DBpia

빅 데이터의 분석과 활용

Analysis and Utilization of Big Data

저자 (Authors)
강만모, 김상락, 박상무
Man-Mo Kang, Sang-Rak Kim, Sang-Moo Park

출처 (Source)
정보과학회지 30(6), 2012.6, 25-32(8 pages)
Communications of the Korean Institute of Information Scientists and Engineers 30(6), 2012.6, 25-32(8 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01879803

APA Style
강만모, 김상락, 박상무 (2012). 빅 데이터의 분석과 활용. 정보과학회지, 30(6), 25-32

이용정보 (Accessed)
이화여자대학교
211.48.46.***
2020/01/08 16:38 (KST)

저작권 안내
DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information
Copyright of all literary works provided by DBpia belongs to the copyright holder(s)and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s)for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

빅 데이터의 분석과 활용

울산대학교 ▌강만모 · 김상락 · 박상무

1. 서 론

최근 IT 분야의 화두가 무엇인지 물어본다면, 빅 데이터가 대답들 중 하나일 것이다. 20년 전의 PC의 메모리, 하드디스크의 용량과 최신 PC, 노트북 사양을 비교해보면 과거에 비해 데이터가 폭발적으로 늘어났다는 것을 실감할 수 있을 것이다. 특히 스마트 단말기 및 SNS 등으로 대표되는 다양한 정보 채널의 등장과 이로 인한 정보의 생산, 유통, 보유량의 증가로 인하여 데이터가 기하급수적으로 증가하고 있다.

빅데이터란 무엇인가? 빅데이터는 일반적인 데이터베이스, 소프트웨어로는 관리하기 어려운 정도의 큰 규모로서, 현재 수십 테라바이트에서 향후 페타바이트, 엑사바이트 정도 크기의 대용량 데이터를 의미하며, 최근 빅데이터는 대용량데이터의 수집, 저장, 분석, 체계화를 위한 도구, 플랫폼, 분석기법 등을 포괄하는 용어로 변화하고 있으며, 대용량데이터를 활용・분석하여 가치 있는 정보를 추출하고 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술을 말한다. 사실 빅 데이터에 대해서 구체적이고 정량적인 정의가 합의된 바는 없다. 세계적인 컨설팅 기관인 McKinsey지는 2011년 5월에 발간한 보고서에서 [1] “빅 데이터의 정의는 기존 데이터베이스 관리 도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 데이터셋 규모로, 그 정의는 주관적이며 앞으로도 계속 변화될 것이다.” 하지만 빅데이터 중에서 가치 있는 데이터는 소수에 불과하다. 따라서 대용량 데이터를 분석하여 의미 있는 데이터를 발견하는 기술이 필요하다.

본 고에서는 2장에서는 빅데이터의 요소 및 분석 기술에 대하여 알아보고 3장에서는 분석을 위한 인프라에 대하여 설명하고 4장에서는 빅데이터의 국내외 현황을 알아보고 5장은 빅데이터의 활용방안에 대하여 설명하고, 6장과 7장에서는 향후 트렌드를 알아보고 결론을 기술한다.

2. 빅데이터의 요소

빅데이터의 요소기술에는 미디어나 위치정보, 동영상 등을 나타내는 데이터의 크기(Volume), 실시간으로 데이터가 생성되는 데이터 입출력 속도(Velocity), 비구조화(비정형)에 대한 데이터의 형태(Variety)가 있다 [1]. 빅데이터 활용을 위한 3대 요소에는 자원, 기술, 인력이 있다. 데이터 자원 확보를 위한 자원과 데이터 저장, 데이터 관리, 대용량데이터 처리를 위한 기술 그리고 수학, 공학, 경제학, 통계학, 심리학 등에 능통한 인력이 필요하다[1,2].

2.1 빅데이터의 3대 요소

빅데이터의 요소기술에는 그림 1과 같이 크기, 속도, 형태(다양성)가 있다.

- **데이터의 크기(Volume)**

미디어나 위치정보, 동영상 등과 같이 데이터의 크기를 나타내는 것으로, 물리적인 크기뿐만 아니라 현재의 기술로 처리가능한 양인지, 불가능한 양인지에 때라 빅데이터인지 판단한다.

- **데이터 입출력 속도(Velocity)**

실시간으로 데이터가 생성될 때 데이터를 처리하는

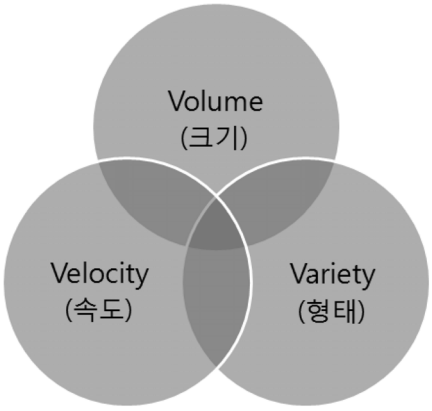

그림 1 빅데이터의 3대 요소

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

속도를 나타내는 것으로 실시간 처리, 스트림 처리 등이 있다.

