빅 데이터 활용과 관련기술 고찰

김정숙 (삼육대학교)

차 례

1. 서론
2. 시대의 화두 빅 데이터
3. 국내·외 빅 데이터 활용 현황
4. 빅 데이터 관련기술 고찰
5. 결론

1. 서론

IT 융합, 소셜 미디어, 서비스 산업 고도화, 기업들의 고객 데이터 수집활동, 멀티미디어 콘텐츠의 폭발적 증가와 스마트폰 보급, SNS 활성화, 사물통신망의 저변확대로 데이터량은 그 종류와 수 또한 급격히 증가하고 있는 추세이다.

따라서, 모든 기업이 보유한 빅 데이터가 '거대한 가치 추출이 가능할 만큼' 충분한 규모에 도달해 누가 먼저 그 가치를 추출해 내느냐가 향후 기업의 성패를 가늠할 상황에 직면하고 있다. 산업혁명에서는 철과 석탄이, IT 혁명에서는 인터넷이 세계 경제 변화를 지탱하는 핵심 요소였듯이 다가올 모바일 스마트 혁명에서는 빅 데이터가 경제 변화의 핵심 자원 역할을 할 것이다. 실제로, 빅 데이터의 '양적 거대함'은 많은 분야에서 불가능을 가능으로 전환할 것으로 기대되면서 구글의 빅 데이터 솔루션이 IBM의 실패 프로젝트를 성공으로 변신시킨 Magic 신화로 확인되었다.

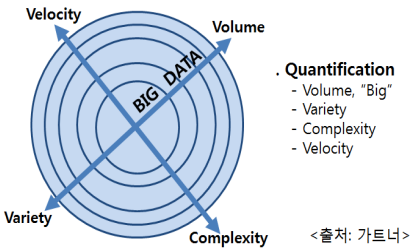
하지만 볼륨의 폭증에 대한 기회를 확보할 수 있는 실질적인 방법은 미흡한 상황이다. 빅 데이터 분석은 기업의 가치 극대화를 위해 활용될 수 있는 만큼 기존 환경과의 연계, 고가용성, 사용의 용이성, 보안, 시각화 및 적시 분석 등이 가능해야 한다. 이미 전통적인 분석 기술로 많은 성과를 거뒀지만 빅 데이터 방식의 분석을 통해 더 많은 성과를 거둬야 한다는 압박감은 증가하고 있다. 분석 기술과 관리 능력에 초점을 맞춰 심층적인 개선에 주력해야 할 것이다.

본고에서는 빅 데이터의 등장과 더불어 효율적인 활용으로 새로운 비즈니스 창출, 기업 경쟁력 극대화 등을 이룰 수 있는 빅 데이터 관련기술을 고찰하고자 한다.

2. 시대의 화두 빅 데이터

2.1 빅 데이터 개요

빅 데이터는 '현재 시스템으로 처리 가능한 범위를 넘어서는 데이터'로 정의된다. 또한, 페타(Peta: 10^{15}), 엑타(Exa: 10^{18}), 제타(Zeta: 10^{21})바이트 등 기존의 데이터 단위를 넘어서는 엄청난 양(Volume), 데이터의 생성과 흐름이 매우 빠르게 진행되는 속도(Velocity), 사진, 동영상 등 기존의 구조화된 데이터가 아닌 다양한(Variety) 형태의 정보 등 3가지 속성을 가진 데이터가 '빅 데이터'라는 게 대다수 전문가들의 공통된 의견이다. 가트너는 3V에 복잡성을 추가해 3V+C로 정의하기도 한다[1].



▶▶ 그림 1. 빅 데이터 정의

이처럼, 빅 데이터는 의미 있는 결과 도출이 가능한 수십~수천 TB에 달하는 거대 데이터 집합을 의미했으나 관련도구, 플랫폼, 분석기법까지 포괄하는 용어로 변화되고 있다. 그 결과, 빅 데이터는 다음과 같은 영역으로의 활용을 기대할 수 있다. Social Graph와 패턴을 통해 Network 구조와 정보 패턴을 파악하는데 활용할 수 있고, 트랜드의 감지와 예측을 통해 사건의 징후와 전개과정을 감지하는데 활용도 가능하며, 데이터에 근거한 의

