빅데이터(Big Data) 활용단계에 따른 요소기술별 추진동향과 시사점

Korea Communications Agency ▌2013.12.10

KCA 한국방송통신전파진흥원
Korea Communications Agency

개요

세계 주요기관에서는 향후 수년을 주도할 ICT 분야 핵심 트렌드로 '빅데이터'를 주시하고 있다. 빠르게 변화하는 ICT 환경에 적극 대응하고 새로운 가치를 찾기 위해서는 가공되지 않은 많은 양의 데이터를 특성화된 처리·분석과정을 통해 유의미한 정보를 생성해낼 필요가 있으며, 과거보다 더욱 빠르고 다양하게 생산·소비되는 현대 사회에서 데이터의 효과적인 가공 과정을 통해 가치 있는 정보를 만드는 변화를 빅데이터 기술이 주도할 것이다. 본고에서는 차세대 ICT 융합분야의 핵심 트렌드인 빅데이터 요소기술을 활용단계에 따라 수집, 공유, 저장, 분석, 처리, 시각화의 6개 세부 요소기술로 분류하고 각 기술의 구성과 특성, 추진동향을 살펴보려한다.

1. 서론

SNS를 활용해 고객의 행동 맥락과 감정상태까지 파악할 수 있을 것인가? 국가의 안보, 보건, 에너지, 경제 정책을 수립하기 위해 대중의 여론을 수집할 수 있는 방법은 없을까? 구글은 대체 어떻게 신종 플루의 발병을 예측했을까? 이러한 고민을 해결하기 위해 최근 떠오르는 분야가 바로 빅데이터(Big Data)이다. 이런 질문에, 그리고 훨씬 더 많은 질문에 답하기 위한 핵심 열쇠가 바로 빅데이터이다. 데이터는 우리 일상생활과 언제나 같이 하는 것이며 우리의 생각도, 읽고 접하고 느끼는 것도 모두 데이터라고 할 수 있다. 이러한 데이터에 '지식'이라는 옷을 입히는 것이 바로 빅데이터 기술이다.

빅데이터는 빅 머니가 될 수 있다. 미래학자 앨빈 토플러는 '부의 미래'에서 다가올 '제4 물결'의 시대에 부를 창출하기 위한 핵심 요소는 '지식'이며, 거대하게 유통되는 정보에서 유용한 지식을 구별해 내기 위한 과학적 방법의 중요성을 강조한 바 있다. 현재 스마트 기기 등 단말기나 정보통신기술 서비스를 제공하는 기업들은 데이터를 수집하고 축적해 비즈니스 플랫폼을 만든다. 아마존, 구글, 애플, MS, 페이스북, 트위터도 궁극적으로는 대량의 데이터 분석과 알고리즘으로 비즈니스 모델을 만들고 기업 운영을 첨단화해 세계 최대 일류 기업으로 우뚝 선 것이다. 데이터와 데이터 분석력을 갖추면 어떤 것도 첨단이 될 수 있다.

빅데이터 기술은 방대한 양의 정보를 고속 처리하여 즉시 분석하고, 그로부터 놀라운 결론을 도출해내는 것으로 이를 통해 항공권 가격부터 웹에 있는 사람들의 의견까지 무수한 현상을 검색 가능한 형태로 바꿀 수 있다. 사람들이 만들었던 수많은 데이터를

기반으로 사람과 유사한 결론을 유추해낼 수 있는 것이다[1].

한편, 소셜 네트워크 서비스(SNS)와 스마트폰, 태블릿 PC 등 모바일 기기의 빠른 보급은 사람 사이의 소통 방식과 정보 전달 방식을 변화시켰고 이는 다양한 데이터를 수집, 저장, 처리하는 빅데이터 기술을 촉발하는 계기가 되었다. 또한 기술적으로는 저장 시스템의 비용이 저렴해지고 컴퓨팅 자원이 풍부해지면서 보다 정확한 데이터, 사용자 맞춤형 데이터를 원하게 된 것도 빅데이터 기술의 등장 요인이라 할 수 있다. 앞으로 수 년 혹은 수십년 동안 빅데이터는 비즈니스, 정치, 교육, 건강 등 전반에 걸쳐 사람들이 생각하는 방식 자체를 바꾸어놓을 것이다. 본고에서는 차세대 ICT 융합분야의 핵심 인 빅데이터 요소 기술을 수집, 공유, 저장, 분석, 처리, 시각화의 6개 세부 요소기술로 분류해 세부적인 특징과 기술 동향을 살펴보고 향후 나아갈 방향에 대해 다루고자 한다.

2. 빅데이터의 특징 및 활용 과정, 요소 기술

가. 빅데이터의 특징

빅데이터는 단순히 큰 데이터가 아니라 부피가 크고, 변화의 속도가 빠르며, 속성이 매우 다양한 데이터라는 세 가지 특징을 가진 큰 데이터를 빅데이터로 정의할 수 있다. 원시 데이터는 가공되지 않은 수집 상태의 데이터를 의미하며, 빅데이터에서는 분석하는 분야 및 대상의 특징에 따라 원시 데이터를 가공하기 때문에 가공되지 않은 그대로의 상태로 저장하는 경우가 많다. 빅데이터는 원시 데이터에서 사용자가 원하는 것이 생기면 거기에 해당하는 데이터를 찾아서 분석해 제공하는 것이며 규모(Volume), 변화 속도(Velocity), 다양성(Variety)이라는 3가지 특징, 즉 3V를 가지고 있는 데이터에서 효과적인 결과물을 창출해 낼수 있다. ([그림 1] 참조).

<그림 1> 빅데이터의 3가지 특징(3V) [2]

※ 출처 "빅데이터:산업 지각변동의 진원", 삼성경제연구소, 2012.05.

최근에는 기존의 3V 요소에 데이터의 진실성(Veracity), 시각화(Visualization), 가치(Value) 등을 포함해 빅데이터 기술 요소를 "6V"로 확대해 규명하고 있다.([표 1] 참조). 결국 빅데이터의 핵심적인 특징은 '가공되지 않은 많은 양의 데이터가 빠르고 다양하게 생산·소비되는 분야에서 효과적으로 데이터를 분석·가공하여 가치 있는 정보를 만들어낼 수 있는 데이터'라 할 수 있다. 빠르게 변화하는 ICT 기술에 적극적으로 대응하고 새로운 가치를 찾는 기술을 빅데이터에 적용하여 의미 있는 정보, 가치 있는 정보를 찾아 활용하는 것이 중요하다는 것이다.