• **데이터의 형태(Variety)**

빅데이터의 정형화 정도에 따른 분류방법으로 정형은 고정된 필드에 저장된 데이터(관계형 데이터베이스)를 말하고, 반정형은 고정된 필드는 아니지만 스키마를 포함하는 데이터(XML, HTML 등)를 말하며, 비정형은 고정된 필드에 저장되어 있지 않은 데이터(텍스트, 이미지, 동영상 등)를 말한다.

2.2 빅데이터 활용을 위한 3대 요소

빅데이터 활용을 위한 요소 기술에는 그림 2와 같이 자원, 기술, 인력이 있다[2].

• **자원**

빅데이터를 위한 자원 확보, 빅데이터 품질 관리를 위한 자원 확보를 말하며 빅데이터를 관리, 처리하는 측면과 함께 활용할 수 있는 기업의 내부, 외부 빅데이터 자원을 수집하는 전략이 필요하다.

• **기술**

빅데이터 프로세스와 새로운 기술을 의미한다. 국가, 기업의 혁신 전략으로 사용할 수 있도록 빅데이터 인프라, 플랫폼, 분석기술 등을 말한다. 빅데이터 플랫폼으로는 하둡(hadoop), 데이터 저장, 관리 기술에는 NoSQL, 분석기술에는 자연처리, 의미분석, 데이터 마이닝 등이 있다. 또한, 분석한 데이터를 보여주는 시각화(Visualization) 기술 등도 있다.

• **인력**

국가, 기업 등은 데이터 사이언티스트같은 인재를 확보하기 위해 내부 역량 강화, 외부 협력 전략을 수립할 필요가 있다. 수학, 공학적인 능력과 경제학, 통계학, 심리학 등에 능통한 인재가 필요하다. 또한, 비판적 시각과 커뮤니케이션 능력, 스토리텔링 등 시각화 능력을 갖춘 인재도 필요하다.

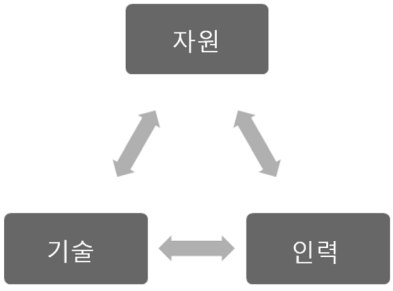

그림 2 빅데이터 활용을 위한 3대 요소

2.3 빅데이터 분석 기술(분석 기법)

데이터의 분석 기술에는 텍스트 마이닝, 평판 분석, 소셜 분석, 클러스터 분석 등의 크게 4가지로 나누어 볼 수 있다[3].

• **텍스트 마이닝(Text Mining)**

텍스트 마이닝은 비/반정형 텍스트 데이터에서 자연어처리 기술에 기반하여 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 텍스트 마이닝 기술을 통해 방대한 텍스트 뭉치에서 의미 있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트가 가진 카테고리를 찾아내거나 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있다. 컴퓨터가 인간이 사용하는 언어(자연어)를 분석하고 그 안에 숨겨진 정보를 발굴해 내기 위해 대용량 언어자원과 통계적, 규칙적 알고리즘이 사용되고 있다. 주요 응용분야로 문서 분류, 문서 군집, 정보 추출, 문서요약 등이 있다.

• **평판분석(Opinion Mining)**

텍스트 마이닝의 관련 분야로는 오피니언 마이닝, 혹은 평판 분석이라고 불리는 기술이 있다. 소셜미디어 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별하는 기술이다. 오피니언 마이닝은 특정 서비스 및 상품에 대한 시장규모 예측, 소비자의 반응, 입소문 분석(Viral Analysis) 등에 활용되고 있다. 정확한 오피니언 마이닝을 위해서는 전문가에 의한 선호도를 나타내는 표현 및 단어 자원의 축적이 필요하다.

• **소셜 네트워크 분석(Social Network Analytics)**

소셜 네트워크 분석은 간단히 소셜 분석으로 나타내며, 수학의 그래프 이론에 뿌리를 두고 있다. 소셜 네트워크 연결구조 및 연결강도 등을 바탕으로 사용자의 명성 및 영향력을 측정하여, 소셜 네트워크 상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾는데 주로 활용된다. 이렇게 소셜 네트워크 상에서 영향력이 있는 사용자를 인플루언서(Influencer)라고 부르는데, 인플루언서 모니터링 및 관리는 마케팅 관점에서 중요하다고 할 수 있다.

