조사보고서 16-08

KCMI
Korea Capital
Market Institute

해외 자본시장의 빅데이터 도입 현황 및 시사점

이성복

자본시장연구원

해외 자본시장의 빅데이터 도입 현황 및 시사점

2016. 12.

연 구 위 원 이 성 복

자본시장연구원
Korea Capital Market Institute

序 言

2007년 나심 탈레브(Nassim N. Taleb)는『The Black Swan』책에서 빅데이터를 이렇게 평가하였다: 빅데이터는 더 많은 정보를 뜻할 수 있지만, 더 잘못된 정보를 뜻하기도 한다(Big data may mean more information, but it also means more false information). 이는 빅데이터에서 유용한 정보와 지식을 추출하는 것이 매우 어려운 작업이 될 수 있다는 점을 시사한다. 또한 1990년대말부터 유행하기 시작했던 데이터 마이닝(data mining)이 그 유용성을 증명하지 못한 채 2000년대 중반에 사라진 것처럼 빅데이터도 한 때의 유행으로 그칠 수 있다는 점을 시사하기도 한다.

그럼에도 불구하고 각 국에서는 빅데이터를 새로운 경제성장의 동력으로 인식하고 빅데이터를 활성화시키는 정책을 꾸준하게 추진하고 있다. 금융권에서도 핀테크(fintech)를 접목하는 과정에서 빅데이터를 활발하게 활용하고 있는 추세이다. 빅데이터가 새로운 수익 창출의 기회를 확장시키고 비용을 크게 절감시킬 수 있다고 기대하기 때문이다. 2016년 들어 국내 금융권에서도 빅데이터의 구체적인 활용방안을 고심하기 시작하였다. 그러나 그 논의는 주로 은행과 보험권역에 한하여 이루어지고 있으며, 자본시장과 금융투자업계는 상대적으로 뒤쳐져 있다.

본 보고서는 이러한 배경을 바탕으로 해외 자본시장의 빅데이터 도입 현황을 살펴보고 국내 자본시장과 금융투자업계에 주는 시사점을 제시하였다는 점에서 의의가 있다. 이와 함께 본 보고서는 자본시장과 금융투자업계가 빅데이터의 기술적인 측면뿐만 아니라 법제도적 환경를 쉽게 이해할 수 있도록 빅데이터의 개념, 분석기법, 정책적 논의, 파급효과 및 한계도 간략히 소개하였다.

빅데이터는 그 용어에서 살펴볼 수 있는 것처럼 모호성이 매우 강하다. 그만큼 엄격한 접근이 필요하다. 본 보고서에서 지적한 바와 같이 단기적으로는 빅데이터 도입에 따른 편익은 비용에 비해 크지 않을 수 있다.

그러나 향후 빅데이터 기술은 자본시장과 금융투자업계의 경쟁력을 좌우하는 잣대가 될 가능성이 높다. 본 보고서에서도 이 점을 고려하여 국내 자본시장과 금융투자업계가 장기적인 관점에서 빅데이터를 효과적으로 활용할 수 있도록 구체적인 빅데이터 전략을 수립하고, 이에 필요한 지원을 아끼지 않을 것을 제안하고 있다.

끝으로 본 보고서를 작성하는 데 노고를 아끼지 않은 이성복 연구위원에게 깊은 감사의 뜻을 전한다. 또한 지정논평을 맡아주신 조성훈 선임연구위원과 원내 세미나에서 유익한 논평을 해 주신 여러 연구위원들에게도 감사드린다. 아울러 자료조사 및 원고정리를 위해 수고한 김현숙 선임연구원과 이수련 연구조원에게도 감사한다. 참고로 본 보고서의 내용은 필자의 개인적인 의견이며 본 연구원의 공식적인 의견이 아님을 밝혀둔다.

2016년 12월

자본시장연구원

원장 안동현

목 차

표 목 차

그 림 목 차

약 어 표

AI	Artificial Intelligence
AML	Anti-Money Laundering
API	Application Programming Interfaces
BOA	Bank of America
BSA	The Software Alliance
CDS	Credit Default Swap
CIO	Chief Information Officer
CRM	Customer Relation Management
CSC	Computer Sciences Corporation
ELS	Equity-Linked Securities
ETF	Exchange Traded Funds
EU	European Union
FRB	Federal Reserve Board
FTC	Federal Trade Commission
GDP	Gross Domestic Product
HFT	High Frequency Trading
ICO	Information Commissioner's Office

KYC	Know Your Customer
MDC	Market Data Cloud
NLP	Natural Language Processing
OECD	Organization for Economic Cooperation and Development
PCAST	President's Council of Advisors on Science and Technology
SEC	Securities and Exchange Commission
SNS	Social Network Services
SQL	Structured Query Language

≪ Executive Summary ≫

본 보고서는 빅데이터에 대한 기본적인 이해를 위해 빅데이터 개념, 분석기법, 정책적 논의, 파급효과 및 한계를 살펴보았고, 해외 자본시장의 빅데이터 도입 현황을 빅데이터 도입 효과, 빅데이터 기업 및 분석기법 사례, 빅데이터 도입 특징으로 나누어 살펴보았다.

빅데이터는 기존 데이터베이스 시스템으로는 저장, 처리, 분석하기 어려운 크기의 정형, 반정형, 비정형 형식을 가진 실시간 데이터를 처리하는 기술이다. 즉 빅데이터는 규모(volume), 다양성(variety), 속도(velocity) 측면에서 기존 데이터 기술과 확연히 다르다고 평가받는다. 빅데이터 기술의 핵심은 데이터 분산저장과 분산처리에 있다. 이 때문에 빅데이터 기술은 대용량 데이터를 저장하고 관리할 수 있으며, 분산저장된 데이터를 빠른 속도로 처리하고 분석할 수 있다. 따라서 빅데이터는 데이터 자체의 특성에도 의미가 있지만 데이터를 수집하고 저장하고 분석하는 기술에 더 큰 의의가 있다. 빅데이터 기술은 최근 실시간 데이터 처리기술의 진전으로 일괄처리 방식(batch-processing)에서 실시간처리 방식(real-time processing)으로 진화하고 있다.

빅데이터 분석의 궁극적인 목적은 경제주체의 의사결정 효율성을 높이는 데 있다. 전통적인 분석기법은 과거 데이터에 초점이 맞추어져 있기 때문에 이에 대한 통계적 의미를 찾는 데 집중하였으나, 빅데이터 분석기법은 빠르게 생성되는 데이터의 특성을 살려 예측분석에 초점을 맞추는 것으로 조사된다. 또한 빅데이터 분석은 정형뿐만 아니라 비정형 형식의 대규모 데이터를 분석하기 때문에 기존 데이터 분석과 달리 인과관계(causation)보다는 상관관계(correlation)에 초점을 두고 있다. 이 때문에 명확한 분석목적을 수립하고 이에 맞는 분석기법을 선택하는 것이 빅데이터

분석에서는 매우 중요할 수 있다.

빅데이터가 경제 전반의 효율성을 높일 수 있으려면 빅데이터 자체의 유통이 활성화되어야 한다. 이 때문에 각 국에서는 빅데이터 유통의 활성화를 위해 공공데이터 개방 정책, 민간데이터 공유 정책, 빅데이터 윤리 정책을 적극적으로 추진하고 있다. 참고로 이러한 정책적 논의는 이전에도 존재했으나 전 세계적으로 빅데이터의 중요성이 크게 부각되면서 새로운 관점에서 재논의되거나 강화되고 있는 추세이다.

먼저 각 국에서는 정부, 공공기관 등이 보유하고 있는 공공데이터를 민간에게 개방하기 위해 공공데이터 포털을 개설하고, 공공데이터 활용의 성공사례를 민간에게 적극적으로 전파하고 있다. 공공데이터가 민간의 새로운 상품 및 서비스 개발에 중요한 역할을 할 것으로 기대하고 있기 때문이다. 각 국의 공공데이터 개방 정책은 대체적으로 2013년 6월 G8 정상회담이 채택한 공공데이터 선언문(Open Data Charter)에 기초하고 있으며, 법과 제도 안에서 민간이 공공데이터를 자유롭게 이용할 수 있도록 허용하고 있다.

또한 각 국에서는 공공데이터뿐만 아니라 민간이 보유한 데이터가 원활하게 공유되도록 관련 정책을 적극적으로 추진하고 있다. 다만 개인정보가 포함된 민간데이터가 무분별하게 공유될 경우 개인의 사생활이 침해될 수 있기 때문에 민간데이터 공유 정책은 아직까지 개인정보 보호에 초점을 맞추고 있다. 그러나 복잡한 개인정보에 대한 동의절차가 빅데이터의 경제적 가치를 훼손시킬 수 있다는 우려가 지속적으로 제기되고 있으며, 최근에는 개인정보 비식별화와 Open API가 민간데이터 유통을 활성화시킬 수 있는 새로운 대안으로 제시되고 있다.

빅데이터 윤리의 정립도 개인정보가 포함된 빅데이터가 원활하게

유통되기 위해 매우 중요한 정책적 과제로 대두되고 있다. 빅데이터 기술 자체는 옳고 그르거나 좋고 나쁘다고 판단할 수 있는 대상이 아니기 때문에 빅데이터를 어떻게 활용하느냐는 사람과 기업의 가치기준에 따라 달라질 수 있다. 빅데이터 윤리 문제는 주로 개인의 사생활 보호와 함께 데이터 불법유출, 개인정보 조작, 부당한 차별과 관련된다. 각 국에서는 이러한 문제가 발생되지 않도록 데이터 관리자인 기업에게 적정한 데이터 관리체계(data governance) 및 보안체계(data security)를 구축하도록 요구하고 있다. 또한 빅데이터에 포함된 개인정보가 부당한 차별수단으로 활용되지 못하도록 기업윤리 기준을 제고하고 있다.

빅데이터는 자원 창출이 아닌 자원 배분의 방법으로 경제성장과 생산성 향상에 기여할 것으로 평가된다. 향후 빅데이터가 글로벌 경제 전반의 생산성을 약 1% 향상시킨다면 전 세계 GDP는 2030년까지 15조달러가 증가할 것으로 추정된다. 이는 앞으로 15년내 글로벌 경제에서 미국 경제가 하나 더 생기는 것과 같은 효과이다. 그러나 빅데이터에 대한 긍정적인 시각과 평가만 존재하는 것은 아니다. 과거 데이터 마이닝(data mining)과 같이 빅데이터도 그 유용성이 제대로 증명되지 못할 수 있다는 회의적인 시각도 존재한다. 최근에는 구글의 독감 트렌드, 2016년 미국 대선 결과 등에서와 같이 빅데이터 활용의 실패 사례도 꾸준하게 보고되고 있다. 그럼에도 불구하고 빅데이터는 데이터 마이닝과 같이 하나의 현상에 머물지 않고 하나의 조류로 자리잡아 가는 추세이다. 또한 최근 보고된 빅데이터 분석의 실패 사례도 빅데이터의 유용성 자체를 부정하기 보다는 빅데이터 구축과 분석에 더 엄격한 접근이 필요하다는 점을 강조한다는 점을 유의할 필요가 있다.

자본시장에서의 빅데이터 활용 논의는 복잡한 시장과 데이터 구조 때문에 다른 금융권역에 비해 활발하지 못했던 것으로 조사된다.

그러나 최근들어 실시간 빅데이터 기술의 발전 등으로 자본시장에서도 빅데이터 활용방안에 대한 논의가 활발해지고 있다.

자본시장에서 빅데이터 도입은 새로운 수익창출 기회를 확장시키고, 리스크관리 및 법규준수의 효율성을 증진시키며, 시간과 비용을 대폭 절감시킬 수 있는 것으로 기대된다. 자본시장에서 빅데이터가 중요하게 부각되는 이유중 하나는 자본시장 자체가 수많은 데이터를 직접 생성하고 그 데이터에 의해 민감하게 반응하기 때문이다. 따라서 자본시장은 어느 산업보다 빅데이터에서 의미있는 정보를 추출할 수 있고 이를 통해 새로운 수익기회를 창출할 가능성이 매우 높은 것으로 평가받는다. 예를 들면, 상장기업 분석, 투자전략 수립, 알고리듬 매매, 매매 tick 또는 로그에 대한 빅데이터 분석을 통해 더 효과적인 트레이딩 전략을 수립할 수 있고, SNS 및 뉴스에 대한 빅데이터 분석을 통해 더 정확한 주가예측을 할 수 있다. 또한 빅데이터는 금융투자회사의 리스크관리 및 법규준수에 획기적인 변화를 가져올 것으로 평가받는다. 빅데이터의 실시간 처리능력과 예측 분석기법이 기존 데이터 분석과는 달리 리스크관리의 예측력과 효율성을 높여줄 것으로 기대되기 때문이다. 예를 들면, 빅데이터는 스트레스테스트, 시장감시, 사기탐지, 자금세탁방지, 고객알기정책, 규제보고 등에 활용될 수 있다. 마지막으로 분산저장과 분산처리 기술에 기반한 빅데이터 기술은 자본시장에서 데이터 분석시간과 데이터 관리비용을 대폭 단축 또는 절감시킬 수 있는 것으로 평가받는다.

단기적으로 보면 자본시장에서 빅데이터 도입에 따른 편익은 비용에 비해 크지 않을 것으로 예상된다. 자본시장은 사회, 경제, 정치, 문화 변수뿐만 아니라 기후변화에도 민감하게 영향을 받기 때문에 정형뿐만 아니라 비정형 데이터에 대한 분석 수요가 이전부터 높아왔다. 또한 자본시장은 불특정 다수의 시장참여자가 동시에 실시간으로 다양한 형식의 데이터를 생산하고 축적하기 때문에 시장과

데이터 구조가 매우 복잡하다. 그만큼 자본시장에서 빅데이터 기술을 적절하게 활용하기 위해서는 복잡한 시장과 데이터 구조를 먼저 이해할 필요가 있다. 따라서 자본시장에서 빅데이터를 효과적으로 활용하는 것은 매우 난해한 작업이 될 수 있다. 해외에서나 국내에서도 자본시장에서 빅데이터에 대한 논의가 은행 및 보험권역에 비해 상대적으로 활발하지 못한 것도 이 때문일 수 있다.

중장기적으로 보면 자본시장에서 빅데이터의 유용성은 매우 높을 것으로 판단된다. 기존 데이터 기술로는 자본시장의 복잡한 시장과 데이터 구조를 이해하고 활용하는 데 상당한 한계가 존재했지만, 빅데이터 기술은 이를 효과적으로 극복하는 데 일조할 것으로 판단된다. 해외 자본시장의 빅데이터 기업과 분석기법 사례에서 그 근거를 찾아볼 수 있다.

첫째, Xignite, Quandle과 같은 빅데이터 유통플랫폼은 금융투자회사 등 자본시장 참여자의 빅데이터에 대한 접근성을 완화시킬 수 있다. 자본시장 참여자가 개별적으로 빅데이터를 직접 구축하고 이를 분석할 수 있는 역량을 갖추는 데는 많은 자원과 시간이 소요될 수 있다. 빅데이터 유통플랫폼은 이를 대신할 수 있고, 자본시장 참여자가 이를 활용하는 것이 더 효율적일 수 있다.

둘째, 자연어처리와 기계학습으로 구성된 알고리듬 기술은 자본시장 참여자가 자본시장의 복잡한 시장과 데이터 구조를 학습해야 하는 부담을 완화시킬 수 있다. Kensho, Alphasense와 같은 조사분석 검색엔진이 대표적인 사례이다. 이들 검색엔진은 검색자의 요구에 따라 시장 및 기업에 대한 유의미한 분석보고서를 자동으로 작성하여 제공한다. 이와 같은 빅데이터 기술을 적극 활용하면 자본시장 참여자는 시간과 자원을 크게 절약할 수 있다.

셋째, Dataminr, StockTwits과 같은 빅데이터 기업의 감성분석은

자본시장의 복잡한 시장구조 때문에 과거에 활용하지 않거나 추출하기 어려웠던 정보를 유의미하게 생산할 수 있다는 가능성을 보여주고 있다. 이들 빅데이터 기업은 트윗정보를 자연어처리 및 감성분석을 통해 자본시장의 투자동향과 투자심리를 파악할 수 있게 하고, 자본시장 참여자가 이전보다 더 합리적이고 시의적절하게 의사를 결정할 수 있도록 돕는다. 이를 통해 자본시장의 효율성은 더 제고될 수 있을 것으로 기대된다.

넷째, 최선집행을 최우선 전략으로 삼는 Deep Value와 같은 알고리듬 전문 빅데이터 기업은 자본시장의 복잡한 시장구조를 빅데이터 기술로 극복할 수 있다는 가능성을 보여준다. 빅데이터 기술은 기존 데이터 기술과 달리 알고리듬의 유효성 검증뿐만 아니라 분산되어 있는 다수의 자본시장 데이터를 알고리듬에 반영하는 데 소요되는 시간을 크게 단축시킬 수 있기 때문이다.

다섯째, Scaled Risk와 같은 빅데이터 기업은 빅데이터 기술이 자본시장에서 가중되고 있는 리스크관리 및 법규준수 부담을 완화시킬 수 있다는 가능성을 보여준다.

본 보고서는 빅데이터에 대한 기본적인 이해와 해외 자본시장의 빅데이터 도입 현황에 대한 조사결과를 토대로 다음과 같이 국내 자본시장에 주는 시사점을 제시하고자 한다.

첫째, 금융투자업계는 단기적인 관점에서 빅데이터 도입에 따른 편익이 비용에 비해 크지 않더라도 중장기적인 관점에서 빅데이터 전략을 수립하고 향후 데이터 환경 변화에 적극 대응할 수 있는 역량을 갖출 필요가 있다. 또한 금융투자회사 각자가 보유하고 있으나 활용하지 못하는 데이터를 자체적으로 파악하고 이를 활용하는 노력도 지속할 필요가 있다.

둘째, 자본시장 유관기관 및 금융투자업계는 자본시장에 특화된

빅데이터 전문기업이 출현할 수 있도록 지원할 필요가 있다. 해외 자본시장의 경우에도 자본시장에 특화된 빅데이터 전문기업의 출현으로 자본시장에서 빅데이터 활용이 촉진되었기 때문이다. 또한 금융투자업계는 빅데이터 관련 업체들이 자본시장에 빅데이터 기술을 전파할 수 있도록 이들 업체들과 긴밀한 협업체계를 구축해 나가는 노력도 병행해야 한다.

셋째, 자본시장 유관기관 및 금융투자업계는 자본시장 참여자가 이용할 수 있는 빅데이터 유통플랫폼을 공동으로 구축하는 노력을 지속할 필요가 있다. 국내 자본시장의 환경을 고려할 경우 빅데이터 유통플랫폼이 민간에서 자생적으로 출현하는 것을 기대하기 어렵다고 판단되기 때문이다.

넷째, 자본시장 유관기관 및 금융투자업계는 빅데이터와 관련된 법제도적 환경을 제대로 이해하고, 공공데이터 활용, 데이터 공유, 개인정보 보호, 비식별화 기술, 빅데이터 윤리 등에 대해서도 꾸준한 관심을 갖고 빅데이터 환경 변화에 적극적으로 대응할 수 있어야 한다.

끝으로 빅데이터에 대한 다양한 시각과 평가가 존재하지만 자본시장에서 빅데이터의 중요성과 유용성은 무시될 수 없다고 판단된다. 특히 최근 인공지능에 대한 관심이 높아지는 추세에서 빅데이터의 역할은 매우 광범위해지고 있다. 빅데이터가 인공지능의 가장 기초적인 요소이기 때문이다. 또한 빅데이터는 자본시장의 경쟁력을 가늠하는 하나의 잣대가 되고 있다. 따라서 국내 자본시장도 국제적인 경쟁력을 향상시키고 유지하기 위해 빅데이터 기술과 역량 개발에 지속적으로 노력할 필요가 있다.

≪ Abstract ≫

Big Data Trends in Global Capital Markets and Implications

This report researches how capital markets in the world have incorporated big data technologies into their business in order to find some implications to Korea capital markets and financial investment companies. Applying big data technologies to capital markets is expected to extend opportunities of generating new revenues, improve efficiencies of risk management and regulatory compliance, and reduce time and costs of business operation. However, from a short-term perspective, benefits from big data technologies would not be enough to cover costs of applying them to capital markets, partly because structures of capital markets and their related data are complicated so that they are not easy to be analyzed even by using advanced big data technologies.

Nonetheless, there have been substantial and effective progresses in capital markets as Fintech firms specialized in big data technologies have emerged in capital markets. The report finds five characteristics of big data trends in global capital markets. First, big data sharing platforms to provide various sources of data relating to capital markets are enhancing market participants' accessibilities to big data technologies. Second, applications of advanced natural language processing and machine learning have helped to understand complex structures of markets and data in more efficient ways. Third, applications of sentiment analysis to

capital markets are improving in analyzing investors' investment activities and understanding their phycological changes in investment decision, which has been regarded as information not easily extracted in market data or social media data due to complicated market structures and limits of data analytics. Fourth, big data technologies are promoting effectiveness of strategical algorithm analytics for capital markets, speically by saving time in testing algorithm robustness. Finally, Fintech firms specialized in providing services of risk management and regulatory compliance are mitigating regulatory burdens of financial investment companies.

Based on these findings described above, the report suggests some implications to Korea capital markets and financial investment companies. First of all, it is important to build adequate capabilites of keeping up with changes in data environments and to establish firm-level big data strategies from a middle or long-term perspective. Moreover, incumbents in capital markets need to help stimulate emergence of Fintech start-ups specialized in big data techonologies, as global capital markets are utilizing them in saving costs and enhancing efficiencies of big data applications. Furthermore, it is desirable for them to cooperate each other in sharing big data and also building a big data sharing platform. Finally, environmental or legal issues relating big data have to be followed in a continuous and timely manner.

Ⅰ. 조사 배경 및 목적

Ⅰ. 조사 배경 및 목적

빅데이터(big data)는 '21세기 원유(crude oil)'로 불릴만큼 향후 경제 전반에 큰 영향을 미칠 수 있는 중요한 자원으로 평가받고 있다(Gartner, 2011). 원유는 이전부터 존재했지만 1900년대 초반 이전까지는 제대로 활용되지 못했다. 그러나 이후 정유기술의 발달로 새로운 유형의 제품과 기업을 탄생시키며 경제 전반의 패러다임을 바꾸었다. 빅데이터도 마찬가지일 수 있다. 기존의 데이터 저장 및 처리 기술로는 다양한 형식의 대규모 데이터를 제대로 활용할 수 없었다. 그러나 데이터 분산저장 및 분산처리 기술의 발전으로 다양한 형식의 대규모 데이터를 분석하고 활용할 수 있게 되었다. 이 때문에 빅데이터도 원유처럼 경제 전반의 혁신, 경쟁, 생산성을 새롭게 제고시킬 수 있는 것으로 평가받게 되었다(McKinsey Global Institute, 2011; PCAST, 2014).

빅데이터는 금융산업과 자본시장의 생산성과 효율성을 증진시키는 데도 큰 영향을 미칠 것으로 예상된다(McKinsey Global Institute, 2011; Aite, 2014). 금융산업과 자본시장의 데이터 보유량은 산업 전체의 50%를 차지하며, 자본시장의 데이터 보유량은 은행과 보험의 2배에 달한다. 특히 자본시장은 불특정 다수의 시장참여자가 동시에 실시간으로 데이터를 생산하고 축적하는 특성을 지니고 있기 때문에 빅데이터 활용가치는 다른 금융권역보다 자본시장에서 더 높게 나타날 수 있다. 그동안 자본시장에서 빅데이터 활용 논의는 복잡한 시장과 데이터 구조 때문에 다른 금융권역에 비해 활발하지 못했다(Aite, 2014). 다행히도 최근 들어 빅데이터 기술이 실시간 데이터도 처리할 수 있을 정도로 빠르게 발전하고, 자본시장에 특화된 빅데이터 전문기업이 출현하면서 자본시장에서도 빅데이터 활용방안에 대한 논의가 활발해지고 있다.

본 보고서는 이러한 배경을 바탕으로 해외 자본시장의 빅데이터 도입 현황을 빅데이터 도입 효과, 빅데이터 기업 및 분석기법 사례, 빅데이터 도입 특징으로 나누어 살펴보고 국내 자본시장과 금융투자업계에 주는

시사점을 제시하였다. 또한 본 보고서는 해외 자본시장의 빅데이터 도입 현황을 소개하기에 앞서 빅데이터에 대한 개념, 분석기법, 정책적 논의, 파급효과 및 한계도 살펴보았다. 자본시장에서 빅데이터가 어떻게 도입되고 활용될 수 있는가를 이해하기 위해서는 빅데이터와 관련하여 기술적인 측면뿐만 아니라 법제도적 환경을 먼저 이해할 필요가 있다고 판단하였기 때문이다. 특히 본 보고서에서는 빅데이터가 과거 데이터 마이닝과 같이 일시적인 현상에 그칠 수 있다는 점을 고려하여 빅데이터에 대한 긍정적인 평가뿐만 아니라 회의적인 시각과 한계에 대해서도 논의하였다.

본 보고서는 국내 자본시장 유관기관 및 금융투자업계에게 빅데이터 기술 도입의 중요성을 환기시킬 목적으로 작성되었다. 단기적으로는 자본시장에서 빅데이터 도입에 따른 편익은 비용에 비해 크지 않을 것으로 예상된다. 이 때문에 금융투자업계는 빅데이터 기술의 도입을 주저할 수 있다. 실제 국내의 경우 은행 및 보험권역에 비해 자본시장의 빅데이터 도입 노력은 상대적으로 미흡한 것으로 보고된다. 본 보고서의 또다른 목적은 국내 자본시장 유관기관과 금융투자업계가 향후 빅데이터 구축 및 활용을 위한 사업전략을 수립하는 데 일조하기 위함이다. 이를 위해 본 보고서는 해외 자본시장의 빅데이터 도입 현황과 특징을 조사하여 소개하였다. 예를 들면, 해외의 경우 증권회사 등이 직접 빅데이터 기술을 도입하기 보다는 빅데이터 전문기업이 출현해 증권회사 등의 빅데이터 구축 및 활용을 지원하는 추세이다.

본 보고서는 다음과 같이 구성되어 있다. 제Ⅱ장에서는 빅데이터에 대한 개념, 특징 및 분석기법을 간략히 소개하고, 빅데이터 활성화를 위한 각 국의 정책적 노력, 빅데이터 도입에 따른 파급효과 및 빅데이터에 대한 회의적인 시각을 논의하였다. 제Ⅲ장에서는 해외 자본시장에서의 빅데이터 도입 현황을 빅데이터 도입효과, 자본시장의 빅데이터 기업 및 분석기법 사례, 빅데이터 도입 특징으로 나누어 살펴보았다. 제Ⅳ장에서는 국내 자본시장 및 금융투자업계에 주는 시사점을 제시하였다.

Ⅱ. 빅데이터에 대한 이해

1. 빅데이터 개념 및 핵심기술

2. 빅데이터 관련 정책적 논의

3. 빅데이터 파급효과와 한계

Ⅱ. 빅데이터에 대한 이해

1. 빅데이터 개념 및 핵심기술

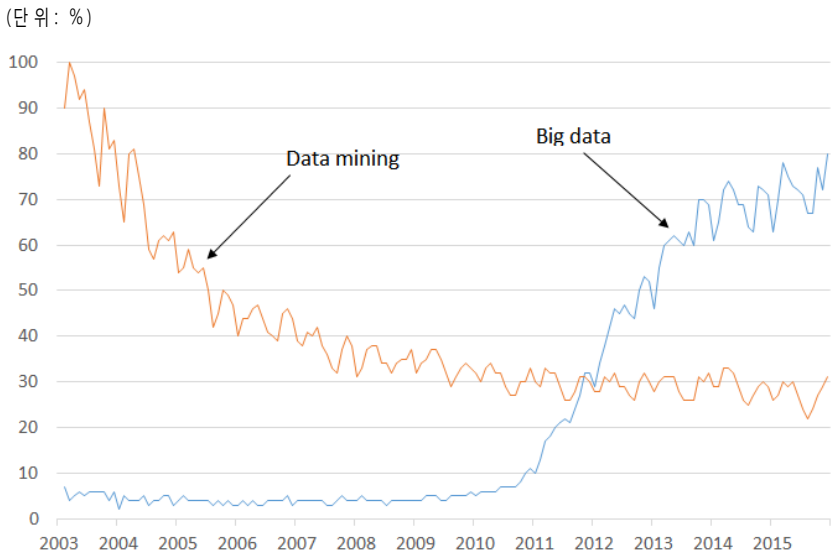
가. 빅데이터 등장배경

1990년대말 인터넷이 보급된 이후부터 데이터 규모는 급속도로 증가하고 그 형태는 계속해서 복잡해지는 추세이다. 이와 함께 컴퓨터 처리속도와 정보통신기술(Information Technology: IT)의 발전으로 기존에 활용되지 못한 데이터의 경제적 가치가 강조되어 왔다. 데이터 마이닝(data mining)도 이 때부터 하나의 유행처럼 빠른 속도로 퍼지며 새로운 데이터 기술로 부각되기 시작했다. 데이터 마이닝은 말 그대로 데이터 속에서 의미있는 정보를 추출하는 일련의 작업을 의미한다. 그러나 당초 기대와는 달리 데이터 마이닝은 광범위하게 활용되지 못하였다. 예를 들면, 2003년 미국은 대표적인 데이터 마이닝 프로그램 중에 하나인 국방부의 Total Information Awareness(TIA)에 대한 국고지원을 중단하였다. 그 주된 이유는 무작위적인 데이터 마이닝이 개인의 사생활을 상당히 침해할 수 있다는 사회적 우려 때문이었다. 이후 그 유용성에 대한 회의적인 평가가 사회 전반으로 확산되었으며, 데이터 마이닝은 점차적으로 쇠퇴하기 시작하였다.