<표 1> 빅데이터 기술의 규정 요소(6V) [3]

구 분	내 용
크기(Volume)	방대한 양의 데이터(페타바이트(1PB=10^{15}Byte) 수준)
다양성(Variety)	정형데이터+비정형데이터(소셜 미디어의 동영상, 사진, 대화내용 등)
속도(Velocity)	실시간으로 생산되며 빠른 속도로 분석, 유통
진실성(Veracity)	의사 결정이나 활동의 배경을 고려하여 이용됨으로써 신뢰 제고
시각화(Visualization)	사용자 친화적인 시각적 기능을 통해 빅데이터의 모든 잠재력 활용
가치(Value)	비즈니스에 실현될 궁극적 가치에 중점을 둠

나. 빅데이터의 활용 과정

빅데이터로부터 지식을 발굴해 활용하기까지는 여러 단계를 거치게 된다. 먼저 발굴하고자 하는 지식과 관련된 다양한 데이터 소스로부터 데이터를 수집하고, 수집한 데이터에서 필요 없는 데이터를 필터링하거나 적절한 형태로 가공하는 등 전처리 단계를 거친다. 그 후에 정보를 체계적으로 저장하고 관리하면서 유용한 지식이나 내재된 지식을 얻기 위한 정보 처리 분석 과정을 통해 의미 있는 지식을 발굴·가시화 하게 된다. 빅데이터 활용 단계는 데이터 소스, 지식을 활용하는 서비스 분야가 무엇인지에 따라 일부 단계를 건너뛰거나 반복 수행되기도 한다. 이처럼 빅데이터는 지식을 최종 활용하기까지 데이터 수집, 데이터 전처리, 정보 저장 관리, 정보 처리·분석 및 지식 가시화 등을 통한 여러 단계를 거치고 있다.

다. 빅데이터의 6가지 요소 기술

앞서 설명한 빅데이터의 활용 단계를 바탕으로 빅데이터 세부요소 기술을 분류하면 수집 기술, 공유기술, 저장·관리 기술, 처리 기술, 분석 기술, 지식 시각화 기술의 6가지로 분류할 수 있다.([표 2] 참조). 내·외부 데이터를 수집하여 정제되지 않은 데이터를 확보할 수 있으며, 이를 통해 지식과 지혜를 추출하여 활용하기 위해서는 빅데이터를 효과적으로 저장·관리해야 한다 (여기에서 말하는 데이터 저장·관리란 추후 데이터를 사용할 목적으로 원본 데이터를 안전하게 영구히 보관해 두는 것을 의미한다. 데이터 저

장·관리는 원하는 데이터에 접근하여 그 내용을 읽어올 뿐 아니라 데이터의 수정·삭제도 제공해야 한다)

이렇게 저장된 엄청난 양의 데이터를 처리하는 기술이 빅데이터 처리 기술이다. 과거와는 달리 현재는 모바일 기기와 다양한 ICT 기기를 이용하여 누구나 데이터를 생성하고 접근할 수 있고 이에 따라 생성되는 데이터의 양은 가히 엄청나기 때문에 이러한 빅데이터를 가공하고 분석하는 처리 기술이야말로 빅데이터 산업에 매우 중요한 역할을 차지하고 있다고 할 수 있다. 이렇게 처리된 데이터를 분석하고 시각화하면 수많은 데이터 속에서 인간의 삶과 사회에 의미 있는 정보와 가치들을 쉽게 얻어내고 인식할 수 있게 된다.

<표 2> 빅데이터 요소 기술 구성 및 분류

요소 기술	설명	해당 기술
빅데이터 수집	조직내부와 외부의 분산된 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술로 단순 데이터 확보가 아닌 검색/수집/변환을 통해 정제된 데이터를 확보하는 기술	ETL/크롤링 엔진/로그 수집기/센싱/ RSS, Open API 등
빅데이터 공유	서로 다른 시스템간의 데이터 공유	멀티 테넌트 데이터 공유/ 협업 필터링 등
빅데이터 저장	작은 데이터라도 모두 저장하여 실시간으로 저렴하게 데이터를 처리하고, 처리된 데이터를 더 빠르고 쉽게 분석하여, 이를 비즈니스 의사 결정에 바로 이용하는 기술	병렬 DBMS/하둡(Hadoop)/ NoSQL 등
빅데이터 처리	엄청난 양의 데이터의 저장·수집·관리·유통·분석을 처리하는 일련의 기술	실시간 처리/분산 병렬 처리/ 인-메모리 처리/인-데이터베이스 처리
빅데이터 분석	데이터를 효율적으로 정확하게 분석하여 비즈니스 등의 영역에 적용하기 위한 기술로 이미 여러 영역에서 활용해 온 기술임	통계 분석/데이터 마이닝/ 텍스트 마이닝/예측 분석/ 최적화/평판 분석/소셜 네트워크 분석 등
빅데이터 시각화	자료를 시각적으로 묘사하는 학문으로 빅데이터는 기존의 단순 선형적 구조의 방식으로 표현하기 힘들기 때문에 빅데이터 시각화 기술이 필수적임	시간시각화/분포시각화/관계 시각화/비교시각화/공간시각화/인포그래픽

3. 빅데이터 요소 기술별 특성과 동향

앞서 설명했듯이 빅데이터를 구성하는 요소를 충족시킬 수 있는 기술로는 수집 기술, 공유 기술과 저장 기술, 처리기술, 분석 기술, 시각화 기술이 있다. 이 기술들을 통하여 빅데이터 시스템은 외부로부터 필요한 데이터를 수집하고 이를 저장하며, 저장된 데이터를 공유할 수 있다. 따라서 이러한 기술들은 빅데이터의 6V 요소를 충족시키며 보다 더 빠르게 다양한 데이터를 수집, 저장, 공유할 수 있도록 개발되어야 한다.

가. 빅데이터 수집 기술

빅데이터 수집기술은 조직내부와 외부의 분산된 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술로 단순 데이터 확보가 아닌 검색, 수집, 변환을 통해 정제된 데이터를 확보하는 기술을 의미한다.

