을 것인가? 생각해보면 숙성시키는 것과 새 부대에 담는 것은 연속적인 행위이다. 새 부대에 담았다가 숙성시키고 이를 꺼내서 음미한다. 변화하는 현재의 지형에서 우리가 새 부대에 담는 것들도 숙성되어야

할 것이고 이들을 숙성시키는 기술은 우리가 현재 가지고 있는 우리의 숙성된 기술이다. 이 둘의 조화는 당연히 필요할 것이고, 이 둘 사이의 끊임없는 긴장 관계를 위해서 지속적인 배움과 개발이 필요한 건 언제나 우리의 숙제이다.

참고 문헌

김경모 (2005). 커뮤니케이션 연결망 분석의 이론적 기초에 관한 탐색적 접근. 〈커뮤니케이션 이론〉, 1권 2호, 162-207.

김영욱 · 함승경 · 김영지 · 최지명. (2017). 사회 쟁점에 대한 비판적 담론 분석. 〈커뮤니케이션 이론〉, 13권 4호, 40-91.

박종희 (2016). 세월호 참사 1년 동안의 언론보도를 통해 드러난 언론매체의 정치적 경도. 〈한국정치학보〉, 50권 1호, 239-269.

송현주 (2011). 사회 연결망을 통한 정치 커뮤니케이션 연구의 필요성과 전망. 〈커뮤니케이션 이론〉, 7권 2호, 75-104.

윤호영 · 길우영 · 이종혁. (2017). 다중미디어 시대의 뉴스 미디어 레퍼토리와 주제 관심도 관계 분석. 〈한국방송학보〉, 31권 1호, 107-148.

이태준 · 이승배 · 오창동. (2017). 원자력 이슈에 대한 정부와 언론의 커뮤니케이션 전략 비교연구. 〈언론과학연구〉, 17(3), 172-229.

최선영 · 고은지. (2017). 미디어 멀티태스킹과 게이미피케이션을 통해 발현한 실시간 참여 민주주의. 한국방송학보, 31(3), 78-113.

황용석 · 이준웅 · 손동영 · 김상민 · 유경한 · 최세정 · 김익현 · 김대원 · 박령주 · 백영민 · 이상엽 · 성윤택. (2017). 〈데이터시대의 언론학 연구〉. 커뮤니케이션북스

Aelst, P. V., Strömbäck, J., Aalberg, T., Esser, F., Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T & Stanyer, J. (2017). Political communication in a high-choice media environment: a challenge for democracy? *Annals of the International Communication Association*, 41(1), 3-27.

Bennett, W. L., & Iyengar, S. (2008). A new era of minimal effects? The changing foundations of political communication.

Journal of Communication, 58(4), 707-731.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662-679.

Cappella, J. N. (2017). Vectors into the future of mass and interpersonal communication research: Big data, social media, and computational social science. *Human Communication Research*, 43(4), 545-558.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996), 1194-1197

Centola, D., & Macy, M. (2007). Complex contagions and the weakness of long ties1. *American Journal of Sociology*, 113(3), 702-734.

Conn, D., & Ramirez, C. M. (2017). Random Forests and Fuzzy Forests in Biomedical Research. In R. Michael Alvarez, (Eds.) *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press, 168-198

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2), 2053951715602908.

Epstein, J. M., Cummings, D. A., Chakravarty, S., Singha, R. M., & Burke, D. S. (2004). *Toward a containment strategy for smallpox bioterror: An individual-based computational approach*. Brookings Institution Press.

Fieldhouse, E., Lessard-Phillips, L., & Edmonds, B. (2016). Cascade or echo chamber? A complex agent-based simulation of voter turnout. *Party Politics*, 22(2), 241-256.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*. 4(5), 4-46.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267-297.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324), 486-488.

Iyengar, S. (2017). A Typology of Media Effects. In Kate Kenski & K. H. Jamieson, (Eds.), The *Oxford Handbook of Political Communication*. New York: Oxford University Press.

Kramer, A. D., Guillory, J. E., &Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790.

Kohs, G. (2017) AlphaGo Documentary [Movie]. New York, NY: Moxie Pictures..

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science*, 323(5915), 721.

Lieberson, S. (1987). *Making it Count: The Improvement of Social Research and Theory*. Oakland, CA: University of California Press.

Macy, M. W., & Willer, R. (2002). From factors to factors: computational sociology and agent-based modeling. *Annual Review of Sociology*, 28(1), 143-166.

Minfer, G., Delen, D., Elder, J., Fast, A., Nisbet, R. A.. The history of text mining. *Practical Text Mining and Statical Analysis for Non-structured Text data Applications.* Waltham, M.A: Elsevier, 3-28

Monge, P. R., & Contractor, N. S. (2003). *Theories of*

Communication Networks. Oxford University Press, USA.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin′words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis,* 16(4), 372-403.

Newman, M. E., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68(3), 036122.

King, G. (2017). Preface.. In R. Michael Alvarez, (Eds.) *Computational Social Science: Discovery and Prediction*. New York: Cambridge University Press,

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*.. 2nd Eds., University of Chicago Press.

Piedrahita, P., Borge-Holthoefer, J., Moreno, Y., & Gonzlezàlez-Bailó, S. (2017). The contagion effects of repreated activation in social networks. Social Networks}, In Press.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The Annals of the American Academy of Political and Social Science*, 659(1), 6-13.

Shehata, A., & Strömbäck, J. (2013). Not (yet) a new era of minimal effects: A study of agenda setting at the aggregate and individual levels. *The International Journal of Press/Politics*, 18(2), 234-255.

Tilly, C. & Tarrow, S. (2015). Contentious Politics. 2nd (Eds). Oxford University Press.

Watts, D. J. (2004). The "new" science of networks. *Annual Review of Sociology*, 30, 243-270.

Watts, D. J. (2013). Computational social science: Exciting

progress and future directions. *The Bridge on Frontiers of Engineering,* 43(4), 5-10.

Vallacher, R. R., Read, S. J., & Nowak, A. (Eds.). (2017). *Computational Social Psychology*. New York: Routledge.

투 고 일 자: 2018년 01월 10일

심 사 일 자: 2018년 01월 30일

게재확정일자: 2018년 02월 22일

Abstract

Communication Study in Flux

: Trend and Prospect of Computational Social Science

Yoon Ho Young
Senior Researcher, Communication Research Institute, Yonsei University

This paper discusses how Computational Social Science(CSS) intertwines with communication study. CSS is reviewed with three main areas: Agent-Based Model, Network Analysis, & Text mining. By discussing current development and prospect, it is elaborated that how the field of communication conjoins with CSS and what is its implications. Research and Educational aspect of CSS is also discussed.

KEYWORDS Computational Social Science, Data Science, Agent-Based Model, Network Analysis, Text Mining