빅데이터가 등장하게 된 배경은 2009년 전후로 알려져 있다. 2006년부터 데이터 분석(data analytics)이라는 용어가 나타나기 시작했으며, 2009년부터 그 개념이 빅데이터(big data)로 대체되었다(Piatetsky-Shapiro, 2012). 이 점에서 빅데이터는 아주 새로운 개념이 아니라는 점을 주지할 필요가 있다. 구글 트렌드(Google Trends) 서비스를 이용해 2004년 이후 빅데이터와 데이터 마이닝에 대한 키워드 검색 추이를 살펴보면, 빅데이터에 대한 관심이 2010년을 기점으로 급속히 증가하였고 2012년부터

데이터 마이닝에 대한 관심보다 앞선 것으로 나타난다.

<그림 Ⅱ-1> 빅데이터와 데이터 마이닝 키워드 트렌드



자료: 구글 트렌드(Google trends)

한편 데이터 마이닝이란 용어는 상대적으로 그 사용빈도가 줄었지만 빅데이터 분석의 기초를 제공했다는 점에서 역사적인 의의가 있다. 데이터 마이닝은 크게 여섯 가지 작업으로 분류된다. 첫째, 특이점 발견(anomaly detection)이다. 데이터에 포함된 특이한 관측치로부터 새로운 경제적 의미를 찾는 작업이다. 둘째, 연계성 학습(association learning)이다. 변수들 간의 상호 연계성을 찾아 일정한 행태적 규칙을 도출하는 작업이다. 셋째, 군집효과(clustering effect) 발견이다. 공통적인 현상을 보이는 집단을 탐색하여 이들의 공통적인 행태적 특성을 발견하는 작업이다. 넷째, 분류 작업(classification)이다. 이미 알려진 데이터 구조를 새로운 데이터에 적용하여 새롭게 분류하고 이를 통해 경제적 의미를 찾는 작업

이다. 다섯째, 회귀분석(regression)이다. 데이터 내 일정 변수간의 인과관계를 분석하고 이에 대한 경제적 의미를 해석하는 작업이다. 여섯째, 요약분석(summarization)이다. 데이터에 포함된 개별 관측치에 대한 통계적 요약을 산출하고 이에 대한 경제적 의미를 찾는 작업이다.

나. 빅데이터 기본개념

빅데이터는 대량의 정형(structured) 또는 비정형(unstructured) 데이터로부터 가치있는 정보를 추출하고 그 정보를 토대로 개인, 특정 그룹 또는 사회 전반의 성향을 파악하고 변화를 예측하는 기술로 정의된다. 즉 빅데이터라는 용어는 데이터 자체의 특성뿐만 아니라 그 데이터를 수집하고 저장하고 분석하는 기술이라는 의미로 주로 사용된다.

빅데이터는 규모(volume), 다양성(variety), 속도(velocity), 정확성(veracity), 가치(value) 측면에서 기존 데이터 기술과는 크게 다른 특징을 갖는다(Laney, 2001). 본 절에서는 공통적으로 논의되고 있는 규모, 다양성, 속도라는 세 가지 측면의 특징을 중점적으로 살펴보았다.

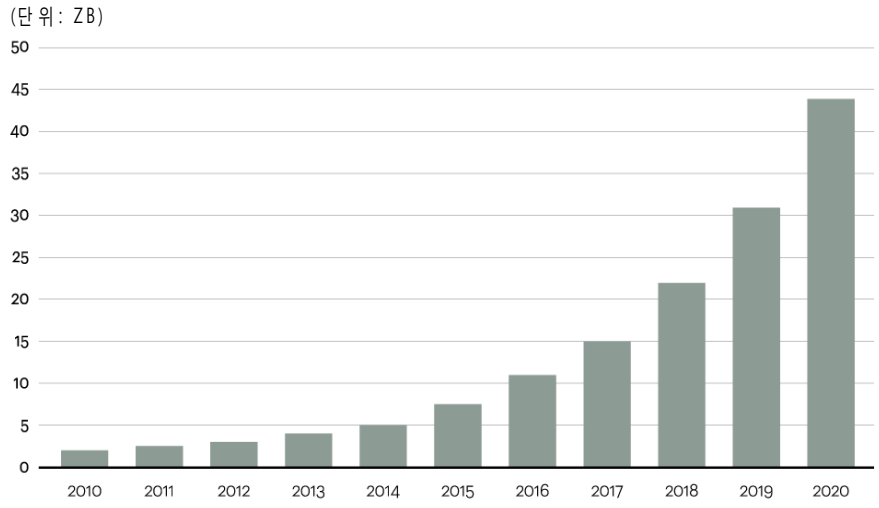
첫째, 빅데이터 기술은 기존 데이터베이스 시스템으로는 저장, 처리, 분석하기 어려운 크기의 대규모 데이터를 처리할 수 있는 기술적 특징을 가진다. 2015년말 기준 전 세계의 데이터 총량은 약 8ZB(Zetabyte)로 추정된다. 1KB(1000byte) 데이터를 밥 한 공기와 같다고 가정할 경우 1ZB(10^{15}KB)는 태평양을 채울 수 있는 데이터의 크기와 같다(Wellman, 2013).[1] 또한 매일 전 세계적으로 2.5EB(Exabyte) 데이터가 생성되고 있는 것으로 추정되며, 매년 데이터가 40%씩 증가할 경우 2020년 전 세계의 데이터 총량은 약 45ZB가 될 것으로 추정된다(Oracle, 2012). 특히

1) 1MB(Megabyte)는 쌀포대 8개에 담긴 데이터의 크기, 1GB(Gigabyte)는 대형트럭 3대에 실린 데이터의 크기, 1TB(Terabyte)는 컨테이너선 2척에 실린 데이터의 크기, 1PB(Petabyte)는 맨하탄을 덮는 데이터의 크기, 1EB는 대한민국의 8.6배를 덮는 데이터의 크기, 1YB(Yottabyte)는 지구를 채울 수 있는 데이터의 크기와 같다.

전 세계 데이터의 90% 이상이 최근 2년 안에 생성된 것으로 조사된다.

한편 빅데이터는 그 자체만으로도 경제적 가치가 있다고 평가받는다. 이전에는 데이터를 어떻게 활용할 수 있는냐(data utilization)가 더 중시되었다. 이 점에서 데이터의 가치(value) 자체를 빅데이터의 또다른 특징으로 보기도 한다. 그러나 2000년대 중반 이후 빅데이터 기술이 정형화된 데이터뿐만 아니라 기존에 사용하지 못했던 모든 유형의 데이터를 활용할 수 있다고 알려지면서 데이터 자체에 경제적 가치가 부여되기 시작하였다. 그러나 빅데이터 전체 규모가 아무리 크더라도 이를 효과적으로 활용할 수 없다면 그 잠재적 가치가 인정되더라도 실질적 가치가 과대평가되어서는 안된다. 또한 데이터 기술뿐만 아니라 법제도적 제약하에서 그 잠재적 가치를 실현하는 데도 많은 시간이 소요될 수 있기 때문에 빅데이터 자체의 경제적 가치는 빅데이터를 얼마나 효과적으로 활용할 수 있느냐와 병행되어 평가될 필요가 있다.

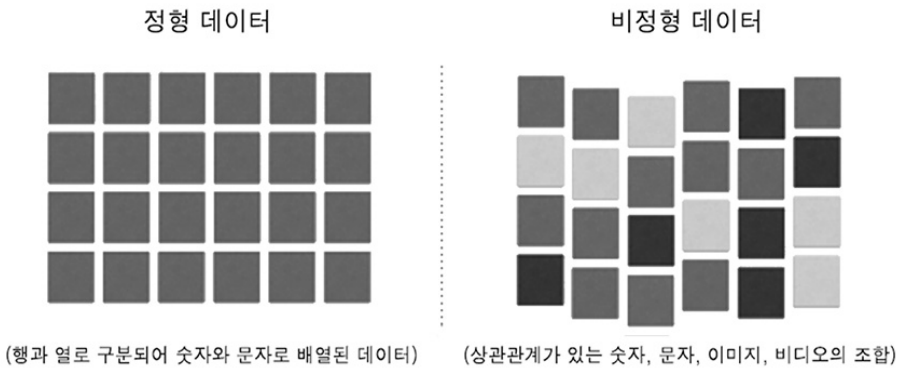
<그림 Ⅱ-2> 빅데이터의 성장 추세



자료 : Oracle(2012)

둘째, 빅데이터 기술은 정형 데이터(structured data), 반정형 데이터(semi-structured data), 비정형 데이터(unstructured data) 등 다양한 형식의 데이터를 저장하고 처리할 수 있는 특징을 가진다. 전 세계 데이터의 90% 이상이 비정형 데이터인 것으로 조사되며, 90% 이상이 기존에 이용되지 않은 데이터로 알려져 있다(Zicari, 2014). 그만큼 빅데이터 기술은 처리하고 분석할 수 있는 데이터 형식이 이전과는 확연히 다르다. 이 때문에 빅데이터를 비정형 데이터라고 정의하는 이도 있다. 빅데이터 기술은 비정형 데이터를 정형 데이터로 전환하기도 하고, 비정형 데이터 그 자체를 분석할 수도 있다.

<그림 Ⅱ-3> 정형 데이터와 비정형 데이터의 차이

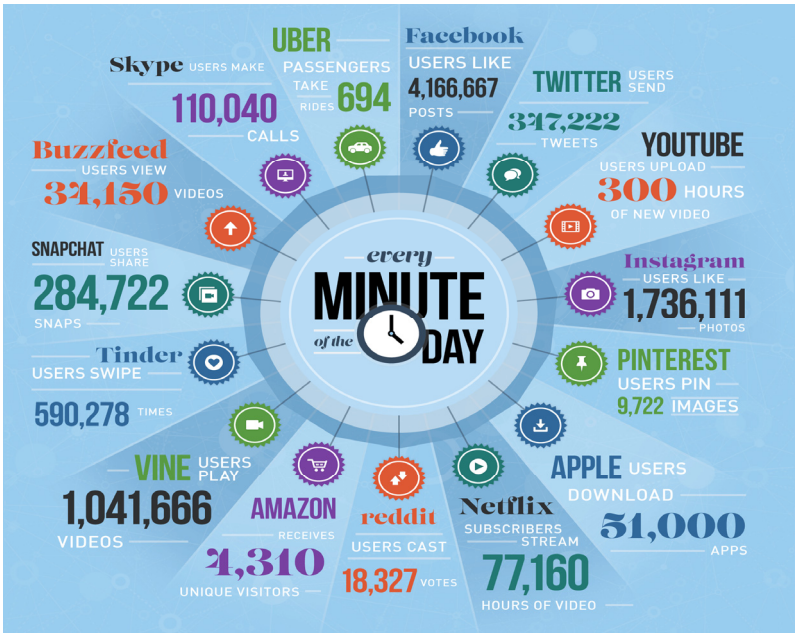


자료 : Delima(2015)

정형 데이터는 숫자와 문자로 구성된 데이터가 행과 열로 배열되어 있는 데이터를 의미한다. 즉 정형 데이터는 마이크로소프트사의 엑셀(Excel)에서와 같이 행과 열로 참조되어 처리할 수 있는 데이터이다. 센서데이터(Radio Frequency ID, GPS, 의료기기 측정값 등), 로그파일(웹사이트, 서버, 애플리케이션 등), POS(Point Of Sale) 데이터, 금융데이터(증권 매수도 호가 및 거래량, 환율, 입출금, 자금이체, 자금조회 등), 고객 입력데이터,

클릭(click) 데이터 등이 정형 데이터에 해당된다. 빅데이터 기술은 이미 생성된 정형 데이터를 활용하기도 하지만 이전에는 기술적 한계로 생성하지 못한 데이터를 정형 데이터로 재구성하여 활용할 수 있기 때문에 이전보다 정형 데이터의 범위를 확장시켜온 것으로 평가받는다. 비정형 데이터는 숫자, 문자, 이미지, 비디오 등 별도의 형식 없이 나열되어 조합할 수 있는 데이터를 의미한다. 위성사진, 지진발생 정도, 대기 데이터, 사진 또는 비디오, 레이더 또는 수중 음파 데이터, 각종 사내 문서, 각종 공시자료, 회의록, 이메일, 고객전화 기록, 소셜미디어 데이터, 웹사이트 내용 등이 비정형 데이터에 해당된다. 개인 컴퓨터에서 작성되거나 저장된 모든 문서파일도 비정형 데이터에 해당된다.

<그림 Ⅱ-4> 빅데이터 생성속도 사례



자료: Hillebrand(2015)

셋째, 빅데이터 기술은 IT기술의 발전으로 데이터의 생성속도가 빠른 만큼 이를 저장하고 분석하는 속도도 빠르다는 특징을 가진다. 최근 빅데이터의 활용 추세를 살펴보면, 빠른 데이터 분석속도가 빅데이터 기술의 가장 유용한 장점으로 평가받는 것으로 조사된다. 특히 최근 실시간 데이터 처리기술이 빠르게 진전되면서 이전에 처리하지 못했던 실시간 데이터의 가치도 덩달아 더 커지고 있다. 예를 들면, 사회관계망서비스(Social Network Service: SNS) 등에서 생성되는 데이터는 빠른 속도로 축적되고 있다. 2015년 기준 트위터(twitter) 이용자는 매 1분마다 약 35만건의 트윗(tweet)을 게시하는 것으로 조사된다. 페이스북(facebook) 이용자도 매 1분마다 약 416만건의 좋아요(like) 반응을 표시하는 것으로 조사된다. 빅데이터 기술은 이렇게 빠르게 생성되는 정보를 활용해 시시각각으로 변하는 시장 트렌트를 파악하는 데 활용되는 것으로 조사된다.

한편 빅데이터 기술은 정확성(veracity)과 같은 특징들로도 설명된다. 데이터에 대용량의 비정형 데이터가 포함될 경우 그 특성상 데이터 자체가 불확실할 수 있고 잘못된 분석기법의 선택으로 잘못된 결과나 해석이 산출될 수 있다. 이 점에서 빅데이터 기술은 앞으로 데이터 분석의 정확성을 제고시키는 방향으로 발전될 필요가 있다. 최근 제기되고 있는 빅데이터에 대한 회의적인 시각도 빅데이터 특성상 정확성이 부족하기 때문인 것으로 파악된다.

다. 빅데이터 핵심기술

빅데이터 기술의 핵심은 하둡(Hadoop)의 데이터 분산저장과 분산처리에 있다. 2006년 하둡 기술이 소개되기 이전에는 처리할 수 있는 데이터 크기가 제한되었고, 데이터 크기에 따라 데이터 분석비용이 크게 증가하였다. 이 때문에 수많은 데이터가 활용되지 못하였다. 그러나 하둡 기술의 소개로 대용량의 정형 데이터뿐만 아니라 비정형 데이터도 저렴한 비용으로 수집, 저장, 분석할 수 있게 되었다. 참고로 빅데이터 기술은 하둡뿐만

아니라 NoSQL(Not only Structured Query Language), 몽고DB(MongoDB), 키산드라(Cassandra) 등과 같은 기술을 포괄하는 개념이다. 그러나 대부분의 빅데이터 기술이 하둡의 분산저장과 분산처리 기술에 기반하고 있다는 점에서 본 보고서는 하둡을 중점적으로 살펴보았다.

하둡은 2003년과 2004년 구글(Google)이 발표한 GFS(Google File System)와 맵리듀스(MapReduce) 기술을 기반으로 개발되었다. 하둡은 당초 구글과 같이 검색서비스에 적용할 기술로 개발되었으나, 기술 개선을 거듭하면서 빅데이터 핵심기술로 발전하였다. 하둡의 데이터분산 저장시스템인 HDFS(Hadoop Distributed File System)는 대용량 데이터(terabypes or petabytes)를 64MB 블록으로 쪼개어 복수의 저장장치(node)에 분산하여 저장하는 파일저장 시스템이다. 이 때문에 데이터 저장장치의 저장용량에 구애받지 않고 대용량 데이터를 저장하고 관리할 수 있다. 단수의 서버에 데이터를 저장하고 처리하는 기존 관계형 데이터베이스 관리시스템(Relational Database Management System: RDBMS)이 파일형식뿐만 아니라 저장용량에도 제약을 받는 것과는 대조적이다.

하둡의 맵리듀스(MapReduce)는 분산저장된 데이터를 분산처리하는 기술이다. 기존 데이터베이스 관리시스템은 데이터 값의 행과 열의 정보를 대상으로 데이터를 분석한다. 이 때문에 분석대상 데이터는 정형 데이터이어야 한다. 그러나 하둡의 맵리듀스는 일차적으로 분산되어 있는 데이터의 저장위치 정보인 노드(node)를 기반으로 맵핑(mapping) 작업을 하고 이를 다시 셔플링(shuffling) 작업을 통해 데이터를 재배열한 후 리듀싱(reducing)으로 의미있는 데이터를 추출하는 과정을 통해 데이터를 분석한다. 이 때문에 맵리듀스는 데이터 형식에도 크게 구애받지 않는다는 특징을 가진다.[2)]

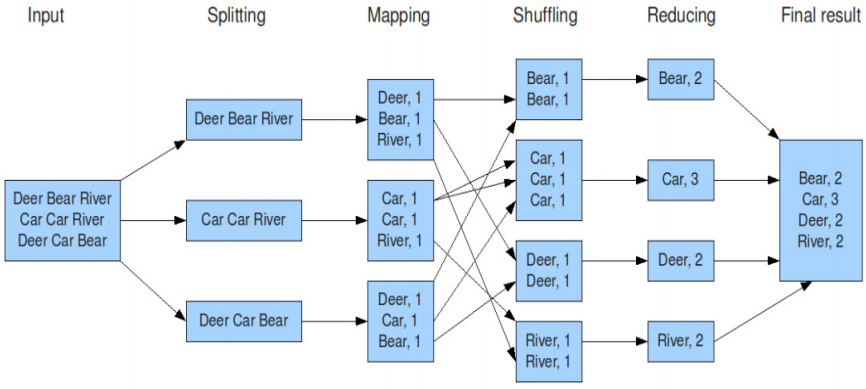
예를 들면, 하둡의 맵리듀스 작업은 <key, value>형식의 정보를 이용해 한 문서 안에 특정 단어가 얼마나 자주 사용되었는지를 통계적으로 쉽게 요약할 수 있다. 이 경우 key는 특정 단어의 값이고, value는 특정 단어의

2) 이에 대한 구체적인 설명은 https://en.wikipedia.org/wiki/MapReduce를 참고하기 바란다.

빈도수이다. 우선 대용량 데이터를 쪼개어 각 노드에 분산저장한다. 이를 데이터 분산처리를 위한 쪼개기(splitting) 작업이라고 한다. 다음 단계에서는 각 노드 안에서 <key, value>를 연산한다. 이를 중간 결과물 산출을 위한 맵핑(mapping) 작업이라고 한다. 그리고 각 노드에 저장된 <key, value>를 재배열해 동일한 <key, value>를 동일한 노드에 저장한다. 이를 최종 결과물 산출을 위한 셔플링(shuffling) 작업이라고 한다. 마지막으로 각 노드에서 재배열된 <key, value>를 통계적으로 요약한다. 이 때 중복 저장된 <key, value>가 새로운 <key, value> 값으로 대체된다. 이를 리듀싱(reducing) 작업이라고 한다. 최종적으로 생산된 <key, value>는 특정 단어(key)와 빈도수(value)이다.

<그림 Ⅱ-5>은 하둡의 맵리듀스 과정을 이해하기 쉽게 사슴(deer), 곰(bear), 강(river), 자동차(car)라는 단어만 있는 문서에서 각각의 사용빈도수를 산출하는 과정을 보여주고 있다. 그러나 실제 하둡의 맵리듀스는 다수의 원시적인 문서에서도 자연어처리(Natual Language Processing: NLP) 등을 통해 단어를 추출하고 각각의 사용빈도수를 추출할 수 있다. 즉 하둡의 맵리듀스는 비정형 데이터를 효과적으로 짧은 시간안에 처리하고 분석할 수 있는 핵심기술이라고 할 수 있다.

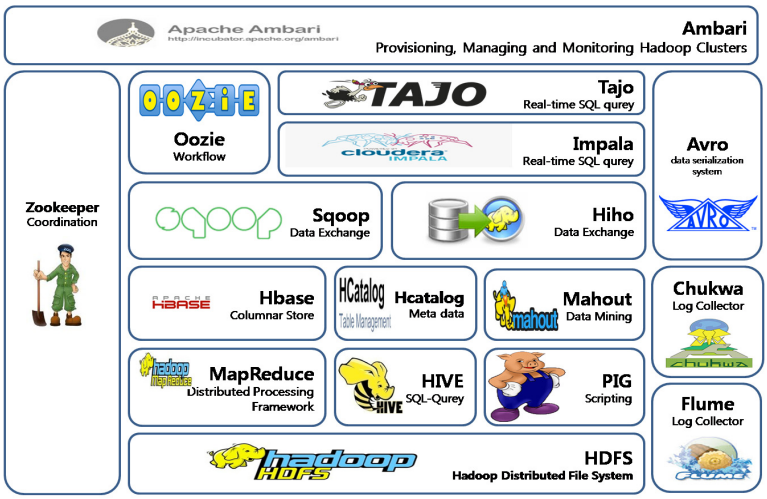
<그림 Ⅱ-5> 하둡의 맵리듀스 처리 예제



자료: MacLean(2011)

이렇게 산출된 결과는 RDBMS, SQL(Structured Query Language), NoSQL, R, Ruby, 검색엔진 등과 같은 분석도구를 이용해 데이터 분석에 활용된다. 하둡의 맵리듀스로 산출된 결과는 새로운 형태의 정형 데이터일 수 있고, 이전보다 함축적인 비정형 데이터일 수도 있다. 예를 들면, 검색엔진은 구글처럼 전 세계에 분산되어 있는 웹페이지에 담겨있는 정보를 맵리듀스로 다시 검색자가 원하는 정보로 요약하여 검색결과를 내놓는다. 2008년 하둡은 RDBMS를 이용할 경우 14년이 걸리는 뉴욕타임스의 130년(1851~1980년)의 신문기사를 24시간만에 저장하고 분석하여 데이터베이스로 구축하는 데 성공하였다(IDG Korea, 2014). 이처럼 이전에는 불가능하거나 상당한 시간과 비용이 소요될 수 있는 작업도 하둡 기술을 이용할 경우 더 저렴하고 빠르게 처리할 수 있다.

<그림 Ⅱ-6> 하둡(Hadoop)의 생태계



자료: http://1004jonghee.tistory.com

하둡의 HDFS와 맵리듀스는 데이터의 분산저장과 분산처리를 가능하게 하는 가장 기본적이고 핵심적인 기술이다. 그러나 더 복잡한 구조의

데이터를 더욱 효율적으로 저장하고 분석하기 위해서는 이를 지원할 하부적인 데이터 관리시스템이 필요하다. Zookeeper, Oozie, Hbase, chukwa, Flume, Scribe, Sqoop, hiho, Pig, Hive, Mahout, Hcatalog, Avro 등이 이러한 목적으로 개발된 빅데이터 분석도구이다. 일반적으로 이를 통틀어 하둡의 생태계(Hadoop's Ecosystem)라고 부른다.[3)]

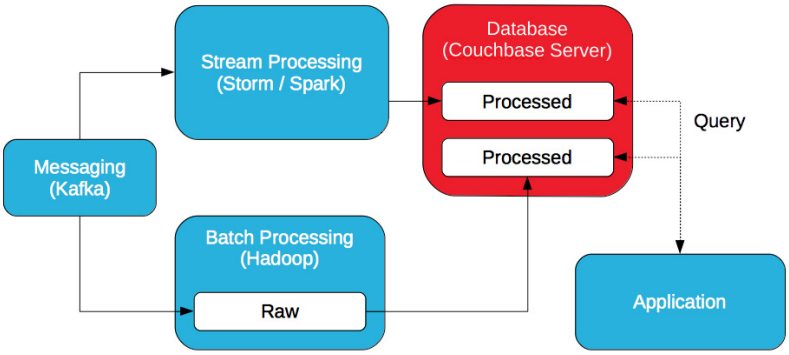
빅데이터 기술은 최근 실시간 데이터 기술의 진전으로 일괄처리 방식(batch-processing)에서 실시간처리 방식(real-time processing)으로 진화하고 있다. 기존의 하둡은 데이터를 저장해야 처리할 수 있는 일괄처리 방식을 채택하였다. 이 때문에 실시간 검색과 분석에 상당한 제약이 따랐다. 그런데 최근 실시간처리 방식에 기반한 데이터 처리 기술인 Storm, Spark 등이 개발되면서 실시간 빅데이터의 활용도가 높아지고 있다. 또한 이들 하부시스템은 인메모리(in-memory) 기술도 지원한다. 인메모리 기술은 각 노드에 데이터를 저장하고 각 노드를 이용해 데이터를 분석하는 것이 아니라 대용량 메모리 자체에 데이터를 저장하여 실시간 분석을 지원하는 기술이다.

빅데이터의 실시간처리 기술은 자본시장에서 빅데이터의 활용 가능성을 매우 높였고 자본시장에서 빅데이터 활용에 대한 논의를 촉진시켰다는 점에서 큰 의미를 갖는다. 예를 들면, 자본시장에서 빅데이터를 통해 투자전략을 추출할 경우 과거 데이터보다 실시간 데이터를 이용하는 것이 더 효과적일 수 있다. 투자전략중 하나는 투자자가 적정한 매수도 시점을 결정하는 것이다. 이를 위해서는 투자자별 매수도 동향, 시세변화, 뉴스, 거시경제 이벤트, 투자자 동향 등을 실시간으로 분석할 수 있어야 한다. 일괄처리 방식으로 추출된 투자전략은 시장상황이 이미 바뀐 상태에서 유효하지 않을 수 있으나, 실시간처리 방식은 변화하고 있는 시장상황을 적시에 반영할 수 있다는 장점이 있다.

---

3) 각 하부 시스템에 대한 자세한 설명은 https://hadoopecosystemtable.github.io/를 참고하기 바란다. 하둡은 개발자 Doug Cutting 아들의 코끼리 장난감의 이름을 본 뜬 것이며, 이후 하부 시스템의 상징이나 이름도 이와 같이 동물을 본 뜨는 것이 관례가 되었다.

<그림 Ⅱ-7> 빅데이터 일괄처리와 실시간처리 기술



자료: softwareengineeringdaily.com

라. 빅데이터 분석기법

빅데이터 분석의 궁극적인 목적은 경제주체의 의사결정 효율성을 높이는 데 있다. 그러나 빅데이터 분석은 다양한 형식의 대규모 데이터를 분석대상으로 하기 때문에 분석목적에 맞는 분석기법을 선택하는 것이 매우 중요할 수 있다. 즉 빅데이터 분석목적에 맞게 데이터를 처리하고 분석하는가에 따라 분석결과가 달라질 수 있고 잘못된 의사결정도 유도될 수 있다. 따라서 빅데이터 분석기법을 제대로 이해하는 것은 빅데이터 활용에 있어서 매우 중요한 선결조건이라고 할 수 있다. 다만 빅데이터 분석기법은 기존에 활용하지 못했던 다양한 형식의 대규모 데이터를 처리하고 분석할 수 있다는 점만으로도 일각에서는 매우 유용하다고 평가받는다.

