Internet and Information Security 제2권 제2호(2011년 11월) pp. 47~64

빅 데이터와 공공 데이터 활용

Big Data and the Utilization of Public Data

이만재*
Manjai Lee

초록 인터넷 사용이 급격히 증가하면서 우리는 정보의 홍수를 경험하고 있다. 이러한 데이터의 급격한 증가는 이제 기존 처리방식으로는 증가하는 데이터를 감당할 수 없으며 정보의 처리에 새로운 패러다임을 필요로 한다는 것을 알려주고 있다. 본 연구에서는 데이터의 증가에 따른 여러 분야에서의 대응 사례를 살펴보고 최근 빅 데이터라는 이름으로 이슈화된 데이터 처리의 중요성을 공공 데이터의 입장에서 검토한다. 공공 데이터는 일반 데이터에 비해 가치가 높으며 정부는 공공 데이터에 포함된 정보를 분석함으로서 효율적인 정책을 생산할 수 있다. 미국과 영국 정부는 공공 데이터를 마이닝(mining)하고 시각화해서 국민에게 대용량의 데이터를 쉽게 이해할 수 있도록 제공하여 정부 정책의 홍보에 활용하고 있다. 본 고에서는 링크드 데이터 방식을 이용하여 부처마다 별도로 관리되고 있는 공공 데이터를 연결하고 적절한 마이닝 기법과 시각화 도구를 사용한다면 디지털 데이터를 제대로 활용하는 차세대 전자정부의 구현이 가능함을 제시한다.

주제어: 공공 데이터, 빅 데이터, 링크드 데이터, 데이터 마이닝, 시각화

Abstract In this Internet age, we are experiencing a flood of information. This radical increase in the amount of data indicates that we need a new paradigm to handle such big data. 'Big data' is a new term that highlights the fact that this increase in the amount of data is not a problem, but a new challenge and opportunity. If we can handle the complexity of data, the competitive power of the organization will be increased. The governments of the US and the UK use public data mining and visualization as a tool to promote government policies. Linked data is one approach used to connect government data sets that span multiple organizations. By utilizing data mining and visualization techniques, it is possible to implement an efficient data driven e-government.

Keywords: public data, big data, linked data, data mining, visualization

논문접수일: 2011. 8. 24, 논문수정일: 2011. 10. 3, 게재확정일: 2011. 10. 11

* 서울대학교 차세대 융합기술연구원 특임연구위원
(Distinguished Fellow, Advanced Institutes of Convergence Technology, Seoul National University, manjai@snu.ac.kr)

Ⅰ. 서 론

인터넷과 웹이라는 기술이 일반에게 알려진지 대략 20년이 지난 지금 우리는 정보의 홍수 속에 살고 있다. 사무실에서 일하는 대부분의 지식 근로자는 업무 시간의 대부분을 정보의 검색에 보내고 있다. 그러나 원하는 정보를 발견하지 못하고 한참 뒤에 다른 경로를 통하여 자신이 원하는 정보를 발견하게 되는 경우가 허다하다. 이러한 경험은 인터넷을 통하여 많은 정보를 얻을 수 있지만 정말로 원하는 시간에 원하는 정보를 얻지 못할 수 있음을 일깨워주고 있다. 여기에서 인류는 실제로 얼마나 많은 정보를 갖고 있고 또 이를 제대로 활용할 수 있으려면 어떻게 해야 하는가에 대한 의문이 제기된다.

인터넷 이전의 정보는 주로 책이나 문서 형태로 생산되었다. TV방송이나 라디오와 같은 전파 매체에 의한 정보는 극히 일부를 제외하고는 방송 후 보관되지 않고 소멸되었다. 디지털 기술과 인터넷의 발달로 모든 사람이 정보의 소비자면서 동시에 정보를 생산할 수 있게 되었다. 여기에 스마트폰과 디지털 카메라와 같은 대량의 정보를 생산할 수 있는 기기가 추가되고 TV매체가 디지털 방식으로 전환되면서 정보 내지 데이터는 무서운 속도로 증가하고 있을 것으로 추정된다.

이러한 데이터의 증가는 이를 처리하는 컴퓨터 시스템에도 영향을 주어 개별적인 데이터의 보관 대신 클라우드 컴퓨팅으로의 변화를 유도하고 있다. 본 연구는 막연히 증가하고 있을 것으로 예측되는 데이터의 양이 구체적으로 얼마나 되며 얼마나 빠른 속도로 증가하는 것을 파악하는 것으로 시작하였다. 그리고 데이터의 증가에 따라 기존 정보처리 방식으로는 이러한 문제의 처리에 있어 한계에 도달하였으며 새로운 처리방식의 패러다임이 등장하고 있음을 확인하였다. 빅 데이터라는 이슈는 이러한 변화를 대표하는 용어로 대규모 정보처리를 필요로 하는 모든 기관에 이러한 패러다임의 변화를 알리는 역할을 하고 있다.

데이터의 증가에 대처하기 위해서는 데이터로부터 값진 정보를 얻는 것이 중요하다. 본 논문에서는 데이터의 증가라는 문제점을 해결하려고 노력한 각국의 사례를 살펴보도록 한다. 민간이나 기업이 보유한 빅 데이터의 문제는 기업 경쟁력의 문제로 기술개발과 비즈니스모델을 변화시킴으로 가능할 것이다. 정부나 공공기관이 보유한 모든 데이터를 하나의 데이터베이스 형태로 통합하는 것은 기술적으로 적절하지도 않고 법과 제도의 문제로 현실적으로 가능하지 않으며, 링크드 데이터와 같은 데이터 세트를 연계하는 방식이 현실적인 대안이 될 수 있음을 살펴보도록 한다.

Ⅱ. 데이터의 증가

1. 정보의 총량 증가

정보의 디지털화가 진행되며 정보량을 디지털 정보단위인 바이트 단위로 표시할 수 있게 되었다. 모든 정보의 디지털화가 완성되기 이전에는 인류가 보유한 정보의 총량을 파악하기 어려웠으며 가치 있는 정보 위주로 정보의 양을 파악하는 연구가 시작되었다. 미 의회 도서관의 자료를 디지털로 변환할 경우 약 3PB(페타바이트 = 10^{15}바이트)의 용량이 필요하다는 연구결과가 발표된 바 있다(Lesk, 2007). 이는 공식적으로 출판된 서적이나 문서만을 포함한 것으로 여기에 사운드, 영화 등 멀티미디어 데이터를 합친다면 인류가 보유한 모든 데이터는 대략 12,000PB로 추정되며, 전 세계가 보유한 이와 같은 모든 정보를 디지털로 변환할 수는 없다는 결론을 얻는다.