• **클러스터 분석(Cluster Analysis)**

클러스터 분석은 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 그룹을 발굴하는데 사용된다. 예를 들어 트위터 상에서 주로 사진/카메라에 대해 이야기하는 사용자 그룹이 있을 수 있고, 자동차에 대해 관심 있는 사용자 그룹이 있을 수 있다. 이러한 관심사나 취미에 따른 사용자 그룹을 군집분석을 통해 분류할 수 있다.

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

3. 빅데이터 분석을 위한 인프라

빅 데이터의 부상과 함께 하둡은 빅 데이터 처리의 핵심 엔진으로 많은 관심을 받고 있다[4,5]. 기존 데이터베이스와 비교하면 초기 단계의 하둡이지만, 가능성을 알아 본 많은 업체들이 하둡 지원에 나섰으며, 기업들 역시 하둡 도입을 적극적으로 검토하고 있다. 하지만 아직 하둡은 한창 성장 하고 있는 기술인 만큼 보완해야 할 점도 많고, 기업이 고려해야 할 점도 많다. 하둡은 기업들이 이전에는 비용과 복잡성 그리고 도구가 없어 폐기했던 데이터를 저장하고 처리할 수 있게 해준다. 엄청나게 많은 양의 데이터를 저장, 처리, 분석할 수 있는 하둡은 IT 조직의 현업에서 자리를 잡아가고 있다. 하지만 이 오픈 소스 플랫폼에 대한 상대적인 생소함과 경험있는 하둡 인재의 부족은 IT 부서에게 새로운 기술적 과제를 제기하고 있다.

3.1 하둡(Hadoop)

하둡은 아래 그림 3과 같이 오픈소스 분산처리기술 프로젝트로, 현재 정형 및 비정형 빅 데이터 분석에 가장 선호되는 솔루션이라고 할 수 있다[5]. 실제로 야후와 페이스북 등에 사용되고 있으며, 채택하는 회사가 늘어나고 있다. 주요 구성요소로 하둡 분산 파일 시스템인 HDFS(Hadoop Distributed File System), Hbase, 맵리듀스(MapReduce)가 포함된다.

HDFS와 Hbase는 각각 구글의 GFS와 BigTable의 영향을 받았다[4]. 기본적으로 비용효율적인 x86 서버로 가상화된 대형 스토리지를 구성하고, HDFS에 저장된 거대한 데이터셋을 간편하게 분산처리 할 수 있는 Java 기반의 맵리듀스 프레임워크를 제공한다. 이외에도 하둡을 기반으로 한 다양한 오픈소스 분산처리 프로젝트가 존재한다[5,7]. 하둡은 텍스트 검색 라이브러리로 폭넓게 사용되고 있는 아파치 루씬의 창시자인 더그 커팅에 의해 제작되었다. 하둡은 오픈소스 웹 검색엔진인 웹 검색엔진인 아파치 너치(Nutch)를 탑재하였으며, 너치는 2002년에 처음 만들어졌으며, 실행 가능한 크롤러와 검색 시스템이다. 2003년에 GFS의 아키텍처가 서술된 논문 출판 이후 분산파일 시스템에 관한 기술이 더욱 발전하게 되었다.

아파치 하둡 프로젝트에 대하여 좀 더 자세히 알아보면 다음과 같다[5,9].

- **코어**

분산 파일시스템으로 일반적인 입출력(직렬화, 자바 RPC, 영속데이터 구조)을 위한 컴포넌트와 인터페이스의 집합이다.

- **에이브로**

교차언어 RPC와 영속적인 데이터 스토리지를 위한 데이터 직렬화 시스템이다.

- **맵리듀스**

범용 컴퓨터들의 커다란 클러스터에서 수행되는 분산 데이터 처리 모델과 실행환경이다.

- **HDFS**

범용 컴퓨터들로 된 커다란 클러스터에서 수행되는 분산 파일 시스템이다.