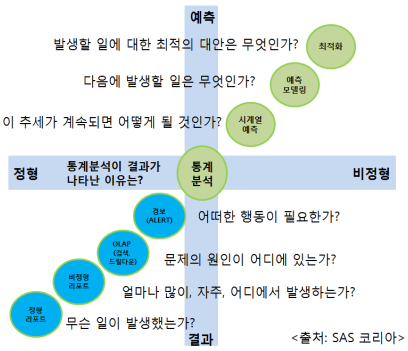
사결정으로 경영자의 직관을 보완하는 효과적인 의사결정을 지원할 수 있다. 또한, 예측 모형과 시뮬레이션을 통해 전략실행 효과의 최적화를 이루는데도 활용할 수 있다. 이러한 빅 데이터의 종류는 정형화 정도에 따라 표 1.과 같이 분류된다.

표 1. 빅 데이터 종류

정의	설명
정형 (Structured)	고정된 필드에 저장된 데이터 (예) 관계형 DB, 스프레드시트
반정형 (Semi-Structured)	고정된 필드에 저장되어 있지 않지만, 메타 데이터나 스키마 등을 포함하는 데이터 (예) XML, HTML 텍스트
비정형 (Unstructured)	고정된 필드에 저장되어 있지 않은 데이터 (예) 텍스트 분석이 가능한 텍스트 문서 및 이미지/동영상/음성 데이터

2.2 빅 데이터 요구사항

다양한 분야에서 이용되는 빅 데이터를 도입하기 위해서는 단순히 많은 데이터의 분석뿐만 아니라 그림 2.와 같은 대처방법의 시스템, 서비스, 조직 등 전사적으로 빅 데이터를 분석할 수 있는 기반 조성이 필요하다.



▶▶ 그림 2. 비즈니스 대처방법

또한, 빅 데이터를 분석하여 활용 가능하게 하기 위해서는 표 2.의 요구사항들을 만족시켜야 한다[2].

표 2. 빅 데이터 활용을 위한 요구사항

구 분	요구사항
Fast Search, Indexing	.빠른 검색과 접근 가능 .빅 데이터 활용 전제조건
Rapid Analysis/Response	.분석 및 처리결과 신속 .효율성 향상 전제조건
Multi/Parallel Processing	.computing power 한계 극복을 위해 다중작업에 의한 동시/병렬처리 가능
TCO Reduction	.전통적 접근방법으로는 빅데이터의 분석 및 활용에 막대한 비용으로 주목받지 못함.

3. 국내 · 외 빅 데이터 활용 현황

최근 빅 데이터는 트위터, 페이스북 등 SNS에서 수집되는 정보를 분석하여 소비자 마음을 읽는 기법으로 기업의 마케팅은 물론 위기 관리 수단으로 급부상하고 있다. 따라서 빅 데이터 활용을 통한 정부의 효율적 운영 등으로 국가 경쟁력 제고에 대한 각국의 노력은 표 3.과 같이 점점 늘어나고 있다.

3.1 각국의 활용 현황[3]

표 3. 각국의 빅 데이터 활용사례

국가	활용분야	내용
미국	국토보안	o 9.11 이후 국토안보부를 중심으로 테러·범죄 방지를 위한 범정부적 빅 데이터 수집, 분석 및 예측체계 도입 - 부시행정부의 국토안보부 장관인 Michael Chertoff는 국토보안을 위한 빅데이터 추진 현황 언급 - 국내외 금융 시스템의 개인, 기관의 금융거래 감시로 자금 세탁 및 테러 자금 조달 색출 강화
	치안	o FBI의 종합 DNA 색인시스템(CODIS) - DNA포렌직, 클라우드DNA분석 등 "빅DNA데이터"의 활용을 통해 2007년 45,400건의 범인 DNA Hit rate 달성 - 1시간 안에 범인 DNA 분석을 위한 주정부 데이터 연계 및 빅 데이터 실시간 분석 솔루션 확보
	의료	o 오바마 Health.20 - 필박스 프로젝트(Pillbox) - 국립보건원(National Library of Medicine)의 사이트로 약 검색을 서비스 - Pillbox를 통해 수집된 빅 데이터를 통해 후천성 면역결핍증(HIV) 등 관리대상 주요 질병의 분포, 연도별 증가 등에 대한 통계치 확보 가능