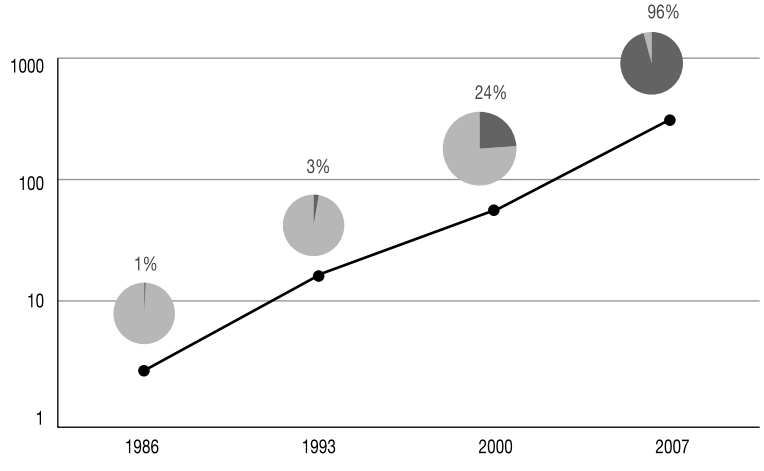
1987년부터 2007년 사이의 매체의 분류, 디지

털 변환비율, 디지털 매체의 압축 가능성, 자료의 중복 등 디지털 자료의 모든 특성을 고려한 1,120개의 정보소스를 종합하여 전 세계 정보의 저장, 통신 및 계산용량을 7년 주기로 추정한 바 있다(Hilbert & Lopez, 2011). [그림 1]은 정보의 용량을 로그 스케일로 표시한 것으로 21년의 기간 중 약 100배의 증가를 보여준다. 이 중 1993년부터 2000년의 기간에 비해 2000년부터 2007년의 증가가 높은 것은 데이터의 증가와 함께 여러 매체가 디지털로 변환됨에 따른 현상이다. 그래프의 증가추세를 감안한다면 2011년 현재 정보의 총량은 1ZB(제타바이트 = 10^{21}바이트)를 넘어선 것으로 추정된다.

이미 디지털로 전환된 데이터 외에도 잠재적인 디지털 데이터인 방송과 통신 데이터를 계량한 결과는 〈표 1〉과 같다. 2007년의 경우 방송방식은 약 1/4만이 디지털로 전환되었기에 디지털 데이터는 여전히 높은 증가를 보일 것으로 추정된다.

2. IDC의 정보의 홍수 분석

앞선 연구로 정보 처리와 관련된 지표의 증가는 기본적으로 무어의 법칙과 유사하게 지수함수를 따르는 것을 볼 수 있다. IDC는 2008년 발표한 백



단위: Exabyte ; 원그래프: 데이터의 디지털비율

*출처: Hilbert & Lopez(2011) 재구성

[그림 1] 전 세계 정보량의 변화(로그 스케일)

〈표 1〉 1986년 이후 정보증가

(단위: EB = 10^{18}바이트)

구분	1986	1993	2000	2007
인터넷 데이터	2.64	15.7	54.5	295
통신 데이터	0.28	0.47	2.24	65.3
방송 데이터	432	71.5	1150	1890
방송 데이터의 디지털 비율	0 %	0 %	7.2 %	24.8 %

*출처: Hilbert & Lopez(2011) 재구성

서를 통해 2007년 이후 정보의 생성이 급격히 증가하고 있고 저장용량은 이에 응하지 못해 균형이 깨진다는 내용을 발표하였다(Gantz et al., 2008). 이러한 현상은 과거의 예상보다 더 급격하게 변한 것으로 관측되었으며 개인의 정보보관을 위한 외장하드의 증가와 노트북이나 모바일 기기의 등장을 주요 원인으로 지적하였다. 연구 결과, 2009년 0.8ZB의 정보용량은 2020년 35ZB로 11년 동안 약 44배의 증가를 예측하였다. 지난 21년간 100배 증가도 매우 급격한 것이라고 생각하고 있는 지금, 이보다 더 급격한 증가가 이루어질 것이라는 예측으로 이러한 변화에 어떻게 대처해야 하는가에 대한 연구가 필요함을 보여주고 있다.

같은 IDC에 발표한 2009년의 백서에서는 40%의 응답자가 원하는 시간에 원하는 정보를 얻은 경우가 75% 이하라는 내용을 제시하고 있다(Gantz et al., 2009). 또한 응답자의 60%는 단순히 정보가 증가하는 것보다 너무나 다양한 다른 형태의 정보를 다루어야 하는 것이 더 큰 문제라는 의견을 피력하였고 정보의 증가에 따른 어려움으로 이메일과 인쇄된 문서를 우선적으로 꼽고 있다.

본 논문에 사용되는 큰 단위의 수는 일반인이 그 큰 정도를 이해하기 어렵다. 구체적으로 35ZB라는 값이 얼마나 큰 것인가를 살펴보자. 우리가 사용하는 정보 또는 데이터는 대부분 파일 단위로 저장되고 검색된다. 35ZB라는 용량은 파일의 개수로 바꾸면 대략 $25x10^{15}$개로 한글 표현으로는 2.5경으로 표시된다. 만약 이러한 정보의 홍수에서 원하는 내용을 찾으려 한다면 2.5경 개의 파일을 검색할 수 있는 검색엔진이 필요하다. 이러한 데이터가 데이터베이스로 검색할 수 있는 정형화된 데이터라면 현 기술로 문제를 해결할 수 있을 가능성은 존재한다. 그러나 대부분의 데이터는 비정형인 형태로 파일의 내용을 확인하기 전에는 어떤 내용이 포함되어 있는지를 파악할 수 없기 때문에 단순히 정보의 양이 증가한다는 것은 오히려 정보의 활용을 방해하는 현상이 될 수 있다. 또한 정보처리에 소요되는 비용은 거의 고정되어 있다고 알려져 있으며, 따라서 데이터당 처리 비용은 급격히 감소할 것이고 이는 실제로 데이터 관리에 이전보다 적은 비용으로 처리해야 한다는 것을 일깨워준다.

3. 매킨지의 빅 데이터 분석

매킨지에서는 2011년 5월 인터넷 데이터의 증가가 전 세계의 경제에 미치는 영향을 분석한 보고서를 빅 데이터(big data)라는 이름으로 발표하였다(Manyika et al., 2011). 빅 데이터를 데이터 용량에 따른 분류가 아니라 기존의 데이터베이스 처리방식으로 해결할 수 없는 데이터의 세트로 정의하고 이러한 데이터를 처리할 수 있는 기술이나 역량을 보유한 기업이나 국가가 미래에 경쟁력을 갖게 되며, 따라서 국가나 기업은 빅 데이터에 관심을 기울여야 한다고 역설하였다. 매킨지의 분석에 따르면 전 세계 인구의 60%에 해당하는 40억 명이 모바일폰을 사용하고 있으며 인구의 12% 수준이 보유한 스마트폰은 수년 내에 모든 모바일폰을 대체할 것이다. 또한 현재 3천 만 개의 네트워크 센서는 연간 30%라는 급격한 성장을 보이고 있으며 이러한 성장세가 언제 멈출지를 예측하기도 어려운 상태이다.