데이터의 수집이란, 단순히 기업의 내부와 외부의 분산된 다양한 데이터를 검색하는 것뿐만 아니라 스크라이브(Scribe), 척와(chukwa), 플룸(flume) 등의 다양한 데이터 수집 기술을 통해 수집하고 이렇게 수집한 데이터를 변환, 저장하는 과정까지 포함된다.

또한, 더욱 가치 있는 데이터를 산출하기 위해 변환, 저장, 분석 과정을 반복하게 된다. 일반적인 조직 내부에 존재하는 정형 데이터는 로그 수집기를 통해 수집하며 조직 외부에 존재하는 비정형 데이터는 크롤링, RSS Reader, 또는 소셜 네트워크 서비스에서 제공하는 Open API를 이용한 프로그래밍을 통해 수집한다.

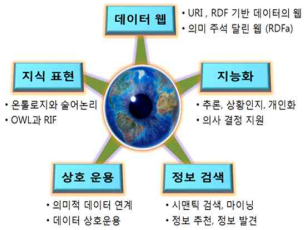
<그림 2> 데이터 수집 방법 [4]

※ 출처 "Big Data Capture", http://www.citsoft.net/?page_id=523

나. 빅데이터 공유 기술

기업 내 운영 환경에서 한 데이터베이스 시스템에서 발생하거나 변경된 데이터를 다른 시스템에 적용하려는 분산 및 복제 환경은 보편화 되어 있으며, 서로 다른 시스템 간의 데이터 공유는 빅데이터 시스템에 있어서도 중요하고 필요한 이슈 중의 하나이다. 데이터 공유를 위한 가장 일반 적인 형태로서는 운영계 시스템의 데이터 복제(Replication) 기술과 정보계 시스템을 위한 데이터 웨어하우스의 ETL(Extract, Transformation, Load) 프로세스가 대표적이라 할 수 있다. 빅데이터 공유 기술로는 복제(Replication), 시맨틱 기술, 멀티 테넌트 데이터 공유 기술, 협업 필터링을 들 수 있다.

이중 시맨틱 기술은 시멘틱 웹으로 대변되며 정보의 표현을 넘어 인간 지식을 명시적으로 표현, 공재활용 할 수 있는 기술이다. ([그림 3]참조)[5]. 데이터 상호 운용 및 데이터 모델 관점에서 시맨틱 웹은 새로운 산업적 가능성을 제시하고 있다. 기존 ER모델 기반 table 구조의 데이터베이스는 그래프 형태의 데이터를 표현하기 적합하지 않은 면이 있으나, 의미 메타데이터 및 그래프 구조의 메타데이터 체계로서 시맨틱 웹 기술은 데이터 상호 운영과 SEA(Semantic Enterprise Architecture)와 같은 시스템 아키텍처 구현에 중요한 기반이 되고 있다.

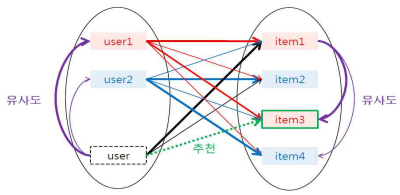


<그림 3> 시맨틱 기술 개념도

멀티테넌트 데이터 공유기술은 멀티테넌트 환경의 데이터 관리 기술로써 데이터와 데이터스키마를 분리 혹은 공유하여 멀티 테넌트 환경의 데이터 공유를 가능케 하고, 공유에 의한 보안 요소 검증 기술이 포함된다. 해당 기술을 통해 테넌트 별 분산 저장된 데이터를 공유하여 각각의 테넌트가 가지고 있는 데이터의 양이 줄어들고 효율적인 데이터 관리가 이루어진다[6].

협업 필터링은 많은 사용자들로부터 얻은 기호정보(taste information)에 따라 사용자들의 관심사들을 자동적으로 예측하게 해주는 방법으로 사용자들의 과거의 경향이 미래에서도 그대로 유지 될 것이라는 전제가 있어야 한다. 예를 들어, 음악에 관한 협력 필터링 혹은 추천시스템(recommendation system)은 사용자들의 기호(좋음, 싫음)에 대한 부분적인 목록(partial list)을 이용하여 그 사용자의 음악에 대한 기호를 예측하게 된다. 이 시스템은 특정 사용자의 정보에만 국한 된 것이 아니라 많은 사용자들로 부터 수집한 정보를 사용한다는 것이 특징이며 단순히 투표를 한 수를 기반으로 각 아이템의 관심사에 대한 평균적인 평가로 처리하는 방법과는 다른 것이다. 이 기술은 비슷한 취향을 가진 고객들에게 서로 아직 구매하지 않은 상품들을 교차 추천하거나 분류된 고객의 취향이나 생활 형태에 따라 관련 상품을 추천하는 형태의 서비스를 제공하

기 위해 사용 된다[7].



<그림 4> 협업 필터링의 예

다. 빅데이터 저장 기술

빅데이터 저장 기술은 작은 데이터라도 모두 저장하여 실시간으로 저렴하게 데이터를 처리하고, 처리된 데이터를 더 빠르고 쉽게 분석하도록 하여 비즈니스 의사 결정에 바로 이용하도록 만드는 기술이다. 이러한 빅데이터들을 저장하기 위해 많은 요소기술 및 솔루션이 존재하고 개발 중에 있다.

빅데이터 저장기술은 구글이나 애플, 야후 등에 의해 요소기술로서 상당한 완성도에 도달했으며 오픈 소스로 Hadoop의 HDFS/Hbase, Cassandra, MongoDB 등이 대표적이다. 또한, 우리나라에서도 ETRI의 GloryFS등과 같은 많은 솔루션도 존재한다.

병렬 DBMS와 NoSQL은 대량의 데이터를 저장하기 위해 수평 확장 접근 방식을 사용하는 기술이며 동일한 분산처리 기술인 하둡(Hadoop)은 저렴한 비용으로 빅데이터 시스템을 구축할수 있는 장점 때문에 현재 빅데이터의 대표적인 기술로 자리잡고 있다.