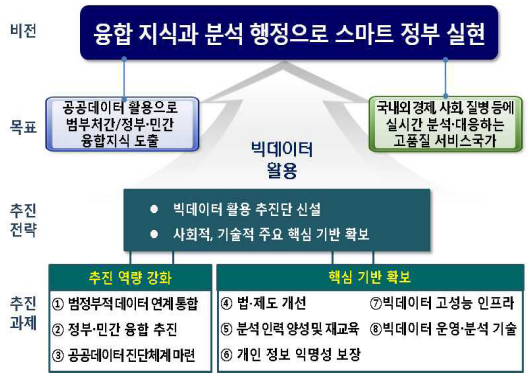
국가	활용분야	내용
영국	정보공개	o 영국은 정부 사이트(data.gov.uk)를 통해 공공부문의 정보 공유 및 활용을 위한 데이터 원스톱 서비스 제공 - 정부의 투명성 제고, 국민의 권리 향상, 데이터의 공개를 통한 경제적 사회적 가치 증대, 차세대 웹(web of data)에서 주도권 획득 목표 - 일반인들의 참여를 장려하고 아이디어 수렴, 앱 개발, 데이터 공개 등의 주제에 대한 커뮤니티 제공

국가	활용분야	내용
싱가포르	국가 위험관리	o 싱가포르 정부는 빈번히 발생하는 테러 및 전염병으로 인한 불확실한 미래 대비를 위하여 2004년부터 빅 데이터 기반 위험 관리 계획을 추진 - RAHS(Risk Assessment & Horizon Scanning) 시스템을 통해 질병, 금융위기 등 모든 국가적 위험을 수집 및 분석하여 위험을 선제적으로 관리 - 수집된 위험 정보는 시뮬레이션, 시나리오 기법 등을 통해 분석되어 사전에 위험을 예측하고 대응 방안을 모색함

국가	활용분야	내용
호주	정보공개	o 호주 정보관리청은 정부 2.0을 통한 정보 개방 - 방대한 양의 정보를 검색하고 분석 및 재사용할 수 있도록 자동화된 툴을 활용하여 시간과 자원을 절감 - AGIMO 산하 정부 2.0 전략/서비스팀에서는 정부 데이터에 대한 리포지터리 및 검색 툴을 서비스하는 data.gov.au 웹사이트 운영

3.2 국내 활용 현황

국내의 경우 빅 데이터와 같은 데이터의 증가보다 인터넷 트래픽의 증가를 더 심각하게 생각하였으며 데이터의 활용이라는 측면의 연구는 상대적으로 부족하다. 또한 고급 정보의 검색을 구글과 같은 외국기업의 솔루션에 의존하기 때문에 데이터의 증가에 대한 문제는 기업의 문제이며 국가 경쟁력의 문제로 인식되지 않은 상황이다. 하지만 빅 데이터가 국가 경쟁력 함양을 위한 국가적, 사회적 기반을 정비하고, 산·학·연 협력의 산업원천 기술 개발로 빅 데이터 주요 핵심 기반 확보를 위해 반드시 필요하다는 인식을 하고, 빅 데이터 활용을 위한 비전 및 목표를 그림3.과 같이 수립하였다.



▶▶ 그림 3. 빅 데이터 활용 추진 방안

4. 빅 데이터 관련기술 고찰

빅 데이터를 데이터 용량에 따른 분류가 아닌 기존 데이터베이스 처리방식으로 해결할 수 없는 데이터의 집합으로 정의하고 이를 처리할 수 있는 기술이나 역량을 보유한 기업이나 국가가 미래의 경쟁력을 갖게 될 것이다. 매킨지의 분석에 따르면 전 세계 인구의 60%에 달하는 40억 명이 모바일 폰을 사용하고 있으며 인구의 12% 수준이 보유한 스마트폰은 수년 내에 모든 모바일 폰을 대체할 것이다. 이에 빅 데이터를 제대로 활용할 경우 가장 큰 효과를 얻을 수 있는 분야로 표4.에 제시된 다섯 가지 도메인을 예시하였다[4][7].

