



빅 데이터(Big Data) 핵심체크

+ 박성환 미래방송연구회 부회장

소셜미디어와 빅 데이터 시대



그림 1. 인터넷 주요서비스의 실시간 이용현황
출처 : <http://pennystocks.la/internet-in-real-time>

빅 데이터라는 이름은 컨설팅회사인 매킨지의 2011년 보고서에서 처음 사용되었다고 한다. 쌓여있는 데이터를 ‘빅 데이터’라고 최초로 언급하여 새로운 이슈로 떠올랐다. 데이터는 우리 주변에 늘 존재하는 것이며, 관심을 가지고 활용되던 것은 구조화된 정형 데이터들이었다. 인터넷과 모바일 시대가 본격적으로 확산되면서 SNS(Social Network Service), 이메일, 블로그 등에서 사용자가 올리는 비정형화된 데이터들이 가치 있는 자원이 될 수 있다는 점에서 주목받고 있다.

물론 해석 능력에 따라서 그 가치는 달라진다. 지금 이 시각 실시간으로 인터넷상에서 얼마나 많은 데이터가 생성되고 있는지 살펴보자. [그림 1]은 인터넷에서 단 1분간 측정된 자료 수치이다. 페이스북 이용자는 310만 명 이상이 좋아요!를 클릭하였으며, 320만 이상이 포스팅을 했다. 유튜브에는 1분에 120시간의 비디오가 업로드되었으며 13만 시간 이상의 비디오가 시청되었다. 트위터 이용자의 경우 660개의 계정이 새롭게 생성되었고, 34만 2천 건의 트윗이 발생하였다. 조사된 사이트들의 1분간 활동을 데이터양으로 추산하면 대략 1,354,440GB(기가바이트), 즉 1.35PB(1페타바이트=10¹⁵ 바이트)의 데이터 용량에 이른다.

빅 데이터의 정의는 매킨지, 가트너, IDC 등 기관별로 약간 상이하게 상세 정의를 하고 있다. 국내의 네이버 지식백과에 따르면 “데이터의 생성 양 · 주기 · 형식 등이 방대한 데이터”라고 정의하고 있다. 즉, 기존 데이터에 비해 너무 크기 때문에, 종래의 방법으로는 수집 · 저장 · 검색 · 분석이 어려운 방대한 데이터를 말한다. 물론 크기만으로 분류해서는 안 되며 특징들을 정리하면 다음과 같다.

빅 데이터의 특징과 소셜미디어

빅 데이터의 특징은 일반적으로 가트너 그룹에서 정의한 3V를 기준으로 한다. 여기에 IBM에서는 1V를 추가하여 4V로 소개하고 있다.

첫 번째 특징은 데이터의 크기를 말하는 Volume(크기)이다. 이와 관련하여 2010년 8월 에릭 슈밋 구글 회장은 인류 문명 탄생 이후에서 2003년까지의 총 정보량과 2010년 8월 당시에 이틀마다 생겨나는 데이터양이 동일하다고 하여서 우리를 놀라게 했다. 그 분량은 5EB (1엑사바이트 = 10^{18} 바이트)라고 했다. 이후 2011년 전 세계 디지털 정보량은 1.8ZB(제타바이트, 1제타바이트 = 10^{21} 바이트 = 1조GB)로 2020년에는 관리해야 할 데이터양이 50배 이상 증가할 것으로 IDC & EMC 보고서에서는 예측했다.

두 번째 특징은 데이터 형태가 각각 다른 Variety(다양성)를 가졌다는 것이다. 우리가 늘 관리하던 고객 데이터, 매출 및 회계 데이터 등 정형화된 것과 최근 그 생성량이 급속도로 증가하고 있는 소셜미디어, 개인 블로그나 멀티미디어 등에서 생산되는 텍스트, 이미지, 동영상 등 다양한 비정형 데이터들이 여기에 속한다.