빅 데이터가 제공하는 기회는 데이터가 증가하는 특정 분야의 기업에만 해당되는 것이 아니며 스마트폰을 사용하는 대부분의 국가에서 발견할 수 있다. 빅 데이터가 가지고 있는 내재적인 가치를 관계자에게 빠르게 제공함으로 새로운 가치를 창조할 수 있는데, 이러한 가치의 창조는 기업이나 기관의 투명성을 확보함과 동시에 소비자의 수요를 적시에 발견할 수 있어 새로운 서비스를 먼

〈표 2〉 매킨지에서 제시한 빅 데이터 활용분야

도메인	분석 대상 데이터	예상효과
미국의 의료산업	제약사 연구개발 데이터, 환자 치료 /임상 데이터, 의료산업의 비용 데이터	연간 $3조 연간 0.7% 생산성 향상
유럽의 공공행정	정부의 행정업무에서 발생하는 데이터	연간 € 2.5조 연간 0.5% 생산성 향상
소매업	고객의 거래 데이터, 구매경향	$1조 + 서비스 업자 수익 $7조 소비자 이익
제조업	고객 취향 데이터, 수요 예측 데이터, 제조과정 데이터, 센서 활용 데이터	60% 마진 증가 0.5~1.0% 생산성 향상
개인 위치 데이터	개인, 차량의 위치 데이터	개발 및 조립비용 50% 감소 운전자본 7% 감소

*출처: Manyika et al.(2011) 재구성

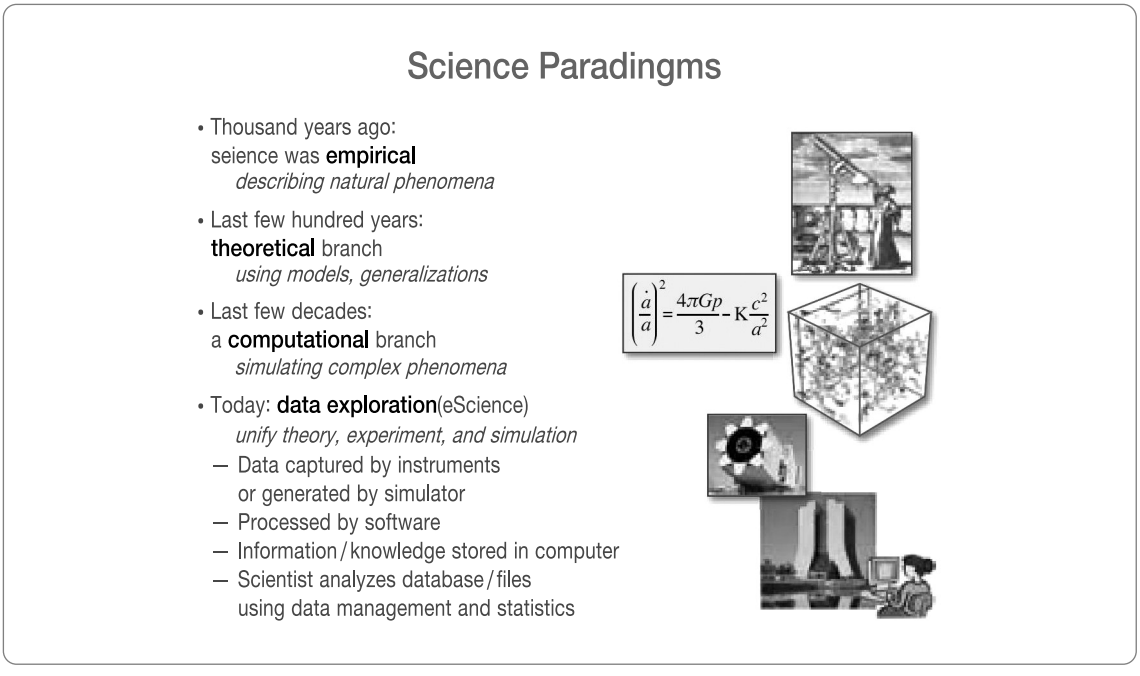
저 제공함으로서 가능하다. 이러한 기회를 얻기 위해서는 자동화된 데이터의 분석이 필요하다. 특히 세금징수나 재고관리에 있어 문제가 되는 특이한 사항을 쉽게 발견하여 담당자에게 제공함으로서 위험을 분산할 수 있는 것은 빅 데이터 처리에 있어 매우 중요한 기능이며 이를 위해 조직의 변화까지를 검토해야 할 단계이다.

빅 데이터 처리기술로는 데이터 마이닝, 머신 러닝, 자연어 처리와 같은 전통적인 데이터 분석 기술과 빅 테이블, 카산드라와 같은 새롭게 등장하는 빅 데이터 처리 기법, 그리고 분석된 데이터를 사용자에게 이해하기 쉽게 시각화하는 기술을 들 수 있다. 보고서에서는 빅 데이터를 제대로 활용할 경우 가장 큰 효과를 얻을 수 있는 분야로 〈표 2〉에 제시된 다섯 도메인을 예시하였다.

빅 데이터에 대처하기 위해서 미국의 경우 데이터를 분석할 능력을 갖춘 14만 명에서 20만 명의 인력이 필요할 것이며 이에 미리 대비할 것을 지적하고 있다. 또한 데이터 관리정책, 기술개발, 데이터의 접근 방식 개선을 구체적인 액션 아이템으로 제시하고 있다.

4. 짐 그레이의 제4 패러다임

데이터의 증가는 일반 기업의 문제가 아니라 과학계에도 영향을 미치고 있다. 마이크로소프트의 짐 그레이는 과학의 방법론 자체를 바꾸어야 한다는 제4 패러다임(The Fourth Paradigm)을 주장하였다(Hey et al., 2009). 그의 주장에 따르면 최초의 과학은 자연현상을 관찰함으로 시작되었고, 다음으로 자연을 이해하기 위한 이론적 모델을 만드는 방식으로 과학 문제를 해결하였으며 뉴턴의 법칙이 좋은 예이다. 이후 컴퓨터가 등장한 지난 수십 년 동안 과학은 복잡한 현상을 시뮬레이션하는 방식을 사용하였다. 제4 패러다임은 이제는 측정 장치에 의해 얻어진 데이터를 소프트웨어로 분석하여 새롭게 얻은 지식을 발표하는 방식으로 바뀌었음을 주장한다. 실제 우주과학 분야에 있어 연구자는 천체망원경을 조작하며 데이터를 얻는 것이 아니라 이미 얻어진 수천 개 이상의 천체 관측 데이터를 소프트웨어를 이용하여 분석하고 이로부터 새로운 이론을 추출하는 방식을 사용하고 있다.



*출처: Hey et al.(2009)

[그림 2] 제4 패러다임 개념 요약

짐 그레이는 미래의 연구방식은 이러한 데이터 분석 기법이 다양한 과학연구의 일반적인 현상이며 과학자는 데이터를 다루는 능력을 가져야 하고 이를 위해서는 데이터 분석 소프트웨어 기술을 더욱 발전시켜야 함을 강조한다. 또한 인터넷은 단순히 연구결과를 배포하며 필요한 자료를 검색하여 사용하는 수단으로 그치지 말고, 검색된 연구논문에 포함된 실험결과뿐 아니라 실험에 사용된 데이터를 함께 제공하는 방식으로 바뀌어야 함을 역설하고 있다. 연구결과의 검증 내지는 추가적인 연구는 이러한 데이터를 활용함으로써 발전할 수 있으며 이러한 방법을 사용할 경우 과학기술을 한 단계 격상할 수 있을 것이라고 주장한다. 빅 데이터는 많은 실험을 거쳐야 하는 생명과학 분야에서 우선적으로 주목받고 있으나 모든 과학 연구에서 도입되고 있다.

Ⅲ. 빅 데이터 문제 접근 사례

앞장에서 정보 내지는 데이터가 급격히 증가하고 있고 이에 대한 대책이 필요함을 제시하였으며 데이터의 증가에 따른 문제는 빅 데이터라는 용어로 이슈화되었음을 보였다. 위키피디아의 빅 데이터의 정의는 매킨지 보고서의 정의와 거의 같다.[1] 가트너 그룹의 정의는 데이터 볼륨의 증가, 데이터 입출력 속도의 증가, 데이터의 다양성의 증가의 3가지 특징을 빅 데이터의 문제로 정의하고 있다(Gartner, 2011). 본 논문에서는 데이터의 활용측면을 함께 고려하여 현재 정부의 정보화와 관련된 빅 데이터의 정의인 '대용량 데이터를 활용

1) 위키피디아, http://en.wikipedia.org/wiki/Big_data

분석하여 가치 있는 정보를 추출하고 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술'을 사용하기로 한다(국가정보화전략위원회, 2011). 이 정의는 빅 데이터 문제가 단순히 새로운 기술 개발의 문제가 아니라 정보의 연계의 문제일 수도 있음을 내포하고 있다. 빅 데이터와 관련된 각국의 활동과 공공 데이터 활용 사례를 살펴보자.