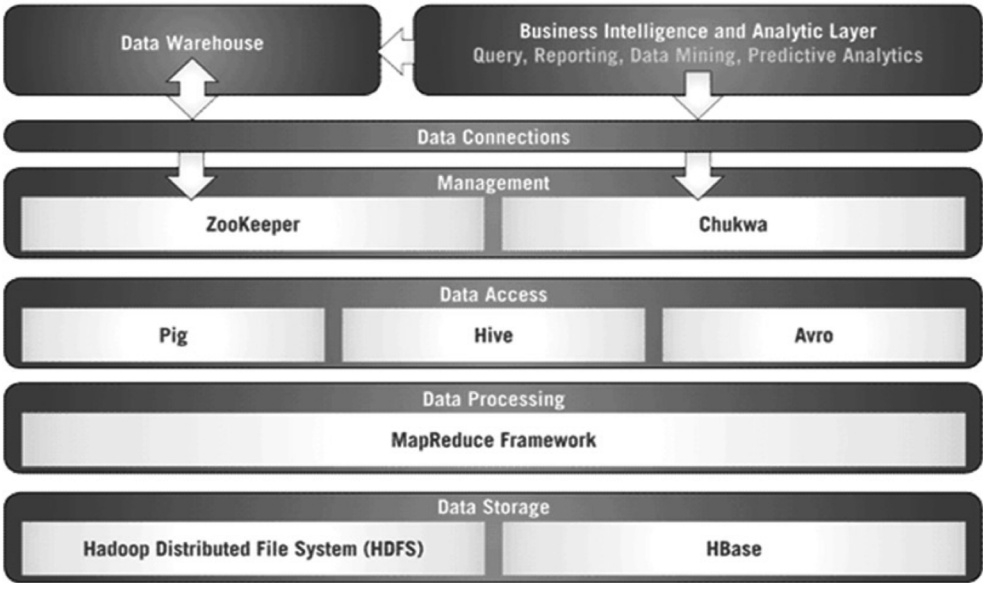

그림 3 하둡 시스템

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

• **피그**

대규모 데이터셋 탐색용 데이터 흐름 언어와 실행 환경이다. HDFS와 맵리듀스 클러스터에서 수행된다.

• **HBase**

분산 컬럼 지향 데이터베이스 스토리지로 HDFS를 사용한다. 맵리듀스를 이용한 일괄처리 방식의 계산과 랜덤 읽기가 가능한 포인트 쿼리 방식 모두를 지원한다.

• **주키퍼**

다수 컴퓨터로 분산 처리되는 고가용성 조정(available coordination) 서비스로, 분산 응용 프로그램들을 구축하기 위하여 사용될 수 있는 분산 락(Lock)같은 프리미티브를 제공한다.

• **하이브**

분산 데이터웨어하우스, HDFS에 저장된 데이터를 관리하고, 데이터 쿼리를 위하여 SQL 기반 쿼리 언어(런타임 엔진에 의해 맵리듀스로 변환되는)를 제공한다.

• **추콰**

분산 데이터 수집 및 분석 시스템, HDFS에 데이터를 저장하는 수집기를 수행하고, 보고서를 생성하기 위해서 맵리듀스를 사용한다.

3.2 R

오픈소스 프로젝트 R은 통계계산 및 시각화를 위한 언어 및 개발환경을 제공하며, R 언어와 개발환경을 통해 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현/개선이 가능하다[6]. 이렇게 구현한 결과는 그래프 등으로 시각화할 수 있으며, Java나 C, Python 등의 다른 프로그래밍 언어와 연결도 용이하다. Mac OS, 리눅스/유닉스, 윈도우 등의 대부분의 컴퓨팅 환경을 지원하는 것도 장점이다. 위의 장점들로 인해 R은 통계분석 분야에서 인지도를 높여왔으며, 하둡 환경에서 분산처리를 지원하는 라이브러리 덕분에 구글, 페이스북, 아마존 등의 빅 데이터 분석이 필요한 기업에서 대용량 데이터 통계분석 및 데이터 마이닝을 위해 널리 사용되고 있다.

3.3 NoSQL

NoSQL은 Not-Only SQL, 혹은 No SQL을 의미하며, 전통적인 관계형 데이터베이스와 다르게 설계된 비관계형 데이터베이스를 의미한다. 대표적인 NoSQL 솔루션으로는 Cassandra, Hbase, MongDB 등이 존재한다[5,7]. NoSQL은 테이블 스키마가 고정되지 않고, 테이블 간 조인 연산을 지원하지 않으며, 수평적 확장이 용이하다는 특징을 가진다. 관계형 데이터베이스의 경우, 일관성과 유효성에 중점을 두고 있는 반면, NoSQL 기술은 분산 가능성에 중점을 두고 일관성과 유효성은 보장하지 않는다. 이것은 일관성, 유효성, 분산가능성 중 2가지만 보장이 가능하다는 분산 데이터베이스 시스템 분야의 이론에 따른 것이다. 따라서 대규모의 유연한 데이터 처리를 위해서는 NoSQL 기술이 적합하지만, 안정성이 중요한 시스템에서는 오랫동안 검증된 관계형 데이터베이스를 채택할 필요가 있다.