세 번째 특징은 스트리밍 데이터의 분석 속도를 의미하는 Velocity(속도)이다. 이전에는 수집된 데이터의 활용 주기가 몇 시간에서 수 주에 이르는 시간이면 충분했으나 지금은 실시간 분, 초 단위로 분석되어 활용되고 있다. 그래서 산업군에 따라서는 2초 먼저 아는 기업이 승리한다는 이야기가 나오고 있다. 이러한 세 가지 특징만 가지고는 무언가 유의미한 결과를 얻기에는 부족하다. 그래서 핵심은 데이터의 가치를 찾아서 활용하는 것이 관건이다.

네 번째 특징은 데이터의 불확실성 속에서 가치 있는 진실성을 찾아내는 Veracity(진실성) 혹은 Value(가치)를 네 번째 특징으로 소개한다.

엄청난 규모의 데이터양, 정형 · 비정형의 다양성, 빠른 분석으로 대응해야 하는 속도, 데이터의 진실성 등을 고려하여 가치창출의 중요

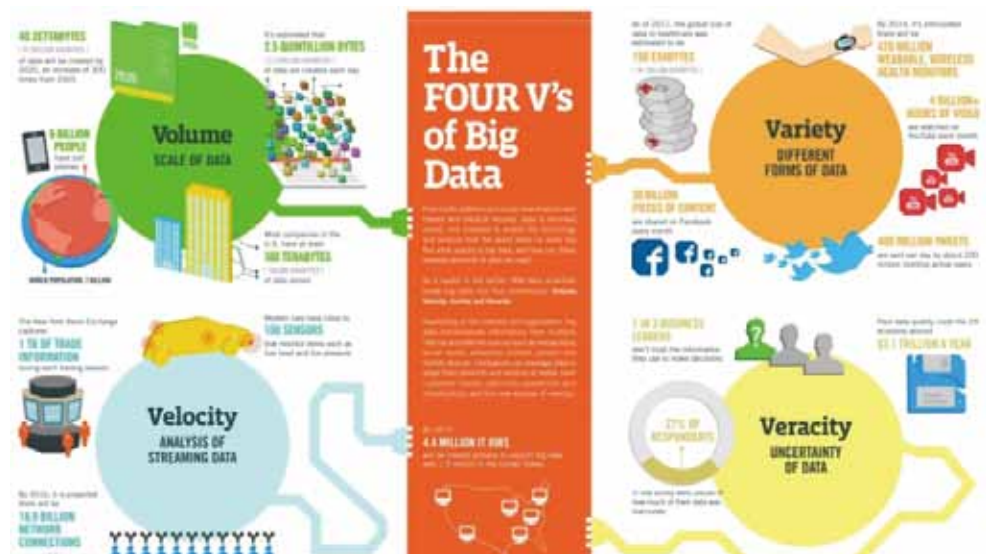


그림 2. The Four V's of Big Data(BM)
출처 : <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>

성이 강조되면서 '빅 데이터 애널리스트'가 새로운 직업군으로 부각되고 있다. 빅 데이터는 사물인터넷(IoT) 시대 도래로 각종 센서 데이터, 인터넷과 모바일 시대의 소셜미디어 데이터의 축적, 클라우드 컴퓨팅과 같은 기술 발달 등으로 자연스럽게 나타난 현상이다. 다음 [그림 3]은 전 세계 수많은 소셜 미디어들이 참여·공유·개방을 통해서 지구촌을

하나의 집단지성 사회로 묶어나가고 있는 상태를 보여준다. [그림 4]는 소셜미디어와 빅 데이터의 교차점을 형상화하였다.

빅 데이터 처리 기술

빅 데이터의 내용은 페이스북·트위터 등에 올린 텍스트부터 유튜브의 영상 콘텐츠와 같은 각종 멀티미디어 콘텐츠까지 다양하게 존재하고 그 형태도 정형·반정형·비정형 데이터로 구성되어 있으므로 이러한 데이터를 스트리밍으로 취득하고 정형·반정형·비정형 등에 따라 구조화하고 분석하는 절차를 거친다.