1. 미국 국토안보부의 비주얼 애널리틱스

미국의 경우 9.11 테러는 정치적으로 중요할 뿐 아니라 데이터의 중요성에 경각심을 불러일으킨 사건이다. CIA, FBI와 같은 여러 안보를 담당하는 기관이 있었지만 테러를 미연에 방지하지 못한 것에는 정보 분석이 뒤따르지 못한 것이 하나의 문제점으로 지적되었다. 테러의 사후대책으로 국방성과는 별도로 국토안보부가 신설되었으며 정보 분석의 결과를 매일 파악할 수 있도록 하는 것을 임무 중 하나로 부여하였다.

국토안보부의 2008년 직제에 따르면 정보 분석(intelligence & analysis) 담당 차관보 직책을 두고 비주얼 애널리틱스(visual analytics) 분야의 연구개발 지원을 업무의 일부로 하고 있다(Borja, 2008). 비주얼 애널리틱스는 기존의 정보 시각화에 분석적인 이론을 결합한 것으로 전반적인 사건의 진행상황을 바로 파악할 수 있고 새로운 대처에 따라 결과가 어떻게 변하는지를 볼 수 있는 인터랙티브 기능을 지원한다. 이러한 시각화 도구를 사용하여 기존에 파악하지 못하던 안보의 위협이나 감시대상의 변화를 쉽게 인지하도록 하여 새롭게 발생할 가능성이 있는 문제에 적시에 대처하는 것을 목표로 삼고 있다(Lavigne & Goulin, 2011). 미국의 이러한 움직임은 안보 분야에서 미국과 행동을 같이 하는 캐나다, 영국 등 국가에서도 전개되고 있다.

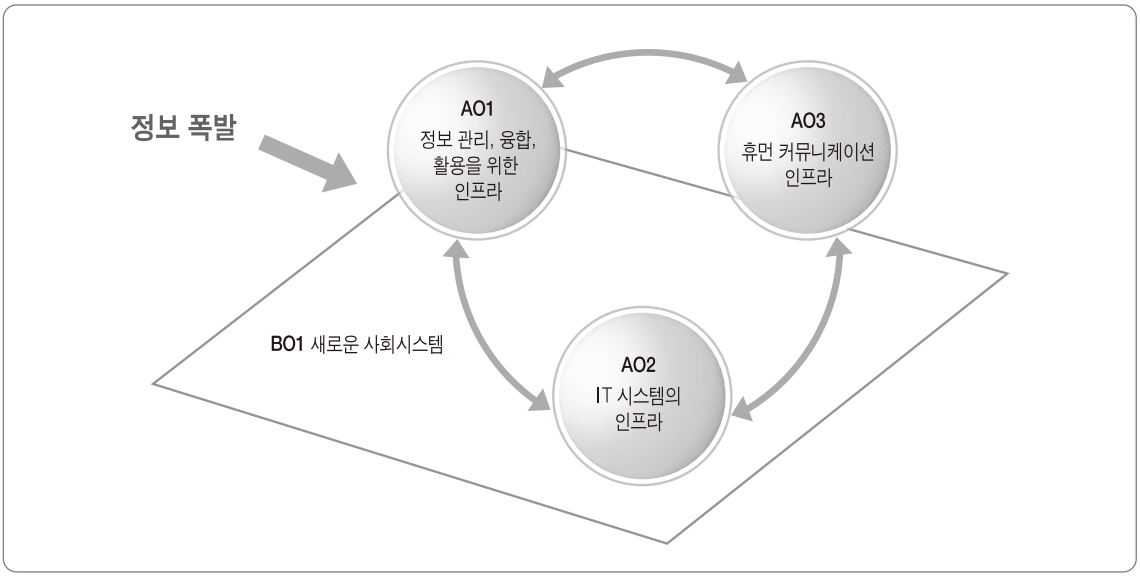
미국 정부는 비주얼 애널리틱스 분야의 연구를 구체화하기 위한 방편으로 2004년 NVAC(National Visualization and Analytics Center)라는 연구센터를 설립하였으며 비주얼 애널리틱스 연구의 내용 및 접근방식에 대한 저서(Thomas & Cook, 2005)를 통하여 분야의 학문적 중요성을 알렸다. NVAC의 기본적인 목표는 국가의 안보문제의 복잡성을 해결하고자 하는 것이다. 그러나 연구결과는 데이터 마이닝과 분석결과의 시각화에 있기 때문에 연구결과의 상당부분은 복잡한 데이터를 분석하는 미국 정부의 모든 부서에서 활용할 수 있다. 이러한 사례는 국내의 안보나 재해대책 분야의 빅 데이터 처리에 좋은 참고가 된다.

2. 일본의 정보폭발 프로젝트

일본은 9.11 테러와 같은 사건을 겪지는 않았지만 정보의 폭발적인 증가가 문제가 될 것이라는 것을 일찍 인지하였으며 2005년부터 2011년까지 6년의 기간 동안 '정보폭발에 대비한 인프라스트럭처' 라는 대규모 연구 프로젝트를 추진하였다(Kitsuregawa, 2007; Kitsuregawa & Nishida, 2010). 프로젝트는 크게 세 그룹으로 나누어 진행되었다. 프로젝트의 연관구조는 [그림 3]과 같다.

(1) 정보폭발시대의 정보관리, 융합, 활용을 위한 인프라스트럭처(A01)

정보폭발이 진행되면 대량의 정보를 다루어야 하는 검색의 문제가 가장 먼저 제기된다. 현재의 검색방식에서는 검색어와 가장 관련이 높다고 생각되는 10개 내외의 문서만을 보여주기 때문에 다수의 의견만 제시되고 의미 있는 소수의 의견을 확인할 수 없으며 이러한 문제를 해결할 수 있는 새로운 검색엔진을 개발한다.



*출처: Kitsuregawa(2007) 재구성

[그림 3] 정보폭발 프로젝트 개념도

(2) 정보폭발 대비 IT시스템의 인프라스트럭처(A02)

정보폭발이 진행되면 대규모의 데이터를 저장할 IT시스템을 필요로 한다. 이 그룹은 대용량 데이터 저장 시스템 정보를 모니터링하면서 관리할 수 있는 그리드 시스템을 연구한다. 시스템의 문제는 인간의 개입 없이 자동적으로 컴퓨팅 자원을 재조정하는 방식으로 유지하도록 한다.

(3) 정보폭발시대의 휴먼 커뮤니케이션 인프라스트럭처(A03)

이 그룹은 정보의 복잡에 대해 인간이 어떻게 대처할 수 있는가를 연구한다. 기본적으로 멀티모달 상호작용을 통하여 정보시스템과 인간과의 정보의 갭을 줄이려는 노력과 미래의 일상생활에서 등장할 로봇과의 상호작용 연구를 포함한다.