4.1 빅 데이터 분석 기법

중세 서양에서는 광산에서 금을 채굴(Mining)하는 것이 아니라 다른 금속을 변환시키고자 하는 연금술이 성행했다. 정보화 사회인 현재는 다량의 적재된 데이터에서 숨겨진 패턴과 관계 등을 파악해 광맥을 찾아내듯, 가치 있는 의사결정의 근거 자료로 제시되는 데이터 마이닝의 '연금술'이 필요하게 되었다.

빅 데이터 처리는 기존 데이터 처리와 어떻게 다를까? IT 시장조사기관 Gartner는 2011년 1월 발간한 보고서 'Big Data Analytics'에서 기존 데이터 처리와 빅 데이터 처리에 대해 표5.와 같은 차이점을 설명하였다[8][9][10].

표 4. 매킨지에서 제시한 빅 데이터 활용분야

도메인	분석대상 데이터	예상효과
미국의 의료산업	제약사 연구개발 데이터 환자 치료 임상 데이터, 의료산업의 비용 데이터	연간 $3조 연간 0.7% 생산성 향상
유럽의 공공행정	정부의 행정업무에서 발생하는 데이터	연간 £2.5조 연간 0.5% 생산성 향상
소매업	고객의 거래 데이터, 구매 경향	$1조 + 서비스업자 수익 $7조 소비자 이익
제조업	고객 취향 데이터, 수요 예측 데이터. 제조과정 데이터, 센서활용 데이터	60% 마진 증가 0.5~1.0% 생산성 향상
개인 위치 데이터	개인, 차량의 위치 데이터	개발 및 조립비용 50% 감소, 운전자본 7%감소

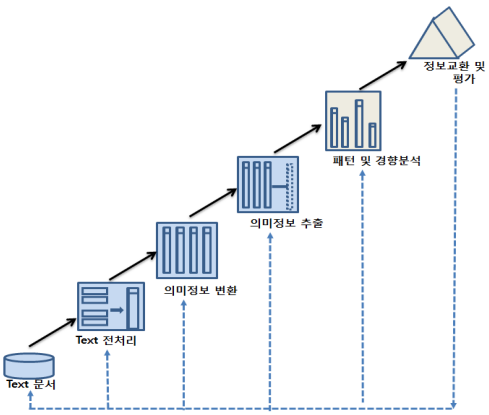
표 5. 빅 데이터의 처리 성격

구분	처리 특징
Speed of Decision making	빠른 의사결정이 상대적으로 덜 요구된다. .장기적/전략적 접근이 필요하다. .즉각적인 의사결정이 상대적으로 덜 요구
Processing Complexity	처리 복잡도가 높다. .다양한 데이터 소스, 복잡한 로직 처리, 대용량 데이터 처리로 처리 복잡도가 높아 분산처리 기술이 필요하다.
Data Volumes	처리할 데이터량이 방대하다. .고객 정보수집 및 분석을 장기간에 걸쳐 수행해야 하므로 처리해야 할 데이터량이 방대하다.
Data Structure	비정형 데이터의 비중이 높다. .소셜 미디어 데이터, 로그 파일, 클릭스트림 데이터, 콜 센터 로그, 통신 CDR 로그 등 비정형 데이터 파일의 비중이 높다.
Analysis flexibility	처리/분석 유연성이 높다. .잘 정의된 데이터 모델/상관관계/절차 등이 없어, 기존 데이터 처리방법에 비해 처리/분석의 유연성이 높다.
Throughput	동시처리량이 낮다. .대용량 및 복잡한 처리를 특징으로 하고 있어, 동시에 처리가 필요한 데이터량이 적다. .실시간 처리가 보장되어야 하는 데이터 분석에는 부적합하다.

빅 데이터 처리의 특징을 만족시키기 위해 다양한 스토리지, 컴퓨팅 기술 및 분석기법들이 개발되었다. 본 절에서는 텍스트/오피니언 마이닝, 소셜 네트워크 분석, 군집 분석 등에 대하여 소개한다[5][14][15].