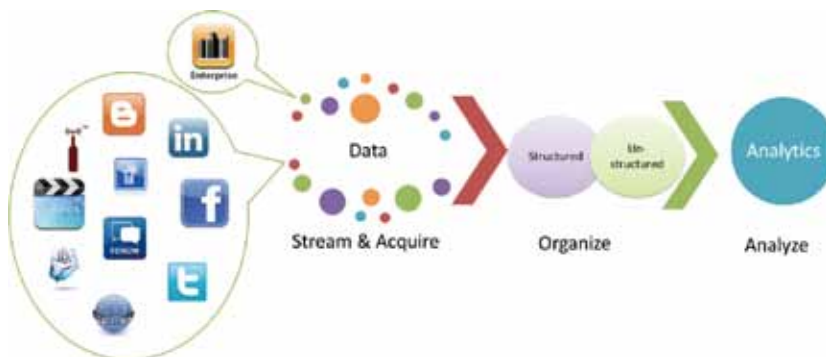
데이터 분석기술은 크게 4가지로 나뉜다. 첫 번째로 반정형·비정형 데이터에서 자연어 처리 기술을 기반으로 정보를 추출하고 가공하여 의미 있는 정보를 추출하는 텍스트 마이닝(Text Mining) 기법이 있다. 다음은 소셜미디어상에서 정치인, 연예인 등 알려진 사람들의 긍정·부정 등의 선호도를 판별하여 서비스를 하는 소셜메트릭스 인사이트와 같은 서비스 사이트에서 이용하는 평판 분석(Opinion Mining) 기법이 있다. 이러한 기법은 마케팅에서 상품의 시장규모, 소비자의 반응 분석 등에 많이 사용되고 있다. 다음은 소셜 네트워크 분석(Social Network Analytics)으로 소셜 그래프(소셜 네트워크를 표현하기 위해 사용되는 구조를 말한다. 소셜 그래프는 노드(Node)와 링크(Tie)로 이루어져 있다.) 상에 나타나는 소셜 네트워크 연결의 강도를 바탕으로 영향력 있는 사용자를 찾아내는 기법이다. 소셜 상에서도 오프라인과 동일하게 영향력이 작용함을 알 수 있다. 다음으로 관심사와 취미, 특기 등에 따라서 군집하여 유사한 특성을 가진 그룹을 찾아낼 수 있는 클러스터 분석(Cluster Analysis) 기법이 있다. 데이터는 수집, 관리를 넘어서 분석, 처리를 통해서 그 가치가 무궁무진한 만큼 IBM, MS, Oracle, EMC, SAP 등 세계 유수의 IT 기업과 Vertica, Splunk, Cloudera 등의 빅 데이터 전문기업들이 경쟁력을 키우고 있다. 국내에서는 소셜 분석 서비스 분야가 가장 활발하게 움직이고 있으며 소셜 분석 전문기업으로



출처 : <http://www.fredcavazza.net>



그림 4. 소셜미디어와 빅 데이터의 교차점
출처 : <http://www.zdnet.com/blog/hinchcliffe>



출처 : <http://www.gatewaygss.de/consultingservices/big-data/overview.html>

는 다음소프트, 솔트룩스, 골든 플래닛, 그루터 등이 조사되고 있다.

빅 데이터 분석 인프라 기술로는 첫 번째로 빅 데이터 처리 기술의 대표적인 프레임워크로 알려져 있는 하둡(Hadoop)이 있다. 두 번째로는 통계계산 및 시각화를 위한 언어 및 개발 환경을 제공하는 오픈소스 프로젝트로 소개된 R언어를 들 수 있다. R은 하둡 환경에서 분산 처리를 지원하는 라이브러리를 제공하여 대용량 데이터 통계 분석 및 데이터 마이닝을 위해 널리 사용되고 있다. 세 번째는 NoSQL(Not Only SQL)로 기존의 관계형 데이터베이스의 수평적 확장성의 한계를 해결하고자 하여 부각되고 있다. 마지막으로 대용량의 데이터를 처리하기 위한 분산 데이터 처리 프레임 워크인 맵리듀스(Map-Reduce) 기술이 있다.