전통적인 분석기법은 과거 데이터에 기반하기 때문에 어떤 일이 발생했고 왜 발생했는가에 대한 답을 찾는 것을 데이터 분석의 목적으로 한다. 즉 전통적인 분석기법은 변수간 인과관계(causation)를 파악하는 것을 데이터 분석의 주요 목적으로 한다. 이와 달리 빅데이터 분석기법은 빠르게 생성되는 데이터의 특성을 고려하여 예측분석에 초점을 맞춘다. 또한

이론적 기반에 따른 데이터간의 인과관계보다는 대규모 데이터안에 존재하는 변수간 상관관계(correlation)를 발견하는 것을 데이터 분석의 주요 목적으로 한다. 물론 빅데이터 분석기법에는 전통적인 데이터 분석기법도 포함된다. 또한 빅데이터 분석기법은 완전히 새로운 것이 아니다. 빅데이터 기술의 발전으로 다양한 유형의 데이터를 통합할 수 있고 대규모의 데이터를 처리할 수 있게 되면서 이미 예전부터 이용해왔던 분석기법들이 다시 주목받게 된 경우가 대부분이다(이명진·김우주, 2012).

<표 Ⅱ-1> 빅데이터 분석기법의 차이점

	전통적인 분석기법	빅데이터 분석기법
분석초점	• 기술적 분석 • 현안진단 분석	• 예측적 분석 • 데이터 사이언스
데이터셋	• 정형 데이터셋 • 정제된 데이터셋 • 간단한 데이터 모델	• 대용량 데이터셋 • 비정형 데이터셋 • 원시 데이터셋 • 복잡한 데이터 모델
분석목적	• 인과관계 분석	• 상관관계 분석

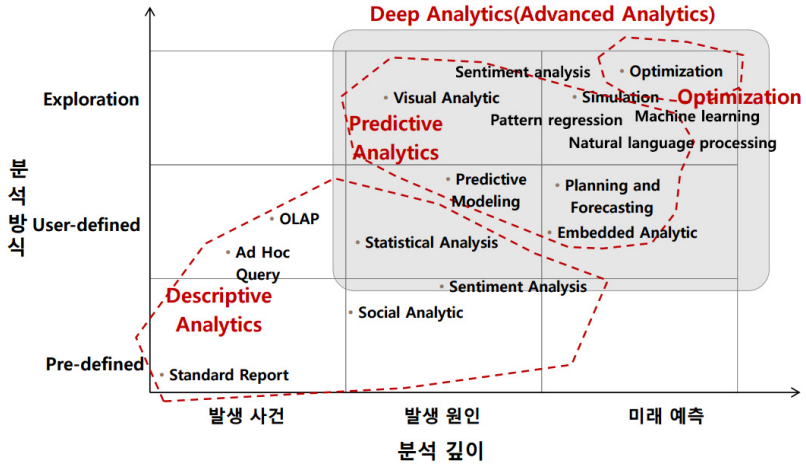
자료: SoftServe(2014)

한편 빅데이터 분석기법은 기능적 목적 측면에서 이전과 크게 달라졌다(양지현, 2011). 전통적인 데이터 분석기법은 현재의 현상 분석과 단순한 시각화에 초점을 맞추어 왔다. 이는 활용할 수 있는 데이터 유형과 처리할 수 있는 데이터 용량이 제한적이었기 때문이다. 예를 들면, 전통적인 데이터 분석기법은 고객이탈 현상을 파악하고자 할 경우 통계적으로 고객이탈 비율을 산출하고 이를 평면적인 그림이나 도표로 산출하는 데 그쳤다. 그 원인을 파악하고 처방을 제시하는 작업은 별도로 진행되었다. 이와 달리 빅데이터를 활용한 분석기법은 이전보다 한층 발전되어 효과적인 의사결정을 돕는 역할에 초점을 맞춘다. 빅데이터 분석기법은 다양한 유형의 데이터를 통합하고 이렇게 통합된 대규모 데이터를 분석하여 숨겨진 패턴을

발견하는 작업부터 미래 상황을 예측하고 처방하는 작업까지 포괄하고 있다. 따라서 빅데이터 분석기법을 활용할 경우 고객이탈 현상의 원인을 파악하는 작업뿐만 아니라 적절한 대응방안을 찾는 작업까지 수행할 수 있는 것으로 평가된다.

빅데이터 분석기법은 크게 현재의 상황을 분석하기 위한 기술분석(descpritive analytics), 선제적 의사결정을 지원하기 위한 예측분석(predictive analytics), 그리고 가능한 결과로부터 가장 최적의 결과를 도출하기 위한 최적화(optimization) 작업으로 분류될 수 있다(Raden, 2010; 이명진·김우주, 2012). 이는 다시 분석방식과 분석깊이에 따라 세부적으로 분류되기도 한다(투이컨설팅, 2013). 분석방식은 미리 정의된 방식, 사용자 정의 방식, 탐색 방식으로, 분석깊이는 분석대상 또는 분석목적에 따라 발생사건, 발생원인, 미래예측으로 구분될 수 있다.

<그림 Ⅱ-8> 빅데이터 분석기법 분류



자료: 투이컨설팅(2013) 수정인용

기술분석은 빅데이터에서 과거에서부터 현재까지 주어진 데이터로 현재의 상황을 분석하고 사용자의 이해를 돕기 위해 표현하거나 설명하기 위해 가장 관심있는 정보를 단순하게 추출하는 분석기법이다(이명진·김우주, 2012). 기술분석 방법으로는 Ad Hoc Query, Association Analysis, Classification, Clustering, OLAP(Online Analytical Processing), Social Analytic, Sentiment Analysis, Statistical Analysis 등이 있다. 예를 들면, 금융회사는 어떤 금융상품에서 고객불만이 가장 높은지를 빅데이터에서 추출하여 살펴보거나 신규 금융상품 중에서 고객의 문의 또는 조회가 높은 금융상품이 무엇인지를 찾아내는 작업 등을 할 수 있다. 이처럼 기술분석은 문제의 원인이나 해답을 찾는 데 가장 초기적인 분석기법으로 알려져 있다. 이 때문에 기업의 빅데이터 분석 중에서 기술분석이 약 80%를 차지하는 것으로 알려져 있다.

예측분석은 현재와 과거의 데이터를 활용해 미래에 발생할 수 있는 이벤트를 예측하는 분석기법이다. 예측분석 방법으로는 Decision Tree, Predictive Modeling, Embedded Analytics, Planning and Forecasting, Visual Analytics, Pattern Regression or Recognition, Sentiment Analysis, Machine Learning, Natural Language Procession 등이 있다. 예를 들면, 금융회사는 예측분석기법을 이용해 회사내 빅데이터에 포함된 고객의 금융상품 및 서비스 이용실태를 이전보다 효과적으로 분석할 수 있다. 이 경우 금융회사는 고객에게 효율적인 금융자산 배분을 자문할 수 있으며 이를 통해 교차판매의 효율성을 제고할 수 있다. 또한 금융회사는 회사내 빅데이터를 예측분석기법으로 분석해 금융사기 또는 사고를 적발할 수 있다. 예측분석을 실행하기 위해서는 데이터 탐색(data exploration), 모델 개발(model development), 모델 사용(model deployment), 모델 관리(model management)의 과정을 거쳐야 한다(이명진·김우주, 2012). 데이터 탐색은 예측모형에 입력된 적합한 데이터를 결정하는 작업이다. 빅데이터 분석은 다양한 형태의 데이터를 수용할 수 있기 때문에 입력될 데이터의 형태에 따라 예측모형의 선택도 달라질 수 있다. 이 때문에 데이터 탐색 작업은 매우 중요한 단계로 평가된다. 모델 개발은 분석 목적에

적합한 모델을 선택하고 검증하는 작업이다. 모델 이용은 실제 데이터를 입력하고 의사결정에 필요한 정보 또는 지식을 추출하는 작업이다. 모델 관리는 이러한 작업을 통해 모델을 수정하는 등 모델의 유효성을 유지하고 향상시키는 작업이다.

최적화는 의사결정에 따라 다르게 발생할 수 있는 미래 이벤트를 예측하고 가장 최적의 의사결정을 도출하는 분석기법으로, 처방분석(prescriptive analytics)으로도 알려져 있다. 최적화를 통해 다양한 시나리오별 의사결정을 도출하기 위해서는 actionable data와 feedback system을 갖추어야 한다. 이 점에서 최적화는 하나의 시나리오를 가정하고 실시하는 예측분석과 다르다고 평가된다. 여기서 actionable data라 함은 고객의 피드백을 통해 무엇이 어디에서 언제 왜 발생하는가를 직접적이고 자동화된 방식으로 분석할 수 있는 데이터를 의미한다. 이를 통해 기업은 최적의 의사결정을 도출할 수 있다. 예를 들면, 빅데이터를 활용한 최적화 분석기법은 투자전략 수립, 자산배분, 리스크관리 등 금융의 다양한 분야에서 더 효과적인 의사결정을 도출하는 데 일조할 수 있다.

그 밖에 비정형인 각종 컨텐츠 데이터를 분석하는 컨텐츠 분석(contents analytics), 실시간으로 변화하는 현상을 이해하고 이에 따라 대응할 수 있도록 의사결정을 돕는 리얼타임 분석(real-time analytics) 등이 차세대 빅데이터 분석기법으로 소개되고 있다(양지현, 2011; 이명진·김우주, 2012). 컨텐츠 분석은 주로 웹문서, SNS, 블로그 등에 포함된 텍스트, 이미지, 비디오 등과 같은 컨텐츠를 분석하여 새로운 트렌드나 패턴을 발견하는 분석기법이다. 실시간 분석은 말 그대로 의사결정 시점 이전까지 생성된 데이터를 모두 활용하는 분석기법이다. 그만큼 실시간 분석은 빠르게 변화하는 상황에서 적시에 대처할 필요가 있는 경우에 요구되는 분석기법이다. 예를 들면, 자본시장의 금융시세는 빠르게 변한다. 이에 따라 투자자는 적시에 대응할 필요가 있다. 또한 다른 투자자의 동향도 빠르게 파악할 필요가 있다. 실시간 분석은 이러한 일련의 작업을 가능하게 한다.

빅데이터 분석기법은 독립적으로 이용되는 것이 아니라 분석 목적과 단계에 따라 상호 보완적으로 활용된다. 예를 들면, 기술분석의 임의분석은

예측분석에 앞서 데이터 탐색을 위해 활용될 수 있다. 또한 최적화를 위해서는 예측모형을 수립해야 하고 이를 검증해야 한다. 다양한 시나리오에 따라 최적화된 의사결정을 도출하기 위해서는 각 시나리오에 대해 예측분석을 실시해야 한다.

2. 빅데이터 관련 정책적 논의

빅데이터는 그 존재 자체로도 경제적 가치를 인정받지만, 이를 제대로 활용하지 못하면 그 가치는 실현될 수 없다. 그만큼 빅데이터의 가치를 극대화하려면 빅데이터 자체를 원활하게 활용할 수 있는 법제도적 환경이 갖추어질 필요가 있다. 2000년대 초반 이후 데이터 마이닝이 쇠락한 이유도 개인정보가 포함된 데이터를 활용할 수 있는 법제도적 환경이 제대로 갖추어지지 않은 상태에서 데이터 마이닝이 무작위적으로 개인의 사생활을 침해할 수 있다는 사회적 우려가 형성되었기 때문이다.

빅데이터는 분산되어 있는 데이터를 한 곳에 집중시켜야 그 가치가 생성되는 특징을 가진다. 그만큼 정부 또는 민간이 보유한 데이터의 유통과 공유가 원활해야 그 가치가 생성된다. 이 점을 고려하여 본 절에서는 빅데이터 활용과 관련된 주요 정책적 논의를 정부 또는 공공기관이 보유한 공공데이터(open data) 개방, 민간이 보유한 데이터 공유, 데이터 유통과 공유로 발생할 수 있는 빅데이터 윤리 문제로 나누어 간략히 살펴보았다. 참고로 이러한 정책적 논의는 빅데이터 기술이 소개되기 이전부터 존재했으며, 최근 전 세계적으로 빅데이터에 대한 중요성이 강조되면서 각 국에서는 그 정책적 논의를 빅데이터 관점에서 새롭게 추진하거나 강화하고 있다.

가. 공공데이터 개방

1) 공공데이터 정의 및 분야

빅데이터에서 공공데이터가 차지하는 의미는 매우 크다. 공공데이터는 정부 또는 공공기관이 독점하고 있는 데이터이다. 즉 공공데이터에는 민간에서 쉽게 접근할 수 없는 정보를 많이 포함하고 있다. 이를 민간이 빅데이터의 일부로 자유롭게 활용할 수 있을 경우 새로운 상품이나 서비스 개발이 가능한 것으로 평가받는다(OECD, 2015a). 민간에게 개방되는 공공데이터는 최소 세 가지 요건을 갖추어야 한다(OECD, 2015a). 첫째, 공공데이터는 누구나 무료로 이용할 수 있어야 한다. 정부 또는 공공기관이 보유한 데이터가 공개될 경우 이에 대한 접근을 차별해서는 안된다는 것을 의미한다. 둘째, 공공데이터는 누구나 재활용할 수 있어야 하고 이를 다시 유통시킬 수 있어야 한다. 이는 공공데이터가 정부 또는 공공기관이 수집한 원시 데이터(raw data) 형태로 공개되어야 한다는 것을 의미한다. 그래야 민간이 공공데이터를 가공하여 의미있는 새로운 정보를 추출할 수 있다. 그렇지 않고 정부 또는 공공기관이 자의적으로 가공한 데이터를 공개할 경우 데이터의 효용가치는 크게 낮아질 수 있다. 셋째, 공공데이터는 가공 또는 활용이 쉬운 형식으로 제공되어야 한다. 단순히 많은 건수의 정보가 공공데이터 명목으로 공개된다고 해서 공공데이터가 새로운 부가가치를 창출할 수 있는 것은 아니며, 일반에게 공개된 공공데이터가 그 정보를 활용하여 새로운 정보를 추출할 수 없는 형식으로 공개된다면 아무리 뛰어난 빅데이터 기술을 구비하였더라도 새로운 부가가치를 창출하는 데 많은 비용이 소요될 수 있다. 참고로 미국의 경우 공공데이터의 요건을 더 자세히 제시하고 있다(U.S. Executive Office of the President, 2013).

정부 또는 공공기관이 기존에 발표한 통계자료는 공공데이터의 요건을 상당수준 충족시키지 못하는 것으로 평가된다(공공데이터전략위원회, 2014). 민간에게 공개된 공공데이터가 대부분 원시 데이터가 아닌 집계

데이터이기 때문이다. 이 때문에 최근 공공데이터는 공급자 중심이 아닌 수요자 중심으로 개방되는 추세로 전환되고 있다. 즉 최근에는 정부 또는 공공기관도 법적인 테두리 안에서 가능한 수요자가 원하는 수준의 원시 데이터를 공개하는 추세이다. 또한 정부 또는 공공기관이 기존에 발표한 공공데이터는 MS워드, PDF, MS엑셀과 같은 폐쇄된 형식으로 공개된 경우가 대부분이었다. 이 때문에 공공데이터를 가공하고 처리하는 데도 여러 가지 비효율적인 문제들이 야기되었다. 최근의 공공데이터는 이러한 문제를 개선하기 위해 오픈포맷인 XML, JSON, RDF 형식으로 공개되고 있다.

각 국은 공공데이터에 대한 접근성을 제고하고 그 활용가치를 극대화하기 위해 공공데이터 포탈을 개설하여 운용하고 있으며, 민간의 성공적인 활용사례를 수집하고 전파하여 민간의 공공데이터 활용을 촉진하고 있다. 특히 금융분야의 공공데이터는 각 국마다 법제도적 환경에 따라 공개 정도와 내용에서 차이가 나는 것으로 조사된다. 미국은 2009년 5월 공공데이터의 원활한 공개와 유통을 위해 공공데이터 포탈(data.gov)을 개설하고, 금융(finance) 데이터를 14개 공공데이터 분야중 하나로 구분하여 개방하고 있다. 뿐만 아니라 미국 연방준비제도이사회(Federal Reserve Board: FRB), 증권거래위원회(Securities and Exchange Commission: SEC) 등 금융당국도 관련 공공데이터를 적극적으로 개방하고 있다. 미국은 이를 통해 새로운 금융상품 및 서비스 개발이 촉진될 수 있을 것으로 기대한다. 유럽연합(European Union: EU)은 공공데이터가 경제 전반의 혁신, 성장과 투명성에 기여할 수 있다고 판단하고, 2012년 EU 공공데이터 포털(europeandataportal.eu)을 개설하여 운영하고 있다. EU 공공데이터는 EU가 생산한 문서에 국한하고, EU 차원에서 관리감독하는 영역에 대한 데이터를 소개하는 방식으로 개방되고 있다. 이 때문에 EU 공공데이터 포탈은 각 회원국의 세부적인 공공데이터를 각 회원국의 공공데이터 포탈에서 찾아볼 수 있도록 안내하고 있다. 영국은 2010년 1월 납세자의 공공데이터 접근성을 향상시키기 위해 공공데이터 포탈(data.gov.uk)을 개설하였다. 한 가지 특이한 점은 영국의 경우 미국이나 EU와 달리 공공

데이터 분야 중에 금융분야를 별도로 구분하지 않고 있다. 뿐만 아니라 영국의 금융당국도 금융관련 공공데이터를 별도로 개방하지 않고 있다. 특히 영국의 경우 공공데이터의 범위에 정부의 데이터만 포함되고 이에 준하는 공공기관의 데이터는 포함되지 않은 것으로 조사된다.

2) 공공데이터 가치 및 개방 전략

공공데이터는 정부의 투명성을 제고하는 역할뿐만 아니라 경제 전반의 정보 비대칭성, 자원배분 효율성, 네트워크 효과를 증대시키는 역할을 할 수 있는 것으로 평가받는다. 또한 공공데이터 개방 정책이 정부의 투명성을 높일 뿐만 아니라 국가적인 성장전략의 일환으로 추진되어야 한다는 주장도 있다(Shakespeare, 2013). 공공데이터를 활용할 경우 민간부문의 생산성이 제고되고 새로운 상품이나 서비스가 출현할 수 있다고 보기 때문이다. McKinsey Global Institute(2013)는 공공데이터를 통해 매년 3.2조달러에서 5.4조달러의 경제적 가치가 창출될 수 있다고 추정한다. 분야별 경제적 가치창출 규모를 비교하면, 금융분야 공공데이터의 경제적 가치창출 효과가 가장 낮은 것으로 추정된다. 이는 소비자금융 부문만을 고려하고 자본시장 부문을 고려하지 않았기 때문일 수 있고, 엄격한 금융규제로 금융분야의 공공데이터가 상대적으로 낮은 수준에서 개방되기 때문일 수 있다.

미국은 공공데이터의 경제적 가치를 자체적으로 산출하지 않은 것으로 조사된다. 다만 McKinsey Global Institute(2013)는 미국이 전 세계 공공데이터의 경제적 가치에서 약 29.7%을 차지할 것으로 추정한다. 이를 토대로 미국은 공공데이터를 통해 약 0.9조달러에서 1.6조달러의 경제적 가치를 창출할 것으로 추정된다.

유럽위원회는 2016년부터 2020년까지 공공데이터가 약 1.1조유로에서 1.2조유로의 경제적 가치(누적)를 창출할 것으로 평가한다(European Commission, 2011). 2020년중 각 분야별 경제적 가치창출 효과를 살펴보면, 유럽위원회는 금융분야의 경제적 가치창출 규모를 약 231.1억유로로

추정하였다. 이 경우에도 자본시장 부문은 포함되지 않았다.

영국은 공공데이터가 2011년 가격을 기준으로 매년 62억파운드에서 72억파운드의 경제적 가치를 창출할 것으로 평가한다(Deloitte, 2013). 그러나 영국은 향후 공공데이터가 경제적 가치를 창출할 수 있는 역량이 이보다는 더 클 것으로 내다보고 있다. 특히 영국은 연결된 데이터(linked data)를 적극적으로 이용하고, 더 많은 재화와 용역에 지리정보와 위치정보를 잘 결합하며, 공공데이터를 정책 수립에 잘 활용할 경우 공공데이터의 경제적 가치 창출 효과는 극대화될 수 있을 것으로 예상한다.

2013년 기준 국내총생산(Gross Domestic Product: GDP) 기준으로 미국, EU, 영국 공공데이터의 경제적 가치를 비교해보면, 미국이 가장 높은 국가로 조사된다. 이러한 결과는 두 가지 측면에서 해석될 수 있다. 첫째, 공공데이터 개방 정도에 따라 그 파급효과도 달라질 수 있다. 이 점에서 미국은 EU나 영국보다 공공데이터 개방에 더 적극적인 것으로 조사된다. 둘째, 정보산업의 발달 정도에 따라 공공데이터의 활용 정도도 달라질 수 있다.

<표 Ⅲ-2> 공공데이터 경제적 가치 비교(2013년)

(단위: 달러, 유로, 파운드, %)

구분[1]	GDP	공공데이터 부가가치		비중	
		최소	최대	최소	최대
전 세계	75,467	3,220	5,390	4.27	7.14
미국[2]	16,663	957	1,602	5.74	9.62
EU[3]	14,635	220	240	1.50	1.64
영국	1,713	6	7	0.36	0.42

주 : 1) 전 세계 및 미국은 달러, EU는 유로, 영국은 파운드 기준임
2) 미국은 McKinsey Global Institute(2013)를 참고하여 추산함
3) EU는 2015년 기준임

자료: McKinsey Global Institute(2013), European Commission(2015), Deloitte(2013)

2013년 6월 G8 정상회담이 채택한 공공데이터 선언문(Open Data Charter)은 각 국의 공공데이터 개방에 있어서 매우 중요한 정책적 방향을 제시하고 있다(UK Cabinet Office, 2013). 첫째, 정부 또는 공공기관이 보유한 공공데이터는 기본적으로 개방되어야 한다. 다만 개방이 불가한 합법적인 이유가 있는 공공데이터의 경우 예외적으로 개방대상에서 제외시킬 수 있다. 둘째, 공공데이터는 고품질이고, 적시에, 포괄적으로, 정확하게 개방되어야 한다. 또한 개방된 공공데이터에 대한 자세한 설명도 제공되어야 한다. 공공데이터의 품질을 관리하기 위해 가능한 빨리 개방하고, 이용자의 피드백을 받아야 하며, 계속해서 관리하고 필요한 경우 수정해야 한다. 셋째, 공공데이터는 모두에게 제공되어야 하며, 모두에게 재사용을 허용해야 한다. 이를 위해 공공데이터는 무료로 제공되어야 하며, 어디에서도 사용할 수 있도록 다양한 형식으로 제공되어야 하고, 가능한 많은 정보가 담겨져 있어야 한다.[4] 넷째, 정부 또는 공공기관의 투명성과 효율성을 높일 수 있는 공공데이터가 공개되어야 한다. 이를 위해 기술전문가와 축적된 경험을 서로 공유하고, 데이터 수집, 표준, 개방 절차 등을 함께 마련해야 한다. 다섯째, 혁신을 촉진할 수 있는 공공데이터가 공개되어야 한다. 이를 위해 공공데이터를 개방하고 활용하는 과정이 민간과 긴밀한 협조를 통해 이루어져야 하며, 기계가 읽을 수 있는 형식으로 제공되어야 한다. 이에 따라 미국, EU, 영국 등은 G8 공공데이터 선언문을 기초로 공공데이터 개방 전략을 수립하고 이를 추진하고 있는 것으로 조사된다(U.S. Government, 2014; European Commission, 2013; UK Government, 2014).

나. 민간데이터 공유

1) 민간데이터 공유 방법

4) EU의 경우 공공데이터 개방에 따른 행정비용이 수반될 수 있는 점을 고려하여 공공데이터 개방에 따른 한계비용을 수혜자 원칙에 입각하여 이용자에게 부담할 수 있도록 허용하고 있다.

데이터 공유(data sharing) 정책은 일반적으로 정부와 민간을 별도로 구분하지 않는다. 예를 들면, 미국의 경우 데이터 공유 정책은 정부부처간 데이터 공유를 활성화하는 데 초점을 맞추고 있다. 그러나 대체적으로 데이터 공유 정책은 민간이 보유한 데이터에 초점이 맞추어져 있다. 이 때문에 데이터 공유 정책은 주로 민간데이터에 포함된 개인정보를 보호하면서 민간의 데이터 공유를 효율적으로 촉진시킬 수 있느냐를 논하고 있다.

최근 데이터 공유 방식은 일방적으로 가공한 데이터를 공개하던 이전 방식과 크게 다른 양상을 보이고 있다. 이전 데이터 공유 방식은 일종의 자체 검열을 통해 집계된 데이터를 공유하는 방식이 주를 이루었다. 예를 들면, 회사 수준(firm-level) 또는 산업 수준(industry-level)의 데이터가 제한적으로 제공되었다. 이 때문에 한 산업을 완전히 이해하는 데 많은 제약이 뒤따랐다. 그러나 최근 데이터 공유 방식은 개인 수준(individual-level)의 데이터까지 확장되고 있다. 이를 통해 한 산업의 세세한 부분까지 데이터 분석을 실시할 수 있게 되었다. 다만 이전과 달리 완전 개방보다는 개인정보 보호 차원에서 데이터 공유 방법에 따라 데이터의 개방성을 제한하고 있다.

데이터 공유 방법은 그 데이터 안에 개인 식별가능 정보(personally identifiable information) 또는 개인정보가 포함되어 있느냐에 따라 크게 달라진다. 개인정보란 어떤 개인을 특정할 수 있는 구체적인 정보 또는 다른 정보와 결합할 경우 개인을 식별할 수 있는 정보를 의미한다. 참고로 미국의 경우를 제외하고는 대부분 국가에서는 포괄적인 의미에서 개인정보라는 용어를 사용하고 있다. 각 국의 개인정보보호 체계에 따르면, 개인정보가 포함된 데이터를 제3자와 공유하기 위해서는 공통적으로 데이터 주체인 개인의 동의(opt-in consent)를 얻거나, 개인에게 거절(opt-out consent)을 받거나, 개인정보 자체를 비식별화(de-identification)할 것을 요구받는다. 전자의 경우 개인 동의 또는 거절 절차에 대한 제도적 기반이 마련되어 있지 않을 경우 데이터에 포함된 모든 개인의 동의를 얻어야 하기 때문에 상당한 비용이 유발될 수 있다. 반면 후자의 경우 개인정보를 비식별화하거나 익명화하기 때문에 데이터에 포함된 모든 개인의 동의가

반드시 필요하지 않을 수 있다. 이 점에서 데이터 공유를 활성화하는 데 후자의 방법이 더욱 비용 효율적이고 법적인 다툼의 소지가 적은 것으로 알려져 있다.

한편 제3자에게 개인정보를 제공하기 위해 개인의 동의를 받거나 개인에게 고지하는 것은 개인정보에 대한 보호의 부담을 개인에게 지우는 것과 같다. 이는 다시 민간이 보유한 데이터의 원활한 유통을 제한하는 요인으로 작용할 수 있다. 제3자가 개인정보를 남용해 직접 마케팅에 활용하거나 사생활을 침해할 우려가 상존하기 때문에, 이를 꺼려하는 개인이 동의하지 않을 수 있기 때문이다. 또한 기업은 이러한 문제를 회피하기 위해 개인이 동의하지 않으면 가입이나 거래를 위한 다음 단계로 넘어가지 못하게 막는 경우(take-it-or-leave-it setting)가 많다(PCAST, 2014). 더욱 문제될 수 있는 것은 개인이 이러한 절차에 동의하더라도 그 내용을 충분히 숙지하지 못할 가능성이 높다는 점이다. 이러한 문제들 때문에 미국의 경우 개인정보에 대한 동의나 고지 절차가 민간의 데이터 공유를 활성화하는 데 시장실패의 원인으로 작용할 수 있다고 보고 있다. 이 문제를 해결하기 위해서는 정부 또는 제3자가 시장에 개입해 개인의 부담을 경감시키는 것이 필요할 수 있다. 이에 대한 대안으로 최근 각 국에서는 민간의 데이터 공유 방법으로 비식별화가 추천되고 있다.

개인정보의 비식별화는 식별가능한 개인정보를 제거하거나 식별가능하지 못하도록 모호하게 만드는 처리과정이다. 익명화(anonymization)는 비식별화와 유사한 개념이나, 개인정보에 대한 비식별화의 절차로 이해할 필요가 있다. 즉 비식별화가 익명화보다 더 넓은 개념이라고 볼 수 있다. 예를 들면, 성명은 개인의 중요한 식별가능 정보이다. 그러나 성명을 익명화했다고 해서 그 개인을 식별할 수 없는 것은 아니다. 비식별화는 개인정보가 포함된 데이터에서 개인을 전혀 식별할 수 없도록 만드는 일련의 모든 과정을 일컫는다(고학수·최경진, 2015; Nelson, 2015). 개인정보의 비식별화가 필요한 이유중 하나는 데이터에 포함된 개인의 모든 동의를 얻지 않아도 비식별화 조치를 통해 개별 데이터의 통계적 효용가치를 그대로 유지할 수 있기 때문이다.