1) Text Mining

텍스트 마이닝은 비·반정형 텍스트 데이터를 자연어 처리 기술에 기반하여 유용한 정보를 추출하여 가공하는 것을 목적으로 하는 기술이다. 이 기술을 통하여 방대한 텍스트 뭉치에서 의미있는 정보를 추출해 내고, 다른 정보와의 연계성을 파악하며, 텍스트가 가진 카테고리를 찾아내는 등, 단순한 정보 검색 그 이상의 결과를 얻어낼 수 있다. 데이터로부터 정보를 추출 및 분석하여 정보를 재생산하는 텍스트 마이닝 과정은 그림 4.과 같은 단계를 거친다.



▶▶ 그림 4. 텍스트 마이닝 과정

텍스트 마이닝 애플리케이션의 성공 사례가 늘어남에 따라 정형 및 비정형화된 데이터를 동시에 분석하는 텍스트 마이닝 기술은 전 세계 모든 곳에서 건전한 조직의 필수 요소로 자리잡게 될 것으로 전망되고 있다.

주요 응용분야로는 문서 분류(Document Classification), 문서 군집(Document Clustering), 정보 추출(Information Extraction), 문서 요약(Document Summarization) 등이 있다.

2) Opinion Mining

최근 새로운 여론 분석 기술로 각광받고 있는 오피니언 마이닝은 웹사이트와 소셜 미디어에 나타난 여론과 의견을 분석하여 유용한 정보로 재가공하는 기술을 말한다. 오피니언 마이닝 기술을 활용하면 네티즌들이 각각의 사건에 대하여 이야기하는 댓글이나 포스팅 등을 긍정 또는 부정으로 분류하여 더 객관적이고 정확하게 평판을 파악할 수 있다.

최근에는 블로그와 커뮤니티 같은 웹사이트뿐 아니라 트위터, 페이스북과 같은 SNS의 중요성이 점차 커지면서 소셜 미디어 분석시장에 뛰어든 국내 기업들도 늘어나고 있는 추세로 온라인에 숨겨진 의미를 발굴하고, 통계화를 통해 의미있는 정보로 재가공하는 오피니언 마이닝은 다가오는 소셜 미디어 시대의 중요한 기술로 자리매김하고 있다. 오류를 피하기 위하여 사용되는 오피니언 마이닝 연구의 3단계는 다음과 같다.

첫째, '주관성 분석'으로 주어진 텍스트가 주관적인지 객관적인지 결정하는 것으로, 주어진 텍스트에 나타난 저자의 태도를 판단하는 단계이다.

둘째, '극성 분석'으로 텍스트가 주관적인 의견을 갖고 있을 경우 긍정인지 부정인지 분류하는 단계이다.

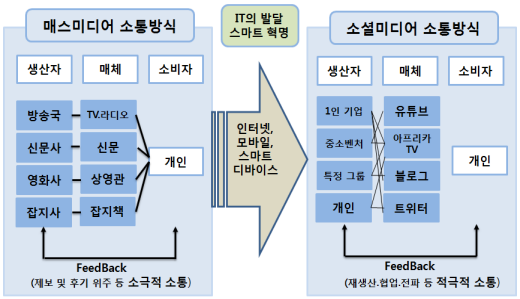
셋째, '극성의 정도 분석'으로 주관적인 텍스트에 대하여 긍정적인 정도와 부정적인 정도를 측정하는 단계이다.

오피니언 마이닝은 특정 서비스 및 상품에 대한 시장 규모 예측, 소비자의 반응, 입소문 분석 등에 활용되고 있으며, 공공분야의 경우 민원의 원인이나 문제점 등을 파악하는 것이 용이해 서비스를 개선할 수 있다. 또한 기업의 경우에는 특정 제품에 대한 고객의 반응을 빠르게 파악하고 선호도를 역으로 추론하는데 효과적으로 활용할 수 있다. 정확한 오피니언 마이닝을 위해서는 전문가에 의한 선호도를 나타내는 표현·단어 자원의 축적이 필요하다.

3) Social Network Analytics

소셜 네트워크 분석은 수학의 그래프 이론에 기초하고 있다. 스마트폰, 태블릿 PC 등의 이동성, 편의성, 실시간성이 강화된 스마트 디바이스의 발달은 소셜 미디어의 확산 및 활성화를 촉진시키고 있으며, 개인의 주관적인 생각이나 경험을 바탕으로 한 정보를 공유 및 재가공하는 등 '참여, 소통, 공유' 기반의 뉴미디어인 소셜 미디어 네트워크의 연결 구조를 바탕으로 사용자의 명성 및 영향력을 측정하여 활용하는 기술이다.

