빅데이터 동향 및 시사점

◉ 목 차 ◉

1. 서론
2. 빅데이터 정의와 동향
3. 빅데이터 기술
4. 빅데이터 2.0
5. 결론 및 시사점

1. 서 론

조지 오웰의 소설 '1984'에는 모든 사람을 감시하고 통제하는 존재인 '빅 브라더'가 등장한다. 빅 브라더는 사람들의 모든 대화와 위치를 기록할 수 있고, 이를 기반으로 사회를 통제, 권력을 유지한다. 영화 '다크 나이트'에서 주인공은 악당을 추적하기 위해 도시 내의 모든 휴대폰들의 음성 신호를 분석하여 시각화하는 장치를 이용하고, 미국의 드라마 '퍼슨 오브 인터레스트'에는 테러리스트를 추적하기 위해 CCTV들의 영상신호를 통합 분석하는 시스템이, '뉴스 룸'에는 휴대폰의 대화 내용을 도청하는 시스템이 주요하게 등장하기도 한다. 사실 소설 '1984'가 출간된 시점에서는 단순한 공상에 불과했을 이 거대한 데이터의 힘은 이제 현실이 되었다. 뉴욕시는 현재 테러 방지를 위해 극도로 정교한 통합 CCTV 분석 시스템을 운영하고 있고, 168년을 이어오던 미디어 '뉴스 오브 더 월드'는 런던 경찰의 시스템을 이용해 수천명의 휴대폰을 도청한 사건으로 2011년 폐간되었다. 이것이 우리가 이 거대한 데이터를 어떻게 다루어야 하는지 논의해야 하는 이유이다.

2013년 9월 20일에 출시된 애플의 신형 아이폰이 첫 3일간 900만대가 팔려나가는 등, 스마트폰의 확산 속도는 점점 더 빨라지고 있고, 이들 디바이스를 통해 생산되는 데이터의 양 역시 점점 증가하고 있다. IDC는 2011년 한 해 동안 생산된 데이터가 1.8ZB에 이른다고 발표하였고, 앞으로 급격하게 늘어날 것으로 전망된다. 이 막대한 양의 데이터를 단순히 저장하는 것만으로도 작지 않은 이슈이지만, 이를 효과적으로 활용하여 새로운 가치를 창출하는 것은 전혀 새로운 차원의 문제이다. 이처럼 거대한 양의 데이터를 다루는 일련의 과정을 통틀어 빅데이터라 하며, 분야를 막론하고 주요하게 다루어져야 하는 유망 기술로 떠오르고 있다. 맥킨지, 가트너, PWC 등 세계적인 컨설팅 기업들은 2011년부터 빅데이터를 유망 기술로 선정하고 이에 대한 보고서를 발표하며 다가오는 기회에 선제적으로 대응하고 있고, 2012년 다보스 포럼 역시

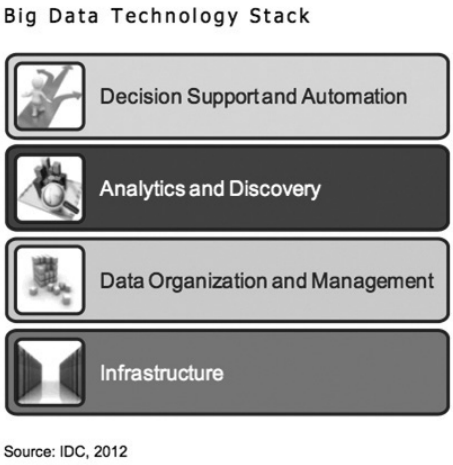
빅데이터를 주제로 선정하며 주요한 이슈로 다룬 바 있다.

하지만 현재의 연구들은 빅데이터의 구현에 필요한 기술적인 부분에 중점을 두고 있다. 따라서 본 연구에서는 빅데이터 연구의 결과가 실제로 널리 활용될 수 있는 방안을 탐색하려 한다.

2. 빅데이터(Big data)

2.1 빅데이터의 정의

새롭게 대두된 개념인 빅데이터를 정의하기 위한 노력들이 있었다. 빅데이터의 개념을 초창기에 제시한 매킨지는 빅데이터를, 그 큰 크기로 인해 기존의 데이터베이스 도구로는 다룰 수 없는 데이터의 모임으로 정의하였다[1]. 기존의 데이터베이스는 규격과 형태가 정해진 정형 데이터에 제한되었던 것에 반해, 빅데이터는 여기에 텍스트, 영상, 음성 등의 내용을 포함하는 비정형 데이터를 포괄한다는 의미이다. 때문에 LG경제연구원에서는 이 비정형 데이터의 활용에 주목하여, '생각'을 만드는 기술로 빅데이터를 특징지었다[2]. 사람들의 자연스러운 일상 대화에 내포된 감정이나 상태 등, 기존의 기술로는 활용이 불가능했던 데이터를 분석하여 컴퓨터의 단순한 반사작용이 아닌 '생각'에 의한 반응을 이끌어내는 기술이라는 의미이다. 또한 가트너와 IDC는 규모가 크고(Volume), 다양한 형태로(Variety), 빠르게 처리되는(Velocity) 빅데이터의 속성을 제시하였고, 여기에 복잡성(Complexity)과 가치(Value)를 추가하여 빅데이터를 정의하였다[3, 4]. 일련의 정의들을 볼 때, 빅데이터는 단순히 규모가 방대한 데이터를 뜻하는 것이 아니며, 이 데이터를 적시에 분석해 가치를 창출할 수 있도록 활용하는 일련의 작업 흐름을 포괄하는 개념으로 생각할 수 있다.