4. 빅데이터의 국내외 현황

글로벌 기업들인 “EMC, IBM, 오라클, SAP, 테라데이타, HP, 구글, MS” 등의 빅데이터 준비현황을 살펴보면 다음과 같다.

4.1 국외 현황

• **EMC**

데이터 저장부터 관리, 분석까지 빅데이터에 관한 모든 것을 제공하기 위해 그린플럼, 이이실론 등 빅데이터 솔루션 업체 및 데이터 관련 다수업체를 인수하였다. 빅데이터 스토리지 솔루션, 콘텐츠 관리 솔루션 등을 제공한다. EMC 애널리틱스 랩을 운영하여 데이터 사이언티스트를 육성하고 있다.

• **IBM**

지난 5년간 140억 달러 이상을 투자하여 비즈니스 분석 관련업체를 인수하여 분석용 데이터 저장관리 업체(네티자), 데이터 통합 업체(에센셜), 분석 솔루션 업체(코그리스)등을 인수하였다. 지속가능한 지구를 만들기 위해 지구 데이터(기온, 토양산태, 교통 흐름 등)를 분석하는 ‘스마트 플래닛’ 프로젝트를 전개하고 있다.

• **오라클**

세계적인 DB 업체, 하이페리온을 인수로 분석 기술을 확보하였으며 오라클 빅데이터 어플라이언스 제품을 출시하였다.

• **SAP**

업무용 애플리케이션 업체에서 최근 DB 전문업체로 변신하였으며 메모리 기반 DB 어플라이언스인 HANA를 제시하였으며 BI 소프트웨어, 플랫폼을 제공하는 비즈니스 오브젝트 회사를 인수하였다.

• **테라데이타**

데이터웨어하우스 및 BI 전문업체이며 비정형 데이터의 고급분석 및 관리 솔루션 업체 애스터데이터를 인수하여 ‘애스터 맵리듀스 플랫폼’을 제시하였다.

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

표 1 데이터의 시대흐름

	PC시대	인터넷시대	모바일시대	스마트시대
패러다임 변화	디지털화, 전산화	온라인화	소셜화, 모바일화	지능화, 개인화, 사물정보화
IT 이슈	PC, PC통신, 데이터베이스	초고속 인터넷, WWW, 웹 서버	모바일 인터넷, 스마트폰	빅데이터, 차세대 PC, M2M
핵심분야	PC, OS	포털, 검색엔진, Web 2.0	스마트폰, 앱서비스, SNS	미래전망, 상황인식, 맞춤형 서비스
대표기업	MS, IBM 등	구글, 네이버, 다음	애플, 페이스북, 트위터	?
IT 비전	1인 1PC	e-Korea	손안의 PC, 소통	새로운 가치창출

• **구글**

대용량 데이터 처리 기술 발표(GFS, MapReduce, Sawzall, BigTable)를 발표하였으며 빅쿼리 서비스를 2011년 공개하였다.

• **MS**

윈도 애저와 윈도 서버 플랫폼용 아파치 하둡 개발 계획하고 있으며 하둡 기술 전문업체 '호튼웍스'와 협력하고 있다.

4.2 국내 현황

그러면 우리나라 상황은 어떠한가? 최근 국내 빅데이터 시장에서 가장 활발하게 부각되는 것은 소셜 분석 서비스이다. 다음소프트, 그루터 등과 같은 소셜 분석 전문업체들은 소셜 빅데이터를 기반으로 마케팅 분석뿐만 아니라 사회정치적 현상까지도 분석하는 등 서비스를 적극 제공하고 있다. 특히 선거와 같은 정치적 여론시장의 분석 수요가 크게 늘어남에 따라 올해 선거가 '빅데이터 선거'가 될 것이라는 전망까지 나오기도 한다.

이미 오라클, EMC, IBM, SAP, MS 등 몇몇 글로벌 IT기업들이 국내업체 및 학계와 제휴하여 국내 빅데이터 시장 진출을 선언하면서 빅데이터 경쟁 무대가 형성되고 있다[8]. 국내 기업들도 글로벌 데이터 기업, 대형의료기관 등과 공동으로 비즈니스 시장분석, 유전자 정보 분석 등 빅데이터 사업에 뛰어들기 시작했다. 이러한 빅데이터 각축전 속에서 우리나라 정부도 지난해 11월 빅데이터를 활용한 스마트 정부 구현을 목표로 중장기 정책방향을 제시한 바 있다. 그야말로 빅데이터를 국가경쟁력의 핵심 자원으로 인식하기 시작한 셈이다.