이외에도 플랫폼 기술로 대용량 데이터 저장을 위한 기술 및 스토리지 시스템에 대한 고려, 병렬 DBMS 기술 등이 응용, 사용되고 있다.

빅 데이터 산업 생태계와 사물인터넷

빅 데이터 생태계는 빅 데이터 서비스를 공급하는 서비스 공급자와 서비스를 구매하고 소비하는 사용자, 공급자와 사용자를 연결해주는 애플리케이션 공급자로 구분할 수 있다. 빅 데이터 서비스 공급자는 다양한 하드웨어, 소프트웨어 플랫폼을 통해서 서비스를 제공하고 사용자는 애플리케이션을 통해서 쉽게 데이터의 가치를 찾도록 활용하는 구조이다.

빅 데이터 산업 생태계의 구분은 회사별로 약간의 차이점을 보이고 있다. 이 글에서는 Vitria사에서 제시한 기술, 인프라, 앱으로 구분한 내용을 소개한다. 사용자 관점에서는 이러한 빅 데이터 생태계의 각 요소를 이해하고 각국의 빅 데이터 활용 사례를 통해서 보다 나은 인류의 미래 생활을 위해서 어떤 활용 분야를 찾을 것인가에 관심을 가져야 한다. 그리고 사물인터넷(IoT)은 빅 데이터의 한 축으로 발전