개인정보에 대한 비식별화 처리가 데이터 공유를 활성화시킬 수 있는 방법임에는 틀림없다. 그럼에도 불구하고 비식별화 처리는 개인정보를 보호하는 데 궁극적인 수단이 될 수 없다는 것이 일반적인 인식이다(Garfinkel, 2015; PCAST, 2014). 비식별화 처리는 빠르게 발달하는 빅데이터 기술에 의해 쉽게 재식별화될 수 있기 때문이다. 예를 들면, 비식별화 처리에 사용된 일정한 규칙이나 알고리듬을 역공학(reverse engineering)을 이용해 추론할 수 있다. 결국 개인정보가 포함된 데이터를 비식별화하더라도 개인정보를 완전히 삭제하지 않는 한 완벽하게 숨길 수 없다. 따라서 개인정보 보호와 데이터의 활용 가치가 서로 충돌한다면 이에 대한 균형적인 사회의 통념이 먼저 성립되는 것이 필요할 수 있다.

개인정보가 포함된 데이터를 민간이 공유하는 방식중 또다른 방식으로 Open API(Application Programming Interfaces)도 새롭게 부각되고 다. API는 소프트웨어 컴포넌트끼리 각 데이터에 상호 접속하고, 데이터를 서로 교환할 수 있도록 미리 짜여진 규약이기 때문에 웹상에서 다양한 목적으로 여러 분야에서 활용되고 있는 기술이다. Open API는 API의 기술적 특성을 활용하여 개인정보가 포함된 데이터에 접속할 수 있는 권한을 외부 또는 제3자에게 허용하는 표준화된 규약이라는 점에서 API와 구별된다.

Open API가 각광받는 이유중에 하나는 제3자가 데이터에 포함된 모든 개인의 동의를 받지 않더라도 데이터에 접근할 수 있는 권한을 갖는다는 점이다. 다만 대부분의 경우 그 데이터를 사용할 수 있는 권한은 해당 개인이 제3자의 개인정보 이용을 동의할 경우로 제한되거나, 적법한 절차에 따라 업무위탁 계약을 체결한 경우로 한정된다. 또한 Open API는 무료로 제공되기도 하며, 설계 목적에 따라 데이터를 대량으로 다운받을 수 있도록 허용할 수 있다.

이러한 장점 때문에 Open API는 민간데이터 공유의 중요한 방식으로 활용되고 있다. 특히 빅데이터를 활용해 새로운 서비스를 제공하고자 하는 신생기업들이 Open API를 개발해 빅데이터 유통을 촉진시키는 역할을 하고 있다. 예를 들면, 핀테크 신생기업들은 환율, 시세, 금리 등 금융시장

데이터를 Open API로 공급하고 있다.

다만 Open API가 활성화되기 위해서는 개인정보가 포함된 데이터에 대한 소유권이 누구에게 있는가가 명확하게 정의될 필요가 있다.[5] 예를 들면, 유럽의 경우 데이터 관리자와 데이터 주체를 명확하게 구분하고 있으며, 데이터에 대한 일반적인 소유권은 데이터 주체에게 있으며 데이터 사용과 수익에 대한 권리는 데이터 관리자가 갖는다고 보고 있다. 이 때문에 데이터 주체인 개인이 제3자에게 자신의 개인정보를 제공하기를 원하면 데이터 관리자는 이를 이행해야 한다. 그런데 현실적으로는 데이터 관리자가 개인정보가 포함된 데이터에 대해 소유권을 주장하는 경우가 많다. 또한 이 과정에서 어느 범위까지 정보 주체인 개인이 데이터에 대한 소유권을 주장할 수 있는가에 대한 분쟁도 발생할 수 있다.

2) 민간데이터 공유 정책

각 국의 민간데이터 공유 정책은 개인정보 및 그 동의 절차를 데이터 주체인 개인의 입장에서 더 명확하게 수립하는 데 초점을 맞추고 있다. 미국의 경우 미국 대통령 과학기술 자문위원회(President's Council of Advisors on Science and Technology: PCAST)는 2014년 발표한 'Big Data And Privacy: A Technological Perspective'에서 민간이 보유한 데이터의 개인 식별가능 정보를 보호하는 방안을 구체적으로 제시하고 있다. EU의 경우 2012년부터 논의를 거쳐 2013년 12월 유럽의회의 승인을 받고 2016년 4월 '일반적인 데이터 보호에 관한 규정(the general data protection regulation)' 최종안을 발표하였다(인하대학교 법학연구소, 2014; Allen & Overy, 2016). 참고로 동 규정은 2018년 5월 25일에 시행될 예정이다. 영국의 경우 개인정보 보호체계가 1998년 정보보호법(data protection act 1998)에 기반하고 있으나, 정보위원회 사무국(Information Commissioner's Office: ICO)이 개인정보 보호와 관련된 지침을 빅데이터 환경을 고려하여 계속해서 수정하고 있다(ICO, 2010;

5) 일반적으로 소유권은 사용권, 수익권, 처분권으로 구분된다.

ICO, 2011; ICO, 2012; ICO, 2016a; ICO, 2016b).

개인정보를 보호하는 궁극적인 목적은 개인의 사생활이 제3자에 의해 침해되는 것을 방지하기 위함이다. 이 측면에서 보자면 제3자가 개인의 사생활을 침해하지 않는다면 개인정보가 포함된 데이터를 광범위하게 활용하는 것을 허용할 여지가 있다. 왜냐하면 빅데이터 분석 자체가 개인의 사생활을 직접적으로 침해한다고 보기는 어렵기 때문이다. 한편 개인의 사생활 침해를 어떻게 정의하느냐도 중요한 관건이 될 수 있다. 제3자가 불특정 다수의 개인정보를 보유하고 있는 것 자체만으로도 특정한 개인의 사생활을 침해하는 것으로 볼 수 있는지에 대한 판단이 필요하다. 아니면 제3자가 자기의 이익을 위해 개인정보가 포함된 데이터를 오용하거나, 남용하거나, 악용하여 특정한 개인의 사생활을 직접적으로 침해하는 것으로 한정하는 것이 바람직한지에 대해서도 정책적으로 판단할 필요가 있다.

최근 새롭게 논의되고 있는 개인정보 정책으로는 오프라인에서 Do-Not-Call 규제와 온라인에서 Do-Not-Track 규제가 있다. Do-Not-Call 규제는 전화권유판매 수신거부 의사를 밝힌 개인에게 전화하지 않는 것을 주요 골자로 하고 있다. 그러나 전화권유판매를 받는 개인은 전화가 무작위로 자신에게 걸려온 것인지 아니면 불법적인 개인정보의 유통에 따른 것인지를 확인할 수 없다. 따라서 수신거부 의사를 밝힌 개인만 보호되는 소극적인 규제라고 볼 수 있다. Do-Not-Track 규제도 마찬가지이다. 온라인에 노출된 개인정보를 제3자가 개인의 동의 없이 수집하는 행위를 금지하는 것이 주요 골자이다. 이 때문에 자신의 개인정보가 오프라인에서 유통된 것인지 온라인에서 무작위로 수집된 것인지를 구분하기 쉽지 않다. 예를 들면, 대학 졸업생이 입사지원을 하면서 회사에게 온라인에 노출된 개인 민감정보에 대한 수집에 동의를 하지 않았음에도 불구하고 회사가 이를 토대로 입사여부를 차별할 수 있다. 그러나 개인은 자신의 민감 정보를 근거로 회사가 입사를 거절했다는 명확한 증거를 제시할 수 없다면 사생활 침해에 대해 반론을 제기하지 못할 것이다. 데이터 주체인 개인에게 온라인에 노출된 개인정보에 대해 잊혀질 권리(the right to be forgotten)를 부여하는 제도가 논의되고 있는 이유도 이 때문이다. 이에 대한 찬반의 논란이

존재하지만, 개인에게 잊혀질 권리를 부여하는 것은 개인의 사생활 보호를 위해 데이터주체인 개인이 능동적으로 대처할 수 있도록 권리를 부여하는 것이기 때문에 바람직할 수 있다. EU는 '일반적인 데이터 보호에 관한 규정(the general data protection regulation)'에서 명시적으로 데이터 주체인 개인에게 잊혀질 권리를 부여하고 있다.

다. 빅데이터 윤리 정책

1) 빅데이터 윤리 이슈

윤리(ethics)는 옳고 그른 것 또는 좋고 나쁜 것에 대한 일반적인 가치판단의 기준이다. 윤리는 많은 분야에서 인간 사회가 지켜야 할 판단의 기준점을 제시한다. 이 중, 기업윤리(business ethics), 연구윤리(research ethics), 기술윤리(technology ethics 또는 technoethics)가 빅데이터와 관련될 수 있다. 한편 빅데이터 기술 자체는 옳고 그르거나 좋고 나쁘다고 판단할 수 있는 대상이 아니다(Davis, 2012). 이를 어떻게 활용하느냐는 사람과 기업의 가치기준에 따라 달라질 수 있기 때문이다. 결국 데이터를 이용하는 개인과 기업의 윤리적 가치기준이 어떻게 확립되어 있느냐에 따라 빅데이터 윤리기준도 결정된다. 본 절에서는 빅데이터 윤리와 관련하여 데이터 불법유출, 개인정보 조작, 부당한 차별에 대하여 집중적으로 살펴보았다.

첫째, 데이터 불법유출은 개인의 사생활 침해뿐만 아니라 개인정보 조작으로 이어질 수 있다. 특히 빅데이터의 경제적 가치가 높게 평가되면서 개인정보가 포함된 데이터가 불법적으로 유출될 가능성이 높아졌다. 국내에서는 2013년 카드사가 보유한 1억 400만건의 개인정보가 유출된 사건이 발생했다. 특정 개인의 윤리적 가치의 부재로 발생한 문제이나, 그 파급효과는 해당 개인뿐만 아니라 카드사, 정부까지 영향을 미쳤다. 데이터 불법유출은 외부의 해킹에 의해서도 발생하지만, 내부 관계자의 사익추구 유인 때문에도 발생할 수 있다. 이렇게 불법 유통된 데이터는 주로 마케팅에

활용되며, 이 과정에서 개인의 사생활이 침해될 우려가 높다. 데이터 불법 유출은 2차적 피해를 유발시킬 수 있기 때문에 엄격하게 다루어져한다는 것이 일반적인 인식이다(Finklea, 2014). 데이터 불법유출은 개인정보 조작으로 이어질 수 있기 때문이다. 예를 들면, 정부 보조, 불법 이민, 테러 등에 이용할 목적으로 개인정보가 조작될 수 있다.

둘째, 해킹 등을 통해 개인정보에 대한 적법한 권한 없이 접근해 개인정보가 조작될 경우 해당 개인의 사회적 신분 또는 신뢰가 심각한 수준까지 훼손될 수 있다. 빅데이터는 기업의 수익창출을 위해 개인의 성향이나 행태를 파악하는 데 활용된다. 이를 위해서는 개인정보가 정확해야 하고 제3자에 의해 조작될 수 없어야 한다. 그러나 개인정보는 제3자에 의해 충분히 조작 가능한 것이 현실이다. 예를 들면, 제3자가 SNS에서 특정 개인을 사칭하는 경우이다. 이 경우 개인정보가 의도적으로 조작될 수 있다. 미국의 경우 2015년 6월 인사관리사무국(the office of personnel management)의 데이터에 포함된 22만명에 달하는 연방정부 공무원, 용역 계약자, 고용 신청자 및 가족의 개인정보가 제3자에게 유출된 사건이 발생했다(Davison, 2015). 이 때 미국은 개인정보 유출보다 개인정보가 조작될 위험이 더 높다고 판단하였다. 연방정부 공무원의 신뢰도에 영향을 미칠 수 있고, 데이터 상에 특정 개인이 고의적으로 누락될 수 있기 때문이다.

셋째, 빅데이터는 부당한 차별에 악용될 수 있다. 빅데이터는 교육, 신용평가, 건강보험, 고용 분야에서 이미 활용되고 있다(FTC, 2016). 교육기관은 빅데이터 기술을 활용해 학생의 학습능력을 측정하고 고급교육 기회를 부여할지를 판단한다. 또한 이들 교육기관은 수강취소 위험이 높은 학생을 선별하고 이에 대한 개입 여부를 판단한다. 이 과정에서 교육기관은 고의적으로 학생에게 양질의 교육기회를 박탈하고 개인의 선택권을 과도하게 침해할 수 있다. 신용정보회사 또는 금융회사는 빅데이터를 활용해 비전통적인 방법으로 개인의 신용상태를 평가하고 있는 추세이다. 이 때문에 기존 금융회사에서 대출을 받지 못한 개인이 대출받을 수 있는 기회가 확장될 수 있다. 그러나 빅데이터에 기반한 신용평가는 반대로 개인을

차별하는 수단으로 악용될 수 있다. 빅데이터는 기대수명, 유전병 노출위험, 재입원 확률, 치료준수 여부 등을 예측하는 데도 활용된다. 이를 통해 병원 또는 보험사는 환자에게 적합한 처방을 내리거나 적정한 보험상품을 추천할 수 있다. 반대로 이를 악용하여 병원은 적절하지 않은 시점에서 불필요한 치료를 강요하거나, 보험사는 보험상품 가입을 거절할 수 있다. 빅데이터를 악용한 부당한 차별은 고용분야에서 더 광범위하게 나타날 수 있다. 빅데이터가 개인민감정보를 활용할 경우 개인의 직무능력보다는 다른 요인 때문에 고용이 결정될 수 있으며, 이 과정에서 개인은 부당하게 차별받을 수 있다.

2) 빅데이터 윤리 정책

각 국의 빅데이터 윤리 정책은 크게 데이터 관리체계 및 보안체계 구축과 기업윤리 기준의 확립으로 구분되며, 정부 차원의 구체적인 법제도화보다는 개인정보 보호체계 하에서 정부가 가이드라인을 제시하는 수준에 머무는 것으로 조사된다. 이는 빅데이터 윤리도 빅데이터 자체보다는 이를 활용하는 회사나 개인의 윤리 수준에 의해 결정되기 때문이다.

데이터의 불법유출은 적정한 데이터 관리체계를 구축함으로써 어느정도 해결할 수 있다. 데이터 관리체계는 의사결정의 일관성과 신뢰성을 높이고, 법 위반과 벌금 위험을 줄이고, 데이터 보안성을 개선하고, 수익창출 기회를 극대화하고, 데이터 질을 보장하고, 재작업을 최소화하고, 직원 생산성을 높이기 위한 내부통제 시스템이다(ICO, 2014). 내부자 또는 제3자에 의한 데이터 불법유출을 방지하기 위해서는 데이터관리자가 데이터에 대한 접근 및 이용 권한을 명확하게 수립해야 한다. 빅데이터를 활용하는 기업 대부분은 외부전문가를 활용하는 경우가 많다. 이 때문에 기업은 데이터 외부유출 위험에 더 많이 노출되어 있다. 데이터의 불법적인 외부유출을 통제하기 위해 먼저 데이터관리자와 처리자(processor)의 책임을 명확히 구분할 필요가 있다. 데이터관리자는 데이터의 수집 및 이용 목적과 방법을 결정하는 자이며, 처리자는 데이터관리자를 대신하여

데이터를 처리하는 자이다. 예를 들면, 증권회사의 원장 데이터의 경우 증권회사가 데이터관리자이다. 그러나 증권회사는 IT서비스 회사에 원장 시스템 관리를 위탁할 수 있다. 이 경우 증권회사 대부분은 IT서비스 회사에 상당한 수준의 재량을 허용할 수 있다. 그럼에도 불구하고 원장 데이터에 대한 일차적인 책임은 데이터관리자인 증권회사에 있다. 만약 데이터처리자인 IT서비스 회사가 원장 데이터를 유출한다면 이에 대한 일차적인 책임도 증권회사가 부담해야 한다.

데이터 불법유출을 방지하기 위해서는 철저한 데이터 보안체계를 확립해야 한다. 미국 공정거래위원회는 데이터 보안과 관련하여 10가지 기본 수칙을 발표하였다(FTC, 2015). 첫째, 데이터 보안은 데이터 수집 및 이용 목적을 명확하게 수립하는 것에서부터 시작된다. 우선 불필요한 개인정보를 수집하지 않아야 하고, 또한 타당한 이유가 없으면 개인정보를 보관하지 않아야 한다. 불필요하게 개인정보를 활용하지 않아야 한다. 이를 통해 개인정보의 불필요한 외부 노출을 최소화하고 방지할 수 있다. 둘째, 데이터 접근 권한을 제한해야 한다. 특히 개인민감정보에 대한 접근 권한을 제한할 필요가 있다. 개인정보 관리시스템에 대한 접근도 제한해야 한다. 셋째, 데이터 접근 시 안전한 비밀번호와 인증절차를 요구해야 한다. 비밀번호는 복잡하고 유일하게 설정하도록 요구하고, 안전한 저장소에 암호화하여 보관하여야 하며, 외부 침입을 방지할 수 있는 체계를 갖춰야 하며, 인증절차를 무단통과하지 못하도록 철저한 방어벽을 설치해야 한다. 넷째, 개인민감정보는 안전하게 저장해야 하고 이를 전송할 때 안전한 보호장치를 사용해야 한다. 개인민감정보는 한 곳에 저장하지 않아야 하며, 이를 전송할 때 반드시 암호화해야 한다. 다섯째, 네트워크를 분산시켜야 하고 데이터 접근 정보를 주기적으로 점검해야 한다. 한 곳에서 모든 데이터를 한 번에 접속할 수 없도록 네트워크를 분산시켜야 하며, 데이터 접근 로그 등을 이용해 내부 또는 외부의 비승인된 데이터 접근여부를 점검해야 한다. 여섯째, 데이터에 대한 원격 접속에 대해 높은 보안수준을 유지해야 한다. 또한 원격 접속 권한도 높은 수준으로 제한해야 한다. 일곱째, 새로운 상품 또는 서비스를 개발할 때 건전한 보안조치를 적용시켜야 한다.

여덟째, 정보처리자가 합리적인 보안조치를 수행하는지를 주기적으로 확인해야 한다. 데이터 공유를 위한 계약은 서면으로 체결되어야 하며, 데이터관리자는 이를 정보처리자가 준수하는지를 주기적으로 확인할 수 있어야 한다. 아홉째, 현재 데이터 보안의 유효성과 잠재적 위험을 인식하여야 한다. 열째, 개인정보가 담긴 문서, 이에 접근할 수 있는 매체를 안전하게 관리해야 한다.

부당한 차별 가능성은 기업윤리 기준을 제고하는 방법으로 해결할 수 있다. 빅데이터를 악용한 부당한 차별은 모든 분야에서 발생할 수 있다. 이를 방지하기 위해서는 기업이 윤리적으로 빅데이터를 분석하고 그 결과를 활용할 수 있어야 한다. 미국 공정거래위원회는 기업이 빅데이터를 활용해 고객의 특성 및 행태를 연구하고자 할 때 지켜야할 4가지 기본수칙을 제안하였다(FTC, 2016). 첫째, 기업은 활용하고자 하는 데이터가 얼마나 모집단을 대표하는지를 평가해야 한다. 둘째, 데이터 분석 모델이 편향(biases)을 얼마나 고려하는지를 알아야 한다. 셋째, 빅데이터에 기반한 예측이 얼마나 정확한지를 점검해야 한다. 넷째, 빅데이터 분석의 결과를 활용할 때 윤리 또는 공정성 문제가 없는지를 살펴봐야 한다. 이러한 수칙이 엄격하게 준수된다면 빅데이터가 부당한 차별을 위해 활용되지 않을 수 있다. 그럼에도 불구하고 기업윤리가 잘 지켜지지 않을 경우 여전히 빅데이터는 부당한 차별에 악용될 수 있다.

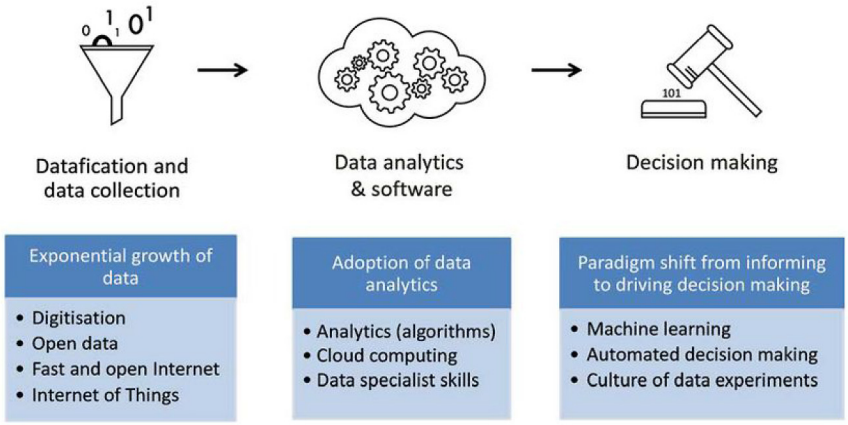
3. 빅데이터 파급효과와 한계

가. 빅데이터 파급효과

빅데이터 산업은 경제 성장, 혁신과 생산성 향상에 촉매제 역할을 할 뿐만 아니라, 경제 패러다임을 바꾸는 촉매제의 역할을 수행할 것으로

평가된다(OECD, 2015a). 다만 빅데이터를 원유와 비교해볼 때, 빅데이터는 자원 창출(resource creation)이 아닌 자원 배분(resource allocation)을 통해 경제 전반의 효율성을 증진시킬 것이라는 점에서 원유와 다르다. 원유는 생산요소를 새롭게 창출하여 경제에 직접적인 영향을 미쳤다. 그러나 빅데이터는 그 자체가 물리적인 요소는 아니기 때문에 자원을 효율적으로 배분하는 방법으로 생산성을 높이고 이를 통해 새로운 부가가치를 창출할 것으로 평가된다. 다만 금융산업 및 자본시장과 같이 정보 자체가 매우 중요한 산업에서는 빅데이터가 새로운 금융상품 및 서비스 개발에 생산요소의 역할을 할 것으로 예상된다.

<그림 Ⅱ-9> 빅데이터 기술의 가치사슬

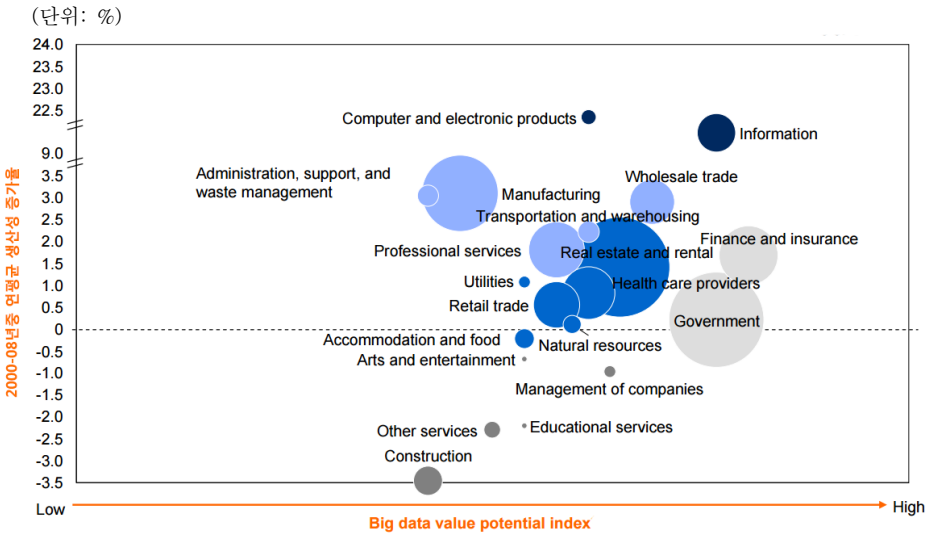


자료: Mckinsey Global Institute(2011)

2014년 8~9월중 CSC(Computer Sciences Corporation)가 전 세계 최고정보책임자(Chief Information Officer: CIO) 590명을 대상으로 설문조사를 실시한 결과, 응답자의 75%가 빅데이터를 잘 활용할 경우 생산성과 효율성을 제고시킬 수 있고, 응답자의 69%가 빅데이터가 더 중요해질 것이라고 평가한 것으로 조사된다(CSC, 2014). 또한 미국의 경우 산업별 생산성이 빅데이터 활용에 따라 평균적으로 연 0.5~1%씩 증가될 것으로

추정한다(McKinsey Global Institute, 2011). 각 산업부문별로는 적게는 1천억달러에서 많게는 7천억달러 규모의 경제적 부가가치가 창출될 것으로 예상되며, 역사적인 생산성 증가율이 가장 낮게 나타난 금융산업에서 빅데이터의 파급효과가 전 산업중에서 가장 높을 것으로 추정된다. 이는 금융산업 자체가 정보의 비대칭성으로 시장실패가 존재하고 정보에 의해 가치가 생성되는 산업이라는 특성 때문이다.

<그림 Ⅱ-10> 빅데이터가 산업별 생산성에 미칠 영향



자료: McKinsey Global Institute(2011)

빅데이터의 파급효과를 산업별 사례로 살펴보면 다음과 같다(BSA, 2015). 빅데이터는 교통혼잡과 교통사고를 줄이는 데 활용될 수 있다. 안전벨트를 착용할 때와 같이 빅데이터를 활용하는 것만으로도 교통상해와 교통사망의 50%를 절감시킬 수 있는 것으로 추정된다. 항공산업에서 빅데이터를 잘 활용해 1%의 생산성을 향상시킬 경우 향후 15년간 약 300억달러의 연료비용을 절감할 수 있을 것으로 추정된다. 비행기가 한번

비행할 때마다 약 0.5TB의 센서 데이터가 생산된다. 이를 잘 활용할 경우 비행성능, 난기류 대응능력, 비행안정 등을 향상시키고 엔진결함을 찾는 데 2천배 더 효과적일 것으로 추정된다. 한편 빅데이터는 에너지 절감과 환경보호에도 상당한 비용 절감의 효과를 가져올 것으로 평가된다. 빌딩 내 설비의 에너지 소비를 효율적으로 관리하는 스마트빌딩을 통해서 연간 250억달러의 에너지비용을 절감할 수 있을 것으로 추정된다. 또한 의료산업에서 빅데이터를 잘 활용할 경우 매년 3억달러의 비용을 절감할 수 있을 것으로 추정된다. 의료산업은 매일 병원당 수백 TB의 데이터를 생산하고 있으며, 이를 잘 활용할 경우 환자의 질병 진단과 처방의 효율성을 제고할 것으로 평가된다.

전 세계 기업이 빅데이터를 활용해 창출할 수 있는 수익규모는 2015~2019년중 약 1.6조달러에 이를 것으로 보인다(IDC, 2014). 지금까지 빅데이터를 의사결정에 활용하는 기업의 경우 5~6%의 생산성이 향상된 것으로 조사된다(Capgemini, 2012). 제조기업이 빅데이터를 잘 활용할 경우 향후 4년간 3,710억달러의 비용을 절감할 수 있는 것으로 보인다(McKinsey Global Institute, 2011). 이처럼 빅데이터는 상품 디자인, 생산공정, 판매채널의 효율성을 높이는 데 기여할 수 있으며, 유럽에서는 빅데이터를 제4차 산업혁명을 이끌 수 있는 핵심요소로 평가하고 있다(EPRS, 2015). 향후 빅데이터가 글로벌 경제 전반의 생산성을 약 1% 향상시킨다면 전 세계 GDP는 2030년까지 15조달러가 증가할 것으로 추정된다(BSA, 2015). 이는 앞으로 15년내 글로벌 경제에서 미국 경제가 하나 더 생기는 것과 같다. 그만큼 빅데이터의 경제적 파급효과는 매우 광범위하게 나타날 수 있으며 그 수준도 매우 크다고 평가될 수 있다.

나. 빅데이터 한계

2000년대 중반부터 각 국에서 새로운 경제성장의 동력으로 빅데이터를 강조하고 있는 가운데 빅데이터에 대한 긍정적인 평가가 주를 이루고 있다

(OECD, 2015a). 특히 각 국의 정부는 2011년부터 다양하고 구체적인 빅데이터 지원정책을 경쟁적으로 내놓고 있다. 그러나 빅데이터에 대한 회의적인 시각도 상당 수준 존재한다. 이들 대부분은 1990년대에 데이터마이닝이 사회적으로 큰 관심을 받았다가 그 유용성이 제대로 증명되지 못한 채 사회적 관심을 점차 잃었던 것처럼 빅데이터도 동일한 경로를 겪게 될 수 있다고 보고 있다. 특히 당초 기대와 달리 빅데이터가 오히려 잘못된 의사결정을 유도하거나 경제적 피해를 초래할 수 있다는 의견도 제시되고 있다.