그런데 글로벌 기업들의 진출과 일부 소셜 분석 서비스의 두각으로 국내 빅데이터 시장이 서서히 성장하고는 있지만, 빅데이터 시장을 주도할 만큼 양질의 전문인력 및 연구역량을 체계적으로 확보하고 있지 못하다는 자성론이 제기된다. 하둡(Hadoop), 카산드라(Cassandra)[7] 등 빅데이터 분석 기술들을 활용하는 서비스들이 점차 늘어나고는 있으나 그 빅데이터를 제대로 읽고 해석하는 능력을 갖춘 국내 전문 인력은 여전히 부족하다는 것이다.

5. 빅데이터의 활용

표 1은 데이터를 처리하는 패러다임이 변화하는 모습을 보여주고 있다. PC 시대에서 인터넷시대로, 현재의 모바일시대 및 스마트시대로 접어들었다. 스마트시대에는 빅데이터가 중요한 이슈이며, 이를 적극 활용하여야 한다.

데이터의 자원이 축적과 공유를 통해 엄청난 규모로 쌓이면서 데이터의 역할은 그림 4와 같이 분석과 추론의 방향으로 진화할 것이며 대규모 데이터를 기반으로 한 자연어처리, 인공지능 기술이 발전하여 서비스의 상용화가 진행될 것이다. 지능형 서비스들은 분석 데이터가 늘어나고 기계학습이 진행될수록 인간의 언어에 대한 이해도가 상승할 것이다.

5.1 빅데이터의 활용을 위한 5대 전략 분야

빅데이터 활용이 정보통신, 교육, 의료, 금융 등 사회

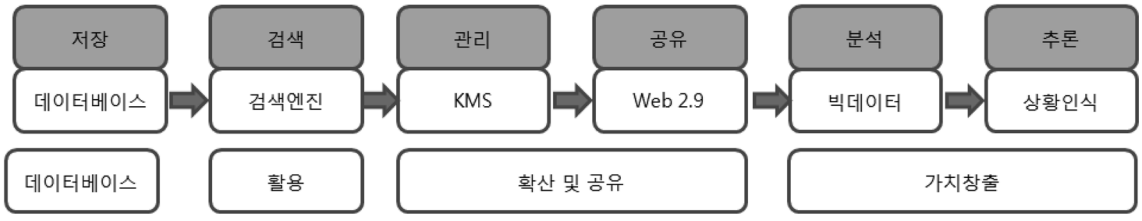

그림 4 빅 데이터의 새로운 가능성과 대응 전략'

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

그림 5 빅데이터의 활용 분야

각 분야로 확산되면서 사회전반의 생산성 향상에 기여할 전망이다[8]. 그림 5에서, EU의 경우, 15~20%의 공공관리 비용 감소했으며, 2~4천 달러의 가치창출 되었으며, 향후 10년간 0.5% 생산성 증가효과 기대된다. 제조업 적용시 상품개발 및 조립비용을 50% 이상 절감 가능하며 운용자본도 7% 이상 절감되고 개인의 LBS 정보는 바이트당 부가가치가 높아 2020년경에는 약 $7,000억의 가치 창출할 것으로 예상된다.

세상에 존재하는 모든 데이터 중 90%가 지난 3년 동안 생성되었다고 할 정도로 급속도로 늘어난 데이터를 활용하는 방안이 정부 및 기업들의 큰 이슈로 등장하고 있다. 빅데이터의 시대가 열리면서 예전에는 유통업체가 고객 반응 데이터를 수집하기 위해서 질문이나 모의실험을 통해 패턴을 예상하고 마케팅 전략을 세웠다면, 지금은 고객의 포인트 카드 및 매장 곳곳에 설치된 CCTV 등의 정보까지 활용하고 있다. 제조업체의 경우도 SNS상에 나온 키워드를 분석해 자사 기업이 출시한 제품에 대한 반응을 분석해 기업 상품 출시 전략 등에 활용하고 있다. 예를 들어 O2(영국의 이동통신사)는 실시간으로 빅데이터를 처리하여 스마트폰을 통한 'SNS + 위치정보서비스'를 결합한 프로모션을 하고 있으며, 구글은 자사 검색통계를 바탕으로 시간 및 지역별 독감 유행정보를 제공하고 있다. 이와 더불어 스마트폰, 사물통신 등을 통한 데이터양이 증가하면서 이를 활용하는 솔루션 기업들도 대거 등장할 것으로 전망된다.

빅데이터의 활용은 산업부문별로 약 0.5~1% 정도의 생산성을 증가시킬 것으로 전망된다. 특히 미국의 의료 부문에서는 연간 3,000억 달러의 가치를 증대시킬 것으로 전망되는데 이는 스페인의 연간 의료지출비의 2배에 이르는 상당한 금액이다. 빅데이터의 잠재적 활용가치는 산업 분야별로 차이가 존재하는데 가장 잠재가치가 높은 부문은 컴퓨터, 전자제품 및 정보통신 분야가 될 것으로 전망된다. 왜냐하면 대용량 데이터에 접근이 용이할수록 혁신 속도가 촉진될 것이기 때문이다. 반면 부동산이 나 음식 관련 업종은 데이터 활용이 제한적이어서 활용가치는 떨어질 것으로 판단된다.