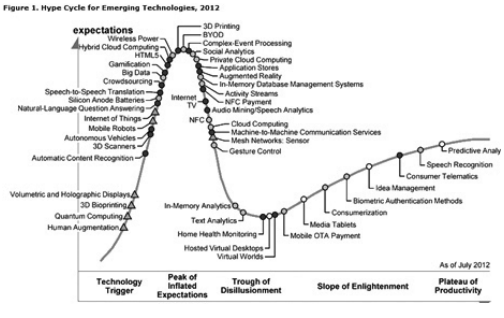
▶ (그림 1) Big Data Technology Stack,IDC

IDC는 빅데이터 산업을 4가지 영역으로 구분하고 있다[4]. 각각 빅데이터를 처리하기 위한 기반 설비 영역(Infrastructure), 정보를 수집, 정리하여 관리하는 영역(Data Organization and Management), 정리된 데이터를 분석하여 유의미한 내용을 발견하는 영역(Analytics and Discovery), 그리고 결과를 활용해 인간의 의사결정을 돕고, 스스로도 결과를 활용하는 영역(Decision Support and Automation)이다. 이는 서버 등의 하드웨어 뿐 아니라 데이터를 분석하는 소프트웨어, 나아가 일반 비즈니스 프로세스를 포괄하는 광대한 영역으로, 기존의 데이터베이스 개념과 빅데이터 간 활용상의 큰 차이점이 바로 여기에 있다. 기존 데이터베이스

는 단순히 데이터를 수집하여 저장하는데 초점을 두고 있었다. 이제는 기존에 수집 및 저장의 대상이 아니었던 종류의 데이터들, 가령 모바일 디바이스, 인터넷 홈페이지, 사무실과 상가 등에서 생성되는 각종 가속도, GPS 센서, 트래픽, 결제모듈 등에서 수집된 데이터들을 분석하여 의미를 파악, 가치있는 정보를 창출하고 이를 시각화하여 효과적으로 전달, 비즈니스에 활용하는 일련의 과정이 모두 빅데이터의 영역에 속한다는 것이다. 빅데이터의 구현을 위해 필요한 기술적 이슈 역시 중요하지만, 이 데이터를 통해 창출될 가치, 그리고 가치를 창출하는 방법에 대해 논의해야 하는 이유이다.

2. 2 빅데이터 현황 및 동향

데이터를 넓은 영역에서 활용하려는 시도들이 빅데이터 시대를 맞이해 갑자기 등장한 것은 아니다. 다만 활용이 가능할 정도로 충분한 데이터의 수집 자체를 포함해, 이를 다룰 수 있는 기술 자체가 충분히 발전하지 못했을 따름이다. 일례로 스마트폰의 폭발적인 보급으로 인한 변화를 들 수 있다. 매킨지는 현재 전 세계 인구의 60%에 달하는 40억의 인구가 모바일 폰을 사용하고 있고, 이 중 12%의 인구가 스마트폰을 사용하고 있으며, 스마트폰이 빠르게 모바일 폰을 대체할 것이라고 전망하고 있다[1]. 모바일 폰의 간단한 음성인식 기술 자체는 이전부터 존재해 왔으나, 자연스러운 대화를 인식하는 기술은 고성능의 스마트폰이 널리 보급되고 애플의 시리가 발표된 시점에서야 가능해진 것이 그 한 예이다. 또한 이 스마트폰을 이용해 SNS 서비스를 이용하는 개인들이 증가하고 있고, 각종 디바이스들에 탑재되는 센서들 역시 증가[5]하며 데이터의 생산에 일조하고 있다. 2013년 9월 발표된 아이폰 5S에는 위치, 가속도 데이터 등을 전담으로 담당하는 M7 칩셋이 추가되어, 이 데이터를 활용하는 새로운 서비스 등장의 기폭제가 될 것으로 예상되고 있다. 이제 막대하게 축적된 데이터를 내재된 가치가 충만한 자원으로 대할 수 있게 되었고, 이를 자산으로 활용해 가치를 창출하려는 시도들이 현실화되고 있는 것이다.



▶ (그림 2) Hype Cycle for Emerging Technology, Gartner

현재의 빅데이터 기술은 공공 데이터, 민간 SW 컨설팅 등 다양한 분야에 적용되어 그 가능성을 인정받고 있는 시기라 할 수 있다. IDC는 빅데이터 산업이 2011년에만 4.8B 달러의 산업 규모를 형성하고 있다고 분석하였으며[4], 가트너는 빅데이터가 현재 기술 도입기로 향후 2~5년 사이에 주류 시장에 편입될 것으로 전망하고 있어[3] 앞으로의 규모는 더욱 더 커질 것으로 예상할 수 있다. SAS는 2017년까지 영국에서만 58,000개의 전문 인력을 포함해 2.16 천억 파운드에 달하는 경제가치가 창출될 것으로 전망하

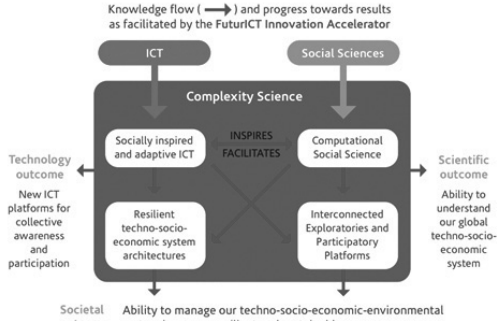
고 있고[6], 매킨지는 빅데이터가 22.3조 달러의 경제가치와 함께 미국에서만 2018년까지 19만명의 전문 인력과 150만개의 데이터 관리 직업이 창출될 것으로 전망하고 있다[1].

맥킨지는 빅데이터가 크게 활용될 수 있는 5개 분야로 의료 및 건강, 공공행정, 위치 등의 개인정보 분야, 소매 및 유통업과 제조업을 제시하였으며 실제로 현재 각국 정부 및 기업들은 빅데이터를 활용하고 있거나 이를 준비하고 있다. 아래에서는 빅데이터의 활용 현황을 크게 공공의 영역과 민간 영역으로 나누어 살펴볼 것이다.