예를 들면, 최근 미국에서는 2016년 미국 대선 결과를 놓고 2012년과 비슷하지만 다른 양상으로 빅데이터에 대한 논쟁이 뜨겁게 일어나고 있다. 2012년 대선에 패한 공화당은 버락 오바마(Barack Obama)의 재선 성공요인중의 하나로 빅데이터에 기반한 선거전략이 유효했기 때문이라고 분석하였다(Scola, 2013). 오바마 캠프는 빅데이터 분석을 통해 할리우드 정치헌금 디너파티에에서 정치헌금을 낼 가능성이 가장 높은 그룹이 40대 여성이라고 파악하고, 이들 여성에게 인기가 많은 영화배우 조지 클루니(George Clooney)를 초대하는 전략을 수립하여 대성공을 거둔 것으로 평가받는다. 또한 오바마 캠프는 박빙일 것으로 예측된 주(state)에서 오바마에 투표할 가능성이 높은 유권자의 개인 성향을 빅데이터로 파악하고 이에 맞는 맞춤형 선거전략을 수립해 부동층 유권자를 흡수하는 데 성공한 것으로 평가받는다. 이와 달리 2016년 미국 대선 이후 빅데이터에 대한 재평가가 필요하다는 논쟁이 일어나고 있다. 클린턴 캠프는 오바마 캠프와 유사하게 빅데이터를 활용해 선거전략이 수립하고 이를 활용했지만, 빅데이터가 너무 과대평가되었다고 판단하고 빅데이터를 무시한 도날드 트럼프(Donald Trump) 캠프에 패배했다. 트럼프 캠프는 대신에 후보자 자신이 SNS에서 직설적인 발언을 내쏟고 막대한 광고비를 지출하는 방법 등으로 소셜 트렌드를 형성하는 전략을 선택한 것으로 알려진다. 이 때문에 트럼프는 빅데이터을 두고 도박한다는 비판을 받기도 했으나, 트럼프 캠프는 2016년 미국 대선에서 주류 언론의 예측과 다르게 승리했다(Crovitz, 2016). 한편 2016년 미국 대선 결과가 빅데이터의 무용성을

입증했다기 보다는 빅데이터를 어떻게 분석하고 해석하느냐가 앞으로 빅데이터 분야에서 중요한 관건이 될 수 있다는 점을 시사한다고 보는 이도 적지 않다(Carpenter, 2016; Enderle, 2016).[6)]

빅데이터에 대한 회의적인 시각은 크게 세 가지 이유 때문인 것으로 조사된다(OECD, 2015b).

첫째, 빈약한 데이터 품질이 잘못된 해석을 야기할 수 있기 때문이다. 이는 전문용어로 'garbage in, garbage out'이라고 표현되기도 한다. 빅데이터 기술은 대규모 데이터를 활용할 수 있다는 점이 큰 장점으로 평가받지만, 처리가능한 데이터의 절대적 규모가 크다고 해서 반드시 데이터의 측정 오류(measurement errors), 유효성(validity), 신뢰성(reliability), 의존성(dependecies)과 같이 데이터 분석에서 근본적으로 제시되는 이슈들이 무시될 수 있는 것은 아니다(Lazer et al, 2014). 따라서 빅데이터이더라도 데이터 품질이 좋지 않으면 잘못된 결과가 도출될 수 있고 이로 인해 경제적 피해가 발생할 수 있다. 예를 들면, 글로벌 금융서비스 기업인 Knight Capital Group은 2012년 알고리듬 트레이딩 시스템이 잘못된 데이터 입력을 걸러내지 못해 4.4억만달러의 손실을 입은 바 있다(Mehta, 2012). 2016년 미국 대선 결과가 빅데이터에 기반한 대선 예측과 다르게 나타난 것도 트럼프를 지지하지만 이를 표현하지 않는 유권자(shy Trump)를 고려하지 못했기 때문이라는 분석도 있다(Enderle, 2016).

둘째, 데이터와 분석기법에 대한 잘못된 선택이 잘못된 결과를 도출할 수 있기 때문이다. 데이터 품질이 좋고 데이터 규모가 크다면 엄밀한 데이터 분석기법으로 편차(biases)를 줄이고 의미있는 정보를 추출할 수 있다는 것이 일반적인 인식이다. 한편 빅데이터 분석기법은 인과관계(causation)

6) 2016년 대선 종료 후 트럼프 캠프도 영국 런던에 소재한 Cambridge Anlytica라는 빅데이터 기업을 활용해 지지율 동향을 파악한 것으로 알려진다. 특히, Cambridge Anlytica는 'hyper-targeted psychological approach'라는 독자적인 분석기법으로 미국 북부의 사양화된 공업지대(Rust Belt) 지역에서 트럼프 지지율이 상승하고 있다는 것을 발견한 것으로 유명하다. 이에 대해 Cambridge Anlytica의 트럼프 캠프 빅데이터 전략팀장인 Matt Oczhowski는 "We're not saying we predicted it right, but we certainly noticed some trends."라고 설명하였다(Mezzofiore, 2016).

보다는 상관관계(correlation) 분석에 크게 의존한다. 이 경우 상관관계 분석 결과가 반드시 두 변수 간의 인과관계를 설명하는 것이 아니기 때문에 통계적으로 유의한 상관관계가 반드시 의미있다고 해석될 수 없다. 즉 잘못된 데이터와 분석기법 선택으로 이러한 오류는 언제든지 발생할 수 있다. 예를 들면, 2006년과 2011년 기간중 미국의 연간 살인사망자와 웹 디자이너 자살사망자는 마이크로소프트의 인터넷 익스플로러(Internet Explorer)의 시장점유율과 높은 상관관계를 보이는 것으로 조사되었다(OECD, 2015b; Marcus & Davis, 2014). 그러나 세 변수 간에 명확한 인과관계가 존재한다고 설명하기는 어렵다. 이러한 허구적인 상관관계(spurious correlation)는 데이터 분석에서 언제든지 나타날 수 있다. 그만큼 상관관계 분석을 중시하는 빅데이터 분석기법에 대해 회의적인 시각이 존재할 수 밖에 없다.

셋째, 데이터 환경이 계속 변화하고 있기 때문이다. 기계학습 또는 인공지능은 빅데이터의 분석기법 중에서 가장 각광받고 있는 분야중 하나이다. 그러나 이러한 분석기법은 사전에 학습된 알고리듬에 기반하거나 이미 축적된 데이터에 기반해 알고리듬을 형성하기 때문에 알고리듬에 영향을 주는 환경적인 요인에 의해 언제든지 쉽게 엄밀성을 상실할 수 있다. 대표적인 요인중 하나가 데이터 환경이 계속해서 변화하기 때문이다.

예를 들면, 2008년 출시된 구글의 독감 트렌드(Google flu trends)가 대표적인 사례이다. 구글의 독감트렌드는 독감과 관련된 검색어에 대한 빅데이터 분석을 기반으로 독감을 예측하는 서비스이며, 처음 출시되었을 때만 해도 미국 질병방제센터보다 1~2주 빠르게 독감을 예측하는 것으로 평가받았다. 그러나 구글의 독감트렌드에서 심각한 예측오류가 발견되었다고 과학학술지인 Science와 Nature에 보고되면서 빅데이터 분석에 대한 회의적인 논의가 촉발되었다(Butler, 2013; Lazer et al., 2014; OECD, 2015b). 구글의 독감트렌트는 구글의 검색엔진에 의해 수집된 데이터에 크게 의존한다. 그런데 구글의 검색엔진 알고리듬은 계속해서 바뀌어 왔다. 2012년만해도 구글의 검색엔진에 86가지 변화가 있었다고 보고된다. 이러한 데이터 환경의 변화가 빅데이터 분석에 제때 반영되지

못하면 데이터에 내재된 패턴도 변할 수 있기 때문에 잘못된 의사결정으로 연결될 수 있다. 구글의 독감트렌드가 미국 전역의 독감 발생률을 과대 예측한 것도 이 때문인 것으로 알려져 있다(Lazer et al., 2014). 또다른 이유는 구글의 독감트렌드가 독감과 관련된 검색어를 대상으로 하기 때문이다(Bulter, 2013). 2012년 전후로 미국 전역에서 독감에 대한 언론보도 비중이 갑자기 높아지면서 독감과 관련된 구글 검색빈도도 갑자기 높아진 것으로 파악된다. 이 때문에 독감에 걸리지 않은 구글 이용자의 독감에 대한 검색이 구글의 독감트렌드의 예측력을 약화시킨 것으로 조사되었다.

빅데이터에 대한 회의적인 시각이 계속해서 제기됨에도 불구하고 각국에서는 빅데이터의 잠재적 역량을 여전히 긍정적으로 평가하고 있다. 왜냐하면 앞서 제시된 회의적인 시각이 빅데이터의 유용성 자체를 부인하기 보다는 빅데이터 분석의 엄밀성을 강조하고 있기 때문이다. 즉 빅데이터의 유용성은 빅데이터를 어떻게 처리하고 분석하느냐에 따라 결정될 수 있다는 점을 시사한다(OECD, 2015b). 따라서 빅데이터는 가보지 않은 길을 가야한다는 점에서 무모한 도전이 될 수 있으나, 가보지 않았기 때문에 오히려 조심하면서 가야하는 길임에 틀림이 없다.

Ⅲ. 해외 자본시장의 빅데이터 도입 현황

1. 빅데이터 도입 효과

2. 빅데이터 기업 및 분석기법 사례

3. 해외 자본시장의 빅데이터 도입 특징

Ⅲ. 해외 자본시장의 빅데이터 도입 현황

본 장에서는 해외 자본시장에서 빅데이터 기술의 도입 현황을 빅데이터 도입 효과, 기업 및 분석기법사례, 해외 자본시장의 빅데이터 도입 특징으로 나누어 살펴보았다. 다만 국내와 같이 해외의 경우에도 자본시장에서 빅데이터의 중요성이 계속해서 강조되고 있지만 이에 대한 구체적인 사례가 소개되지 않고 있다. 이 때문에 자본시장에서 빅데이터 관련 현황을 조사하는 데 많은 제약이 따랐다.

1. 빅데이터 도입 효과

자본시장에서 빅데이터는 새로운 수익창출의 기회를 확장시키고, 이전보다 리스크관리 및 규제준수를 용이하게 할 수 있으며, 데이터 관리 및 사업운영 효율화 등으로 비용을 절감시킬 수 있는 것으로 평가받는다(Aite, 2014; Eyraud, 2016; PwC, 2013; Singh, 2014; Verma & Mani, 2014).

가. 새로운 수익창출 기회 확장

자본시장에서 빅데이터가 중요하게 부각되는 이유중 하나는 자본시장 자체가 수많은 데이터를 직접 생성하고 그 데이터에 의해 민감하게 반응하기 때문이다. 따라서 자본시장은 어느 산업보다 빅데이터에서 의미있는 정보를 추출할 수 있고 이를 통해 새로운 수익기회를 창출할 가능성이 매우 높다. 예를 들면, 빅데이터가 상장기업 분석, 투자전략 수립, 고빈도매매(High Frequency Trading: HFT) 및 프로그램 매매, tick 분석, SNS 및 뉴스 분석을 통한 트레이딩전략 수립, 스마트베타 투자전략 수립 등에

활용될 경우 새로운 수익기회를 창출할 수 있을 것으로 기대된다. 또한 실시간 빅데이터 기술을 활용할 경우 더 정확하고 빠른 거래분석이 가능하며, 이를 통해 투자자에게 더 높은 수익창출의 기회를 제공할 수 있다고 보고 있다. HFT 및 프로그램 매매는 실시간 시장데이터에 크게 의존한다. 그런데 기존 데이터 시스템은 실시간 시장데이터를 수집하고 처리하고 분석하는 데 속도 측면에서 한계가 있다. Spark 및 Impala와 같은 실시간 빅데이터 기술은 이를 극복할 수 있는 하나의 방법으로 소개된다.

빅데이터가 고객행태 분석, 고객분할 분석, 고객서비스 관리, 투자자문 서비스에 활용될 경우 새로운 금융상품 및 서비스를 설계하고 판매채널의 효율성을 제고할 수 있을 것으로 예상된다(Chemitiganti, 2016; PwC, 2013). 예를 들면, 고객의 투자행태 분석을 통해 금융상품 및 서비스에 대한 수요를 예측하고 이에 맞게 신규 금융상품 및 서비스를 설계할 수 있다. 또한 금융상품 및 서비스에 대한 고객의 반응을 수집하여 고객서비스를 개선하고 판매채널의 효율화를 도모할 수 있다. 그러나 현재까지 자본시장에서 이 분야에 특화된 빅데이터 기업은 출현하지 않은 것으로 조사된다. 증권회사의 활용사례도 쉽게 조사되지 않는다.

빅데이터는 양질의 투자자문 및 자산관리 서비스의 자동화에 활용되어 자산관리서비스를 대중화시켜 새로운 수익기회를 창출하고 있는 것으로 평가받는다. 특히 낮은 최소투자한도와 낮은 자문보수를 제시하는 로보어드바이저(robo-advisor)가 이 분야에서 선두를 달리고 있다. 로보어드바이저는 고객의 투자성향 파악, 포트폴리오 추천, 주문, 포트폴리오 리밸런싱(rebalancing)의 과정을 통해 고객에게 자동화된 투자자문 및 자산관리 서비스를 제공하는 자동화된 온라인 플랫폼이다. 이 과정에 빅데이터는 중요한 역할을 하는 것으로 알려져 있다(Mecham, 2016). 우선 해외 로보어드바이저의 경우 고객의 투자성향을 파악하는 데 빅데이터에 기반한 알고리듬을 이용한다. 또한 로보어드바이저가 고객에게 추천하는 포트폴리오도 빅데이터에 기반하여 고객의 투자성향과 투자목적에 따라 맞춤형으로 제공한다. 특히 고객의 포트폴리오 리밸런싱에 빅데이터의 역할이 큰 것으로 조사된다. 포트폴리오 리밸런싱은 주기적으로 시장상황에 맞게

고객에게 제공되기도 하지만, 기대치 못한 시장상황 변화에 빠르게 대응하기 위해서도 제공된다. 이 경우 실시간 빅데이터 분석을 기반으로 사람보다 빠르게 포트폴리오 리밸런싱을 제공한다는 것이 로보어드바이저의 장점으로 알려져 있다.

나. 리스크관리 및 규제준수 효율성 증진

빅데이터는 증권회사의 리스크관리에 획기적인 변화를 가져올 수 있다고 평가된다(Eyraud, 2016). 리스크관리에서 시간(time)은 매우 중요한 변수이기 때문에 얼마나 적시에 의사결정을 내릴 수 있느냐에 따라 그 성과가 좌우될 수 있다(Avantage Reply, 2014). 예를 들면, 빅데이터 기술을 이용할 경우 증권회사는 실시간 데이터 분석을 통해 투자성과 및 리스크 매트릭스를 실시간으로 작성할 수 있으며, 예측분석을 통해 기대위험을 예측하고, 이에 따라 리스크 익스포져를 효과적으로 관리할 수 있다. 또한 비정형 데이터를 포함한 빅데이터를 스트레스 테스트에 활용할 경우 예측력을 더 높일 수 있다. 이를 통해 최적의 헤지전략을 도출하고 관련된 제반비용을 절약할 수 있다.

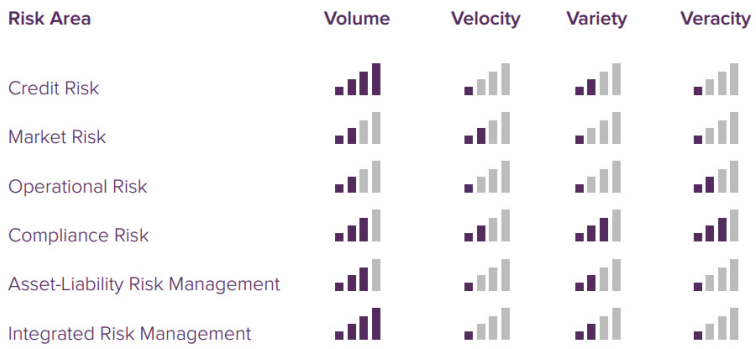
빅데이터는 거래감시(market surveillance), 사기탐지(fraud detection), 자금세탁방지(Anti-Money Laundering: AML), 고객알기정책(Know Your Customer: KYC), 규제보고(regulatory reporting) 등 내부통제 및 준법감시에도 활용될 수 있다. 빅데이터는 회사내 또는 고객간 메시지, 전화기록, 이메일 및 웹페이지 등 다양한 채널에서 생성되는 데이터를 결합하여 빅데이터로 구성하고, 패턴매칭 분석 등의 분석기법을 활용해 비정상적인 행위를 탐지할 수 있으며, 자금세탁방지와 고객알기정책을 더 수월하게 수행할 수 있다. 이를 통해 증권회사는 내부통제의 효율성을 높이고 규제준수 비용을 절약할 수 있다.

증권회사는 기존 시스템(legacy system)에 분산되어 있는 거래를 매칭하거나 조정해야 하며 이 과정에서 무효, 중복, 실패 거래가 발생하는 것으로 알려져 있다.

이에 따라 증권회사의 운영리스크도 증가할 수 있다. 그러나 빅데이터의 데이터 태깅(data tagging) 기술을 활용할 경우 거래를 오류없이 확인할 수 있고 운영리스크 증가를 최소화할 수 있는 것으로 조사된다. 또한 빅데이터는 거래비용 또는 주문집행 성과를 평가할 수 있는 거래후 분석도구를 제공할 수 있다. 뿐만 아니라 빅데이터는 사업부문별로 분리되어 있는 데이터를 통합관리하는 데도 효율적일 수 있다. 마지막으로 빅데이터는 비정형 데이터을 처리할 수 있기 때문에 주문집행 및 거래체결을 더 효과적으로 감시할 수 있다.

리스크관리 분야에서 증권회사가 빅데이터 기술을 활용할 유인이 높아지는 이유중 하나는 증권회사의 기존 시스템이 갈수록 방대해지는 데이터를 관리하는 데 한계에 직면할 수 있기 때문이다. 특히 자본시장에서 데이터 생성속도가 빨라지는 만큼 데이터 축적량도 급증하고 있으며, 금융규제가 복잡해지면서 증권회사가 관리하고 처리해야 하는 데이터량도 비대해지는 추세이다. 이 때문에 대규모 정형 또는 비정형 데이터를 분산저장과 분산처리할 수 있는 빅데이터 기술의 중요성이 더욱 확대되고 있다.

<그림 Ⅲ-1> 리스크관리 분야에서 빅데이터의 중요성



자료: Avantage Reply(2014)

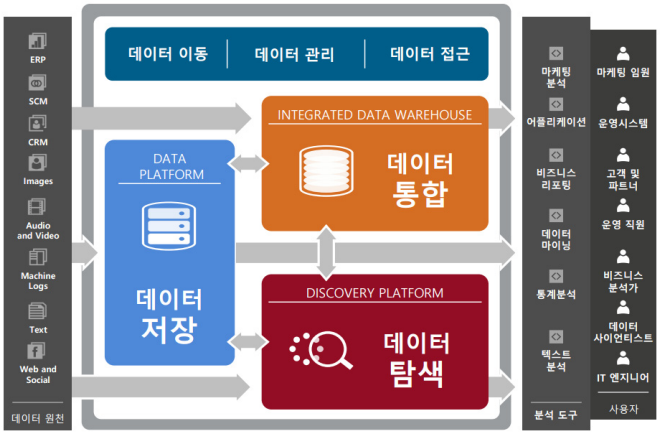
다. 비용절감 및 사업운영 효율성 증대

빅데이터 기술의 장점 중에 하나는 데이터 관리 및 유지비용을 대폭 절감시킬 수 있다는 것이다. 또한 빅데이터 기술은 대규모의 데이터를 저장용량에 구애받지 않고 분산저장할 수 있다. 이처럼 빅데이터가 자본시장에서 생성되는 데이터를 비정형 데이터로 저장하고 처리할 수 있다는 장점 때문에 비용 측면에서 매우 효율적이라고 평가받는다(Eyraud, 2016).

빅데이터 시스템은 크게 데이터 저장, 데이터 통합, 데이터 탐색, 데이터 분석, 데이터 보안 시스템으로 구성된다. 우선 빅데이터 시스템은 분산저장 방식을 채택한다. 예를 들면, 증권회사가 빅데이터 시스템을 도입할 경우 고객의 원장데이터를 한 저장매체에 저장하거나 임의적으로 쪼개서 저장할 필요가 없다. 빅데이터 시스템이 각 저장소인 노드에 자동적으로 분산시켜 저장한다. 또한 빅데이터 시스템이 텍스트 데이터를 수집, 저장, 분석할 수 있기 때문에 빅데이터 시스템을 도입할 경우 사내 이메일, 문서, 회의록, 메신저기록, 콜센터기록, 웹로그 등을 효과적으로 관리할 수 있으며, 사내 인트라넷에서 생성된 데이터도 가치있는 정보로 저장하고 관리할 수 있다.

빅데이터 시스템은 다양한 출처의 데이터를 통합하여 이를 분석할 수 있는 툴을 제공하기 때문에 증권회사 사업전략을 구현하는 데도 효과적일 수 있다. 증권회사 사업전략의 주요목적은 기존고객 유지, 신규고객 확보, 투자금액 증대 등에 있다. 이를 위해 증권회사는 빅데이터을 활용해 고객의 투자성향, 투자목적, 투자실태, 투자행태 등을 이전보다 효과적으로 분석할 수 있다. 예를 들면, 증권회사는 고객의 민원이나 전화상담 등 피드백을 활용하기 위해서는 수작업으로 그 내용을 정형화해야 했으나 빅데이터 시스템을 활용할 경우 다양한 유형의 고객 피드백을 자동적으로 데이터로 변환시키고 분석하여 고객이 원하는 서비스가 무엇인지를 종합적으로 파악해낼 수 있는 것으로 조사된다. 이 경우 데이터 수집 및 분석에 소요되는 시간을 상당 수준까지 단축시킬 수 있다.

<그림 Ⅲ-2> 빅데이터 시스템 구성



자료: 장동인(2015)

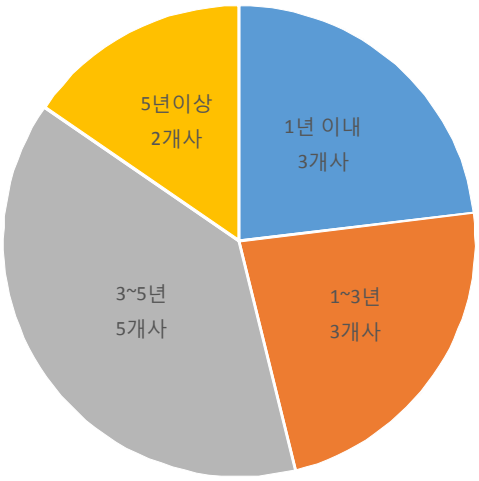
다만 증권회사 등 자본시장 참여자가 빅데이터 시스템을 도입하여 그 분석기법을 자체적으로 활용하는 데는 상당 시간 소요될 것으로 예상된다(Aite, 2014). 초기단계에서 빅데이터 분석에 따른 편익이 빅데이터 시스템 도입비용보다 크지 않을 것으로 기대되기 때문이다. 이 때문에 증권회사 등은 빅데이터 시스템을 도입하는 데 주저할 수 있다. 실제 국내의 경우 은행 및 보험업권에서는 빅데이터 시스템을 도입한 사례가 보고되나, 증권회사가 빅데이터 시스템을 도입한 사례는 없는 것으로 조사된다.

증권회사 등이 빅데이터 시스템을 도입하더라도 단기적으로는 빅데이터 활용 목적보다는 전산시스템을 빅데이터 시스템으로 전환할 목적이 큰 것으로 조사된다. 해외 자본시장의 경우 빅데이터의 미래가치를 인정하고 예전에 축적하지 않거나 활용하지 않았던 데이터를 저장하고 수집하여 관리하는 추세이다. 그만큼 증권회사 등은 데이터를 집중저장 및 집중분석하는 기존 시스템으로는 비정형 데이터가 포함된 데이터를 저장하고 관리하는 데 한계를 느끼고 있다. 이 때문에 증권회사 등은 데이터 저장 및 관리의 효율성을 제고할 목적으로 빅데이터 시스템을 도입할 유인이 높은 것으로 조사된다.

해외의 경우 증권회사 등이 빅데이터 시스템을 직접 도입하지 않더라도 빅데이터 전문기업을 통해 빅데이터를 활용할 수 있는 생태계가 형성되고 있다. 증권회사 등이 개별적으로 빅데이터를 구축하고 이를 직접 분석하는 것보다 빅데이터 전문기업이 제공하는 플랫폼에서 광범위하게 구축된 빅데이터를 이용하고, 빅데이터 전문기업이 제공하는 맞춤형 분석도구를 통해 빅데이터를 분석하는 것이 규모의 경제와 범위의 경제 측면에서 더 효율적일 수 있기 때문이다.

Aite(2014)가 해외 헤지펀드, 증권회사, 자산운용회사 등 13개사를 대상으로 실시한 설문조사에 따르면, 몇 년 이내에 빅데이터을 활용할 계획이냐는 질문에 대해 1년 이내로 응답한 곳이 3개사, 1~3년 이내로 응답한 곳이 3개사, 3~5년 이내로 응답한 곳이 5개사, 5년 이상으로 응답한 곳이 2개사로 조사된다. 그러나 해외 자본시장의 경우 최근 2~3년간 빅데이터를 활용하는 자본시장 참여자는 꾸준하게 증가하고 있는 것으로 조사된다.

<그림 Ⅲ-3> 금융투자회사의 빅데이터 활용 계획



자료: Aite(2014)

2. 빅데이터 기업 및 분석기법 사례

기술적인 제약과 비용적인 부담 등의 문제로 인해 증권회사 등이 독자적으로 빅데이터 시스템을 직접 구축하기까지는 많은 시간이 소요될 것으로 예상된다. 이 틈새 속에서 자본시장의 빅데이터 수요를 충족시킬 수 있는 빅데이터 기업이 계속 출현하고 있는 추세이다. 따라서 본 절에서는 해외 자본시장에서 유망한 빅데이터 기업과 분석기법에 대한 사례를 중점적으로 살펴보았다. 이를 통해 자본시장에서 빅데이터가 어떻게 활용되는지를 간접적으로 파악하고자 하였다.

가. 빅데이터 기업 사례

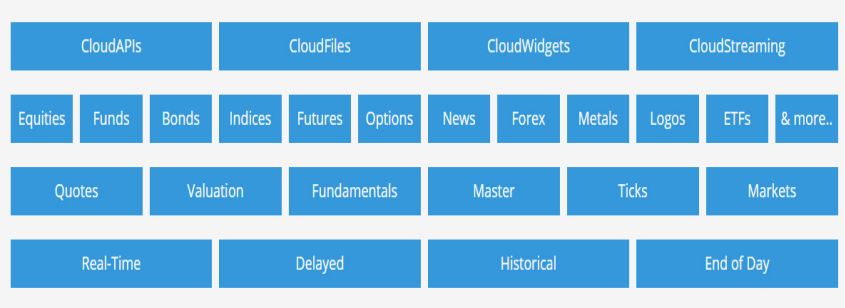
1) Xignite

2003년 미국 캘리포니아에 설립된 Xignite는 시장데이터 클라우드(Market Data Cloud: MDC)와 API를 활용해 1,000개 이상의 금융업자, 기관투자자, 언론매체, 소프트웨어 개발자 등에게 주식시장, 채권시장, 펀드시장, 상장지수펀드(Excahnge Traded Fund: ETF), 금리, 환율, 원자재 등과 관련된 실시간 또는 역사적 시세정보 및 금융정보를 유료로 제공하는 빅데이터 기업이다. Xignite는 데이터 제공업체로 설립되었지만 빅데이터 기술을 도입하면서 자본시장에서 그 활용성이 더 높아진 것으로 평가받고 있다.

Xignite의 시장데이터 공급 구조는 세 단계로 구분된다. 첫째, Xignite는 시장데이터를 수집하고 통합하여 빅데이터를 구축한다. 둘째, Xignite는 MDC에 빅데이터를 탑재한다. 셋째, Xignite는 고객이 API를 통해 MDC에 접속하여 필요한 데이터를 이용할 수 있도록 한다. Xignite는 금융관련 애플리케이션 개발자를 위한 Financial Data API, 거래소 및 데이터 밴더(vendor)를 위한 Xignite Market Data Distribution, 금융기관을 위한

Xignite Enterprise Data Distribution 등 고객 유형에 따라 맞춤형 데이터 공급수단을 제공한다.