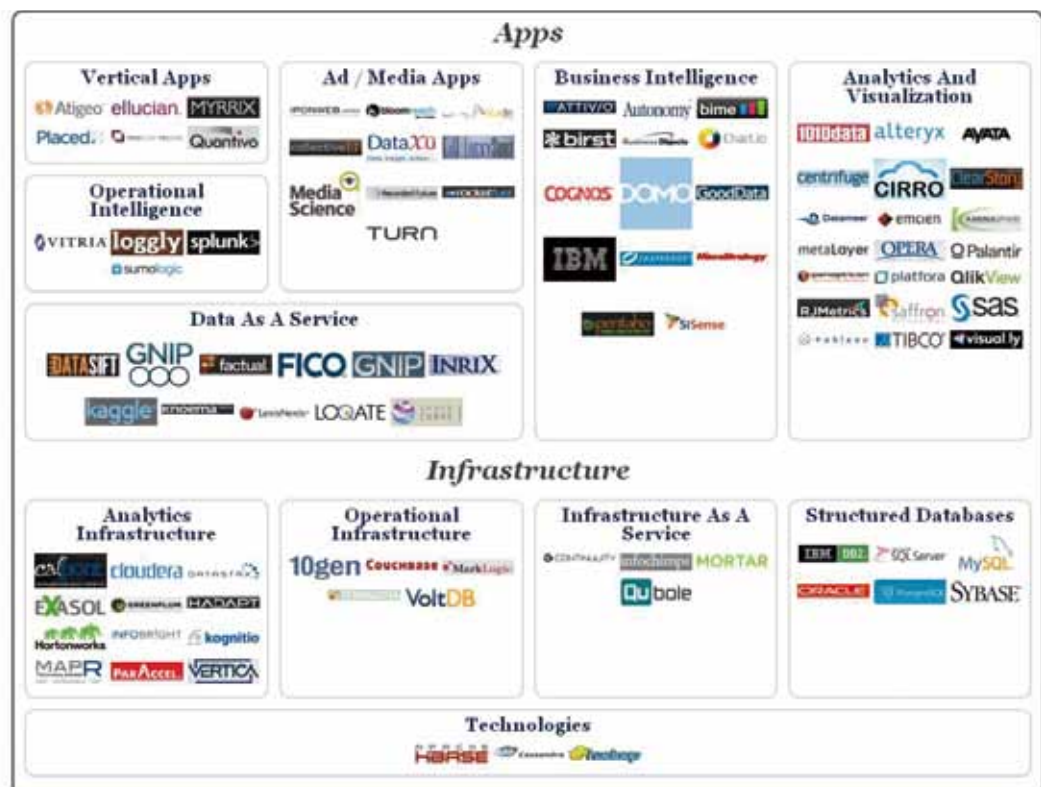


그림 6. 빅 데이터 생태계 지형도, 출처 : Vitria 2012

할 것으로 예측된다.

빅 데이터 활용 사례

빅 데이터의 활용은 정부 차원에서 자연재해, 재난과 같은 안전 대응 문제, 환경 분야, 각종 모니터링과 상황 분석, 보다 나은 미래를 위한 예측 분야 등과 기업 차원에서 새로운 이익 창출을 위한 비즈니스 모델 발굴, 경쟁력 강화 방안, 생산성의 향상 등의 분야를 예측할 수 있으며 개인 차원에서는 개인의 가치 창출 및 자신의 분야에 대한 비즈니스 창출 등으로 무궁무진한 활용이 가능할 것으로 기대된다. 여기서는 주요한 몇 가지 응용 분야를 사례로 제시한다.

공공분야를 먼저 살펴보면 최근 상징적인 빅 데이터 활용 사례는 범죄 수사 분야를 들 수 있다. 2013년 4월 발생한 보스턴 마라톤 폭탄 테러에서 테러범 검거를 위하여 미국의 NSA(National Security Agency)는 현장 주변에 있던 시민들이 소셜미디어에 올린 10TB에 이르는 사진, 동영상 분석했다. 물론 사고 현장 주변의 CCTV 600여 대의 정보도 수집했다. 이러한 자료를 근간으로 용의 추정 인물의 유형을 압축하고, 결정적으로 폭탄 추정 가방을 내려놓는 영상을 확보하였으며, 부상자의 증언으로 사실을 확인하여 범인 검거의 개가를 올렸다. 이처럼 국가 안보 분야, 감수량이나 특정 지역의 기후 변화, 교통 정보 수집 등을 모니터링하여 매우 다양한 분야에 유용하게 활용할 수 있다. 뉴욕 시에서는 가로수 가지치기가 필요한 주기를 예측해서 안전사고를 줄이는 성과를 내고 있다. 미국에서는 Health 2.0 프로젝트로 의료기관, 환자, 정부, 의료 보험회사를 하나로 통합하여 데이터를 활용하는 일을 하고 있다. 동일한 데이터를 통해서 실수 없는 의료 지원 체계를 구축할 수 있을 것으로 기대된다.

기업에서의 활용을 살펴보면 수익증대, 비용 절감 등을 통해서 비즈니스 효율을 높이는 것에 많은 관심을 보이고 있다. 널리 알려진 모범 사례로는 아마존의 상품 추천 서비스이다. 아마존은 자사 회원들의 소비 패턴을 분석하여 구매 관련 상품을 추천하고, 특정 물품 구매 시의 소비성향 데이터를 분석하여 연관되는 제품을 추천하여 수익증대뿐만 아니라 맞춤형 서비스로 좋은 평가를 받고 있다. 여기서 더 나아가 아마존은 킨들을 활용하여 전자책을 읽는 독자들이 표시하는 밑줄 정보, 독서 멈춤 지점 정보 등이 다음 작품의 방향성을 제공할 수 있다는 점을 고려하고 있다. 나도 모르는 사이에 나의 심리 저변을 움직이는 행동이 분석되고 활용되는 것이다. 월마트는 고객의 구매 패턴을 분석하여 전통적인 마케팅과 접목하여 상품 진열에 활용하고 있다. 허리케인이 예보되면 손전등뿐만 아니라 비상식량 구매가 늘어나므로 관련 제품을 함께 진열하여 매출 증대로 연결하는 것이다. 가장 많은 고객 데이터를 가지고 있을 것으로 추정되는 구글도 사용자의 로그 데이터를 활용하여 검색 알고리즘 및 검색 결과를 개선하고 있다. 구글은 앞으로 엄청난 일들을 빅 데이터를 활용하여 해낼 수 있을 것으로 전망된다. 그것은 아마도 새로운 비즈니스 영역을 만들어 낼 것이다. 구글 검색 패턴으로 미국 내 감기의 유행이 지역적으로 어떻게 이동하는지, 대선 판도가 어떻게 되는지를 예측하는 것은 이미 알려진 사실들이다.