5.2 빅데이터의 활용가치

불확실하고 리스크가 존재하는 미래사회에는 통찰력, 대응력이 필요하고, 스마트하고 융복합의 미래사회에는 경쟁력과 창조력을 필요로 한다. 즉, 미래사회의 특성에 맞는 빅데이터의 역할이 존재한다. 표 2는 미래사회의 특징과 빅데이터의 역할에 대한 설명이다.

표 2 미래사회와 빅데이터의 역할

미래사회의 특징	빅데이터의 역할	
불확실성	통찰력	– 사회현상, 현실세계의 데이터를 기반으로 한 패턴 분석과 미래전망 – 여러가지 가능성에 대한 시나리오 시뮬레이션 – 다각적인 상황이 고려된 통찰력을 제시 – 다수의 시나리오의 상황 변화에 유연하게 대처
리스크	대응력	– 환경, 소셜, 모니터링 정보의 패턴 분석을 통한 위험징후, 이상 신호 포착 – 이슈를 사전에 인지, 분석하고 빠른 의사결정과 실시간 대응 지원 – 기업과 국가 경영의 명성 제고 및 낭비요소 절감
스마트	경쟁력	– 대규모 데이터 분석을 통한 상황인지, 인공지능 서비스 등 가능 – 개인화, 지능화 서비스 제공 확대 – 소셜분석, 평가, 신용, 평판 분석을 통해 최적의 선택 지원 – 트렌트 변화 분석을 통한 제품 경쟁력 확보
융합	창조력	– 타 분야와의 결합을 통한 새로운 가치창출 – 인과관계, 상관관계가 컨버전스 분야의 데이터 분석으로 안전성 확보, 시행착오 최소화 – 방대한 데이터 활용을 통한 새로운 융합시장 창출

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

6. 향후 트렌드

6.1 대중화 단계로의 진입

관련 기술이 발달하고 민간기업의 수요가 꾸준히 증가하여 곧 대중화 단계로 접어들 것으로 예상된다. 아직까지는 태동단계로서, 다양한 분야에서의 가능성이 논의될 뿐, 풍부한 적용은 아직 이루어지지 않고 있는 것이 사실이다. 비즈니스 관련 의사결정 과정에서 빅데이터를 효과적으로 활용하고 있는 기업은 아직 미미한 수준이다. 미국을 비롯한 유럽 등의 나라는 정책적 과제로 빅데이터를 추진하고, 일부 민간 기업들의 성공적 사례가 등장하고 있다.

빅데이터 관련 사업을 쉽게 추진할 수 있도록 지원하는 솔루션 및 서비스가 증가하고, 이는 또 다시 대중화에 기여할 것이다. 지난 2011년 11월 구글이 일부 기업이용자를 대상으로만 제공해왔던 빅데이터 분석 서비스인 'Google BigQuery'의 프리뷰버전을 일반에 공개한 것이 대표적 사례로서 구글은 이를 통해 빅데이터 분석을 필요로 하는 대다수 중소기업들이 투자비용에 대한 부담없이 이용할 수 있을 것으로 기대된다.

6.2 '데이터 큐레이터' 사업의 등장 가능성

빅데이터를 활용하려는 기업은 많지만, 이를 직접 수집할 수 있는 여력이 있는 업체는 제한적이고 빅데이터 수집・분석 능력은 서비스 차별화 및 경쟁력과 직결되며 추가 수익원 발굴을 원하는 빅데이터 보유업체가 이를 필요로 하는 기업에게 실시간으로 제공하는 부가서비스를 제공할 것으로 예상된다. 특히 이동통신사, 구글 등 플랫폼업체, 페이스북이나 포털 등의 서비스 업체들이 이 같은 사업을 추진할 가능성이 높다.