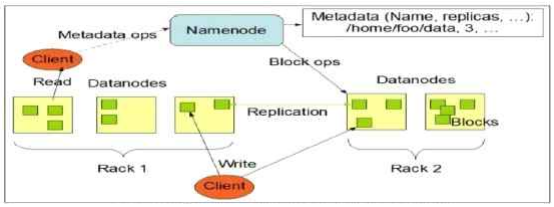
이외에도 SAN(Storage Area Network), NAS(Network Attached Storage)과 같이 기존 저장 기술도 있고, Amazon S3나 OpenStack Swift와 같은 클라우드 파일 저장 시스템, GFS(Google File System), HDFS(Hadoop Distributed File System)와 같은 분산 파일 시스템 등이 모두 대량의 데이터를 저장하기 위한 기술이다.

라. 빅데이터 처리 기술

빅데이터 처리기술은 엄청난 양의 데이터를 저장·수집·관리·유통·분석을 처리하는 일련의 기술로써 수많은 사용자 요청을 실시간으로 처리한 후 처리 결과를 반환하는 기술이다. 대량의 데이터를 분산 환경에서 병렬로 처리하여 수평적인 확장 접근 방식을 취하는 한편 방대한 양의 데이터를 하드디스크가 아닌 메모리에 보관하여 실시간으

로 분석할 수 있도록 하는 인-메모리 기술과 DB 내부에서 직접 분석로직을 실행하는 인-데이터베이스 기술이 대표적인 기술이라 할수있다. 빅데이터 처리기술은 분산처리 시스템인 하둡과 MongoDB, Kafka, 병렬처리 모델인 맵리듀스, 인메모리 기술인 에스퍼와 레디스 등으로 발전하고 있다. SAS, IBM 등 많은 글로벌 기업들이 빅데이터 분석 솔루션을 개발 및 연구 중이며 비즈니스 분석을 위한 통합 솔루션의 기반을 다지고 보다 빠르고 정확하게 데이터를 분석하기 위한 기술들을 개발하고 있다. 대표적인 통합 분석 솔루션인 SAS Analytics의 경우 데이터/텍스트 마이닝, 데이터 시각화, 콘텐트 분류, 예측, 모델 관리 및 전개, 경영 과학(최적화), 품질 개선, 통계 등과 같이 대부분의 분석 기술 통합 제공하기 위한 정책을 추진중이다.

분산처리 기술중 오픈소스인 아파치 하둡(Hadoop)은 구글의 GFS를 대체할 수 있도록 분산 파일시스템(HDFS)과 MapReduce를 구현한 빅데이터 처리 기술의 대표적인 프레임워크로 하둡의 분산 파일시스템은 파일을 블록단위로 나누어 각 노드 클러스터에 저장을 하며, 데이터 유실을 막고 부하처리를 위해서 각 블록의 복사본(Replication)을 생성한다.([그림 5]참조)[8].

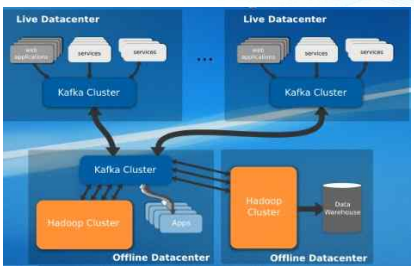


<그림 5> 하둡(hadoop)의 분산 파일시스템(HDFS) 처리구조

MongoDB는 Replica Sets와 Auto-Sharding으로 구성된 데이터베이스로써, Replica Sets으로 안정성과 가용성을 확보하고 Auto-sharding으로 분산 확장을 하는 구조로 이루어져 있다. Config Server에 메타데이터를 저장하고 실제 데이터는 Chunk단위로 나뉘어 Shard서버에 저장하며, Auto-Sharding으로 분산 확장을 통해 데이터의 유실을 막고 가용성을 높이기 위한 각 서버의 Replica를 설정한다.

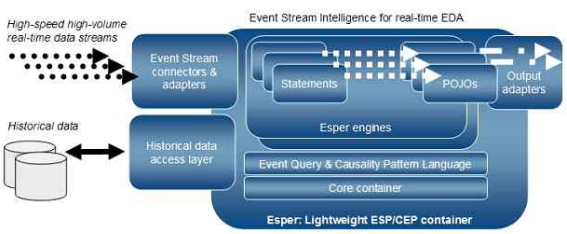
분산 처리 시스템인 "Kafka"는 링크드인에서 공개한 기술로서 데이터가 분할되어 클러스터에 의해 관리되고 메시지 일괄 처리와 전송데이터를 압축하는 기술을 지원한다. 로그 집계 메시징 시스템이 결합된 형태로써 분산 환경에서 대규모, 대용량 데이터를 처리한다.([그림 6]참조). 최근 DEVIEW 2013의 발표에서는 Netflix에서도 카프카를 활

용해 데이터 로그를 처리한다고 공개하는 등 그 활용범위가 점차 확대되고 있다.



<그림 6> 링크드인 카프카(Kafka)의 설계 구조[9]
※ 출처 Richard Park, "SDEC 2012 : Apache Kafka", 2012.09.

에스퍼 (Esper)는 실시간 처리를 위한 인-메모리 기술 중 CEP(Complex Event Processing)처리 기술로 CEP란 여러 이벤트 소스로부터 발생한 이벤트를 대상으로 의미 있는 데이터를 추출하여 대응되는 액션을 수행하는 오픈소스 기술이다. 실시간 스트리밍 데이터를 처리하는 이벤트 기반 아키텍처로 구성되어 있으며 특정 이벤트만 필터링해서 필요한 작업을 수행한다.



<그림 7> Esper의 이벤트 흐름 아키텍처[10]
※ 출처 EsperTech, http://esper.codehaus.org/

레디스 (Redis)는 “Remote dictionary System”의 약자로 BSD 기반의 오프소스로 VMWare가 인수해서 업그레이드를 진행하고 있다. 인-메모리 기술로서 키/값 저장소로 구성되어 있어 NoSQL 데이터베이스로 분류하기도 한다. 메시지 큐(Message Queue), 공유 메모리(Shared Memory) 용도로 사용될 수 있어 인스타그램, Stack Overflow, Digg와 같은 기업들이 실시간 데이터 스트리밍 처리를 위해 활용되고 있다.