콘텐츠 분야의 빅 데이터 활용

세계 최대의 검색엔진 구글에서는 클라우드에 저장된 웹사이트 정보를 활용하여 1분에 200만 번 이상의 검색 결과를 제시하고 있다. 세계 최대의 미디어 공유 솔루션인 유튜브에서 1분에 120시간의 동영상이 업로드되고 있다. 이제 인터넷, 모바일 등의 네트워크상에서 수많은 정형·비정형의 데이터를 언제 어디서 누가 어떻게 활용할지 알 수 없는 상황이 되었다. 구글이 가진 빅 데이터는 현미경과 같이 생물을 확대 관찰하는 역할을 하면서 많은 일들을 해낼 것이다. 구글은 미디어 업계의 전자 현미경이라고 할 수 있다. 하지만 구글 이외에도 신속하게 빅 데이터를 활용한 미디어 기업이 있으니 바로 넷플릭스이다. 넷플릭스는 2500만 명 이용자의 일시정지, 되감기 등의 이용 행태, 하루 평균 3000만 건의 동영상 재생 기록, 3개월간 20억 시간 이상의 동영상 시청 시간 기록 등을 가지고 하루 평균 400만 건에 달하는 이용자의 평가, 300만 건의 검색정보, 위치 정보, 단말 정보, 주중 및 주말의 시청 행태, 시장조사 업체의 메타데이터, 페이스북과 트위터로부터 수집한

소셜 데이터 등을 추적 분석한 것으로 알려져 있다. 이러한 분석을 통해서 작품의 장르, 감독, 배우 캐스팅을 하였다. 스트리밍 업체인 넷플릭스가 '하우스 오브 카드'를 자체 제작하여 2013년 제65회 에미상에서 최우수 감독상을 비롯한 3관왕의 차지하는 돌풍을 일으킨 것은 획기적인 사례이다.

할리우드에서는 소셜미디어 빅 데이터를 활용하여 신작 영화에 대한 잠재적인 영화 관람객의 반응을 추적해 박스 오피스 실적을 예측하고 있다. 월드와이드 모션픽처그룹은 신작 영화 개봉 24주 전부터 SNS에서 데이터를 수집하고, 설문조사 및 포커스 그룹 인터뷰를 통해서 영화 티켓 판매 실적을 예측하는 무비 트래커 서비스를 하고 있다. 인도의 TV 다큐멘터리 프로그램인 '사띠아메르 자야테'의 경우 TV 프로그램에도 빅 데이터 분석을 통해서 시청자의 피드백을 프로그램 제작에 반영하는 사례로 알려져 있다. 게임 분야에서는 게임의 난이도 조절, 게임 콘텐츠 품질 향상, 게임과 SNS를 통해 획득한 데이터를 결합하여 이용자의 선호 경향을 파악하여 맞춤형 광고를 제시하여 매출 증대하기 등과 같은 다양한 시도가 이루어지고 있다. 게임 분야에서 빅 데이터를 가장 잘 활용하는 사례로는 세계 최대 소셜 게임 사업자인 징가(Zynga)가 있다. 음악 분야에는 판도라의 사례에 주목하자. 판도라는 8년간 2억 명 이상 이용자의 프로필과 음악 취취 행태를 수집하고 분석하여 '뮤직 게놈 프로젝트'를 통해 음악의 음조, 템포, 악기 등 최대 450개의 속성을 기준으로 음악을 분석하여 제공함으로써 월평균 7천 만회 이상의 청취 횟수를 확보한 세계 최대의 인터넷 방송 서비스로 성장하였다. 이 외에도 잘 알려진 스마트폰 애플리케이션 샤잠(Shazam)은 친절하게도 우리가 접하는 노래의 원곡을 알려준다. 샤잠은 빅 데이터 기술을 활용하여 광고에 삽입된 음악, 방송 프로그램의 음성 정보까지 데이터베이스화하여 소비자에게 알려주어 콘텐츠 구매 비즈니스로 발전하고 있다.