(4) 지식기반사회의 거버넌스 문제(B01)

공학자보다는 인문사회과학자의 참여로 진행되었으며 기술의 역기능 등 사회에 미치는 파장이 연구의 대상이 된다. 특히 건강정보와 요양지원 등 일본 특유의 고령자 사회에 대한 사회적 문제를 다수 포함하고 있다.

3. 링크드 데이터

빅 데이터 문제를 단순히 대용량 데이터의 문제로 정의할 경우 링크드 데이터와의 연관성은 많지 않다. 그러나 현 기술로 처리할 수 없는 용량은 아니나 어떠한 이유에서 연계가 불가능하여 분석이 불가능한 데이터가 큰 비중을 차지하고 있는 실정을 감안한다면 링크드 데이터는 빅 데이터의 하나의 해결책으로 정의할 수 있다. 인터넷에 의한 정보전달은 대부분 텍스트 기반의 웹 문서형태로 이루어져 있다. 이러한 텍스트 문서는 일반 사용자가 보기 적합하도록 만들어진 것으로 정형화된 데이터베이스 처리를 기반으로 하는 컴퓨터와 같은 기계를 이용한 분석은 매우 어렵다. 매시업(mashup) 애플리케이션을 개발하기 위해 이러한

데이터를 부분적으로 활용한 예는 있으나 근본적인 문제 해결을 위해서는 HTML과 같은 인터넷 프로토콜을 이용한 데이터의 연결이 가능하도록 웹의 변경이 필요하다.

컴퓨터와 같은 기계 스스로 웹에 접근하여 자료를 분석할 수 있도록 하는 기술은 시맨틱 웹이라는 이름으로 시작되었다. 그러나 시맨틱 웹의 연구자의 의도와는 달리 인터넷 커뮤니티가 이를 채택하지 않아 제대로 활용되지 못하였다. 웹의 창시자라고도 할 수 있는 팀 버너스리는 'Raw Data Now'라는 연설을 통하여 기존의 인터넷을 문서의 연결뿐 아니라 데이터의 연결을 가능하게 하자는 링크드 데이터(linked data) 보급을 강조하였다(Berners-Lee, 2009). 이러한 데이터 연결 프로젝트는 2007년 1월 W3C에 의해 발표된 후 많은 연구자와 기관의 지원을 받아 급속히 확산되고 있다(Bizer et al., 2009).

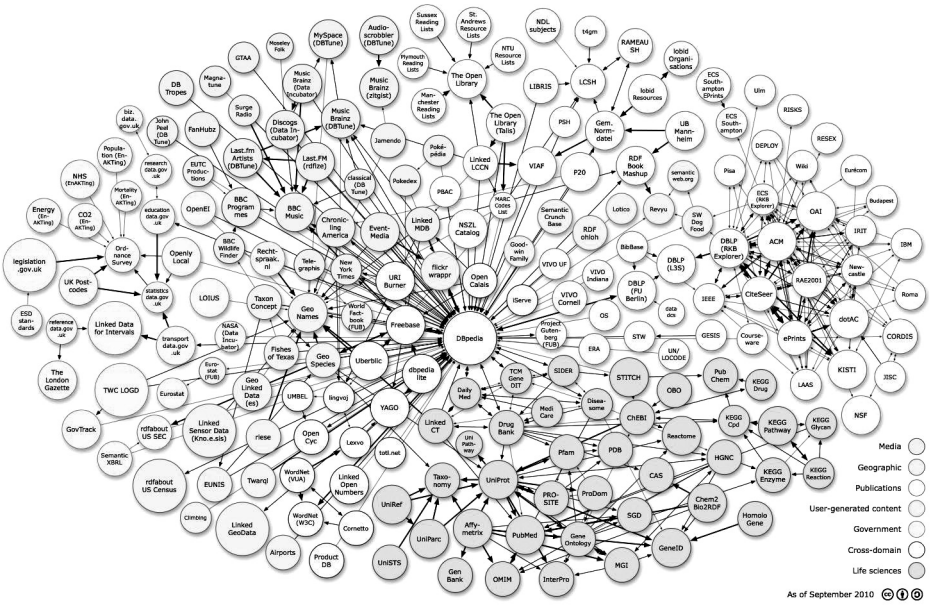
링크드 데이터의 장점을 예를 들어 보자. 평창동계올림픽에 대한 정보를 소개하는 웹사이트를 만든다고 가정한다. 내용 중에 김연아 선수를 소개하는 페이지를 개설해야 하는데 김연아 선수에 대한 소개내용은 김연아 선수가 새로운 경기에 출전할 때마다 바꾸어야 한다. 매번 이러한 내용을 확인하고 수정하는 대신 위키피디아의 김연아라는 문서와 링크시키면 새로운 경기가 진행될 경우 문서를 수정하는 번거로움을 피할 수 있다. 이는 위키피디아에서 링크드 데이터를 제공하기에 가능한 것이다.

링크드 데이터는 트리플이라는 연결 구조를 통해 구현된다. 예를 들어 우리나라 지역에 대한 데이터는 대한민국 지역정보 데이터세트에 보관되어 있고 올림픽 개최도시는 올림픽 데이터 세트에 포함되어 있을 경우 "대한민국-도시이다-평창"이라는 관계와 "올림픽-개최되었다-평창"이라는 두 개의 트리플 구조를 통해 평창에 관한 정보를 찾고자 할 경우 대한민국의 지역정보 데이터세트에 기록된 내용을 이용할 수 있다. 이러한 트리플 방식의 연계는 링크드 데이터 구조의 핵심이다. 2010년 11월 현재 285억 개의 트리플이 보고되었으며 급속히 증가하고 있음을 알 수 있다. 링크드 데이터는 정보의 복잡성을 해결하고 데이터에 직접 접근하여 데이터를 알기 쉬운 형태로 제공할 수 있도록 하는데 꼭 필요한 기능이다. 앞선 비주얼 애널리틱스가 정보의 분석과 시각화를 강조하고 있다면 링크드 데이터는 인터넷에 존재하는 모든 데이터세트를 효율적으로 연결하여 분석을 가능하게 하는 인프라로 볼 수 있다.

[그림 4]는 2010년 11월 현재 링크드 데이터를 이용하여 연결된 주요 웹 사이트를 보여주고 있다. 중심에는 링크드 데이터의 출발 사이트인 DBPedia가 많은 타 사이트와 연결된 것을 볼 수 있다. 링크드 데이터를 활용되는 도메인에 대한 통계는 〈표 3〉과 같다. 정부 관련 정보가 가장 많은 트리플 정보를 차지하고 있으며 실제 사용은 생명과학 분야에서 가장 활발히 전개되고 있음을 볼 수 있다.

4. 공공 데이터 활용

공공 데이터(public data)는 정부 또는 공공기관이 보유하고 있는 데이터를 말한다. 공공기관의 업무과정에서 얻어진 데이터들로 데이터 중에서 상대적으로 가치가 높으며, 현재 기술로 처리하기 어려울 정도의 대용량 데이터는 아니나 분석의 용도로 활용되지 못하고 있다. 앞서 III장의 시작에 빅 데이터를 단순한 데이터 크기의 문제가 아니라 분석의 문제로 정의한 것은 공공 데이터의 문제를 부각시키기 위한 것이다. 미국, 영국 등 선진국에서는 공공 데이터의 활용이 정부의 대민 서비스와 직결되어 있음을 인지하고 공공 데이터의 활용을



*출처: Heath & Bizer(2011)

[그림 4] 2010년 11월의 링크드 데이터 연결 구조

〈표 3〉 링크드 데이터의 연결 현황

도메인	데이터 세트	트리플(단위: 천)	%	링크(단위: 천)	%
미디어	26	2,454,299	8.59	50,377	12.73
지리정보	16	5,907,260	20.68	16,539	4.18
정부	26	11,630,305	40.72	17,944	4.53
출판	68	2,307,901	8.08	77,995	19.70
도메인 간 연결	22	3,541,130	12.40	29,121	7.36
생명과학	42	2,664,119	9.33	200,418	50.64
사용자 콘텐츠	7	57,463	0.20	3,402	0.86
계		28,562,479		395,756	

*출처: Heath & Bizer(2011)

정부 정책으로 추진하고 있다.