마. 빅데이터 분석 기술

현대 사회에서 불확실성이 높고 의사결정이 초래하는 파급효과가 큰 의사결정일수록 실제 데이터 분석을 바탕으로 의사결정을 해야 한다. 실제로 비즈니스 전략으로써 많은 기업에서는 본능적인 비즈니스 감각보다는 빅데이터를 활용하여 주요 의사결정을 내리고 있다. 빅데이터 분석기술로는 빅데이터 통계분석, 데이터 마이닝, 텍스트 마이닝, 예측 분석, 최적화, 평판 분석, 소셜 네트워크 분석 등이 있다.

먼저, 통계 분석은 다양한 분석에서 활용되는 기술로써 통계적 컴퓨팅에 사용되는 R, SAS 등을 통하여 다양한 통계기법으로 분석할 수 있다. 다양한 통계 기법을 통하여 목적에 맞는 가장 정확하고 효율적인 통계 분석을 수행할 수 있게 된다.

<표 3> 빅데이터 분석을 위한 대표적인 통계기법

통계 기법		설명
고차원 회귀/분류 분석	Lasso	모형의 예측 성능 향상과 변수선택을 동시에 할 수 있는 기법
	Ensemble	단순 모형을 결합하여 고성능 모형을 찾는 방법
	SVM(Support Vector MAchine)	주어진 자료들을 분리하는 가장 좋은 초평면을 찾는 방법
군집 분석	K-means clustering	군집의 중심과 자료와의 거리를 최소화
	Hierarchical clustering	계층 구조를 갖는 군집화 방법
	Model based clustering	혼합 모형에 기초한 군집 분석으로 자료를 군집화하는 동시에 각 군집의 분포를 추정

데이터 마이닝은 통계 및 수학적 기술뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미 있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정으로 다양한 분야에서 활용될 수 있으며, KDD, 기계학습, 패턴인식, 통계학, 신경망 컴퓨팅 등과 관련하여 빅데이터 분석에 있어서 가장 기본적인 분석 기술이라 할 수 있다. 데이터마이닝의 주요 기법으로는 OLAP(On-Line Analytical Processing), 군집 분석(Cluster Analytics), 연결 분석(Link Analytics), 사례기반 추론(Case-Based Reasoning), 연관성 규칙 발견(Association Rule Discovery), 인공 신경망(Artificial Neural Network), 의사 결정 나무(Decision Tree), 유전자 알고리즘

(Genetic Algorithm) 등이 있다. 텍스트 마이닝은 구조화되지 않은 대규모의 텍스트 집합으로부터 새로운 지식을 발견하는 과정으로 텍스트 문서 전처리 및 패턴 분석 등의 단계를 가지며, 순환 구조로써 계속적인 피드백을 수행한다.

예측 분석은 과거 자료와 변수 간의 관계를 이용하여 관심이 되는 변수를 추정하는 것으로써 앞서 언급된 통계 분석, 데이터 마이닝 및 텍스트 마이닝 기술들을 기반으로 예측 분석을 수행하게 된다.

<표 4> 예측분석 주요 기법

구분	설 명
추세분석(trend analysis)	기술적 분석의 출발점으로써, 다항 회귀 모형과 유사한 모형을 가정하고 모수의 추정을 통해 예측값을 구하는 분석법
평활법(smoothing method)	현재로부터 가장 최근에 관측된 자료에는 큰 가중값을 주고, 과거로 갈수록 그 가중값의 크기를 줄여나가는 일종의 가중평균을 이용한 예측방법
자기회귀누적이동평균(ARIMA)모형에 의한 분석법	현 시점의 관측값을 과거의 관측값들과 백색잡음이라고 불리는 오차들의 형태로 표현하는 모형으로써, 박스-젠킨스모형이라는 이름으로 가장 많이 사용되고 있음

최적화는 주어진 가능한 결과들에 대한 평가를 수행하여 최적의 결과를 도출하는 것으로 비즈니스 환경에서 취할 수 있는 여러 가지 대안들 중 제시된 전략을 평가하고 최적의 대안을 선택하도록 도와줄 수 있는 필수적인 분석 기술이다.

평판분석은 소셜 미디어 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별하는 분석 기술로 주로 특정 서비스 및 상품에 대한 시장규모 예측, 소비자의 반응, 입소문 분석 등에 활용된다.

소셜 네트워크 분석은 소셜 네트워크 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하는 기술로써, 수학의 그래프 이론에 뿌리를 두고 있다. 주로 마케팅을 위하여 소셜 네트워크상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾는데 주로 활용된다.

그 밖에 실시간 분석, 준 실시간 분석 등과 같은 정확성 보다는 분석 속도에 초점을 두는 분석 기술도 있다. 분석에 필요한 모든 가용한 데이터를 활용하여 사용자의 요청에 대한 분석을 수행하고 빠르고 적시에 지식(분석 결과)을 제공해 줄 수 있는 분석 기법과 실시간 분석 등을 위하여 인-데이터베이스 분석, 인-메모리 분석, 다중 프로세스를 활용하는 MPP 등과 같이 보다 빠른 지원 기술을 이용할 수 있다.

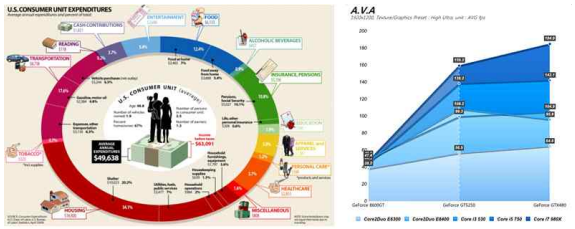
바. 빅데이터 시각화 기술

빅데이터 시각화 기술은 수많은 데이터들이 만들어 내는 무질서한 흐름 속에서 숨겨진 패턴을 발견하여 사람들이 쉽게 알 수 있도록 표현하는 기술로서, 빅데이터를 어떻게 다양한 시각화 도구로 표현하느냐에 따라 사용자들이 얻을 수 있는 직관이 달라진

다. 시각화 기술은 시간 시각화, 분포 시각화, 관계 시각화, 비교 시각화, 공간 시각화, 인포그래픽 등의 기법이 있다.

먼저 시간 시각화 기술은 분절형과 연속형으로 구분되는데 분절형은 데이터는 특정 시점 또는 특정 시간의 구간 값(예:어떤 시험의 평균 통과율)을 막대그래프, 누적 막대그래프, 점그래프 등으로 표현한다. 연속형은 기온 변화 같이 지속적으로 변화하는 값, 시계열 그래프, 계단식 그래프, LOESS 곡선 추정 등으로 표현하는 방식이다.