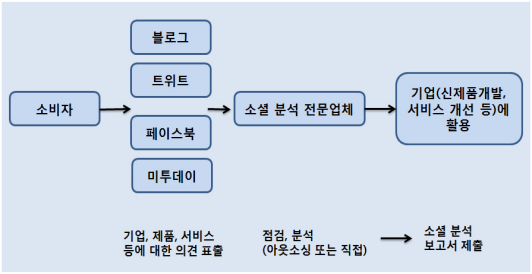
IT의 발달과 스마트 혁명의 본격화로 더욱 활성화된 소셜 미디어는 그림5.와 같이 정부와 국민, 기업과 소비자, 개인과 개인의 소통방식에 혁신적 변화를 가져왔다[그림 5].



자료: 한국정보화진흥원

▶▶ 그림 5. IT의 발달과 소통방식의 변화

소셜 네트워크 분석은 텍스트 마이닝 기법에 의해 주로 이루어져 왔다. 소셜 프로세스 과정은 그림 6.과 같다.



▶▶ 그림 6. 소셜 프로세스

SNS는 개인을 노드(Node), 개인의 사회적 관계를 링크(Link)로 간주하면 소셜 네크워크를 구축할 수 있고 이렇게 형성된 소셜 네트워크에서 다음의 4단계를 통해 정보를 추출 및 분석할 수 있다.

첫째, 소셜 네트워크의 위상학적 구조(Network Topology Structure) 분석으로 네트워크 전반적 특성을 파악한다.

둘째, 네트워크 구조의 시간에 따른 진화를 분석한다.

셋째, 네트워크상의 각 노드(사용자)가 생산, 확산시키는 콘텐츠(포스트, 댓글, 리트윗, 동영상, 링크 등) 흐름을 분석한다.

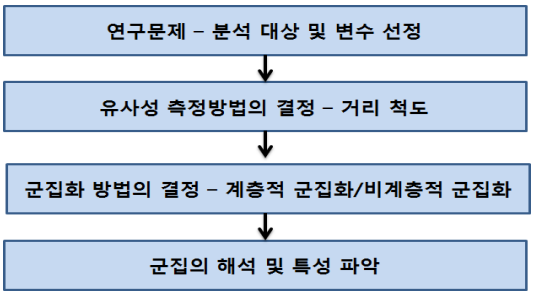
넷째, 종합하여, 각 개인 또는 그룹의 소셜 네트워크 내 영향력, 관심사, 성향 및 행동 패턴을 분석 추출한다.

소셜 네트워크 분석의 활용 효과는 이미 각 기업들이 빅 데이터에 주목하면서 SNS 데이터 분석기술을 통해 방대한 비정형 데이터들을 분석하고 이를 비즈니스에 활용하고 있다. 하지만 그 부작용 최소화를 위한 자체 모니터링, 위험 완화 프로그램 개발 등 관련기업·시장의 사회적 책임이 강조되어야 하며, 프라이버시 보호 등 부작용 대응을 위한 기술개발 및 산업육성이 지원되어야 한다.

4) Cluster Analysis

군집분석은 각 객체(대상)의 유사성을 측정하여 높은 대상 집단을 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체간의 상이성을 규명하는 통계 분석 방법이다.

군집분석을 위하여 가장 흔히 사용하는 자료는 그림7.과 같이 간격척도 혹은 비율척도로 측정된 거리값(Distance measures)의 경우에 따라 서열척도로 측정된 값들로 군집분석이 가능하다.



▶▶ 그림 7. 분석 절차

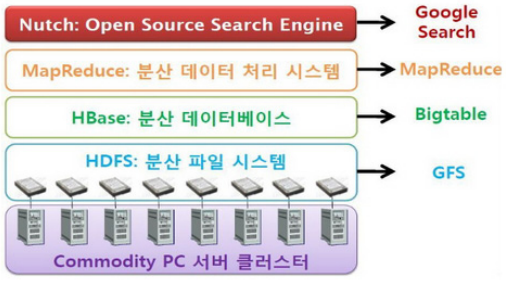
계층적 군집 분석은 문제점은 없지만 군집의 수를 사전에 지정해 주어야 한다. 현실적으로 계층적 방법에 의해 군집화를 한 다음 그 결과로부터 가장 적절한 수의 군집수를 결정하여 다시 비계층적 방법에 의해 분석하면서 그 수를 지정하는 방법을 사용한다. 아울러 계층적 군집분석에서 나타나는 예외값들을 이때 제거하는 것이 바람직하다.