2.2.1 공공기관의 빅데이터 활용 현황

공공의 영역에서 빅데이터를 활용하는 첫 걸음은 자신들이 구축한 데이터를 공개하는 것이다. 공공의 목적으로 축적된 데이터가 사회 구성원 모두의 재산이며, 이를 활용해 비즈니스 기회를 창출할 수 있다는 시각인 것이다. 미국은 이미 2009년부터 연방정부의 공공데이터를 공개하고 있고[7], 영국 정부 역시 2007년부터 Power of Information 계획을 수립하며 공공 데이터의 융합 및 활용 방안을 논의하였으며, Show us a better way 프로젝트를 추진하여 데이터 활용 방안에 대해 폭넓은 아이디어를 확보하고 있다. 2012년 6월에는 Department of Business Innovation & Skills (BIS) 주도로 행정부 주요 부처들이 참여한 Open Data Strategy 2012-2014 보고서[8]를 통해 의료, 교육, 세금, 고용, 기상 등의 주요 데이터들을 2015년까지 공개할 계획을 발표하였다. 실제로 웹을 통해 영국 정부의 공공데이터가 공개되어 기업들에게 활용됨으로써 국가 경제에 긍정적 영향을 미치고 있으며[9], 본 보고서에서는 이를 통해 창출될 가치가 연간 1조원이 넘을 것으로 전망하고 있다.

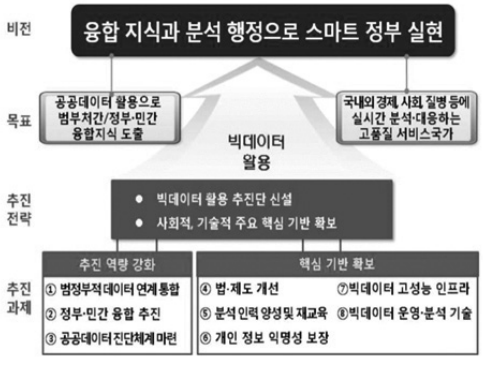
정부 주도로 빅데이터 개념 하에서 공공데이터를 직접 활용하기 위한 전략들 역시 수립되어 시행되고 있다. 국가적 차원으로 축적된 방대한 데이터를 융합하여 행정적 투명성을 재고할 수 있고, 생산성 향상 및 운영 효율화 등, 공공 서비스의 경쟁력을 강화할 수 있다. 2010년 미 대통령 자문기구로(PCAST)는 연방기관들의 빅데이터 대응계획 수립이 필요함을 제기하였고[10], 2012년 3월, 국립과학재단, 국방부를 포함한 6개 연방정부기관 통합으로 Big Data R&D Initiative' 를 추진하며 2억달러를 투자한다고 발표하다[11].



▶ (그림 3) FuturICT Framework,

EU는 오늘날 복잡하게 얽혀있는 이슈들에 대응하기 위해, ICT와 사회과학을 기반으로 FuturICT 프로젝트를 추진하고 있다. FuturICT 프로젝트는 30여개국 수백 개의 연구기관과, EMC, SAS, Yahoo 등을 포함한 수십여 개의 기업, 정부기관들이 참여하

고 있는 거대한 프로젝트로서, 개별적이고 분산되어 있는 현재의 데이터를 넘어 통합적이고 거시적으로 데이터를 다루는 방식의 필요성을 제시하며 금융위기, 재난방지, 교육 등 다양한 분야에 대한 연구를 진행하고 있다[12].



▶ (그림 4) 빅데이터 활용 스마트정부 구현안

대한민국 정부 역시 2011년 국가정보화전략위원회의 '빅데이터를 활용한 스마트 정부 구현(안)' 보고서를 발간하는 등, 공공기관에서의 빅데이터 활용을 모색해오고 있다. 부처간 데이터 공유 및 통합분석을 통해 융합지식을 창출하고, 이에 기반해 실시간으로 변화하는 이슈들에 선대응하는 스마트 정부 실현을 목표로 하고 있다. 이어 2012년 4월에는 한국정보화진흥원의 주도로 '빅데이터 국가전략포럼'이 발족되었고 ETRI는 방송통신위원회와 한국방송통신전파진흥원의 지원으로 공공분야 빅데이터 활용을 위한 지식자산 구축에 대한 개발 과제를 수행하고 있다[13].

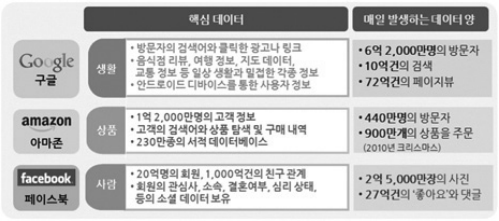
공공데이터의 공개, 전략수립 및 연구단계를 넘어 실제로 국가와 도시 운영에 활용되고 있는 사례 역시 찾아볼 수 있다. 미국의 National Security Agency(NSA)는 빅데이터를 통해 범죄 예측과 테러 방지 등 공공 안전을 꾀하고 있으며, 오바마 행정부의 핵심 공약인 의료보험 정책(Obama Care)은 의료, 질병관련 데이터 축적 및 활용을 기본 골자로 삼고 있는 것 역시 유의할 만 하다[X]. 브라질의 리우데자네이루는 30여개에 이르는 여러 기관의 데이터를 지능형 운영센터(IOC)에 통합해 도시의 총체적인 움직임을 모니터링하는 시스템을 갖추고 있다. 지능형 운영센터는 IBM으로부터 분석 솔루션을 제공받아 교통, 전력, 기상, 자연재해 대비 등을 포괄하는 능동적인 대응 시스템을 운영중이다. 싱가폴 정부는 2004년부터 국가위험관리시스템(RAHS: Risk Assessment and Horizon Scanning)을 구축하여 운영하고 있고, 실시간 교통정보에서 한단계 더 나아간 교통량 예측 시스템을 운영중이기도 하다[14]. 대한민국 역시 빅데이터를 시정에 적용하고 있다. 서울시는 2013년부터 심야에 버스를 운행하여 시민들로부터 높은 호응을 받고 있는데, 버스의 노선을 결정하는데 심야시간에 발생한 30억여건의 통화 데이터와 함께 택시 승하차 500만건의 데이터를 분석한 결과를 반영하였다. 빅데이터를 활용해 도시문제 해결 방안을 제시하였다는 성과로 대통령상을 수상하기도 했다[15].