6.3 산업 활성화를 위한 제도 정비

빅데이터 활성화로 인해 기업간 양극화 현상이 심화될 가능성이 존재하며 현재 빅데이터는 데이터를 창출하는 기업이 독점적으로 활용하는 사례가 많으며, 이는 정보의 비대칭성을 불러와 기업 간 양극화를 가져올 수 있다. 산업 활성화를 위해 빅데이터를 보유할 수 없는 중소기업들도 원하는 데이터를 쉽게 취득할 수 있는 경로를 만들어 주는 것이 중요하며 빅데이터 관련 규제 및 법령 개정을 통해 프라이버시 문제 해결은 물론 데이터의 거래를 활성화시킬 수 있는 방안이 본격적으로 추진될 것으로 예상된다. 빅데이터를 위한 규제완화는 보안측면에서 현행 법률에 어긋날 수 있으며, 일부 기업과 소비자 단체 등에서의 거센 반발이 있을 수 있다. 그럼에도 빅데이터가 원활히 창출 및 유통될 수 있도록 지원하는 것은 빅데이터 산업뿐 아니라 국가의 전체 산업 활성화에 기여할 가능성이 높다. 정부차원의 데이터 수집 및 공개 사업은 물론 빅데이터 분석 툴 제공 등이 증가할 것으로 전망된다.

7. 결 론

빅데이터를 단순히 수집・축적하는 것이 중요한 것이 아니라 구조화되지 않은 대규모 데이터 속에서 숨겨진 패턴을 찾아내고 여러 변수들을 통합적으로 고려하면서 창의적으로 해석할 수 있는 분석능력이 더 중요해지고 있다. 요즘 빅데이터와 관련해서 유행하는 용어인 '데이터 과학자(data scientist)'라고 불리는 연구역량이 필요하다는 뜻이다. 물론 빅데이터 연구역량은 통계물리학, 수학, 컴퓨터 과학, 소셜 네트워크 분석과 같은 다양한 분석기법을 익혀야 하는 등 다년간의 개발과 훈련 기간이 요구된다는 점에서 단시일 내로 확보되기 어렵다. 그런 점에서 최근의 '빅데이터 신드롬'은 ICT와 관련한 교육 및 학계 차원의 인력양성 시스템에 대한 반성적 성찰을 근본적으로 요구한다고 하겠다. 단기적 성격의 특정 서비스에만 함몰되어 중장기적 연구와 개발에는 인색했던 것은 아닌지 반성할 필요가 있다. 공공 및 민간 차원의 빅데이터 활용서비스 모델을 개발하고 이를 체계적으로 연구하기 위한 "빅데이터 연구 및 활용센터" 설립 등의 중장기 계획이 방송통신위원회를 중심으로 활발하게 논의되고 있다는 점은 참으로 다행스런 일이다. 빅데이터를 분석 및 활용하는 인재를 양성하고 정부, 기업 및 학계가 관심을 가지고 대처한다면 빅데이터에 대한 우리의 미래는 밝을 것이다.

참고문헌

[1] McKinsey, "Big Data : The Next Frontier for Innovation, Competition, and Productivity", McKinsey & Company, 2011년 5월

[2] Gartner, "Big Data Analytics", Gartner Group, 2011년 1월

[3] 김정숙, "빅 데이터 활용과 관련기술 고찰", 한국콘텐츠학회지 제10권, 제1호, 2012.3, page(s): 9-116

[4] Vertica, "Managing Big Data with Hadoop & Vertica", Vertica Systems, 2009년 10월

[5] http://hadoop.apache.org/

[6] http://www.r-project.org/

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)

[7] http://cassandra.apache.org/
[8] 이만재, “빅 데이터와 공공 데이터 활용” Internet and Information Security, 제2권, 제2호, 2011년 11월, pp. 47-64
[9] 톰 화이트, “Hadoop 완벽가이드” 한빛미디어, 2010년 5월

약 력

강 만 모

1998 울산대학교 전자계산학과 학사
2000 울산대학교 전자계산학과 석사
2011 울산대학교 정보통신공학 박사
2006~2009 울산대학교 객원교수
2009~현재 대광산업 기술연구소 책임연구원, 울산대학교 전기공학부 강사
관심분야 : 전자상거래, 멀티에이전트, 소프트웨어공학, 빅데이터 분석
E-mail : manmoakng@ulsan.ac.kr

김 상 락

2010 울산대학교 메카트로닉스/IT 석사
2012 울산대학교 정보통신 공학 박사
2000~2009 (주)아이티스타 연구소장
2010~현재 울산대학교 전기공학부 강사, 비케이앤씨 대표
관심분야 : SLA, 분산병렬처리시스템, 인포그라픽스, 빅데이터
E-mail : shem0304@ulsan.ac.kr

박 상 무

1995 울산대학교 컴퓨터공학과 학사
1997 울산대학교 컴퓨터공학과 석사
2010 울산대학교 컴퓨터공학과 박사
2011~현재 울산대학교 전기공학부 객원교수
관심분야 : 인공지능, 신경회로망, 임베디드 시스템
E-mail : cthreepo@ulsan.ac.kr

이화여자대학교 | IP:211.48.46.*** | Accessed 2020/01/08 16:38(KST)