분포 시각화기술은 전체 분포와 시간에 따른 분포로 나누어진다. 최대, 최소, 전체분포를 나타내는 그래프로 전체의 관점에서 각 부분 간의 관계를 보여주는 전체 분포는 파이 차트, 도넛 차트, 누적 막대그래프, 인터랙티브 누적 막대그래프 등으로 표현된다. ([그림 8] 참조). 시간에 따른 분포는 1900년부터 2000년까지의 연령별 한국 인구 분포와 같이 시간에 따라 어떤 변화가 있었는지 나타내는 기술로 누적 연속 그래프, 누적 영역 그래프, 인터랙티브 누적 영역 그래프, 선 그래프 등으로 표현할 수 있다.

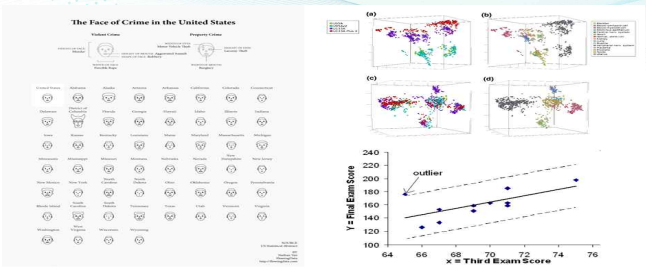


<그림 8> 도넛 차트와 누적 영역 그래프의 예

관계 시각화 기술은 각기 다른 변수 사이에서 관계를 찾는 기술로 상관관계, 분포, 비교로 구분할 수 있다. “상관관계”는 스캐터플롯, 스캐터플롯 행렬, 버블차트 등으로 표현할 수 있다.

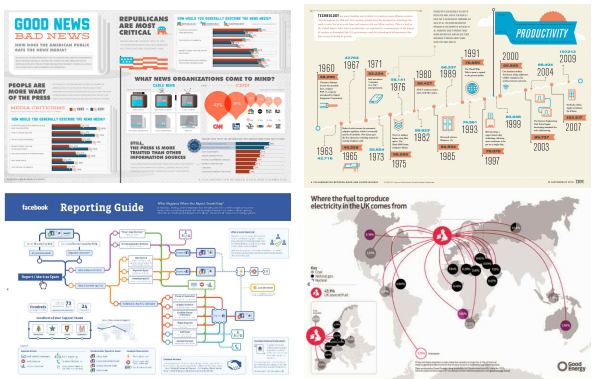
비교 시각화 기술은 여러 변수의 비교하는 히트맵, 체르노프 페이스, 스타 차트, 평행좌표 그래프, 다차원척도법(Multi-Dimensional Scaling)[1], 아웃라이어(Outlier) 찾기[2] 등으로 표현되어 진다. ([그림 9] 참조). 공간 시각화는 위치를 점이 찍힌 지도, 선을 그린 지도, 버블을 그린 지도 등으로 특정하고 색상으로 영역을 구분한다. 시간과 공간에 따라 작은 지도를 하나로 그려 패턴의 변화를 보여주는(예: 실업률 변화 등) 스몰 멀티플이나 애니메이션 확산 지도를 활용한다.

1) 여러 가지 변수에 따라 데이터를 배치할 경우에 사용되는 기술
2) 나머지 데이터와 동떨어져 있는 소수의 데이터를 찾아 원인을 확인해보는 기술



<그림 9> 체르노프 페이스, 다차원 척도법, 아웃라이어 찾기의 예

인포그래픽(infographic)은 인포메이션(information)과 그래픽(Graphic)의 합성어로 다량의 정보를 차트, 지도, 다이어그램, 로고, 일러스트레이션 등을 활용하여 한눈에 파악할 수 있다.([그림 10] 참조).



<그림 10> 통계, 타임라인, 프로세스, 위치 기반 인포그래픽의 예

4. 빅데이터 요소기술의 발전 방향 및 제언

빅데이터는 로그, DB, SQL 등의 구조적 데이터와 소셜미디어, 센서, 멀티미디어 등의 비정형 데이터의 조합이다. 실제로 데이터들은 인덱스(index)나 기타 유기적인 구조를 갖추지 못했으며 다양한 종류의 파일들로 구성되어 있다. 이런 이유로, 빅데이터를 다루는 일은 현재의 기술수준으로 매우 까다로운 작업이다. 데이터의 일관성이 부족하기 때문에 기존의 정보 처리나 저장 기술만으로는 관리가 어렵고, 방대한 데이터의 양 자체만으로도 보통 서버나 SAN 방식[3]을 사용해 처리하는 것이 불가능하다. 이러한 복잡성 때문에 빅데이터 시대를 이끌고 있는 하둡 플랫폼도 최근에는 피그(pig), 하이브(Hive) 등 초창기 하둡의 단점을 보완하는 보조 툴이 등장하며 진화하고 있고, 데이터 분석 시 작업 효율성을 증대시킬 수 있는 아파치 우지(Apache Ooize) 등의 툴도 속속 등장하고 있다. 앞으로도 점차 작업 효율성과 개발 편의성을 증대시킬 하둡 기반의 다양한 툴이 등장할 것으로 예상된다.

데이터 공유와 관리 측면에서 살펴보면 기업이 보관하던 개인 데이터를 사용자가 유지관리하고 제품 및 서비스 이용 시 필요한 데이터를 제공하도록 사용자 권한도 강화될 것이다. 이에 개인 데이터 공유·거래를 위한 기술 개발 및 상호운용성 보장을 위한 표준화도 중요한 역할을 차지할 것으로 전망된다. 개인적 데이터 흐름을 제어하고 통제할 수 있는 VRM[4] 기술을 개발하여 표준화하고 다양한 시범사업을 통해 검증해야 할 것이다.

한편 2010년경부터 국내에 소개된 빅데이터의 개념은 현재까지도 높은 관심을 끌고 있지만 아직까지 기업에서 실질적으로 빅데이터 플랫폼 구축은 쉽지 않은 상황이다.
기업이 보유하고 있는 인프라보다 훨씬 광범위한 데이터와 컴퓨팅리소스를 필요로 하여 비용측면에서 시스템 도입에 부담을 주고 있고 또, 대용량 데이터를 어떻게 활용하고 실제 비즈니스에 어떻게 적용하여 수익을 얻을 것인가에 대한 확신이 분명하지 않은 것도 기업에서 시스템 도입을 주저하는 이유라 할 것이다.
따라서 빅데이터 기술도 향후에는 IT 엔지니어나 데이터 사이언티스의 실험을 위한 성격에서 벗어나 기업에서의 활용을 위해 편리하고 안정적인 사용을 보장하는 쪽으로 진화할 필요가 있다.