상기의 세계 각국의 공공 데이터 전략 및 활용 사례를 볼 때, 공공의 영역에서 빅데이터 시대에 대응하기 위해서는 우선 데이터의 통합이 우선시된다고 할 수 있다. 각 부처에 나뉘어 관리되는 데이터를 효율적으로 통합할 수 있도록 관련 제도와 시스템을 정비하는 것이 선결과제인 것이다. 새로운 데이터의 탐색도 중요하지만, 기존에 장기간 축적되어온 데이터를 연계하여 활용하는 것이 더 현실적이고 효율적인 방법이며, 이를 널리 활용할 수 있도록 공개하는 것이 세계적인 추세이며, 바람직한 방향으로 볼 수 있

다. 다만 대한민국은 공공데이터를 활용한 서비스의 경우 무료로 제공되어야 한다는 인식이 지배적인데, 이 경우 일반 기업에서 데이터를 활용해 서비스를 제공할 동기부여가 떨어지므로 결과적으로 데이터의 활용도가 낮아질 수 밖에 없다. 공개된 데이터를 민간기업이 활용하여 새로운 서비스를 창출할 수 있도록, 공공 데이터의 수익모델에 대한 논의 역시 불가결하다.

2.2.2 민간기업의 빅데이터 활용 현황

상대적으로 분석이 용이한 데이터의 생산이 많고, 관리와 분석에 필요한 기술을 다수 보유하고 있는 ICT 기업들이 중심이 되어 빅데이터 시대에 발 빠르게 대응하고 있다.



▶ (그림 5) 주요 ICT기업들의 데이터 생산 현황, SERI

구글의 하루 동안 순수 방문자수는 6억 2천만명이며, 이들은 10억건의 검색을 통해 72억건의 페이지를 확인한다. 이렇게 생성되는 데이터를 활용하는 구글 트렌드(Google Trends) 서비스가, 소비 지출을 예측하는데 설문을 통해 도출된 지수보다 더 정확한 것으로 분석되었다[16]. 아마존의 1억 5천만명이 넘는 회원 중 하루 440만명이 사이트에 방문하고, 900만개의 상품을 구입하고 있다. 가장 급속도로 성장하고 있는 페이스북의 경우, 20억명의 회원이 평균 50명의 친구관계를 맺어, 1천억건의 연결고리를 구성하고 있다. 이들이 하루 동안 2억 5천만장의 사진을 올리고, 서로 27억건의 'Like'버튼을 클릭하며 상호작용하고 있다[14].

특히 본래의 기업활동을 통해 생산되어온 데이터와 근래 폭발적으로 성장하고 있는 소셜 네트워크 데이터를 결합하여, 자사의 비즈니스에 적용 가능한 새로운 가치를 창출하는 사례가 일반적이다. 아마존은 고객 개인의 데이터뿐 아니라 고객의 지인, 비슷한 성향을 가진 다른 고객의 정보를 통합 분석하여 추천 상품의 정확도를 향상시키는 사용자 기반의 협업 필터링 기술을 적용하고 있다[17]. 사실 아마존은 이전부터 자사가 보유한 데이터를 다방면으로 활용해오고 있는데, 15억개가 넘는 상품들을 효율적으로 관리, 추적하는 기술이 그 한 예이다. 최근 MIT는 아마존이 자사의 데이터를 분석한 결과를 광고 플랫폼으로 활용하도록 하는 신기술을 개발했다고 보고하기도 했다. 아마존 자체 뿐 아니라 일반 웹사이트에 아마존의 데이터를 활용한 마케팅 플랫폼을 제공하는 것이다[18]. 이베이는 고객의 소셜미디어 활동 내용을 분석하여, 기념일 등에 지인에게 선물할 제품을 추천하는 서비스를 시행 중에 있다. 미국의 통신사 T-Mobile은 자사의 데이터를 분석해 활용함으로써 고객의 이탈을 방지할 수 있었다. T-Mobile은 3천만명의 가입자로부터 매일 180억여건의 네트워크 데이터가 발생한다. 이 데이터를 통해 드러나는 개개인의 이용 패턴을 과거 이탈고객이 보였던 이용 패턴과 비교하여 잠재 이탈고객을 특정할 수 있는 것이다. 특히, 통신사 내 할인혜택 등으로 인해 이탈고객과 연결된 다른 고객들도 동반 이탈하는 현상이 소셜 네트워크 분석을 통해 드러났고, 이에 잠재이탈고객에게 혜택을 제공하여 이탈고객을 절반 수준으로 감소시킬 수 있었다[14].