4.2 빅 데이터 분석 인프라 기술

빅 데이터 분석기법들은 테라 바이트 또는 페타 바이트 규모의 데이터에 적용되고 있다. 그렇다면 엄청난 규모의 빅 데이터 분석을 수행할 수 있는 인프라 기술은 어떤 것이 있을까?

1) Hadoop

하둡은 오픈 소스 분산처리기술 프로젝트로, 현재 정형/비정형 빅 데이터 분석에서 가장 선호되는 솔루션이다. 실제로 야후와 페이스북 등에 사용되고 있으며 채택하는 회사가 늘어나고 있다. 하둡의 주요 구성요소는 그림 8.처럼 하둡 분산 파일 시스템인 HDFS(Hadoop Distributed File System), HBase, MapReduce의 3가지이다[11][12].



▶▶ 그림 8. 하둡 구조 & 대응하는 구글 분산처리기술

HDFS와 HBase는 각각 구글의 파일 시스템인 GFS(Google File System)와 빅 테이블의 영향을 받았다. 기본적으로 비용 효율적인 x86 서버로 가상화된 대형 스토리지(HDFS)를 구성하고, HDFS에 저장된 거대한 데이터셋을 간편하게 분산처리할 수 있는 Java 기반의 MapReduce 프레임워크를 제공한다.

2) R

오픈 소스 프로젝트 R은 통계 계산 및 시각화를 위한 언어 및 개발환경을 제공하며, R 언어와 개발환경을 통해 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현/개선이 가능하다. 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 인하여 R은 통계 분석 분야에서 인지도를 높여왔으며, 하둡 환경 상에서 분산처리를 지원하는 라이브러리 덕분에 구글, 페이스북, 아마존 등의 빅 데이터 분석이 필요한 기업에서 대용량 데이터 통계분석 및 데이터 마이닝을 위해 널리 사용되고 있다[13].

3) NoSQL

NoSQL은 Not-Only SQL, 혹은 No SQL을 의미하며, 전통적인 관계형 데이터베이스RDBMS와 다르게 설계된 비관계형 데이터베이스를 의미한다. 대표적인 NoSQL 솔루션으로는 Cassandra, Hbase, MongDB 등이 존재한다. NoSQL은 테이블 스키마(Table Schema)가 고정되지 않고, 테이블 간 조인(Join) 연산을 지원하지 않으며, 수평적 확장(Horizontal Scalability)이 용이하다는 특징을 가진다. 관계형 데이터베이스의 경우, 일관성(Consistency: 모든 노드는 같은 시간에 같은 데이터를 보여줘야 한다)과 유효성(Availability: 일부 노드가 다운되어도 다른 노드에 영향을 주지 않아야 한다)에 중점을 두고 있는 반면, NoSQL 기술은 분산가능성(Partition Tolerance: 네트워크 전송 중 일부 데이터를 손실하더라도 시스템은 정상 동작을 해야 한다)에 중점을 두고 일관성과 유효성은 보장하지 않는다. 이것은 일관성, 유효성, 분산가능성 중 2가지만 보장이 가능하다는 분산 데이터베이스 시스템 분야의 CAP 이론에 따른 것이다. 따라서 대규모의 유연한 데이터 처리를 위해서는 NoSQL 기술이 적합하지만, 안정성이 중요한 시스템에서는 오랫동안 검증된 관계형 데이터베이스를 채택할 필요가 있다.

5. 결론

정보 통신 기술의 발달과 소셜 미디어의 급속한 확산으로 빅 데이터의 수집과 분석, 활용 기술이 전 세계적으로 핫 이슈가 되고 있다. 빅 데이터 분석 기술은 정형 및 비정형의 대용량 데이터를 활용·분석하여 가치있는 정보를 추출하고, 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하는데 있어서 필수 요소가 아닐 수 없다. 이렇듯 빅 데이터를 활용한 기대 가치는 기존 사업의 효율적 지원을 위한 심층적인 데이터 기반 서비스 제공이라는 가치와 빅 데이터를 활용한 신규 수익원 창출이라는 가치를 갖는다.