데이터 수집 기술을 포함해 빅데이터 처리 각 단계별 요소 기술의 고도화를 위한 몇 가지 제언은 다음과 같다.

3) SAN(storage area network): 서로 다른 종류의 데이터 저장장치를 한 데이터 서버에 연결하여, 총괄적으로 관리해주는 네트워크
4) VRM(Visitor Relationship Management): 웹사이트 방문자들의 행동과 성향 분석을 기반으로 마케팅하는 고객관리 시스템

가. 빅데이터 수집시 데이터 규격의 호환성 및 유연성 확보

데이터 수집 기술은 대부분 ftp를 사용하여 로그 데이터 서버 혹은 일반 서버로부터 로그 정보와 기타 데이터 처리에 필요한 데이터를 가져오게 된다. 이 때 전송해 오는 데이터의 경우 무결성 확인을 위해 항상 데이터를 체크해야 하며 빅데이터 시스템의 데이터 수집의 경우에는 짧은 시간에 대용량의 데이터 수집이 일어나게 된다. 또, 모든 데이터가 동일한 규격을 사용하지 않는데 예를 들어, 로그 데이터만 해도 각각의 서비스들에서 자신만의 고유 로그 규격을 사용하여 수많은 로그를 저장하고 있다.
따라서 데이터 수집 기술은 서로 다른 규격의 데이터들에 대한 유연성을 제공해야 하며, 동시에 시스템이 요구하는 최대 지연 시간을 준수할 수 있어야 한다.

현재 실시간 분석 기술과 같은 실시간성이 요구되는 기술들이 실시간 정보를 제공하는 혹은 분석하는 서비스들과 함께 부각되고 있기 때문에 이에 맞추어 데이터 수집 기술 역시 실시간 수집이 이루어질 수 있도록 각종 데이터 규격에 대한 호환성과 유연성, 그리고 무결성이 지켜질 수 있도록 추가적인 개발과 연구가 이루어져야 할 것이다.

나. 데이터 공유·거래 프레임워크 설정

현재 빅데이터 공유 및 활용에 있어 가장 큰 이슈 중의 하나는 정보 보호라 할 수 있을 것이다. 빅데이터 시대의 도래와 함께, 가공한 정보에 대한 소유권은 누가 갖는지, 기업 등에 남아있는 개인정보에 대해 삭제할 권리는 없는지 등 데이터 권리에 대한 다양한 이슈가 불거져 나오고 있다. 특히 정보의 소유와 관리의 분리를 통한 효율성 제고를 표방하는 클라우드 컴퓨팅 서비스의 확대로 데이터 주권에 대한 문제가 부각되고 있다. 특히 빅데이터의 유형중에서 소셜 미디어는 개인 정보와 행동, 배경 등의 자료가 고스란히 노출되고 정보가 공유되어 개인 프라이버시 침해, 감시의 일상화 등의 이슈가 제기된다. 이를 해결하기 위해서는 개인 데이터의 안전한 흐름을 보장하면서, 이를 활용하여 새로운 부가가치를 창출할 수 있도록 데이터 공유·거래 프레임워크를 설정해야 한다. 기업이 보관하던 개인 데이터를 사용자가 유지관리하면서 제품 및 서비스 이용 시 필요한 데이터를 제공하도록 사용자 권한을 강화하고, 제품 및 서비스 제공업체는 사용자가 제공한 직접적인 요구사항과 개인 데이터를 분석하여 고객 맞춤형 서비스를 제공하는 등 혁신을 촉진하여야 한다. 이러한 프레임워크가 구축되면 향후에는 제3자가 개인 데이터 관리 서비스를 제공하는 새로운 시장이 창출될 수도 있을 것이다.

다. 실시간 분석처리를 위한 기술 고도화

데이터 활용 방식의 변화로 현재 발생하는 상황 파악 및 발생 원인의 실시간 분석에

대한 중요성이 커짐에 따라 대규모 스트림 처리 기술에 대한 연구가 필요하다. 실시간 데이터 처리 기술의 활용이 확대되면 더 큰 메모리에 대한 요구 및 메모리 소비 전력 절감에 대한 필요성이 커질 것으로 예상되므로 요구되는 처리 속도 및 처리 로직 등을 고려하여 차세대 비휘발성 메모리 활용에 대한 고려가 필요할 것으로 보인다. 또한 빅데이터 처리 기술의 연구는 대규모 분산 처리 및 빅데이터 저장, 연동을 위한 분산형 데이터 클라우드와의 연계를 고려한 연구방향 설정이 필요하다. 이를 위해 이종 데이터를 동시 처리하는 새로운 데이터 관리 소프트웨어 개발, 저전력·고성능화를 위한 새로운 고성능 컴퓨팅 기술 개발, Foresight·Insight 발굴을 위한 복합형 인공지능 기술 개발 등을 통한 데이터 분석처리 기술 개선이 필요하다.

라. 빅데이터 시각화 기술 확대 필요

빅데이터 시각화는 데이터에서 통찰을 얻는 가장 효과적인 방법인데 원시 데이터를 가공 및 분석하는 전통적인 방법론을 기반으로 한 통계분석 등의 고급분석을 위해서는 전문가가 필요하나 시각화된 데이터는 일반 대중들도 쉽게 이해할 수 있다.

따라서 빅데이터의 대중적인 활용도를 높이기 위해서 일반 사용자들도 쉽게 이해할수 있는 시각화 기술 및 서비스를 더욱 연구하고 개발할 필요가 있다. 또한, 데이터 사이언티스트나 IT 전문가가 아닌 일반 현업 사용자들도 직접 손쉽게 데이터를 탐색하고 보고서를 생성 및 공유할 수 있도록 친화적인 시스템으로의 대중화가 필요하다.