사물인터넷(IoT)의 발전으로 기존에 존재하지 않았거나, 국한적으로 사용되었던 데이터를 네트워크를 통해 수집하여 활용할 수 있게 되었다. 특히 자동차와 ICT 기술의 융합, 각종 센서 정보를 통합 관리하는 전자 시스템을 통해 정교한 운전 경험을 제공함과 더불어, 이를 통해 가치를 창출하는 사례를 참고할 수 있다. 혼다는 기후와 도로의 통행량 등을 분석해 최적운전경로 및 다양한 운전자 지원 정보를 제공하는 Internavi를 개발하였다. 이를 통해 목적지에 도달하는데 걸리는 시간을 20% 단축하여 온실가스의 발생 역시 16% 감소시키는 효과를 거두고 있다[14]. SK 플래닛 역시 자사의 T-map 네비게이션 서비스에 680만여대의 차량으로부터 수집되는 교통량 데이터에 이용자의 검색빈도 등을 융합하는 빅데이터를 적용하고 있다[19]. 볼보는 자사의 제품으로부터 얻어지는 데이터를 분석하여 품질을 관리하고 있다. 종래에는 50만대 이상의 차량이 판매된 이후에 알게 되었을 결함을 1,000대 판매 시점에 포착할 수 있게 되었고, 사후관리 비용을 크게 줄일 수 있게 되었다[20]. 포드는 400만대의 차량에 설치된 센서를 통해 운전자의 주행 습관뿐 아니라 주행 환경에 대한 데이터를 수집하여 운전자 행태를 분석하고, 신제품에 반영할 수 있는 고객의 숨은 니즈를 발견하는데 빅데이터를 활용하고 있다[17].

상기의 사례들을 볼 때, 민간기업은 공공기관에 비해 직접적인 수익을 창출해야 하는 만큼 빅데이터 활용에 보다 적극적이고 직접적인 방식을 택하고 있다고 볼 수 있다. 다만 보다 적극적인 빅데이터의 활용을 유도하기 위해 선결되어야 하는 이슈 두 가지를 제기하려 한다. 첫째로, 데이터의 가치를 보다 유의미할 수 있도록 재평가할 필요가 있다. 2011년 세계 최대의 소셜 데이터를 보유한 기업인 Facebook의 기업공개가 이루어졌다. 기업공개 시 평가된 Facebook의 회계 상 가치는 63억 달러로, 기업공개 이전의 1040억 달러가 넘는 잠정 평가가치에 비해 지나치게 낮았다. 이유는 Facebook이 보유한 데이터의 가치를 포함하지 않은, 하드웨어 및 구성원의 가치만 평가되었기 때문이다[21]. 또한 핀터레스트, 인스타그램, 카카오톡 등 데이터가 가장 큰 자산인 ICT 스타트 업 기업의 경우, 잠재력이 지나치게 고평가되거나, 명확한 수익모델의 부재로 인해 지나치게 저평가될 우려 역시 존재한다. 따라서 데이터를 통해 창출될 기업의 가치를 평가할 방법을 탐구할 필요가 있음을 알 수 있다. 빅데이터 활용에 대한 법적인 이슈 역시 해결될 필요가 있다. 빅데이터는 그 규모의 특성상, 이용자 개개인이 데이터의 원천이라는 사실을 크게 인지하지 못한 채 자신의 정보를 제공하게 될 가능성이 크다. 이러한 데이터의 활용에 있어서 개인정보의 보호 등이 이슈가 될 수 있고, 또한 법적인 규제로 인해 새로운 서비스가 제한되기도 한다. 애플은 자사 고객의 계정으로 자신들의 디바이스를 추적하는 서비스를 제공하고 있는데, 국내 위치데이터 관련법으로 인해 국내 소비자들은 해당 서비스에 제한을 받고 있다. 본 서비스는 휴대폰의 도난, 분실 등으로 인해 개인정보의 노출이 우려될 때 유용한 서비스이나, 아이러니하게도 개인정보를 보호하기 위한 관련법과 충돌하고 있는 것이다. 빅데이터로 인한 기업환경의 변화는 앞으로 더욱 가속화될 것으로 전망되는 만큼, 규제와 관련된 논의가 시급하다 할 수 있다.

3. 빅데이터 기술

빅데이터 기술은 상당부분 세계 최대의 검색엔진 기업인 구글에 기반한다. 구글은 최근 10년간 GFS(Google File System), MapReduce, BigTable 등의 기반기술을 발표하였고, 이것들이 현재 빅데이터를 다루는데 표준화된 하둡(Hadoop), NoSQL 등의 기술로 발전하였다. 하둡과 NoSQL 등은 야후 등을 포함해 다양한 기업에서 빅데이터를 다루기 위한 목적으로 채택되어 사용되고 있다. 핵심은 데이터의 분산처리이며, 크기가 수 테라바이트(Tb)에 이르러 하나의 저장장치로 다룰 수 없는 대용량의 데이터를 여러 컴퓨터에 효율적으로 나누어 처리하는 것에 초점을 맞추고 있다.

하지만, 하둡과 NoSQL은 빅데이터의 정보처리 영역에 제한된 기술이다. 빅데이터는 다양한 분야를 포괄하는 만큼 한 분야에 한정된 기술이 아닌, 기존 사회과학 분야에서 정교하게 다듬어져 온 여러 통계기법 등을 포함해, 기계학습 등의 인공지능, 안전을 담보할 보안, 데이터를 이해하기 쉽게 만들고 인간을 통해 새로운 시각으로 바라볼 수 있게 하는 시각화 기술 등 다양한 영역의 기술의 필수적이다.

3.1 데이터 분석

텍스트 자체는 기존 데이터베이스에서도 널리 활용되고 있지만, 인간이 사용하는 자연어를 컴퓨터에 적용하는 것은 최근 들어서야 실용화되고 있다. 대표적인 예가 애플의 시리인데, 이에 필요한 기술이 텍스트마이닝이다. 실제로 소셜 네트워크 등에서 분석대상이 되는 것은 단순한 클릭 수 뿐 아니라 사용자 간 오고 간 텍스트 기반의 커뮤니케이션인데, 비정형 데이터인 자연어를 분석하여 의미를 추출하고, 유용한 데이터로 가공하는 것이다. 텍스트가 가지고 있는 의미뿐 아니라 그 안에 담긴 의견, 감정 또한 분석의 대상이 될 수 있다. 이를 다루는 분석기술을 오피니언 마이닝이라 하고, 방대한 텍스트 데이터를 재가공하여 +/−, 혹은 수치로 나타낼 수 있다는 점에서 여론조사 등에서 활용될 수 있다. 이용자들의 표현에 담긴 생각의 조각들을 모아 거대한 흐름, 일정한 법칙성을 찾아내 새로운 의견을 형성할 수 있는 것이다.