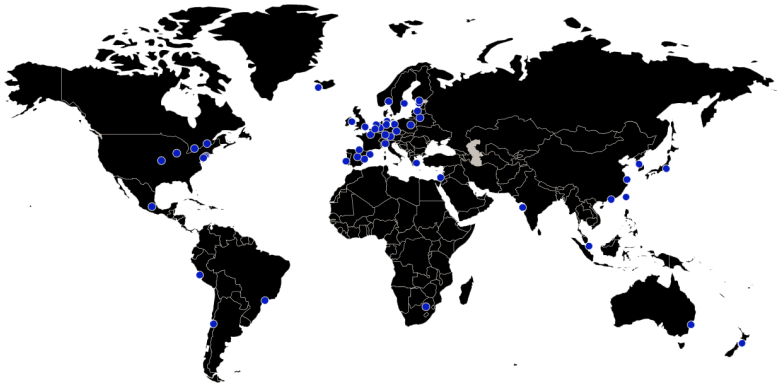
<그림 Ⅲ-4> Xignite의 서비스별 데이터 유형 및 특성



자료: Xignite 홈페이지

참고로 Xignite가 제공하는 시장데이터는 40개 이상 국가의 증권거래소 등에서 수집된다. 한국거래소의 시세정보도 Xignite에 제공되는 것으로 조사된다.

<그림 Ⅲ-5> Xignite의 시장데이터 지역별 현황



자료: Xignite 홈페이지

Xignite의 대표적인 고객 유형은 Wealthfront, Personal Capital, Betterment, FutureAdvisor, Yodlee, SigFig, jemstep, Square, Cashstar, Seeking Alpha, StockTwits, ETF.com, Robinhood, etoro, MotifInvesting, 8securities, Kensho, Ycharts, ChartIQ 등 자산관리, 소매금융, 지급결제, 투자자지원, 투자거래, 투자분석 등을 지원하는 핀테크 기업으로 조사된다. 또한 BlackRock, Charles Schwab, BNY Mellon, Ameritrade, Envestnet, Oppenheimer 등 기존 금융회사도 비용절감 및 시장데이터 구축을 위해 Xignite를 이용하는 것으로 알려져 있다.

<그림 Ⅲ-6> Xignite 이용 핀테크 기업

자료: Xignite 홈페이지

Xignite는 2009년 11월부터 2016년 2월까지 Altos Ventures, QUICK Corp, Startup Capital Ventures, StarVest Partners 등으로부터 약 3,639만달러의 투자를 유치하였다(Cruchbase.com).

2) Quandl

2011년 캐나다 토론토에 설립된 Quandl은 자본시장에서 활용 가능한 빅데이터를 중개하고 빅데이터 분석툴을 제공하는 빅데이터 기업이다. 이를 위해 Qunadl은 ZACLS, Zillow, 세계은행, 나스닥(NASDAQ), barchart, eurostat, 미국 FRB, OECD, ICI, EIA, Sharadar 등으로부터 주식, 선물, 원자재, 외환, 이자율, 옵션, 자산관리 및 펀드, 인덱스, 산업, 경제 등 다양한 분야의 데이터를 제공받아 빅데이터를 구축하고, 이를 Quandl이라는 플랫폼에 탑재하여 고객이 API, R, Python, Excel, Ruby 등 다양한 빅데이터 분석도구를 통해 활용할 수 있도록 관련 서비스를 통합적으로 제공한다. 이는 자본시장에서 빅데이터가 잘 유통될 수 있도록 데이터 유통망을 구축하고 이를 통합해 의미있는 정보를 공급할 수 있는 데이터 유통 플랫폼을 구축하였다는 의미를 가진다.

<그림 Ⅲ-7> Quandl의 빅데이터 서비스 공급 체계

자료: Quandl 홈페이지

Qunadl의 서비스는 오픈 데이터와 프리미엄 데이터 서비스로 대분된다. Qunadl의 오픈 데이터 서비스는 500개 이상의 데이터 제공자로부터 2,000만 금융 및 경제 데이터셋을 오픈 데이터 플랫폼에 탑재하여 제공하고 있으며,

고객이 무료 API를 통해 Excel, CSV, JSON, XML 등 고객이 원하는 형식의 데이터로 다운받아 이용할 수 있도록 제공하고 있다. 또한 다른 데이터 플랫폼에서도 사용할 수 있도록 다양한 데이터 변환 기능도 제공하고 있다. Quandl의 프리미엄 데이터 서비스는 Quandl 소속 데이터 관리자가 일차적 가공을 통해 데이터 품질을 제고시킨다는 점에서 오픈 데이터 서비스와 구별된다.

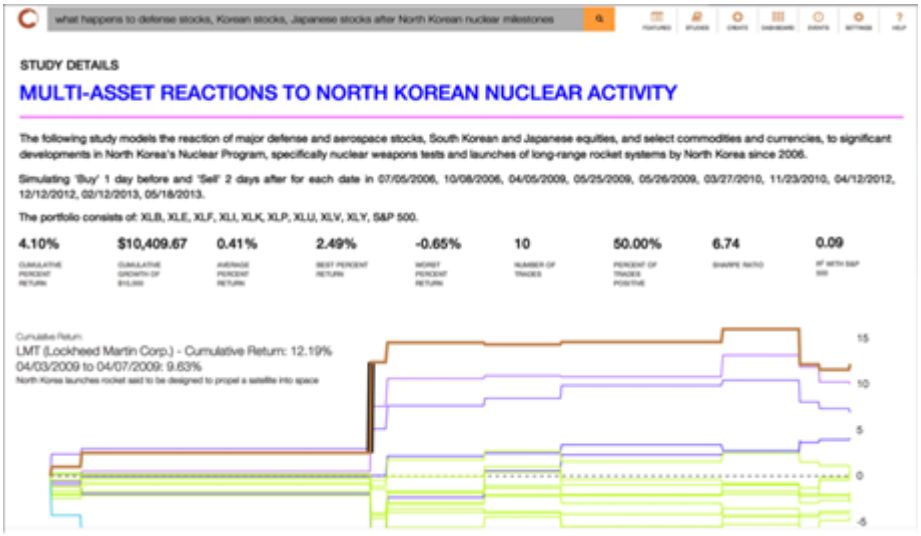
Quandl의 주요 고객은 2016년 9월말 현재 The Economist, HSBC, JPMorgan, Kensho, DOMO, BASF, The University of Chicago, Smithfield, Money.Net, Teza 등으로 조사된다. AmiBroker, Quantopian, Synvero, TrandingView, Wealth-Lab, WooTrader 등 금융 및 투자 플랫폼에서도 Quandl 데이터를 이용하는 것으로 조사된다.

Quandle은 2012년 9월부터 2016년 9월까지 August Capital, iGan Partners, Nexus Venture Partners 등으로부터 약 1,887만달러의 투자를 유치하였다(Cruchbase.com).

3) Kensho

2013년 미국 캠브리지에 설립된 Kensho는 기업실적, 정치 이벤트, 경제 데이터, 정책 변화 등을 빅데이터로 구성하고 이를 금융정보와 결합하여 실시간 이벤트들이 어떻게 자본시장에 영향을 미치는지를 분석할 수 있는 검색엔진을 제공하는 빅데이터 기업이다. Kensho는 자체 구축한 빅데이터와 웹페이지 등에서 설정할 수 있는 9만개 이상의 행동(customizable actions)을 스캔하고 그 결과를 가지고 65만개 이상의 질문을 조합하여 고객이 검색하는 질문에 답변할 수 있는 검색엔진이다(Madrid, 2015). 고객은 질문을 구글과 같이 검색창에 검색하거나 작성할 수 있으며, Kensho는 독창적인 데이터 분석기법을 활용하여 검색 질문에 대해 답변을 제시한다. 예를 들면, 고객이 자신의 투자자산이 북한 핵실험 활동에 어떻게 반응하지를 Kensho에 물어보면, Kensho는 이에 대한 검색 및 분석결과를 보고서 형태로 검색창에 제출한다.

<그림 Ⅲ-8> Kensho의 검색서비스 예시



자료: Kensho 홈페이지

Kensho의 투자분석 검색엔진은 설립 초기부터 애널리스트 15명이 4주에 걸쳐 할 수 있는 투자분석 작업을 5분만에 처리할 수 있는 능력을 가진 것으로 평가받는다. 그만큼 Kensho의 등장은 효과적인 빅데이터 분석이 비용절감뿐만 아니라 수익창출에도 기여할 수 있다는 기대감을 시장에 심어주는 계기가 되었다. 2014년 Goldman Sachs, Google Ventures, J.P. Morgan, Bank of America Merrill Lynch, CNBC, IQT 등이 Kensho의 전략적 투자자(strategical investor)로 나선 이유도 이 때문이다.

또한 Kensho는 2014년 12월부터 미국 방송국인 CNBC에 Kensho Stats Box와 Ask Kensho 서비스를 제공하고 있다. 이 서비스는 방송 진행자가 특정 질문을 Kensho에 제시하면, Kensho가 그 질문에 맞는 최적의 자본시장 관련 통계를 제시하거나 답변하는 방식으로 제공된다. 2016년 8월에는 인공지능 기술을 활용해 상장기업의 성장성을 측정할 수 있는 지수를 개발하는 사업을 추진중에 있다.

<그림 Ⅲ-9> Kensho Stats Box의 서비스 예시

자료 : Kensho 홈페이지

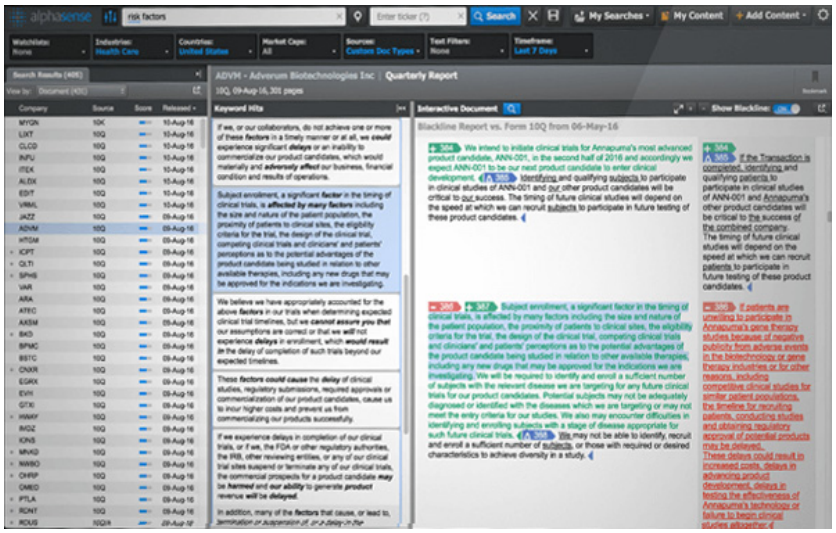
Kensho는 2014년중 Goldman Sachs, Accel Partners, Breyer Capital, GV 등 10곳 투자자로부터 약 5,780만달러 이상의 투자를 유치하였으며, 앞서 나열한 다수의 전략자 투자자가 지속적으로 투자하고 있는 것으로 조사된다(Cruchbase.com).

4) AlphaSense

2010년 미국 캘리포니아에 설립된 AlphaSense는 3,500개 이상 상장기업의 다양한 금융정보를 손쉽게 검색할 수 있는 검색엔진을 제공하는 빅데이터 기업이다. 현재 투자은행, 자산관리회사, 헤지펀드, 사모펀드, 리서치 회사 등 450개 이상의 회사가 AlphaSense를 이용하고 있으며, 상장기업의 재무정보, SEC제출 보고서, 애널리스트 보고서, 화상전화 스크립트, IR 보고서, 실시간 뉴스 및 보도자료 등이 포함된 데이터를 손쉽게 검색하고, 투자자들은 이러한 데이터를 상장기업에 대한 투자전략을 수립하는 데 활용하고 있는 것으로 알려져 있다.

AlphaSense는 1천여곳의 데이터 출처로부터 데이터를 수집하여 한 곳에 저장하고, 자연어 분석 처리를 통해 투자활동과 관련된 키워드를 추출할 뿐만 아니라 연관검색 등과 같이 자체 개발한 Smart Synonyms 엔진을 활용해 상장기업에 대한 최신 정보를 요약해 보여준다. 고객은 이를 통해 상장기업에 대한 분석 시간을 대폭 단축시킬 수 있다. 참고로 AlphaSense는 회원제로 운용되고 있으며, 시험버전도 사전심사를 통해서 제한적으로 공개하고 있다. 이 때문에 AlphaSense의 구체적인 작동원리를 파악하거나 검색결과를 살펴보기는 쉽지 않다.

<그림 Ⅲ-10> Alphasense의 검색 사례



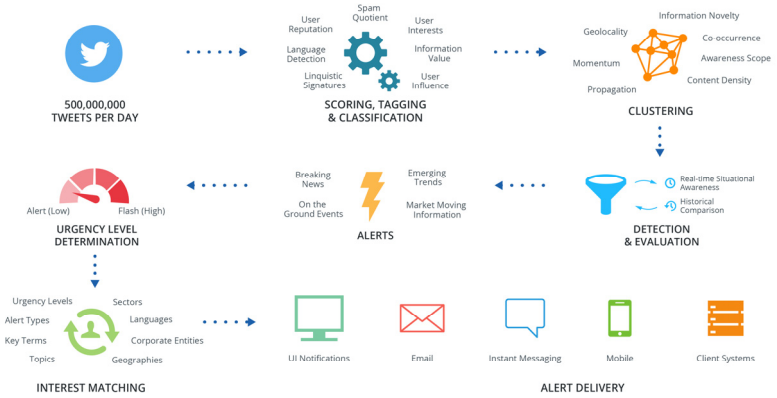
자료: Alphasense 홈페이지

AlphaSense는 2013년 1월부터 2016년 3월까지 First Fellow Partners, Quantum Strategic Partners, Tribeca Venture Partners 등 5곳 투자자로부터 약 3,500만달러 이상의 투자를 유치하였다(Cruchbase.com).

5) Dataminr

2009년 미국 뉴욕에 설립된 Dataminr는 인공지능이 탑재된 알고리듬 엔진을 이용하여 매일 5억개 이상의 트윗, 뉴스, 시장데이터 등 광범위한 데이터를 스캔하고 이를 이용해 금융회사, 통신사, 공공기관 등에 실시간으로 트윗동향 정보를 제공하는 빅데이터 기업이다. 이렇게 제공된 정보는 다시 기관투자자의 투자전략 수립에 영향을 미치게 되고 새로운 수익기회를 창출하는 데 활용된다. 기관투자자는 약간의 정보 시차로도 큰 수익을 낼 수 있기 때문에 Dataminr를 적극적으로 이용하고 있는 것으로 알려져 있다.

<그림 Ⅲ-11> Dataminr의 작동 원리



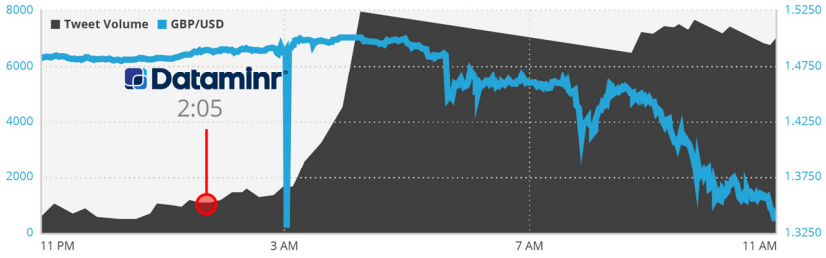
자료: Dataminr 홈페이지

Dataminr의 데이터 처리 및 분석 절차는 총 8단계로 구분된다. 우선 매일 5억개 이상의 트윗 정보뿐만 아니라 시장동향을 파악할 수 있는 뉴스정보, 시장데이터 등을 수집한다. 다음 단계로 트윗 정보를 스코어링(scoring), 태깅(tagging), 분류(classification) 절차를 거쳐 일차적으로 분석한다. 그 다음으로 흩어진 트윗 정보를 지리적 위치(geolocality), 정보 참신성

(information novelty), 동시 발생여부(co-occurrence), 모멘텀(momentum), 인지 범위(awareness scope), 콘텐트 밀도(content density), 전파력(propagation) 등 7개 기준으로 군집화(clustering)한다. 다음 단계에서는 이렇게 군집화된 트윗 동향이 유의미한지를 판단하기 위해 실시간 상황별 인지도를 평가하고 과거 트윗 동향과 비교 분석하여 고객에게 제공할 트윗 동향을 요약한다. 이렇게 추출된 트윗 동향을 신규 트렌트(emerging trends), 시장변동 정보(market moving information), 속보뉴스(breaking news), 중요 이벤트 관련(on the ground events) 등 네 가지 유형으로 구분하여 트윗 동향 정보를 생산한다. 여섯 번째 단계로 트윗 동향 정보의 등급을 결정한다. 일곱 번째 단계로 각 트윗 동향 정보를 수요에 맞게 공급하기 위해 매칭(interest matching) 작업을 실시한다. 마지막으로 고객에게 다양한 채널을 통해 맞춤형 트윗 정보 동향을 전달한다.

Dataminr는 뉴스, 금융, 공공부문, 기업보안 분야에 트윗 동향 정보를 제공하고 있다. 특히 금융분야에서는 금융뉴스 속보, 원자재, 상장기업, 거시경제와 관련된 트윗 동향을 추출하여 제공하고 있다. 예를 들면, Dataminr는 2016년 6월 24일 영국에서 EU 탈퇴여부에 대한 국민투표 선거결과를 공식 발표 1시간 이전에 관련 트윗 정보 분석을 통해 예측하고 긴급 경보 메시지를 고객에게 전달하였다. 이는 파운드가 달러대비 31년만에 최저치로 평가절하되기 1시간 이전과 같은 시간으로 조사된다.

<그림 Ⅲ-12> Dataminr의 영국 EU 탈퇴 경보와 파운드 평가절하



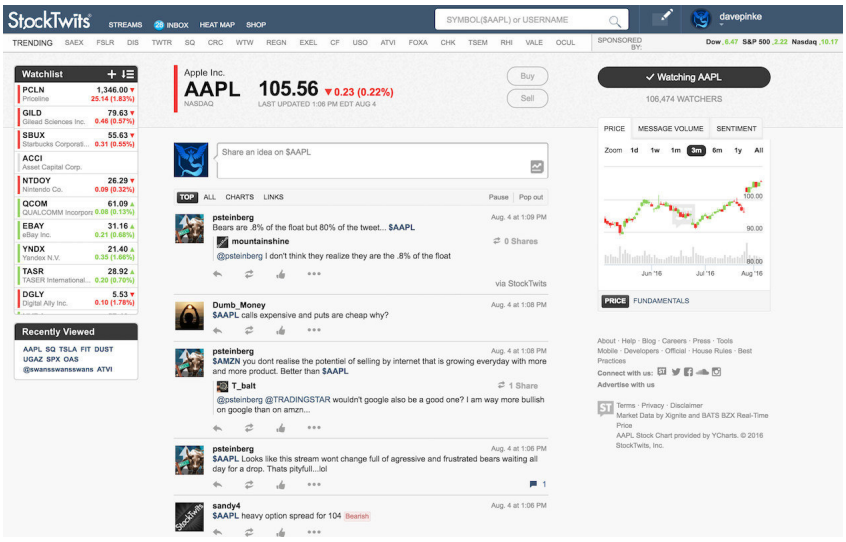
자료: Dataminr 홈페이지

Dataminr는 2010년 8월부터 2015년 9월까지 Credit Suisse, BoxGroup, EquityZen, Fidelity Investments, GSV Capital, Institutional Venture Partners, Richmond Global Ventures 등 19곳 투자자로부터 약 1억 8,344만달러 이상의 투자를 유치하였다(Cruchbase.com).

6) StockTwits

2008년 미국 뉴욕에 설립된 StockTwits는 국내 네이버 포탈의 증권서비스와 같이 주식에 특화된 SNS를 제공하는 빅데이터 기업이다. 이를 통해 고객이 서로 뉴스, SEC제출 보고서, 투자정보, 실시간 매매동향과 투자정보를 공유하고 이를 기반으로 고객이 투자와 관련된 의사결정을 내릴 수 있도록 돕고 있다.

<그림 Ⅲ-13> StockTwits의 애플사 트윗 창



자료: StockTwits 홈페이지

2014년말 현재 50만명 이상이 StockTwits에 가입하여 이중 약 3만명이 매달 130만 메시지를 주고 받는 것으로 보고된다(Lindzon, 2015). 평균적으로 매달 7,277명이 총 7천 종목에 대해 적어도 한 개 이상의 메시지를 보내며, 이중 9%의 메시지만이 중복된 것으로 조사된다. 또한 StockTwits은 고객이 특정 주식에 대해 투자결정을 내릴 수 있도록 메시지를 빅데이터 분석기법으로 분석하여 제공하고 있다. Bloomberg, Thomson Reuters, 헤지펀드, 증권회사 등은 StockTwits 데이터를 활용해 고객에게 투자정보 서비스를 제공하거나 투자전략을 수립하는 것으로 조사된다.

StockTwits는 2008년 9월부터 2016년 7월까지 Betaworks, True Ventures, Foundry Group 등 14곳 투자자로부터 약 1,390만달러 이상의 투자를 유치하였다(Cruchbase.com).

7) Deep Value

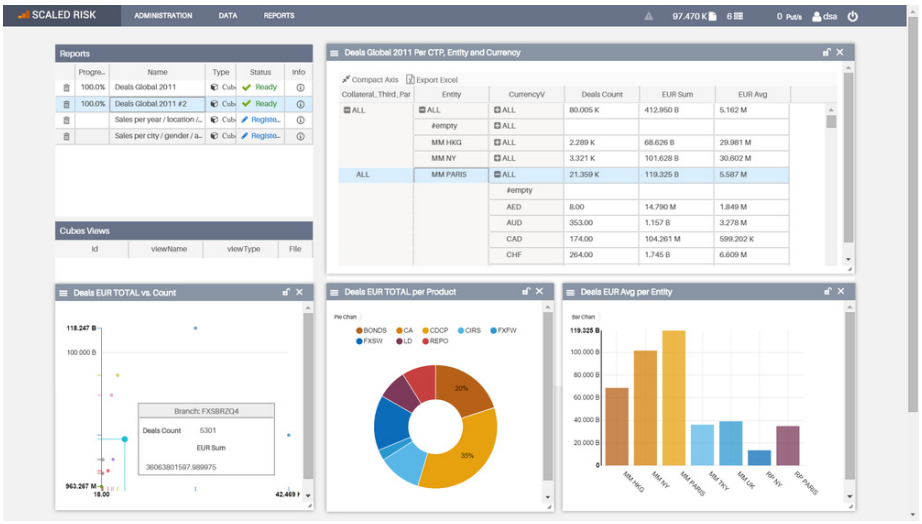
2005년 미국에 설립된 Deep Value는 빅데이터 기술을 접목해 최선집행(best execution)을 최우선 투자전략으로 하는 매매 알고리듬(trading algorithm)을 개발하는 빅데이터 기업이다. Deep Value는 약 2일 정도 소요되는 매매 알고리듬의 테스트 시간을 빅데이터의 분산기술을 활용해 15분으로 단축시켜 매매 알고리듬의 유효성을 제고시킨 것으로 평가받는다(Telx, 2014). 또한 Deep Value는 최선집행을 위해 실시간 시장데이터를 분석해 매매 알고리듬 실행에 적용하는 것으로 알려져 있다. Deep Value의 누적거래 주식은 2015년 4월 기준 2010년부터 590억주이며 누적거래 금액은 총 2.05조달러에 달한다(Telx, 2014). 또한 Deep Value의 거래규모는 2013년 기준 미국 거래소 거래규모의 5.26%에 해당되는 것으로 조사된다.

8) Scaled Risk

2012년 프랑스 파리에 설립된 Scaled Risk는 빅데이터와 인메모리

기술을 기반으로 투자은행과 증권회사가 보유하고 있는 데이터에 대한 검색기능과 분석도구가 탑재된 데이터관리 플랫폼을 제공하는 빅데이터 기업이다. Scaled Risk는 고객이 실시간 거래분석, 리스크관리, 판매관리 및 수익분석, 자금세탁방지, 고객알기정책, 규제보고 등의 업무를 간편화할 수 있는 플랫폼을 제공하는 것으로 조사된다. 또한 Scaled Risk는 거래소에 실시간 및 역사적 데이터 분석을 통한 시장감시 서비스를 제공한다.

<그림 III-14> Scaled Risk의 리스크관리 서비스 사례



자료: Scaled Risk 홈페이지

Scaled Risk의 데이터관리 플랫폼은 금융회사가 보유하고 있는 거래데이터, 시장데이터, 지급결제데이터, 고객관리데이터(Customer Relation Management: CRM), 사내 이메일, 문서, 웹상 문서, SNS 활동을 빅데이터로 구축해 증권회사가 직접 리스크관리에 활용할 수 있도록 다양한 분석툴을 제공하고 있다.

나. 빅데이터 분석기법 사례

자본시장에서 활용할 수 있는 데이터는 매우 다양하기 때문에 데이터 자체를 보유한다는 것보다 자본시장과 관련된 빅데이터를 얼마나 잘 분석하고 이를 통해 가치있는 정보와 지식을 생산할 수 있느냐가 관건일 수 있다. 제II장에서 논의한 바와 같이 빅데이터 분석기법은 매우 다양하다. 그러나 본 절에서 이를 모두 다루는 데 한계가 있으므로 자본시장에서 활용 가능한 핵심적인 분석기법 사례를 중심으로 살펴보기로 한다.

1) 임의분석(ad hoc analysis)

임의분석은 미리 정해진 형식이나 틀 없이 의사결정자의 요구나 필요에 따라 실시하는 분석이다. 임의분석에서 빅데이터가 갖는 의미는 이전과 달리 정형 데이터뿐만 아니라 증권회사 등이 보유하고 있거나 외부에서 구득한 비정형 데이터를 자유롭게 활용할 수 있다는 점이다. 또한 빅데이터는 임의분석에 소요되는 시간도 대폭 단축시킬 수 있다. 예를 들면, 미국의 뱅크오브아메리카(Bank of America: BoA)는 대출계좌 40만건에 대한 신용평가점수를 산출하기 위해 임의분석을 실시했으며, 그 시간을 이전 기술로 걸리는 시간의 3분의 1로 감축시켰다(미래창조과학부 외, 2015). 그만큼 빅데이터 기술은 시간과 비용을 크게 단축시킬 수 있다는 것을 의미한다.

빅데이터를 통한 임의분석은 360도 고객 점검(360-degree customer view) 분야 등에 활용될 수 있다. 360도 고객 점검은 고객알기정책보다 더 확장된 개념이다. 증권회사는 고객이 원하는 금융투자상품 또는 서비스가 무엇인지를 살펴보기 위해 360도 고객 점검을 실시할 수 있다. 이를 위해 증권회사는 고객의 금융자산 및 금융거래 보유현황, 투자실태 및 투자성과, 투자방식, 투자전략, 인터넷활동, 로그기록, 민원 및 전화상담 등에 대한 광범위한 데이터를 통합할 필요가 있다. 빅데이터 기술을 이용하기

이전에는 임의분석을 활용하더라도 고객에 대한 데이터는 파편적으로 분석되었다. 이 때문에 고객 점검은 부분적으로 이루졌으며, 고객의 니즈를 충분히 충족시키는 의사결정을 내리기 못했다는 평가가 일반적이다.

<그림 Ⅲ-15> 임의분석을 통한 360도 고객 점검

임의분석은 빅데이터의 구조를 이해하는 데도 유용할 수 있다. 빅데이터에서 정보가 어떻게 구성되어 있고 그 가운데 어떤 정보를 활용할 수 있는지를 임의분석을 통해 파악할 수 있다. 이를 데이터 발견작업(data exploration or discovery)이라고 말한다. 일반적으로 증권회사가 어떤 의사결정을 내리기 위해서는 이를 뒷받침할 수 있는 데이터가 필요하다. 그런데 대부분의 경우 데이터를 새로 생성하거나 추출하는 데 많은 시간이 소요된다. 뿐만 아니라 데이터 구조를 새롭게 이해하거나 의사결정에 필요한 데이터를 선택하는 데도 시간이 많이 소요될 수 있다. 이 때문에 의사결정의 효율성이 낮아질 수 있다. 빅데이터를 활용한 임의분석은 다양한 각도에서 데이터를 사전적으로 분석하는 데 활용될 수 있기 때문에 이러한 비효율성을 개선하는 데 효과적인 것으로 평가받는다.