5. 국내 빅데이터 기술 동향과 전망

작년 말 발표된 국가정보화 빅데이터 마스터플랜에 따르면 국내 빅데이터 기술은 구글, 오라클, IBM 등 글로벌 기업대비 2~6년의 기술 격차가 있는 것으로 추정된다[11]. 아울러 국내 빅데이터 기업의 기술은 시작단계에 불과해 세부 요소 기술에 대한 체계적인 분석도 부족한 실정이다. 국내 빅데이터 기술력 강화를 위해서는 선진 기술 대비 국내 기업이 경쟁력을 가질 수 있는 빅데이터 핵심 기술 분야에 대한 핵심원천 기술 개발이 우선적으로 이루어져야 한다. 중점 기술은 데이터 수집을 위한 데이터 가상화와 융합형 API, 가상화 기반 분산형 메모리 시스템, 공정오류를 예측하는 기술 등 유망한 국내 핵심 기술을 선정해 집중적인 투자가 필요할 것이다. 핵심 기술개발과 함께 국내 기업들이 상대적으로 더욱 뒤처진 '빅데이터 플랫폼' 부문에도 지속적인 투자 및 기술개발이 병행되어야 할 것이다. 국내 빅데이터 기업은 빅데이터 아키텍처 및 플랫폼 기술력이 미흡해 관련 기술 및 부가적인 IT 서비스 산업 발전을 위한 중적적인 기

술 개발이 시급한 실정이기 때문이다. 또한, 지능형 분석과 예측을 위한 기계학습 등 인공지능 분야는 국내에 기반 기술과 원천 기술 자체가 부재하므로 이에 대한 전략적 접근도 이루어져야 할 것이다.

글로벌 기업들이 국내 빅데이터 비즈니스 분야와 공공 빅데이터 처리 분야에서 이미 초기 시장을 선점하고 있다는 점은 국내 빅데이터 기술경쟁력 강화 정책을 추진함에 있어 걸림돌로 작용하고 있다. 국내 기업은 기술력 뿐 아니라 자금력의 한계로 주로 HP, EMC, SAS, Oracle 등 글로벌 기업이 선전한 빅데이터 기술과 오픈소스 솔루션에 의존하고 있는 실정이다. 실제 국내 대표기업에서 적용중인 소프트웨어 원천기술을 살펴보면 SK텔레콤이 오라클 데이터 분석기술을, LG전자와 LIG 생명보험은 IBM 기술을, 우리은행과 신한지주는 테라데이터 기술을 사용하고 있는 등 외산 기술이 이미 국내 시장을 잠식하고 있는 실정이다. 따라서, 정부에서는 공공부문을 필두로 국내 우수 빅데이터 분야에 대한 SW 활용 연계를 강화할 필요가 있을 것으로 보인다. 얼마전 출범한 빅데이터 분석활용 센터에서 국산 SW를 검증하여 빅데이터 시범사업 등 정부사업에서 활용 할 수 있도록 지원하고, 공공조달 시장에서 국내 기업의 기술력을 적용한 SW 개발이 우선적으로 이루어 질수 있도록 점진적인 국내SW 연계 정책이 이루어져야 할 것이다.

데이터 관리 전문인력 양성과 함께 다양한 빅데이터 요소 기술을 다룰 수 있는 융합형 인재양성도 중요한 과제이다. 우리의 빅데이터 기술 분야 경쟁력 확보를 위해서는 수집, 저장, 분석, 시각화 등 단일 요소 기술 분야의 기술고도화가 우선적으로 필요하다. 하지만 장기적인 관점에서는 ICT와 데이터 분석 기술 전문가, 혹은 통계학과 시각화 기술의 전문가 등 빅데이터와 관련된 다양한 분야와 융합된 인재양성을 통해 중장기 기술경쟁력 확보가 가능할 것이다.

마지막으로 국제적인 데이터 공유 및 활용을 위한 협력관계를 구축하여 빅데이터 공유에 따른 이익을 최대화하면서도 개인정보의 국외 이전에 따른 소비자 편익과 개인정보 보호의 균형을 이룰 수 있도록 해야 할 것이다. 미래에는 데이터의 국외이전에 대한 규제가 실질적인 무역규제의 한 형태로 논의될 가능성이 있으므로 이에 대한 적극적인 준비가 필요하다. 이를 위해, 외국에서 수집하는 데이터를 국내로 이전하는 경우에 발생할 수 있는 문제해결을 위한 제도적 기반도 점차 마련해 나가야 할 것이다.

[참고문헌]

[1] 빅토르 마이어 쉔버거, 케네스 쿠키어, “데이터는 알고 있다 - 빅데이터가 만드는 세상”, (주)북이십일 21세기북스, 2013
[2] 채승병, 안신현, 전상인, “빅데이터:산업 지각변동의 진원”, 삼성경제연구소, 2012.05.
[3] 범지인, 송두한, 최성종, “빅데이터 활용 사례와 시사점”, 농협경제연구소, 2013.08.28.
[4] “Big Data Capture”, http://www.citsoft.net/?page_id=523
[5] “다시 보는 시맨틱 웹 그리고 시맨틱 기술 I”, 한국데이터베이스진흥원, 2013.01.
[6] Frederick Chong, Gianpaolo Carraro, and Roger Wolter, “멀티 테넌트 데이터 공유 기술”, http://msdn.microsoft.com/en-us/library/aa479086.aspx, 2006,06.
[7] 위키피디아, “협업 필터링”, http://ko.wikipedia.org/wiki/%ED%98%91%EC%97%85_%ED%95%84%ED%84%B0%EB%A7%81
[8] “빅데이터 시대 도래에 따른 데이터 처리기술 현황과 전망”, 보안뉴스, 2011.12.12.
[9] Richard Park, “SDEC 2012 : Apache Kafka”, 2012.09.
[10] EsperTech, http://esper.codehaus.org/
[11] “스마트 국가 구현을 위한 빅데이터 마스터플랜”, 교육과학기술부 등 5개 부처 합동, 2012.11.28.

방송통신기술 이슈&전망

방송통신기술 시장·정책 콘텐츠

발 행 호 ▌2013년 제 10 호
발간물명 ▌빅데이터 활용단계에 따른 요소기술별 추진동향과 시사점

본 자료는 http://www.kompass.or.kr에서 확인 하실 수 있습니다.

138-950 서울시 송파구 중대로 135 IT벤처타워 한국방송통신전파진흥원 R&D기획본부