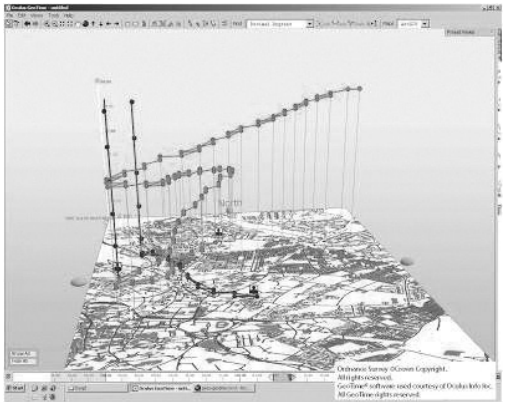
상기 두 분석기술은 최근 각광받고 있는 소셜 네트워크 자체를 분석하는 기술과 쉽게 결합되어, 더 강력하게 활용될 수 있다. 소셜 네트워크 분석은 수학적 그래프 이론에 뿌리를 두고 이용자 간 연결 회수, 연결의 강도와 연결 구조 등을 분석하는데, 이용자의 영향력 등을 측정할 수 있다는 점에서 중요하다. 비슷한 성향의 이용자끼리 뭉치는 경향으로 인해 분석 대상이 될 데이터 자체가 편향될 우려가 있는데, 소셜 네트워크 분석을 통해 집단 간의 연결고리가 되는 이용자를 찾아내 메시지의 전파를 용이하게 하고, 광범위한 데이터를 확보하여 데이터의 맥락을 파악하는데 장점이 있다. 또한 특정 집단의 특성을 집중적으로 파악하려 할 경우 군집분석 기법을 적용할 수 있다. 각 대상의 유사성을 측정하여 유사성이 높은 집단끼리 분류하고 이들간의 차이점을 규명하는 통계기법이다. 이는 명확한 분류기준이 존재하지 않거나, 비공식적인 연결관계를 가지고 있어 소셜 네트워크의 연결구조가 명확하지 않은 경우에 특히 유용하다[22].

3.2 인공지능

인공지능으로 가장 유명한 사례는 20여년 전, 인간 체스 챔피언에게 승리한 IBM의 딥 블루일 것이다. IBM은 딥 블루 이후로도 꾸준히 인공지능 컴퓨터의 개발을 지속했고, 2011년 미국의 퀴즈 프로그램 제퍼디 쇼에서 우승한 왓슨을 선보이기에 이른다. 왓슨은 2억 건이 넘는 자연어 분석을 통해 사용자의 질문을 추론하고 적절한 답안을 제시하는 능력을 갖추고 있는데, 이 자체로도 빅데이터를 활용한 좋은 사례가 될 수 있지만 다른 빅데이터와의 결합을 통해 더 큰 시너지를 창출해낼 수 있기도 하다. 막대한 양의 데이터를 인간이 모두 분석할 수 없기 때문에 이를 패턴화하여 스스로 데이터의 의미를 분석하는 인공지능이 중요해지기 때문이다. 현재 왓슨은 의료 분야에서 활용되고 있고, 향후 금융과 교육 등, 서비스의 범위를 확대할 것이라고 한다. 다양한 분야에 적용되며 추가적인 데이터베이스가 확보된다면 더욱 정교한 추론을 할 수 있게 될 것으로 기대되고 있다.

LG경제연구원은 지능형 서비스로 인해 기대되는 특성으로 의사결정의 효율화, 고도의 개인화, 현행 기술로 해결할 수 없는 근본적인 난제의 해결 가능성을 꼽았다[2]. 기존의 데이터베이스가 정보의 수집 및 저장에 초점을 맞추고, 이를 해석하여 활용하는 것을 숙련된 전문가에게 일임한 것에 반해, 빅데이터 스스로 필요한 대책을 제시할 수 있을 것이다. 또한 사용자 개개인에게 최적화된 서비스를 제공하는 것 역시 기대할 수 있다. 이미 구글은 이용자의 검색 패턴을 파악하여 맞춤형 검색 결과를 제시하고 있으며, 페이스북, 아마존 등의 업체들도 사용자 특성에 적합한 광고를 전달하는 등으로 활용하고 있다. 난치병의 치료, 교통 시스템의 효율화 등 기존 정형화된 방식으로 접근하기 어려웠던 난제들 역시 빅데이터를 통해 실마리를 제공받을 수 있다.

3.3 시각화



▶ (그림 6) Space-Time Cube 활용 데이터 시각화, GeoTime

빅데이터를 활용하는 것은 결국 인간이기 때문에, 데이터를 인간이 쉽게 이해할 수 있고 또 통찰력을 발휘할 수 있도록 시각화하는 기술 역시 중요할 수 밖에 없다. 전통적으로 데이터는 다이어그램, 표, 그래프 등으로 표현되었다. 최근 인포그래픽스와의 결합을 통해 데이터의 의미를 효과적으로 전달하기 위한 수단이 각광받고 있으며 연구 또한 진행되고 있는데, 다만 빅데이터의 효과적인 표현 방법이 디자인 영역에 국한되어 있다는 아쉬움이 있다. 데이터베이스의 특성과 활용 분야에 맞추어 최적화된 시각화 기술을 적용할 필요가 있다.