팀 버너스리는 정부가 갖고 있는 데이터를 링크드 데이터 방식으로 웹에 제공해야 한다는 것을 강조하였다(Berners-Lee, 2009). 이미 정부가 갖고 있는 데이터를 공개하여 투명한 정부를 만들어야 한다는 생각은 미국의 오바마 대통령의 선거공약에도 포함될 정도로 미 국민의 지지를 얻은 정책이다. 일반적으로 데이터의 공개에 대해 부처 내에서는 상당한 거부감을 갖고 있어 이를 가능케 하기 위해서는 최고 책임자의 정책의지가 필요하다(Robinson et al., 2009). 미국의 오바마 대통령은 자신의 선거공약 대로 2009년 정부의 정보를 공개

하는 미 연방정부의 CIO(Chief Information Officer)에 비벡 쿤드라를 임명하여 data.gov라는 웹 사이트를 구축하도록 하고 이를 통하여 정부 데이터를 공개하도록 하였다. 영국 역시 같은 시기에 data.go.uk를 기획하고 팀 버너스리를 책임자로 기용하였다. 팀 버너스리는 웹에 대한 확고한 방향성과 개인적인 인지도를 바탕으로 출발은 미국보다 늦었지만 미국보다 앞선 수준의 공공정보를 제공하고 있어 공공 데이터의 공개를 추진하고자 하는 타 정부의 참고모델이 되고 있다.

팀 버너스리는 링크드 데이터가 제대로 활용되기 위해서는 데이터를 어떻게 제공해야 하는지에 대한 가이드라인을 제시한 바 있다. 링크드 데이터의 본질을 제대로 이해하고 데이터를 구현했는가에 따라 〈표 4〉와 같이 제공되는 데이터를 5등급의 별로 표시하는 방식을 제시하였다. 이는 공공기관의 참여를 자발적으로 유도하는데 큰 효과를 얻고 있다. 별이 많은 데이터는 별이 적은 데이터의 기능을 모두 포함한 상태에서 추가된 기능만을 표시하였다.

각국의 정부는 정부 보유 데이터를 이용하여 각종 정책을 개발하고 대 국민 서비스를 직접 수행한다. 정부가 보유한 데이터를 공공 데이터로 제공할 경우 가능한 서비스가 어떤 것이 있을지는 스마트폰의 앱과 같이 미리 속단하기 어렵다. 현재 부분적으로 제공된 공공 데이터로 가능한 예를 살펴보자. 공공 데이터 중 빅 데이터를 사용한 예는 적절한 예가 없기에 일반인에게 파급효과가 큰 내용 위주로 제시하였다. 또한 공공 서비스를 위해서는 데이터를 이해하기 쉬운 시각화 기술을 사용해야 함을 알려주고 있다.

(1) Where does my money go?

공개지식재단(Open Knowledge Foundation)이라는 민간기관에서 운영하는 웹 사이트에서는 영국 정부가 제공한 공공 데이터를 사용하여 일반인이 이해하기 쉽도록 시각화 기법을 사용하여 보여주고 있다.[2] 그 중 대표적인 내용이 '내가 낸 세금은 어디에 쓰이나?' 라는 이름의 인터랙션이 가능한 애플리케이션이다. 개인의 수입에 해당된 금액을 지정할 경우 세금 납부액과 세금이 사용되는 항목을 복지, 의료, 교육, 국방 등 항목별로 2단계로 도식화하여 보여주고 있어 납세에 대한 이해를 도와주고 있다.

(2) OECD Better Life Initiative

OECD는 동일한 기준으로 회원국 간의 통계를 비교할 수 있도록 다양한 데이터를 생산하여 제공

〈표 4〉 팀 버너스리의 링크드 데이터 5등급 표시

등 급	설 명
★	임의의 형식으로 공개
★★	기계 처리 가능한 정형 데이터 형식(예: 엑셀)
★★★	공개 표준 형식
★★★★	URL을 사용하는 링크드 데이터 형식
★★★★★	타 웹사이트와 연계 제공

*출처: Berners-Lee(2009) 재구성

2) Where does my money go? http://wheredoesmymoneygo.org/dashboard

[그림 5] "내가 낸 세금은 어디에 쓰이나?"화면(연간 수입 10만 파운드 표현)

한다. 그 중 가장 널리 알려진 것은 회원국의 삶의 질을 지표화한 것이다. 이 지표에는 주택, 수입, 직업, 교육 등 전체 11개의 측정치를 제공하며 개인의 선호에 따라 이에 비중을 감안하여 국민의 삶을 평가할 수 있도록 하고 있다.[3)]

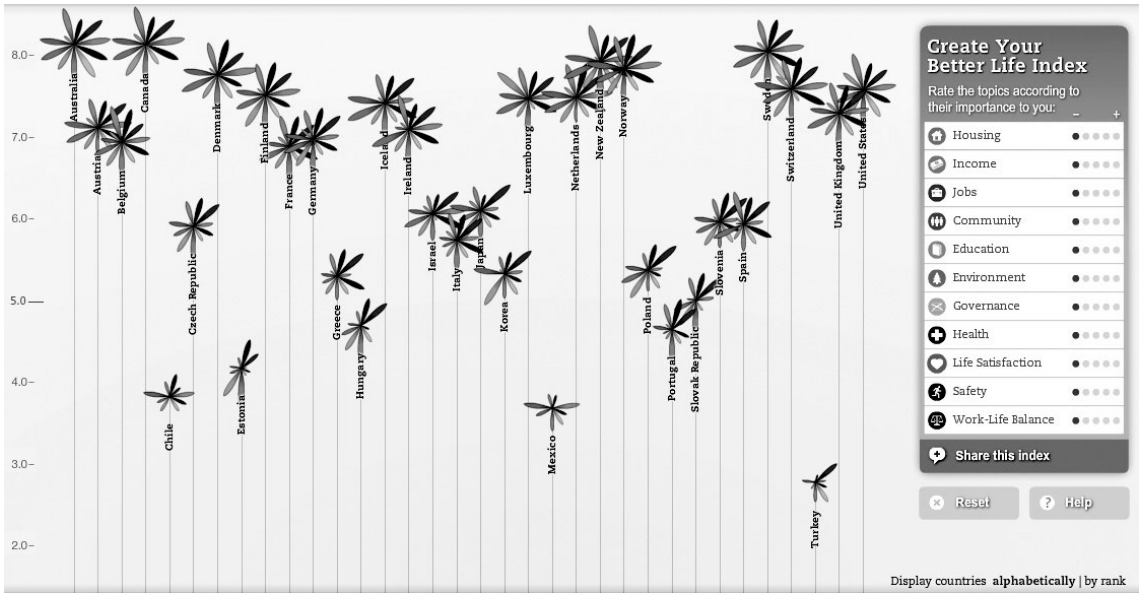
(3) IT Dashboard

미국 정부에서 직접 운영하는 웹 사이트 중 하나로 미 정부에서 발주하는 IT 관련 예산을 종합적으로 보여준다.[4)] 정부의 공식 데이터를 사용하고 있기에 예산 자체, 예산의 증감, 시간의 변화에 따른 변화 그래프, 면적을 사용한 그래프 등 다양한 방식의 시각화 기법을 사용한다.