마무리

빅 데이터의 활용은 각 산업 분야에서 무궁무진한 가능성이 존재한다. 하지만 우리가 관심을 가지는 미디어 콘텐츠 분야에서는 아직 많은 사례가 알려지지 않고 있다. 현재까지 파악된 내용을 근거로 콘텐츠 분야에서 빅 데이터를 활용하여 새로운 기회를 열고자 한다면 다음 두 가지 관점에서 균형 있는 대비를 하여야 할 것으로 생각된다. 먼저 기존의 프로그램과 마찬가지로 콘텐츠의 기획 및 제작 과정에서 이미 확보하여 분석된 데이터를 반영하고, 제작 및 방송 중에도 시청자들의 오프라인 반응과 소셜 데이터를 반영하여 시청자와 호흡하는 콘텐츠를 만들어야 한다는 것이다. 물론 반영의 정도는 기획자의 풍부한 창의력과 상상력에 따라서 달라질 것은 자명하다. 다음으로는 수익 극대화를 위해서 타겟 시장이 국내외 어디인지를 고려하여 연령별, 지역별 그리고 사회의 현재 이슈 현상에 맞추어 서비스 시기를 조절하는 콘텐츠 공급 마케팅 전략의 구사가 필요하다.

이상에서 느낄 수 있는 것은 미디어 분야에서 활동하는 사람들은 새로운 미디어 기술의 등장에 익숙해져야 하며 사회적, 문화적 변화에 어떤 반향을 가져올지에 대한 연구를 선도적으로 행해야 할 것이다. 이 기회에 엔지니어들은 새로운 직업으로 대두되는 '빅 데이터 애널리스트(Bigdata Analyst)' 혹은 '데이터 과학자(Data Scientist)'라고 불리는 전문가에 도전해 보는 것도 추천할만하다. 빅 데이터에 대한 이론적 지식과 융합 기술을 바탕으로 협력하는 능력, 통찰력을 가진 전문가의 미래는 밝을 것이기 때문이다. 아울러 어디로 얼마만큼 확장될지 예측할 수 없는 활용으로 사생활 침해, 기계적 판단으로 일어나는 피해, 보안 이슈 등에 대응하는 알고리즘미스트(Algorithmist) 전문가라는 직업도 부각되고 있으니 참고하시고, 관련하여 미국의 대학에서는 빅 데이터 자격증 과정이 진행되고 있으니 한발 앞서 관심을 가져볼 만한 새로운 분야라고 하겠다. 

*[참고자료]

- 빅 데이터 산업 생태계 분석 동향: 정보통신정책연구원, 2013.7.16
- 빅 데이터 동향 및 시사점. nipa, 어윤봉, 2012.8.16
- 빅 데이터 분석을 통한 한류지도 구축 방안, 한국콘텐츠진흥원, 2014.7.31
- 콘텐츠 분야에서의 빅 데이터 기법 활용 사례, 한국콘텐츠진흥원, 2014년 2월호
- 네이버 지식백과
- 위키백과, <http://ko.wikipedia.org/wiki/>