2) 자연어처리(natural language processing)

자연어처리는 텍스트를 기계적으로 분석해 데이터로 수집하고 이를 통해 의미있는 정보를 추출하는 일련의 언어처리기법이다. 참고로 자연어처리는 텍스트분석(text analytics)으로 불리기도 한다. 자연어처리는 크게 형태소 분석(morphological analysis), 구문 분석(syntex analysis), 의미분석, 화용 분석의 단계를 거친다.[7] 형태소 분석은 텍스트를 분석하여 최소단위의 의미로 쪼개는 작업이다. 일종의 단어 인식과 같은 작업이다. 형태소란 의미의 최소단위로 더 이상 분석하기 어려운 가장 작은 의미의 문자열을 뜻한다. 구문 분석은 문장이 가지는 구조적 형태를 분석하는 작업이다. 즉 문맥을 이해하는 작업이다. 구문은 형태소가 결합하여 문장이나 구절을 만드는 규칙을 의미한다. 의미 분석은 통사 분석을 토대로 문자열의 의미를 해석하는 작업이다. 이 때 형태소가 가진 의미를 표현하는 기법이 요구된다. 예를 들면, '철수가 가라앉는다'와 같이 통사적으로 옳으나 의미적으로 틀린 문장을 걸러낼 수 있어야 한다. 화용 분석은 사람이 이해할 수 있는 자연어로 재구현하도록 분석하는 작업이다. 예를 들면, '이것은 그것이다'라는 말은 새로운 의미로 전달될 수 없는 자연스럽지 못한 문장구조이다.

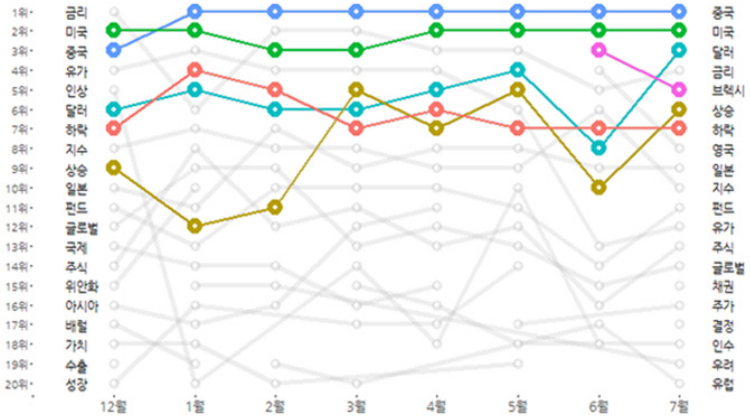
자연어처리중 가장 초기에 나타난 활용사례는 구글 웹페이지 번역서비스와 같은 기계번역 서비스이다. 최근에는 자연어처리가 감성분석, 기계학습, 인공지능 등 빅데이터의 다양한 분야에 활용되고 있는 추세이다. 한편, 국내의 경우 한글에 대한 자연어처리 엔진이 아직 고도화되지 못한 것으로 조사된다. 특히 자연어처리에 가장 기본적인 구성요소인 한글사전이 제대로 구축되지 못해 키워드 추출 과정에서 상당한 오류가 나타나는 것으로 조사된다.

자연어처리는 자본시장에서 투자동향이나 투자정보를 추출하는 데 유용할 수 있다. Dataminr의 트윗 동향 분석도 자연어처리에 기반하고 있다. 미래에셋증권은 2015년부터 신문기사 속에 숨겨진 투자 키워드를 자연어

7) 자세한 내용은 서울대 Biointelligence Lab 홈페이지를 참고하기 바란다.

처리 분석을 통해 고객에게 투자레터 형식으로 제공하고 있다. 예를 들면, 2016년 7월중 미래에셋이 가장 많이 언급된 투자 키워드를 추출한 결과, 중국, 미국, 달러, 금리, 브렉시트, 상승, 하락, 영국 순서로 투자키워드가 제시되었다. 이를 통해 '브렉시트' 키워드가 전월대비 순위 하락하고, '상승' 키워드가 4단계 순위 상승한 것을 발견하였다.

<그림 Ⅲ-16> 미래에셋증권의 투자 키워드 발견 사례



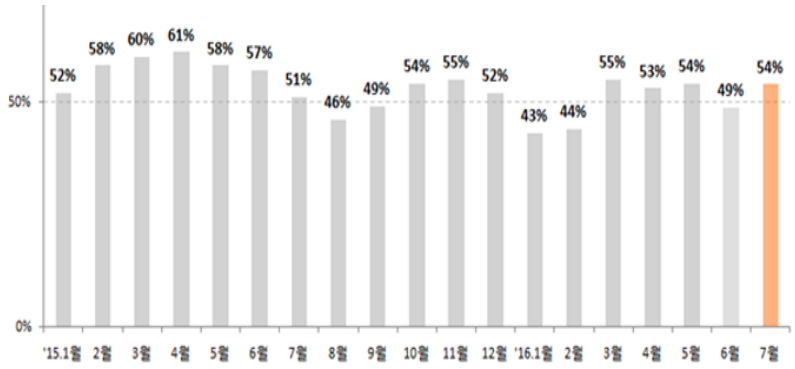
자료: 미래에셋 투자레터

3) 감성분석(sentiment analysis)

감성분석은 자연어처리를 응용한 분석기법이다. 감성분석은 오피니언 마이닝(opinion mining)으로 불리며, 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적 의견을 분석하는 작업이다(신수정, 2014). 예를 들면, 미래에셋증권의 투지심리지표는 감성분석을 토대로 산출한 것이다. 감성분석은 총 세 단계 작업으로 이루어진다. 첫 번째는 분석대상이 될 데이터를 수집하는 작업이다. 두 번째는 감성이 있는 텍스트를 분류하는 작업이다. 감성이 있는 텍스트는 크게 '긍정', '부정', '중립', '객관'으로 분류될 수 있다.

세 번째는 주어진 텍스트에서 감성을 판단하는 탐지 작업이다. 감성분석은 작성자 단위, 문장 단위, 속성 단위로 세분화되어 실시될 수 있다. 동일한 작성자가 긍정과 부정 의견을 모두 제시할 경우 작성자 단위로만 감성분석을 하면 잘못된 결과를 도출할 수 있다. 이 경우에는 문장 단위로 세분화하여 감성분석을 실시해야 한다. 또한 분석대상이 금융상품인 경우 그 상품의 속성에 대해 다른 의견을 제시할 수 있다. 예를 들면, 주가연계증권(Equity-Linked Securities: ELS)은 정상기에 수익률이 좋지만, 위험기에 손실률이 높다. 이 경우 ELS에 대한 투자자의 감성분석은 속성 단위로 분석하는 것이 바람직할 수 있다. 또한 고객의 저축과 투자 성향을 감성분석을 통해 파악하고자 할 경우 고객이 보유한 금융자산의 속성을 원금보장과 원금손실로 분류할 수 있다.

<그림 Ⅲ-17> 미래에셋증권의 투자심리지표 추이



주 : 투자심리값은 투자에 대한 긍정어와 부정어 중 긍정어가 차지하는 비중 값임
자료: 미래에셋 투자레터

StockTwits도 주가 감성분석 서비스를 통해 이용자의 최근 7일간 해당 주식에 대한 트윗 내용을 자연어처리로 분석하여 감성지수를 산출하는 등 주가에 대한 다른 투자자의 심리적 기대를 자신의 투자전략에 참고할 수 있도록 돕고 있다.

<그림 Ⅲ-18> StockTwits의 주가에 대한 감성분석 사례

(1) 애플 주가 (2) 애플 감성지수

자료: StockTwits 홈페이지

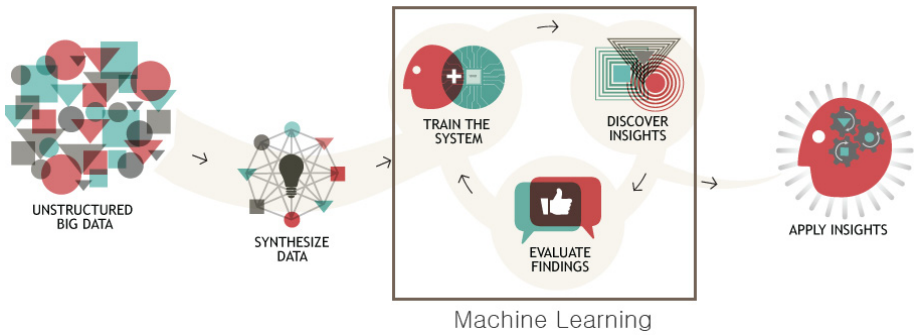
최근 뉴스 및 트윗에 대한 감성분석을 통해 주가를 예측하는 연구가 학계를 중심으로 활발하게 이루어지고 있다. 주가에 영향을 미치는 요인들이 주가에 반영되기 전에 뉴스에 먼저 노출되는 경우가 많기 때문이다. 이 때문에 투자자는 주가에 영향을 미치는 요인을 뉴스로 먼저 접하고 투자결정을 내린다. 즉 뉴스에 대한 투자자의 반응이 주가에 영향을 미칠 수 있다. 예를 들면, 김유신 외(2012)는 시황, 전망, 해외 뉴스에 대한 감성분류를 통해 투자자의 감성이 주가변동을 주가하락시 70.0%, 주가상승시 78.8% 설명하는 것으로 보고한다.

해외에서는 SNS 활동에 대한 감성분석이 다양한 금융투자상품 개발에 활용되고 있다(신경식 외, 2016). 영국의 Derwent Capital Markets은 트윗을 분석해 시장의 투자심리를 파악하고 이를 투자 포트폴리오에 반영하는 '트위터 펀드'를 개발하여 운용중에 있다. 일본 카부닷컴은 트위터와 일본 SNS 믹시에 게시된 반응을 문장단위로 분석하여 주가변동과 관계가 높다고 판단되는 정보를 추출하고 이를 매매거래에 활용하고 있는 것으로 조사된다.

4) 기계학습(machine learning)

기계학습은 기계가 반복적으로 주어진 데이터 또는 직접 구한 데이터를 학습하여 스스로 정보를 추출하고 이를 기반으로 새로운 지식을 축적하는 분석기법이다. 인공지능(Artificial Intelligence: AI)은 가장 고차원적인 기계학습을 수행하는 기계를 일컫는다. 기계학습의 가장 간단한 사례는 이메일을 일정한 규칙에 따라 학습하고 이를 기반으로 스팸메일을 걸러내는 스팸필터이다.

<그림 Ⅲ-19> 기계학습 순서



자료 : 디스코빅스 홈페이지

기계학습은 데이터에서 바로 지식을 추출할 수도 있지만, 대체적으로 하나의 지식을 형성할 수 있는 특징을 추출하는 단계를 거치기도 한다(T-robotics 블로그). 예를 들면, 사진 속에서 사물을 인식하기 위해 픽셀 값에서 먼저 특징적인 선이나 색 분포 등을 추출한 후 이를 기반으로 사진 속의 사물을 '사과' 또는 '고양이'로 지정한다. 이러한 중간 표현단계를 특징 맵핑(feature mapping)이라고 한다. 기계학습은 얼마만큼 정확한 특징들을 추출할 수 있느냐에 따라 그 성능이 크게 좌우된다. 이 경우는 사진뿐만 아니라, 모든 유형의 데이터에도 적용된다.

기계학습 방식은 다양하나, 크게 사람의 지도 여부에 따라 지도학습(supervised learning)과 자율학습(unsupervised learning)으로 구분되며, 이와 별도로 강화학습(reinforcement learning) 방식도 있다.

지도학습은 사람이 각각의 입력(x)에 대해 값(y)을 지정한 훈련 데이터(training data)를 기계에 주면 기계가 그것을 학습하는 것이다. 예를 들면, 기계가 동물사진을 보고 동물을 구분할 수 있도록 학습시킬 때 지도학습 방법은 각 동물사진(x)에 동물이름 값(y)을 지정하여 동물사진을 기계가 학습하게 한다. 구글포토, 페이스북, 애플포토의 얼굴사진별 사진분류 기능도 일종의 지도학습의 사례이다. 이 기능은 사진 앨범에 추출한 대표적인 각 얼굴 사진에 사람이 이름값을 지정하면 기계학습이 자동적으로 사진앨범의 모든 사진에 동일하거나 유사한 얼굴에 각 이름값을 지정하고 각 이름값에 따라 사진을 분류한다. 이러한 지도학습을 분류(classification)라고 구분한다. 자동차 번호판 자동인식도, 애플 시리, 구글 보이스 등 음성인식도 분류 방식의 지도학습 사례이다. 지도학습에는 회귀(regression) 방식도 있다. 분류 방식은 y값이 이산적(discrete)일 때 이용되며, 회귀 방식은 y값이 연속적일 때 사용된다. 이 밖에도 지도학습 방법은 다양하다.

자율학습은 사람의 개입 없이 기계가 스스로 학습하여 y값을 추정하는 학습방식이다. 정답이 주어지지 않기 때문에 학습의 결과가 맞는지를 확인할 수 없지만 x값과 y값이 동시에 있는 데이터를 반복적으로 습득하면 y값을 올바르게 추정할 수 있게 된다. 예를 들면, 자율학습을 하는 기계에게 동물사진을 주고 동물을 스스로 구분하게 하면, 기계는 날개가 있는 동물, 목이 긴 동물, 털이 있는 동물 등 여러 기준을 만들어 동물을 구분할 것이다. 또한 기계는 동물사진을 계속 학습하면서 털 무늬가 단색인 경우, 얼룩 띠가 있는 경우, 얼룩점이 있는 경우 등과 같이 더 구체적인 기준을 만들고 이에 따라 다시 동물을 구분할 수 있다. 그리고 각 세부 기준을 한 동물을 정의할 수 있는 기준으로 취합하여 각 동물을 구분하는 데 활용할 수 있다. 이 경우 실제 동물이름 값을 알지 못하기 때문에 임의값을 지정한다. 마지막으로 사람이 동물이름을 직접 지정하지 않더라도 기계가

웹문서 등에서 자신의 동물 분류와 일치하는 사진에서 실제 동물이름 값을 반복적으로 발견하면 자신이 임의적으로 지정한 동물이름 값을 실제 동물이름 값으로 대체한다. 이러한 자율학습 방식을 군집화(clustering)라고 한다. 이 밖에도 자율학습 방법은 다양하다.

강화학습은 현재의 상태(state)에서 어떤 행동(action)을 취하는 것이 최적인지를 도출하는 학습방법이다. 기계가 반복게임(repeated game)에서 자신의 누적 이득(cumulative payoff)을 극대화하는 전략을 매 게임마다 선택하는 것과 같다. 구글의 알파고(AlphaGo)도 일종의 강화학습 알고리듬에 따라 바둑을 두는 인공지능이다. 이보다 한층 더 복잡한 기계학습 방식으로는 심화학습(deep learning)이 있다. 국내에서도 흔히 딥러닝으로 불린다.

심화학습은 인공신경망(artificial neural network) 기술을 자율학습에 접목시킨 기계학습 방식이다(T-robotics 블로그). 인공신경망 기술은 데이터를 잘 구분할 수 있도록 일종의 선을 긋고 왜곡하고 합하는 작업을 반복하여 최적화된 구분선을 찾는 학습방식이다. 자율학습보다 더 정밀한 방식이라고 할 수 있다. 그런데 인공신경망은 1940년에 개발되어 1980년대까지 각광을 받았지만 이후 여러 한계로 관심을 잃었던 기술이나, 자율학습 방식의 발전으로 최적화의 초기값 문제를 해결하면서 다시 주목받기 시작했다. 빅데이터 기술의 발전과 다양한 형식의 데이터 활용이 가능해지면서 기존에 있었던 최적화 오류도 최소화할 수 있게 되었다. 이 때문에 최근 기계학습은 모두 딥러닝으로 통할 정도로 큰 관심을 받고 있다.

기계학습 활용분야는 매우 다양하다. 예를 들면, 로보어드바이저도 기계학습을 활용하여 고객의 투자자산에 대한 리밸런싱 전략을 제시할 수 있다. 기계학습의 가장 큰 장점은 짧은 시간 안에 방대한 데이터를 학습할 수 있고 이에 대한 기억을 상실하지 않는다는 점이다. 그렇기 때문에 더 많은 규모의 데이터를 학습할수록 더 최적화된 의사결정을 도출할 수 있다. 이는 사람의 데이터 수집, 학습 및 기억 능력의 한계를 고려할 때 획기적인 혁신이라고 할 수 있다. 더구나 사람은 노화하기 때문에 세대에 걸쳐 재학습이 필요하나 기계는 재학습할 필요가 없다. 또한 기계학습은

불공정거래 또는 주가조작 행위를 감지하거나, 금융시장의 이상거래 또는 대량 오류주문을 감시하는 데 활용되는 추세이다.

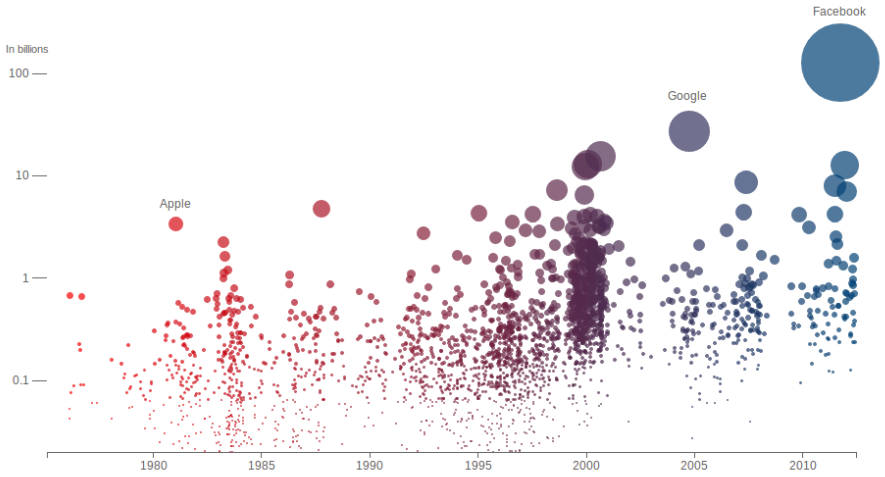
패턴인식은 기계학습의 응용기법으로 데이터에서 일정한 패턴이나 정규성을 발견하고 정의하는 분석기법이다. 예를 들면, 패턴인식은 문자인식, 필체인식, 음성인식, 지문인식, 홍채인식, 얼굴인식, 보행 패턴 분석, 화법 및 발화 습관 분석, 오작동진단, 심전도 신호 분석, X-ray 판독, 지진 패턴, 날씨 예측, 주가 예측, 사기방지, 부정거래 발견 등 다양한 분야에서 활용될 수 있다. 즉 패턴인식은 기계학습의 한 단계라고 볼 수 있다. 예를 들면, 최근 상장기업 CEO의 방송인터뷰 등에서 나타난 얼굴표정을 인식해 향후 해당 기업의 주가를 예측하는 시도가 이루어지고 있는 것으로 알려져 있다(Egan, 2014).

5) 데이터 시각화(data visualization)

데이터 시각화는 의사결정의 효율성과 직결된다. 이 때문에 빅데이터 분석에서 시각화는 매우 중요하게 다루어지고 있다. 빅데이터 기술은 비정형 데이터도 수집, 저장, 처리, 분석할 수 있기 때문에 비정형 데이터에서 추출한 정보를 직관적으로 알아보기 쉽게 보여줄 수 있어야 한다. 빅데이터 정보를 얼마나 빠른 시간에 분석하여 얼마나 효율적으로 전달할 수 있는가에 따라 경영진의 의사결정 시간도 단축될 수 있기 때문이다.

효과적인 정보전달이 가능한 시각화 기법은 빅데이터 분석이 빠른 시간 내에 의사결정에 필요한 정보를 추출할 수 있다는 가능성뿐만 아니라 기업의 빅데이터 활용 유인을 높일 수 있다. 예를 들면, 뉴욕타임즈는 2010년 페이스북의 기업공개에 맞춰 과거 신문기사를 텍스트 마이닝(text mining) 기법으로 분석해 페이스북과 과거 IT기업의 기업공개 가치를 비교하는 그래프를 기사에 게재한 바 있다. 이를 통해 페이스북의 기업공개 가치가 역사적으로 어떤 의미를 가지는지를 독자 및 자본시장 참여자에게 시의적절하게 효과적으로 전달할 수 있었다.

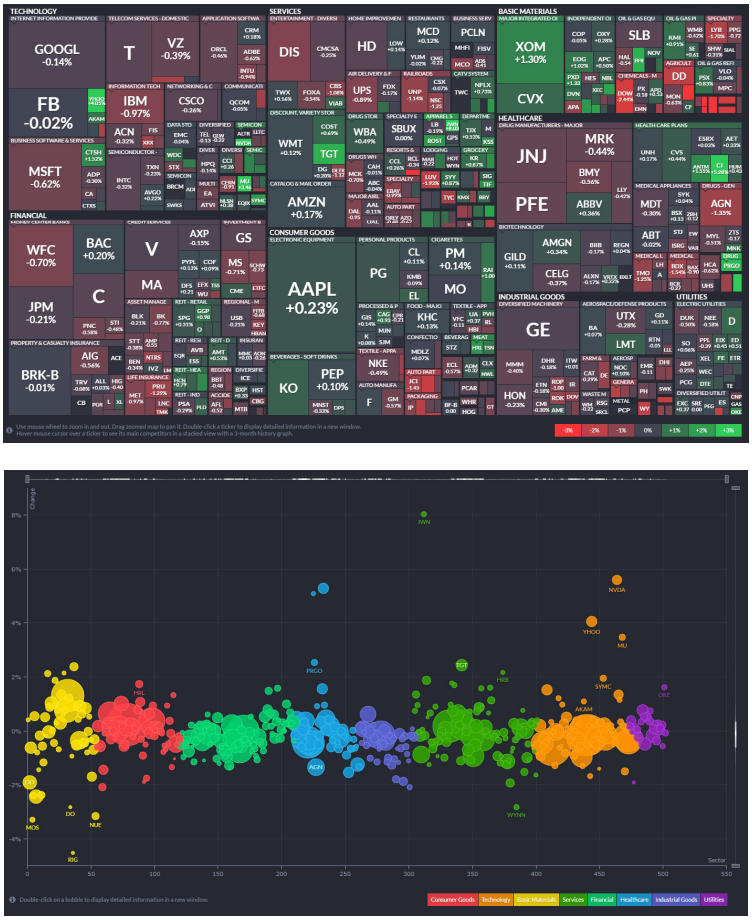
<그림 Ⅲ-20> 뉴욕타임즈 신문게재 기업공개 기사 시각화 사례



자료: New York Times

자본시장에서 빅데이터를 활용한 시각화 기법은 매우 중요한 정보전달의 수단으로 활용될 수 있다. 예를 들면, 미국의 FinViz는 S&P500, 미국 전체, 전 세계 주식, 미국 ETF 시황을 히트맵 또는 버블맵으로 볼 수 있는 시각화 서비스를 제공한다. FinViz는 고객의 수요에 따라 각종 옵션을 선택할 수 있도록 하고 이에 따라 시각화 방법을 선택할 수 있도록 하였다. 이 점에서 빅데이터에 기반한 시각화 기법은 단순히 데이터 분석의 결과물을 보여주는 것뿐만 아니라 이용자와의 상호작용을 통해 이용자의 선택에 따라 필요한 정보를 재추출할 수 있다는 장점이 있다는 것을 보여준다. 또한 각종 시각화 기법에 기반하여 고객의 포트폴리오를 관리하는 서비스를 제공하고 있다. 2016년 6월말 현재 BNP Paribas, Barclays Capital, BNY Mellon, Credit Suisse, Deutsche Bank, Goldman Sachs, Morgan Stanley, MacQuarie, Nomura, RBS, UBS, WellsFargo 등 주요 기관투자자들이 FinViz 유료서비스에 가입하여 사용하고 있다.

<그림 Ⅲ-21> FinViz의 미국 주식시장 섹터별 시각화

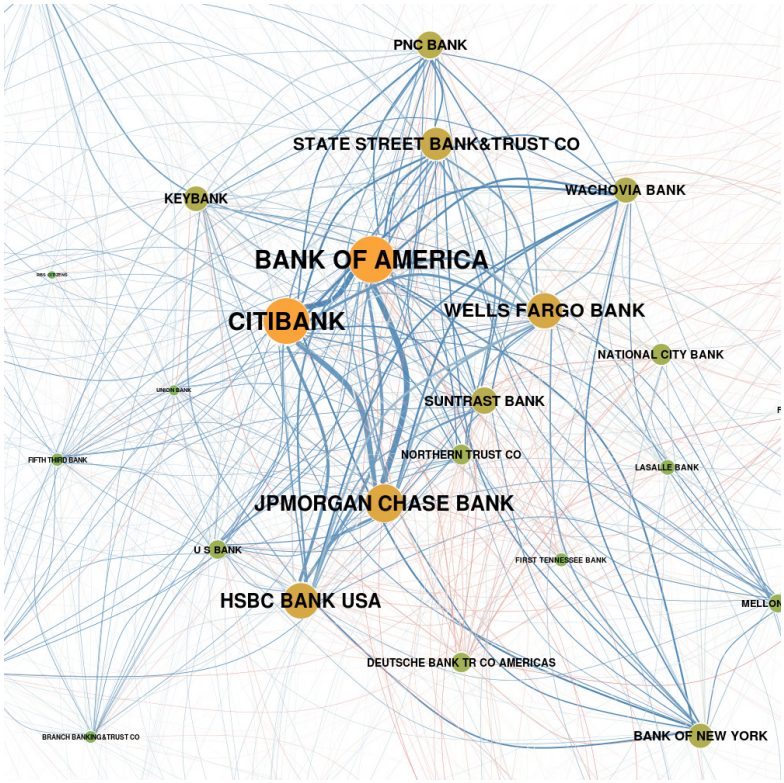


자료: FinViz 홈페이지

앞서 논의한 바와 같이 빅데이터 분석은 인과관계보다 상관관계 분석에 초점을 맞춘다. 이 때문에 빅데이터 분석을 통해 다수의 변수간 상관관계를 얼마나 효과적으로 전달할 수 있느냐는 매우 중요하게 인식되고 있다. 다수의 변수간 상관관계 또는 상호연계성을 나타낼 수 있는 시각화 기법은 자본시장에서도 다양하게 활용될 수 있다. 예를 들면, 고객의 특성별

주식투자 종목의 연계성과 매수도 관계를 일별, 월별, 연도별로 구분하여 시각화할 수 있다. 또한 투자자의 감성분석에도 활용될 수 있다. 금융당국은 금융회사의 상호 익스포져를 이와 같이 시각화하여 시스템리스크를 유발시킬 수 있는 상호연계성 위험(interconnectivity risk)을 측정하는 데도 활용할 수 있다. 예를 들면, Nanumyan et al.(2015)은 미국와 독일 은행의 신용부도스왑(Credit Default Swap: CDS) 포지션의 상호연계성을 빅데이터 시각화 기법을 활용해 시각화하였다.

<그림 Ⅲ-22> 신용부도스왑 거래관계 시각화 사례



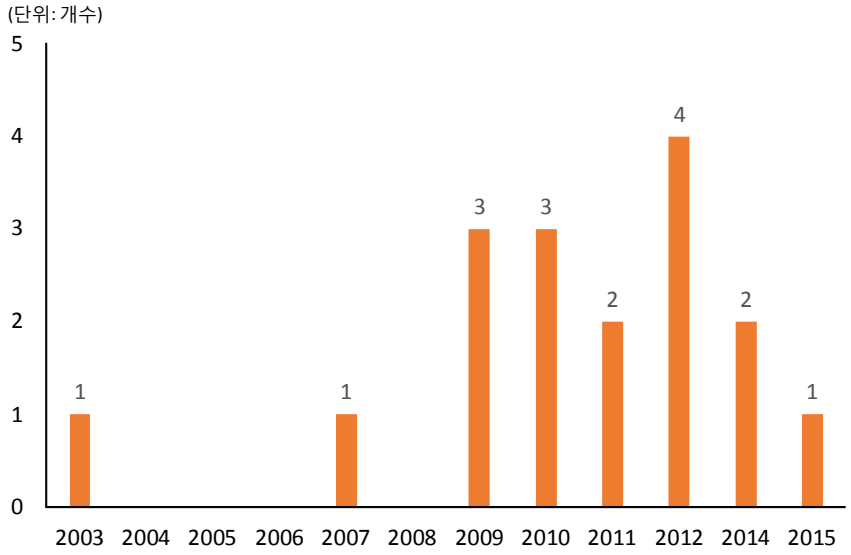
자료: Nanumyan et al. (2015)

3. 해외 자본시장의 빅데이터 도입 특징

1) 자본시장에 특화된 빅데이터 전문기업 출현

증권회사 등 금융투자회사가 직접 빅데이터 시스템을 구축하거나 다양한 출처의 데이터를 계속적으로 수집하고 관리할 경우 시스템 구축 및 전문인력 채용을 위해 상당한 수준의 비용을 부담해야 한다. 이 때문에 금융투자회사는 독립적인 빅데이터 시스템을 구축하는 것을 주저할 수 있다. 이 틈새 속에서 빅데이터 전문기업의 출현은 해외 자본시장에서 빅데이터 활용을 촉진시키는 계기로 작용하는 것으로 조사된다.