빅 데이터를 잘 활용하여 성장한 기업들의 예를 살펴보자[6].

기업	사례
Google	가장 정교한 검색결과 제공 빅 데이터 처리 핵심기술 MapReduce 공개
facebook	빅 데이터 처리 최고 자리를 두고 구글과 경쟁 중
amazon.com	사용자 정보처리를 통해 제안되는 '추천' 시스템에서 전체 매출 30% 발생
NETFLIX	사용자 기호, 이용행태 기반 추천시스템으로 DVD 수요분산과 롱테일 정착

RDBMS에 저장되지 않은 문서, 비정형 텍스트, 웹 로그 데이터 및 기타 콘텐츠에도 소중한 정보가 담겨 있다. 이러한 데이터는 변경되는 일도 드물다. 이러한 정보를 처리하기 위해서는 고도의 분석 기술이 요구된다. 즉, 기존에 데이터베이스로 관리해 오던 정형화된 형태의 데이터뿐만 아니라 증가하는 비정형의 대용량 데이터까지도

비즈니스 효과를 창출하는 데 활용되어야 한다. 이를 위해 정형 데이터와 비정형 데이터를 통합 분석하여 단일 솔루션 내에서 처리하는 기술이 요구된다. 대표적인 기술로는 하둡, 데이터 플랫폼, 맵리듀스 등이 있다. 그리고 이를 처리할 수 있는 기술이나 역량을 보유한 기업이나 국가가 미래의 경쟁력을 갖게 될 것이다. 향후 DBMS 시장은 DB 측면에서 정형·비정형 데이터를 어떻게 통합해 비즈니스에 어떤 통찰력을 제공해 줄 수 있는가가 핵심 이슈가 될 것이다. 이에, 본 논고에서는 빅 데이터의 활용과 분석 기술 및 인프라 기술에 대하여 고찰하였다.

참 고 문 헌

[1] 고수연, '빅 데이터' 분석으로 기업 경쟁력 극대화, 컴퓨터월드, pp.50-54, Feb, 2012.

[2] 이현재, Big Data를 위한 S/W 아키텍처 구현, 2011 한국소프트웨어 아키텍트 대회 논문집, pp.433-438, 2011.

[3] 이각범, 빅 데이터를 활용한 스마트 정부 구현(안), 국가정보화전략위원회보고서, 2011. 10.26

[4] 이만재, 빅 데이터와 공공 데이터 활용, Internet and Information Security, Vol. 2 No.2, 2011.

[5] 조성우, Big Data 시대의 기술, KT종합기술원 중앙연구소, 2011.

[6] 이성춘, 임양수, 안민지, Big Data, 미래를 여는 비밀 열쇠, KT경제경영연구소, 2011.

[7] Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011.

[8] Health, T.,& Bizer, C., Linked data: Evolving the web into a global data space, San Rafael, CA:Morgan & Claypool, 2011.

[9] Hilbert, M., & Lopez, P., The world's technological capacity to store, communicate, and compute information, Science, 332(6025), pp.60-65, 2011.

[10] Lavigne, V., & Goulin, D., Applicability of visual analytics to defense and security operations, Proceedings of the 16th International Command and Control Research and Technology Symposium, 2011.

[11] Managing Big Data with Hadoop & Vertica, Vertica Systems, 2009.

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

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

[14] http://cassandra.apache.org/

[15] http://www.mongodb.org/

저 자 소 개

● 김 정 숙(Jung-Sook Kim) 정회원

- 1994년 2월 : 광운대학교 전자계산학과 (이학사)
- 1999년 2월 : 동국대학교 컴퓨터공학과 (공학박사)
- 2000년 3월 ~ 2001년 2월 : 김포대학 컴퓨터계열 전임강사
- 2001년 3월 ~ 현재 : 삼육대학교 컴퓨터학부 교수

<관심분야> : 웹프로그래밍, 임베디드시스템, 모바일 컴퓨팅, 프로그래밍 언어, 컴파일러