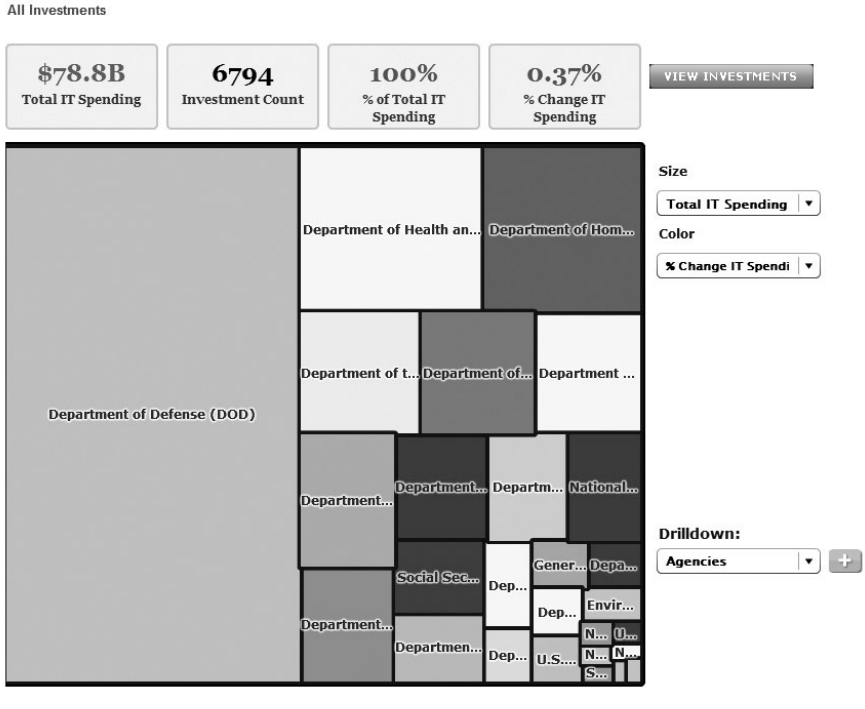
(4) 구글 공공 데이터 익스플로러

구글은 2010년 3월 공공데이터를 활용하여 데이터의 의미를 이해하도록 하는 공공 데이터 익스플로러(public data explorer)라는 웹 사이트를 운영하고 있다. OECD, 미국 정부, 유럽 통계청의 자료를 활용하고 있으며 개인도 데이터세트를 업로드하면 이용이 가능하다. 대부분의 내용은 정형화된 숫자 데이터를 활용한 것이나 빅 데이터와 관계된 애플리케이션으로 자체의 검색엔진을 이용한 분석한 독감의 유행에 대한 데이터세트를 포함하고 있다.[5)] 이는 구글의 지메일에 포함된 내용에서 독감과 관련된 단어가 언제 등장하였는가 하는 패턴을 국가별로 분석한 것으로 미국 질병센터에 보고된 내용과 일치하는 것을 보여주고 있다. 이는 비정형 데이터인 메일 분석으로 이전의 분석 방법으로 얻기 어려운 새로운 현상을 발견할 수 있음을 보여주는 예이다.

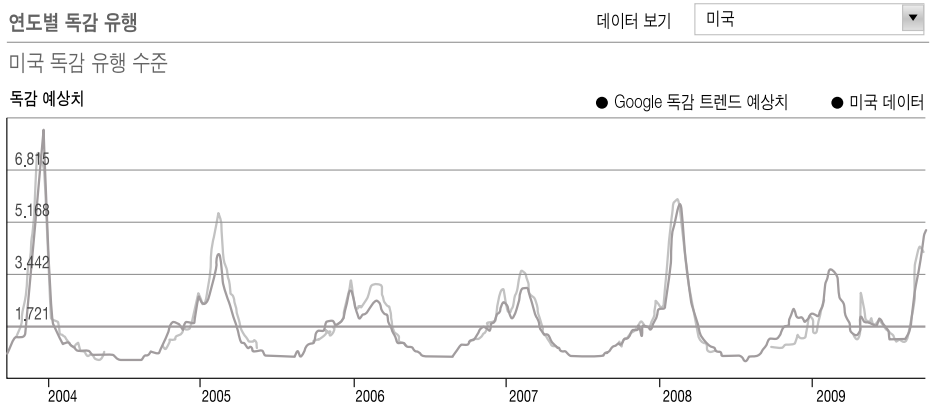
3) OECD Better life Initiative, http://www.oecdbetterlifeindex.org
4) USA Government IT Dashboard, http://www.itdashboard.gov
5) 구글 독감트렌드, http://www.google.org/flutrends/about/how.html



[그림 6] OECD Better Life Initiative



[그림 7] 미국의 IT 관련 예산의 변화 표현(면적 표현 방식)



[그림 8] 구글의 공공 데이터 익스플로러의 독감 유행 분석

Ⅳ. 국내의 빅 데이터 처리

1. 공공 데이터의 활용상황

지금까지 주로 외국의 빅 데이터 대처방안 및 공공 데이터 활용에 대한 사례를 살펴보았다. 국내의 경우 빅 데이터와 같은 데이터의 증가보다 인터넷 트래픽의 증가를 더 심각하게 생각하였으며 데이터의 활용이라는 측면의 연구는 상대적으로 부족하다. 또한 고급 정보의 검색을 구글과 같은 외국기업의 솔루션에 의존하기 때문에 데이터의 증가에 대한 문제는 기업의 문제이며 국가 경쟁력의 문제로 인식되지 않은 상태이다. 정부 내지는 정부 산하기관에서 보유한 데이터는 국내에서 보유한 데이터 중 가장 규모가 크고 가치 있는 데이터 중 하나일 것이며 이러한 데이터가 정책 입안에 어떻게 활용되고 있는지를 살펴보자.

우리나라의 경우 「공공기관의 정보공개에 관한 법」에 따라 모든 공공기관은 국가안보와 같은 법으로 정한 비공개 대상정보를 제외하고는 모두 공개하도록 되어 있다. 특히 전자정부가 활성화되면서 정보목록을 작성하고 이를 정보통신망을 활용한 정보공개시스템을 통하여 공개하도록 되어 있어 인터넷 접속만으로 정부가 어떠한 정보를 공개하고 있는지 확인할 수 있는 상태이다. 그러나 해당법령에서는 정보의 정의에 있어 문서, 도면, 사진, 필름 등 다양한 매체를 포함하고 있으나 앞에서 말한 가장 가치가 있는 원시 데이터에 대한 언급이 없어 대부분의 정보는 기계에 의한 접근이 불가능한 문서로 제한되어 있는 상태이다.

공공정보와 일부 중복되기도 하는 다양한 보고서들은 국가지식 포털을 통하여 접근할 수 있다. 한국정보화진흥원이 운영하는 국가지식포털의 경우 정부 내지는 산하 공공기관에서 작성한 보고서와 같은 문서를 직접 또는 링크를 제공한다. 보고서의 내용 중에 필요한 데이터가 포함되어 있기에 전혀 가치가 없지는 않지만 일반적으로 해당되는 문서를 찾는 특별한 목적을 가진 개인에게만 도움이 될 수 있는 정보공개 형태이다.