<그림 Ⅲ-23> 자본시장 관련 빅데이터 기업의 설립연도별 현황

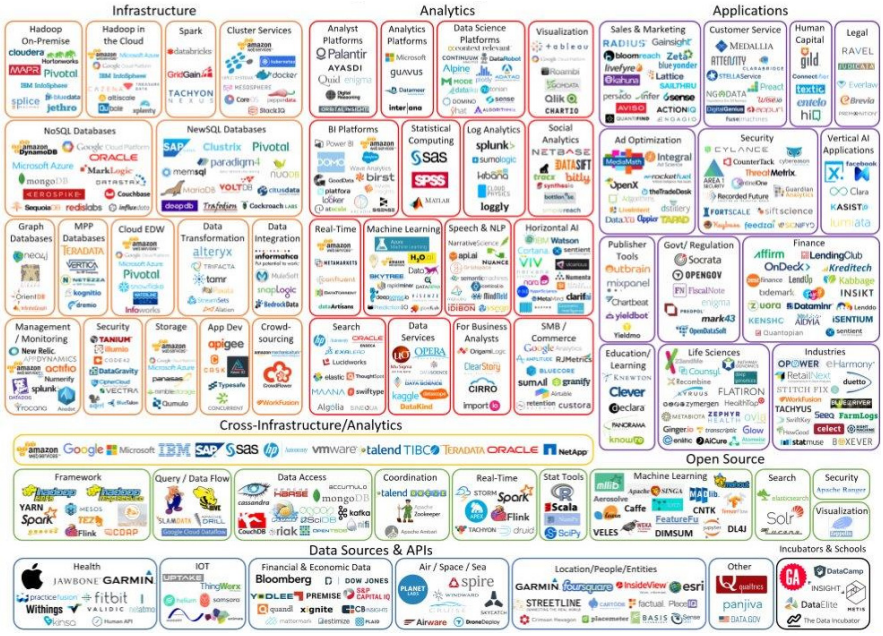


자료: CBInsights 및 CrunchBase 사이트

Eyrdaud(2016)에 나열될 자본시장 관련 빅데이터 기업 17곳의 설립연도 현황을 살펴보면, 자본시장 관련 빅데이터 기업은 2009년을 기점으로 본격적으로 설립된 것으로 조사된다. 이들 빅데이터 기업은 서비스 유형에 따라 크게 빅데이터 유통 플랫폼(Xignite, Quandl 등), 조사분석 검색엔진(Kensho, AlphaSense 등), 투자동향 분석(Dataminr, StockTwits 등), 투자전략 분석(Deep Value 등), 리스크관리 및 법규준수(Scaled Risk) 등으로 구분될 수 있다. 참고로 Eyrdaud(2016)에 나열된 자본시장 관련 빅데이터 기업 리스트가 자본시장에 특화된 빅데이터 기업 전체를 대표하지 않는다는 점을 주지할 필요가 있다.

참고로 해외에서는 <그림 Ⅲ-24>와 같은 다양한 유형의 서비스를 제공하는 빅데이터 기업이 상당수 출현했다는 점은 시사하는 바가 크다. 그만큼 빅데이터에 대한 수요가 높다는 것을 방증하기도 한다.

<그림 Ⅲ-24> 2016년 기준 빅데이터 기업 현황



자료: CBInsights 사이트

2) 빅데이터 유통플랫폼 활용

자본시장에서 활용 가능한 빅데이터는 매우 다양하고 폭넓다. 거시경제와 각 산업에 영향을 미치는 사회, 경제, 정치, 문화 변수뿐만 아니라 기후변화와 같은 현상들도 자본시장에 영향을 미친다. 또한 자본시장은 자체적으로 실시간으로 대규모의 시장데이터, 상장기업에 대한 공시문서, 실시간 국내외 뉴스 및 풍문·소문 등도 생산한다. 이 때문에 금융투자회사가 직접 빅데이터를 구축하는 것은 비효율적일 수 있다. 빅데이터 유통플랫폼은 이와 같은 문제를 해소시킬 수 있다.

빅데이터 유통플랫폼은 데이터 수집 및 관리 측면에서 비용 효율적일 수 있다는 측면에서 경제적인 의미를 갖는다. 모든 금융투자회사가 개별적으로 동일한 데이터셋을 수집하고 관리하는 것보다 하나의 빅데이터 유통플랫폼이 이를 관리하고 공동으로 사용하는 것이 더 경제적일 수 있기 때문이다. 또한 다양한 범위의 데이터를 하나의 빅데이터 유통플랫폼이 수집하고 관리할 경우 금융투자회사는 수요에 따라 활용할 데이터셋을 선택할 수 있기 때문에 비용도 절약될 수 있다. 앞서 논의한 Xignite와 Quandl이 데이터 제공업체에서 빅데이터 유통플랫폼으로 발전한 대표적인 사례이다.

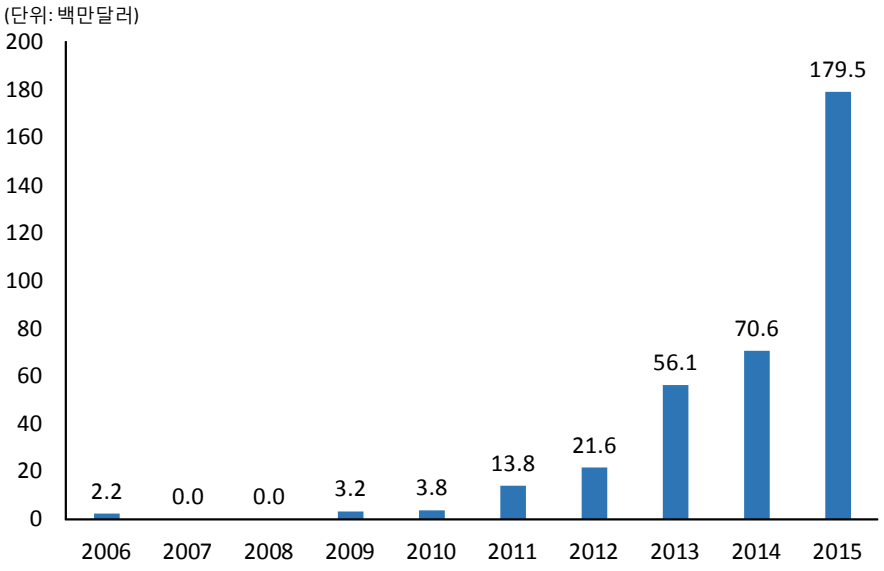
또한 빅데이터 유통플랫폼은 개별 금융투자회사의 해외 자본시장과 관련된 데이터에 대한 접근성을 개선시킬 수 있다. 자본시장의 국제화로 해외 투자활동이 자유로울수록 자본시장 관련 데이터도 국제적으로 유통되고 활용될 것으로 예상되기 때문이다. 예를 들면, Xignite는 40개 이상 국가의 증권거래소 시장데이터를 제공하고 있다.

빅데이터 유통플랫폼은 단순히 빅데이터를 유통시키는 역할에 그치지 않고, 이를 용이하게 분석할 수 있는 분석플랫폼도 함께 제공하고 있다. 이 때문에 금융투자회사는 빅데이터 유통플랫폼을 이용함으로써 빅데이터 시스템 도입과 관련된 비용을 절감시키고 빅데이터 분석기법과 관련된 기술적인 한계도 극복할 수 있다.

3) 빅데이터에 대한 투자 증가

자본시장에서 빅데이터 기술은 자본시장 참여자의 수익창출 기회를 확대시키고, 리스크관리 및 법규준수의 효율성을 증대시키며, 비용 절감의 효과를 가져올 것으로 기대되고 있다. 이 때문에 자본시장의 빅데이터에 대한 투자도 빠른 속도로 증가하고 있는 것으로 조사된다.

<그림 Ⅲ-25> 자본시장 관련 빅데이터 기업에 대한 투자규모



자료: CBInsights 및 CrunchBase 사이트

Eyrdaud(2016)에 나열될 자본시장 관련 빅데이터 기업 17곳의 자본투자 유치실적을 CBInsights 및 CrunchBase 사이트에서 조사하여 취합한 결과, 2015말 기준 약 1억 800만달러의 자금이 자본시장관련 빅데이터 기업에 투자된 것으로 조사된다. 이는 전년말 대비 약 2.5배 증가한 규모로 이러한 추세는 계속 유지될 것으로 전망된다. 다만 향후 2~3년간 빅데이터에 대한 전반적인 투자가 소폭 감소될 것으로 예상됨에 따라

그 증가폭은 다소 완만해질 것으로 보인다(Gartner, 2016). 다만 증권회사 등 자본시장 참여자가 자체적인 빅데이터 시스템 구축 등을 위해 투자한 실적은 포함되지 않았기 때문에, 이를 고려할 경우 자본시장의 빅데이터 투자규모는 상기 조사결과보다 더 높을 수 있다.

4) 자연어처리 및 기계학습 분석기법 중요

자본시장에서 가장 중요하게 활용되는 빅데이터 분석기법은 자연어처리와 기계학습인 것으로 조사된다. 그 이유 중 하나는 자본시장 관련 데이터 대부분이 복잡한 텍스트 데이터이거나 대규모 시장데이터로 구성되어 있기 때문이다. 예를 들면, Kensho와 AlphaSense의 검색엔진은 자연어처리에 기반하고 있으며, 기계학습을 통해 검색자의 요구에 따라 시장 또는 기업에 대한 분석보고서를 작성해주는 서비스를 제공하고 있다. Dataminr와 StockTwits도 자연어처리에 기반한 감성분석을 통해 투자동향 및 투자심리를 분석하는 서비스를 제공하고 있다. Deep Value와 Scaled Risk는 대규모 시장데이터를 기계학습을 통해 실시간으로 학습하고 매매 알고리듬의 유효성 검증, 실시간 거래분석, 불공정거래 탐지 등의 반복적인 업무를 자동적으로 실시하고 있다.

자연어처리와 기계학습은 상호 보완적인 빅데이터 분석기법이며, 인공지능의 기초가 되는 기술이라는 점에서도 매우 중요하다. 자연어처리 능력에 따라 기계학습의 역량의 기준이 되는 알고리듬의 정확성이 결정될 수 있기 때문이다. 예를 들면, Kensho의 검색엔진 알고리듬도 자연어처리와 기계학습의 정교한 결합의 산출물로 평가받는다. Goldman Sachs, J.P. Morgan, Bank of America Merril Lynch 등이 Kensho의 전략적 투자자로 나선 이유도 이 때문이다. 참고로 국내 자연어처리와 기계학습의 기술수준은 해외보다 2~5년 정도 뒤쳐져 있는 것으로 조사된다(석왕헌·이광희, 2015).

Ⅳ. 시사점

Ⅳ. 시사점

본 보고서는 빅데이터에 대한 기본적인 이해를 위해 빅데이터의 개념, 분석기법, 정책적 논의, 파급효과 및 한계를 살펴보고, 해외 자본시장의 빅데이터 도입 현황을 빅데이터 도입 효과, 빅데이터 기업 및 분석기법 사례, 빅데이터 도입 특징으로 나누어 살펴보았다.

빅데이터는 2009년 전후로 데이터 분석의 새로운 기술로 등장하였으며, 하나의 현상에 그치는 데 머물지 않고 사회 및 경제 전반에 영향을 미칠 뿐만 아니라 원유와 같이 새로운 경제성장의 동력이 될 것이라는 평가를 받고 있다. 이러한 배경 가운데 기업은 이미 빅데이터를 생산성과 효율성 증진에 활용하고 있는 추세이며, 각 국의 정부에서는 경제성장 전략의 하나로 공공데이터 개방이 확대되고 민간데이터 유통이 촉진될 수 있도록 빅데이터 환경을 구축하고 있다. 그러나 빅데이터에 대한 긍정적인 시각만 존재하는 것은 아니다. 빅데이터 기술이 기존 데이터 분석기술과 달리 정형, 반정형, 비정형의 다양한 형식을 가진 대규모 데이터를 빠른 시간에 처리하고 분석할 수 있는 장점이 있으나, 데이터 분석의 목적을 달성할 수 있다고 반드시 보장할 수 없기 때문이다.

빅데이터 분석의 궁극적인 목적은 경제주체간 정보의 비대칭성을 해소하고 경제주체의 의사결정의 효율성을 제고하는 데 있다. 특히 정보의 비대칭성이 항상 존재하는 금융산업 및 자본시장에서는 빅데이터 분석이 새로운 수익창출 기회를 확장시키고 비용을 절감시키는 데 유용할 수 있다.[8] 그러나 대규모 데이터를 보유하고 이를 분석할 수 있는 시스템을 갖추고 있더라도 이를 잘 활용할 수 있는 기술과 역량을 갖추지 못하면 빅데이터의 경제적 가치창출 효과는 실현되기 어려울 수 있다. 또한 빅데이터 분석이 기존 데이터 분석과 달리 인과관계보다 상관관계에 중점을 두고 있기 때문에

8) 정보의 비대칭성은 경제주체간 빅데이터 활용 역량에 따라 더욱 확대될 가능성도 배제할 수 없다. 그러나 빅데이터 기술이 이전보다 정보의 효용성을 높인다는 점에서 또한 데이터의 유통이 이전보다 활성화될 수 있다는 점에서 정보의 비대칭성은 빅데이터 기술로 점차적으로 또는 부분적으로 해소된다고 볼 수 있다.

잘못된 데이터를 선택하거나 엄격한 분석기법을 따르지 않을 경우 잘못된 의사결정으로 이어질 수 있다. 따라서 빅데이터에 대한 막연한 기대보다는 빅데이터 활용목적을 명확하게 수립하고 그 목적 하에 적정한 빅데이터 분석기법을 선택하고 이용할 수 있는 역량을 키우는 것이 지금 시점에서는 매우 중요할 수 있다.

자본시장은 사회, 경제, 정치, 문화 변수뿐만 아니라 기후변화와 같은 현상에도 영향을 받기 때문에 정형뿐만 아니라 비정형 데이터에 대한 분석 수요가 이전부터 높아왔다. 또한 자본시장은 불특정 다수의 시장참여자가 동시에 실시간으로 다양한 형식의 데이터를 생산하고 축적하기 때문에 시장과 데이터 구조가 매우 복잡하다. 그만큼 자본시장에서 빅데이터 기술을 적절하게 활용하기 위해서는 복잡한 시장과 데이터 구조를 먼저 이해할 필요가 있다. 따라서 자본시장에서 빅데이터를 효과적으로 활용하는 것은 매우 난해한 작업이 될 수 있다. 해외에서나 국내에서도 자본시장에서 빅데이터에 대한 논의가 은행 및 보험권역에 비해 상대적으로 활발하지 못한 것도 이 때문일 수 있다.

빅데이터가 향후 사회경제 전반에 미치는 파급효과가 클 것이라는 긍정적인 평가에도 불구하고, 단기적으로는 빅데이터 기술의 도입에 따른 편익은 비용보다 크지 않을 수 있다. 그러나 해외 자본시장의 빅데이터 도입 현황을 살펴보면 자본시장에서 빅데이터의 잠재적 유용성은 무시할 수 없다고 판단된다. 따라서 국내 자본시장 유관기관 및 금융투자업계는 해외 자본시장의 빅데이터 도입 추세를 지속적으로 관찰하고 국내에 적용하는 노력을 지속할 필요가 있다.

본 보고서는 빅데이터에 대한 기본적인 이해와 해외 자본시장의 빅데이터 도입 현황에 대한 조사결과를 토대로 다음과 같이 국내 자본시장에 주는 시사점을 제시하고자 한다.

첫째, 금융투자업계는 단기적인 관점에서 빅데이터 도입에 따른 편익이 비용에 비해 크지 않더라도 중장기적인 관점에서 빅데이터 전략을 수립하고 향후 데이터 환경 변화에 적극 대응할 수 있는 역량을 갖출

필요가 있다. 이를 위해서는 무엇보다도 최고경영진의 의지가 중요하다. 또한 금융투자회사 각자가 보유하고 있으나 활용하지 못하는 데이터를 자체적으로 발굴하고 이를 활용하는 노력도 지속할 필요가 있다. 예를 들면, 투자자 니즈에 맞는 새로운 금융투자상품 및 서비스를 개발하기 위해 시장데이터, 원장데이터, 로그데이터 등을 결합하여 빅데이터로 구축하고 이를 빅데이터 분석에 적극 활용할 수 있다.

둘째, 자본시장 유관기관 및 금융투자업계는 자본시장에 특화된 빅데이터 전문기업이 출현할 수 있도록 지원할 필요가 있다. 해외 자본시장의 경우에도 자본시장에 특화된 빅데이터 전문기업의 출현으로 자본시장에서 빅데이터 활용이 촉진되었기 때문이다. 또한 금융투자업계는 빅데이터 관련 업체들이 자본시장에 빅데이터 기술을 전파할 수 있도록 이들 업체들과 긴밀한 협업체계를 구축해 나가는 노력도 함께 해야 한다.

셋째, 자본시장 유관기관 및 금융투자업계는 자본시장 참여자가 이용할 수 있는 빅데이터 유통플랫폼을 공동으로 구축하는 노력을 지속할 필요가 있다. 국내 자본시장의 환경을 고려할 경우 빅데이터 유통플랫폼이 민간에서 자생적으로 출현하는 것을 기대하기 어렵기 때문이다.

넷째, 자본시장 유관기관 및 금융투자업계는 빅데이터와 관련된 법제도적 환경을 제대로 이해하고, 공공데이터 활용, 데이터 공유, 개인정보 보호, 비식별화 기술, 빅데이터 윤리 등에 대해서도 꾸준한 관심을 갖고 빅데이터 환경 변화에 적극적으로 대응할 수 있어야 한다.

끝으로 빅데이터에 대한 다양한 시각과 평가가 존재하지만 자본시장에서 빅데이터의 중요성과 유용성은 무시될 수 없다고 판단된다. 특히 최근 인공지능에 대한 관심이 높아지는 추세에서 빅데이터 역할은 매우 광범위해지고 있다. 빅데이터가 인공지능의 가장 기초적인 요소이기 때문이다. 또한 빅데이터는 자본시장의 경쟁력을 가늠하는 하나의 잣대가 되고 있다. 따라서 국내 자본시장도 국제적인 경쟁력을 향상시키고 유지하기 위해 빅데이터 기술과 역량 개발에 지속적으로 노력할 필요가 있다.

참고문헌

참 고 문 헌

고학수·최경진, 2015,『개인정보의 비식별화 처리가 개인정보 보호에 미치는 영향에 관한 연구』, 개인정보보호위원회 정책연구보고서.

공공데이터전략위원회, 2014, 오픈데이터 5대 강국 도약을 위한 공공데이터 개방 발전전략.

김유신·김남규·정승렬, 2012, 뉴스와 주가: 빅데이터 감성분석을 통한 지능형 투자의사결정모형,『지능정보연구』제 18권 제2호, 143-156.

권선애·조가원, 2014,『빅데이터와 보안』, 한국 IBM big Data 2014.

미래창조과학부 외, 2015,『2015년 빅데이터 글로벌 사례집』.

석왕헌·이광희, 2015,『인공지능 기술과 산업의 가능성』, ETRI Creative Opinion Issue Report 2015-04.

신경식·박정현·김성현, 2016,『금융 산업 빅데이터 도입 방안』, 빅데이터 기획보고서 7호.

신수정, 2014,『글에서 감성을 읽다! 감성 분석의 이해』, IT World IDG Tech Report.

양지현, 2011, Advanced Analytics 경쟁 우위 확보를 위한 차세대 핵심 과제, LG CNS,『Technology Inside』제 12호.

엄태웅, 2015,『쉽게 풀어쓴 딥러닝(Deep Learning)의 거의 모든 것』.

이명진·김우주, 2012, 빅데이터를 위한 고급분석 기법과 지원 기술,『Entrue Journal of Information Technology』 11(1), 45-56.

이성복, 2016, 글로벌 자본시장의 핀테크 혁신 동향 및 시사점,『자본시장 Weekly Opinion』, 2016-03.

인하대학교 법학연구소, 2014,『개인정보의 범위에 관한 연구』, 개인정보보호위원회 최종보고서.

장동인, 2015, 디지털 기회선점을 위해 관점을 바꿔라, CIO Summit 2015 Discussion Paper.

투이컨설팅, 2013,『Big Data Analytic과 기업경쟁력』.

IDG Korea, 2014, 한국의 빅데이터, 어디까지 왔나 - IDG Market Pulse.

Aite, 2014, *Big data in capital markets: At the start of the journey*.

Allen & Overy, 2016, *The EU general data protection regulation*.

A-Team Group, 2012, *Big data solutions in capital markets - a reality check*.

Avantage Reply, 2014, *Applying big data to risk management*.

BSA, 2015, *What's the big deal with data?*

Bulter, 2013, When google got flu wrong? Nature (2013.2.13.).

Capgemini, 2012, The deciding factor: Big data & decision making, Capgemini Business Analytics Report.

Carpenter, J., 2016, Trump win shows limits of big data, power of emotional intelligence, Forbes (2016.11.10.).

Chemitiganti, V., 2016, *Capital markets pivots to big data in 2016*.

Chen, R., Lazer, M., 2013, *Sentiment analysis of twitter feeds for the prediction of stock market movement*.

Chung, S., Liu, S., 2011, *Predicting stock market fluctuations from twitter an analysis of the predictive powers of real-time social media*.

Crovitz, L.G., 2016, Trump's big data gamble, The Wall Street Journal (2016.7.14.).

CSC, 2014, *CSC global CIO survey: 2014-2015*.

Davidson, J., 2015, Manipulation of fed's personal data is a major danger in OPM cyber-heist, Washington Post (2015.8.18.).

Davis, K., 2012, Ethics of big data, O'reilly.

Delima, A., 2015, The new digital era enterprise, Likedin (2015.8.31.).

Delottie, 2013, *Market assessment of public sector information*.

Egan, M., 2014, Looking for a stock tip? Check the CEO's face, CNN Money (2014.11.18.).

Enderle, R., 2016, How Trump defeated Clinton using analytics, CIO Opinion (2016.11.11.).

EPRS, 2015, *Industry 4.0 digitalisation for productivity and growth*.

European Commission, 2011, *Open data - An engine for innovation, growth and transparent governance*.

European Commission, 2013, *EU implementation of the G8 Open Data Charter*.

European Commission, 2015, *Creating value through open data: Study on the impact of re-use of public data resources*.

Eyraud, S., 2016, *Big data and capital markets*.

Finklea, K., 2014, *Identity theft: Trends and issues, congressional research service*.

FTC, 2012, *Protecting consumer privacy in an era of rapid change*.

FTC, 2015, *Protecting personal information: A guide for business*.

FTC, 2016, *Big data: A tool for inclusion or exclusion? Understanding the issues.*

Garfinkel, S.L., 2015, De-identification of personal information, NISTIR #8053.

Heudecker, N., Hare, J., 2016, *Survey analysis: Big data investments begin tapering in 2016*.

Hillebrand, J., 2015, How much big data is generated every minute on top digital and social media? Thinktostart (2015.8.25.).

ICO, 2010, *A complete guide to notification, data protection*, Notification Handbook.

ICO, 2011, *Data sharing code of practice*.

ICO, 2012, *Determining what is personal data*.

ICO, 2016a, *The guide to data protection*.

ICO, 2016b, *Direct marketing, data protection act-privacy and electronic communications regulations*.

IDC, 2014, *Capturing the $1.6 trillion data dividend*.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014, The parable of google flu: Traps in big data analysis, Science Policy Forum, 343.

Lindzon, H., 2015, How to read the new and social human ticker, StockTwits.

MacLean, D., 2011, A very brief introduction to mapreduce, Stanford HCI Group.

Madrid, 2015, *Kensho - King of financial data analytics*.

Mcargus, G., Davis, E., 2014, Eight (no, nine!) problems with big data, The New York Times (2014.4.6.).

McKinsey Global Institute, 2011, *Big data: The next frontier for innovation, competition, and productivity*.

McKinsey Global Institute, 2013, *Open data: Unlocking innovation and performance with liquid information*.

Mecham, S., 2016, *What do big data and robo advisors have in common?*.

Metha, N., 2012, Knight $440 million loss sealed by rules on canceling trading, Bloomberg (2012.8.14.).

Mezzofiore, G., 2016, How a little-known data firm helped Trump become president, Marshable (2016.11.10.).

Mittal, A.M., Goel A., 2011, *Stock prediction using twitter sentiment analysis*.

Nanumyan, V., Garas, A., Schweitzer, F., 2015, *The network of counterparty risk: Analysing correlations in OTC derivatives*.

Nelson, G.S., 2015, Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification, ThotWave Technologies, LLC.

OECD, 2015a, *Data-driven innovation big data for growth and well-being*.

OECD, 2015b, The proliferation of “big data” and implications for official statistics and statistical agencies, OECD Digital Economy Papers (245).

Oracle, 2012, *Big data analytics with oracle advanced analytics in-database option*.

Pavlyshenko, B., 2013, *Tweets mining using NLP can help in goods marketing*.

PCAST, 2014, *Big data and privacy: A technological perspective, office of science and technology policy*.

Pettey, C., 2011, Gartner says worldwide enterprise it spending to reach $2.7 trillion in 2012, Garter (2011.10.17.).

Piatetsky-Shapiro, G., 2012, Big data hype (and reality), *Harvard Business Review*.

Pilar, J., 2015, *The european commission's open data policy*.

PwC, 2013, *Where have you been all my life? How the financial services industry can unlock the value in big data*.

Raden, N., 2010, Get analytics right from the start, Hired Brains Research.

Rhind, D., 2014, What is the value of open data? The Advisory Panel on Public Sector Information (APPSI) Seminar, UK Open Government.

Scola, N., 2013, Obama, the 'big data' president, Washington Post (2013.6.14.).

Shakespeare, S., 2013, An independent review of public sector information, Delottie for The Department for Business Innovation and Skills Report.

Singh, M., 2014, Big data in capital markets, International Journal of Computer Applications 107(5), 42-45.

SoftServe, 2014, The 2014 software development trends, Survey Results.

Telx, 2014, Deep value: Leverages telx to deliver advanced trading algorithms, Telx Case Study.

U.S. Executive Office of the President, 2010, *Sharing data while protecting privacy*.

U.S. Executive Office of the President, 2013, *Open data policy* -

Managing information as an asset.

U.S. Government, 2014, *U.S open data action plan*.

U.S. Government, 2015, *U.S The open government partnership/ third open government national action plan for the united states of America*.

UK Cabinet Office, 2013, Open data charter, Policy Paper.

UK Government, 2012, Open data white paper unleashing the potential, UK for The Stationery Office Report.

UK Government, 2014, Open data strategy 2014-2016, Corporate Report.

Veldhoen, A., Prins, S.D., 2014, Applying big data to risk management: Transforming risk management practices within the financial services industry, Avantage Reply (2014.12.10.).

Verma, R., Mani, S.R., 2016, Use of big data technologies in capital markets, Infosys.

Wellman, D., 2013, What is big data? SlideShare Presentation.

Zicari, R.V., 2014, Big data: A data-driven society? Big Data Lab Frankfurt.

디스코빅스	www.discovix.com
미래에셋 투자레터	www.m-stockblog.com
Alphasense	www.alpha-sense.com
Biointelligence Lab	bi.snu.ac.kr
Capgemini	www.capgemini.com
CBInsights	www.cbinsights.com

Crunchbase	www.crunchbase.com
Dataminr	www.dataminr.com
Deep Value	www.deep-value.com
Kensho	www.kensho.com
Quandl	www.quandl.com
Scaled Risk	www.scaledrisk.com
StockTwits	www.stocktwits.com
T-Robotics	t-robotics.blogspot.com
Wikipedia	www.wikipedia.org
Xignite	www.xignite.com
1004jonghee	1004jonghee.tistory.com