예를 들어 위치와 관련된 데이터의 경우, Space-Time-Cube 개념을 적용해 봄직하다. Space-Time-Cube는 1970년대에 최초로 제시되었던 개념으로, 공간정보와 시간 정보를 결합하여 시각화할 수 있다는 장점이 있다. 이 개념을 활용하여 GeoTime 등 여

러가지 도구들이 소개되었고, 개인의 위치, 운송, 지리 등 시간과 공간 정보가 결합된 다방면에 걸쳐 활용되고 있다[23].

3.4 새로운 데이터베이스

▶ (그림 7) Intel의 사물인터넷 및 웨어러블 컴퓨터 프로세서 Quark, Engadget

Intel은 IDF 2013에서 쿼크(Quark)라는 새로운 프로세서를 발표하였다. 기존의 모바일 프로세서였던 아톰 프로세서에 비해 크기는 1/5에 불과하고, 전력 소비량은 1/10에 불과한 이 칩셋은 사물인터넷, 나아가 웨어러블 컴퓨팅 환경을 지원하기 위한 목적으로 개발되었다. 모바일 디바이스가 스마트폰의 형태를 벗어날 수 있는 계기를 마련했으며, 다양한 디바이스에 탑재될 여러 종류의 센서로부터 생산될 데이터를 보다 효과적으로 관리할 수 있는 길이 열린 것이다.

이렇게 생산되는 데이터의 종류가 늘어날 것이기에, 여러 형태의 데이터를 활용할 수 있도록 형태를 조정해주는 링크드 데이터 기술 역시 중요해지고 있다. 서울시는 분산되어있는 데이터를 연결하는 링크드 데이터기술에 기반한 열린데이터 광장을 운영하고 있는데[24], 이는 공공기관의 데이터 공개 흐름에서 한발 더 나아간 것으로 평가할 수 있다. 단순히 분산되어 있는 데이터를 통합된 형태로 제공하는 것이 아니라, 데이터가 가지고 있는 의미를 분석하여 연관된 데이터를 함께 제공함으로써 데이터의 접근성과 활용도를 높이는 것이다.

4. 빅데이터2.0

보통사람들은 빅이라는 말에 양(量)의 개념으로 받아들여 SNS 등의 수많은 데이터들을 빅데이터와 연관지어 생각하는 경향이 있다. 전혀 무관하지는 않지만 빅이라는 단어에 집착해 수많은 데이터라는 잘못된 인지라고 할 수 있다. 빅데이터란 양적인 문제보다 선택적 데이터의 질에 관한 문제이기 때문이다.

빅데이터 1.0 vs. 빅데이터2.0

단순한 데이터의 양적폭증을 ‘빅데이터 1.0’이라고 본 다면 초고속 인터넷 인프라와 데이터 분 석력에 기반해서 정밀하게 여과되고 통합된 데이터의 증가는 ‘빅 데이터 2.0’이라고 할 수 있다. 단순히 빅데이터를 수집하고 이해하는 빅데이터 1.0 에 머무르지 않고 기업 의사결정 과정에 적극적으로 통합해내는 빅데이터 2.0으로 나아갈 때 기업이 희망하는 가치를 얻어낼 수 있다고 말한다. 실제 의사결정에 반영을 하려면 많은 데이터들을 분석하여 통찰력을 끌어내는게 핵심인데 이를 정보 최적화(Information Optimization)이라 하고 이것이 빅데이터2.0의 핵심적 특징이다. 이 정보최적화는 빅데이터2.0을 특징지우는 핵심적 특징인데, 그 과정자체가 빅데이터1.0의 일방향성이 아닌 쌍방향적 성격을 지니고 있다. 예를 들어 고객과 기업의 관계가 종속적, 일방적인 것 아니라, 상호작용적 관계라는 점이다. 빅데이터 1.0에서는 단순히 고객의

니즈와 소리를 분석해 경영전략을 수립하는 데 그쳤다고 하면, 빅데이터2.0는 고객이 기업의 경영과정 전체에 적극적으로 참여할 수 있어야 한다. 또한 그 참여가 단발성 참여가 아닌 지속적인 과정이라는 것이다. 즉 빅데이터1.0이 특정 문제를 위한 일회적 솔루션(Solution)이라면, 빅데이터2.0는 그런 솔루션을 포함한 하나의 과정 그 자체라는 것이다.

송민정(2012)은 빅데이터2.0를 정의하며 빅데이터1.0과 구분되는 특징을 제시했는데 [25], 이를 데이터 저널리즘의 관점에서 적용하면 크게 3가지로 요약될 수 있다. 첫째는 언론사와 독자, 대중의 관계가 종속적, 일방적인 것 아니라, 대등한 관계에서 상호작용적 관계라는 점이다. 빅데이터 1.0에서는 단순히 독자와 대중의 소리를 분석해 기사작성을 하는 데 그쳤다고 하면, 빅데이터2.0는 독자와 대중이 언론사의 콘텐츠 생산 과정 전체에 적극적으로 참여할 수 있어야 한다. 또한 그 참여가 단발성 참여가 아닌 지속적인 과정이라는 것이다. 즉 빅데이터1.0이 특정 문제를 위한 일회적 솔루션(Solution)이라면, 빅데이터2.0는 그런 솔루션을 포함한 하나의 과정 그 자체라는 것이다.

두 번째는 빅데이터2.0은 새로운 콘텐츠와 서비스를 창출해야 한다는 점이다. 저널리즘의 위기라 불리우는 근간에

이제 저널리즘이 데이터를 단순히 인용하는 차원의 사용이 아니라 데이터를 적극 활용하여 데이터 중심의 보도와 그를 통한 새로운 콘텐츠와 서비스가 창출되어야 한다. 언론이 더 이상 정보를 독점하는 시대가 끝나고 이젠 현장 목격자나 블로그 같은 다양한 정보원이 '네이티브 리포팅'(nativereporting)을 통해 수시로 뉴스를 실시간으로 전해주고 있다. 저널리즘이 이런 네이티브 정보원에게 둘러싸여있기 때문에 이젠 1차적 객관적 사실 이와에 데이터를 분석하고 걸러내며, 2차적 사실과 통찰력을 보여주는 것이 갈수록 더 많은 가치를 갖게 된 것이다. 언론사는 이제 빅데이터2.0을 허브로 하여 정보원으로의 독자, 대중과 긴밀한 관계를 유지하여 새 콘텐츠·서비스 창출의 선순환 구조를 이루어 나가야 한다.

세 번째는 생태계와 같이 데이터 범위를 확장시켜야 한다는 점이다. 빅데이터 분석 가치는 세상에 널려있는 데이터를 결합해 인사이트를 넓혀가는 데 있으며, 데이터 보유 주체간의 윈윈 모델로 데이터 범위 확장이 가능하다. 빅데이터2.0를 통해 미디어 생태계 뿐만이 아니라, 통신, 과학, 문화등 다양한 영역의 생태계에 확장시킬 수 있다. 생태계의 확장은 새로운 시장을 넓히고, 새로운 콘텐츠와 서비스가 나올 기반을 넓혀 주는 것이다.

5. 결론 및 시사점

빅데이터는 분명 새로운 기회가 될 수 있다. 지금까지 데이터의 체계적인 분석과 활용이 제한된 영역에서 이루어졌던 것에 반해, 데이터베이스의 영역 자체가 넓어지는 빅데이터 환경에서는 여러 분야에서 축적되어 온 여러 형태의 데이터가 새로운 힘이 될 수 있기 때문이다. 특히 우리나라의 경우 스마트폰이 널리 보급되어 있고 네트워크 인프라가 잘 갖추어져 데이터의 생산 면에서 매우 발전되어 있다고 할 수 있다. 다만 하드웨어에 치중되어있는 국내 산업구조 상, 빅데이터 산업에서 기회를 찾는 것에는 제약이 있는 것이 사실이다. 빅데이터의 인프라, 대표적으로 저장장치의 경우

가격이 매우 낮아져, 전 세계에 존재하는 음악파일을 모두 저장하는데 필요한 비용은 600달러에 불과하다. 따라서 빅데이터 산업의 주요한 가치는 데이터의 수집, 분석을 용이하게 하는 사회과학적 방법론의 개발과 소프트웨어에 존재한다고 볼 수 있다. 이미 IBM, SAS, EMC 등 세계적인 IT 솔루션 기업들은 개발 및 관련기업 인수 등을 통해 앞다투어 분석 기술을 확보하고 있기에 데이터베이스의 종속 우려가 있는 것이 현실이다. 다만 인터넷, 모바일 서비스 등에서 많은 강점을 활용할 수 있기에 새로운 서비스 창출에 기대를 걸 수 있을 것이다.

참 고 문 헌

[1] McKinsey & Company. Big data : The next frontier for innovation, competition, and productivity. 2011

[2] LG경제연구원. 빅데이터 시대의 한국:갈라파고스가 되지 않으려면. 2012

[3] 이만재. 빅데이터 어낼리틱스와 공공데이터 활용. 정보과학 학회지. 2012

[4] IDC. Worldwide Big Data Technology and Service 2012-2015 forecase. 2012

[5] 김상락, 강만모, 박상무. 빅데이터가 여는 미래의 세상. 정보과학 학회지. 2012

[6] SAS. Data equity : Unlocking the value of big data. 2012

[7] http://www.data.gov

[8] BIS. BIS Open data strategy. 2012

[9] http://www.data.gov.uk

[10] President's Council of Advisors on Science and Technology. Designing a Digital Future : Federally Funded Research and Development in Networking and Information Technology. 2010

[11] US White House. Big data is a Big Deal. 2012

[12] http://www.futurICT.eu

[13] 이강용 외 4명. 공공분야에서의 빅데이터 활용을 위한 지식자산(Knowledge Base) 구축. 정보과학 학회지. 2012

[14] SERI. 빅데이터: 산업 지각변동의 진원. 2012

[15] 서울특별시 보도자료. 서울시, 빅데이터 활용한 정책 사례로 대통령상 수상. 2013. 9. 25

[16] Vosen, Schmidt. Forecasting Private Consumption: Survey-Based Indicators vs Google Trends. Journal of Forecasting. 2011.

[17] SERI. 기업의 신 경쟁력, 빅데이터 큐레이션. 2013

[18] MIT Technology Reviews. Amazon Woos Advertisers with What It Knows about Consumers. http://www.technologyreview.com/news/509471/amazon-woos-advertisers-with-what-it-knows-about-consumers

[19] IDG 컨퍼런스. Business Impact & Big Data 2013

[20] Icio. Converting Data into Business Value at Volvo. http://www.i-cio.com/case-studies/volvo-big-data

[21] 빅토르 마이어 쇤버거 외 2명. 빅데이터가 만드는 세상. pp220. 21세기 북스.

[22] 이진형. 데이터 빅뱅, 빅데이터(BIG DATA)의 동향. Journal of Communications & Radio Spectrum. 2013

[23] Shih-Lung Shaw, Hongbo Yo. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical-virtual space. Journal of Transport Geography. Vol 17. 2009

[24] http://data.seoul.go.kr

[25] 송민정 (2012). 빅데이터가 만드는 비즈니스 미래지도. 한스미디어.

저/자/소/개

신 동 희

1997년 성균관대학교 학사
2001년 미 Syracuse University 석사
2004년 미 Syracuse University 박사
관심분야 : HCI, Social Computing, Informatics

저/자/소/개

이 재 길

2013년 성균관대학교 공학 / 심리학 학사
2013년 성균관대학교 석박사 통합과정
관심분야:HCI,Social Computing, Information Science