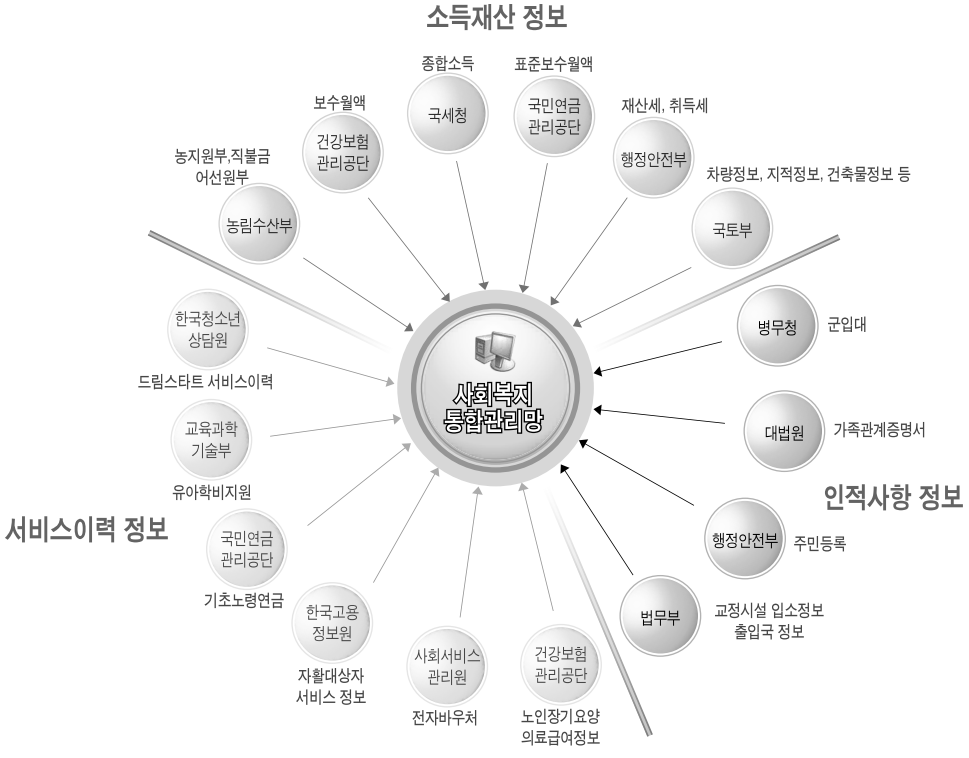
정부에서 제공하는 통계는 숫자로 표시된 정형화된 원시 데이터의 대표적인 예이다. 통계청은 통계처리의 전문성을 발휘한 국가통계포털(kosis.kr)을 운영하고 있다. 통계청에서 제공하는 통계를 대부분 엑셀 파일 형태로 제공하고 있

어 부분적으로는 원시 데이터를 필요로 하는 요구에 부응하고 있다. 또한 국제기관, IMF, 월드뱅크, OECD 등과 같은 국제기관과 제휴하여 제공되는 데이터에 접근할 수 있다. 특히 G20과 같은 행사에 맞추어 G20 통계 상황판과 같은 시각화 기법을 사용한 데이터를 제공하고 있어 공공 데이터 제공의 ★★ 등급으로 평가된다.

2. 빅 데이터 활용을 위한 개선방향

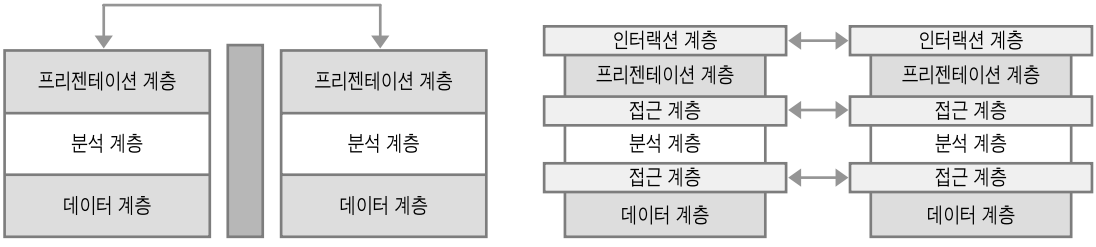
국내의 경우 부처 및 정부 산하기관별로 통계를 생산하고 있으며 부처 개별적으로 정보를 공개하고 있다. 그러나 앞서 설명한 대로 대부분의 정보를 ★★ 등급으로 제공하고 있다고 볼 수 있다. 국가 행정의 복잡성에 따라 한 부처가 정책을 시행하기 위해서는 해당 부처의 보유 통계만으로 정책을 구현하기 어렵다. [그림 9]에서 볼 수 있듯이 보건복지부에서 추진 중인 국민 복지를 위한 사회복지 관리망의 경우 27개 기관에서 215종의 정보를 제공받아 서비스를 제공하고 있다(최현수, 2009). 이는 정부 차원의 빅 데이터 관련 업무로는 가장 발전된 사례라고 할 수 있다. 그러나 이러한 통합망을 구축하기 위해 각 부처에 해당되는 데이터를 구체적으로 파악하고 총리실에서 이를 조정하는 행정적인 노력이 필요했던 것으로 알려지고 있다.

현재와 같은 접근방법을 사용할 경우 새로운 대국민 서비스를 개발하기 위해서는 서비스의 정의가 이루어진 후 정부의 이해관계 부처가 모여 필요한 데이터 세트를 분석하고 이를 제공하기 위해



*출처: 최현수(2009)

[그림 9] 사회복지통합관리망의 데이터 연결 사례



(a) 현 정부의 공공업무처리 (b) 빅 데이터 기반 공공업무처리

[그림 10] 공공 데이터 접근방법 개선

부서 내의 정보시스템을 수정해야 하는 작업을 반복해야 한다. 예를 들면 구제역 종합대책, 도시홍수 방지대책 등 각 부서가 보유한 데이터를 종합적으로 분석해야 하는 새로운 사건이 발생할 경우, 필요한 데이터 세트를 정의하고 부처가 논의를 진행하는 일이 반복될 가능성이 높다.

따라서 국가안보와 같은 특수한 경우를 제외하고는 정보를 제공하는 부처에서는 타 부처에서 활용될 가능성이 높은 데이터를 문서 형태로 제공하는 대신 데이터 계층에서 직접 접근할 수 있도록 하여 정책을 집행하는 부서에서 정책을 개발할 수 있도록 정부 내의 공공 데이터를 개방하는 것이 바람직하다. [그림 10]의 (a)는 현 정부의 부처 간 정보 교류 방식을 표시하고 있다. 특별한 통합 절차를 통하지 않는 경우 완성된 보고서 형태의 프리젠테이션 형태의 인터랙션만이 가능하다. 정부의 분석기능을 강화하기 위해서는 원시 데이터나 분석 결과와 직접 연결되는 [그림 10]의 (b)와 같은 인터페이스가 필수적이다. 앞서 설명한 링크드 데이터는 이러한 인터페이스의 한 방식으로 볼 수 있다.

부처 간에 데이터를 교류하기 위해서는 선결하여야 할 문제가 있다. 정부 부처가 수집하여 사용하는 데이터에는 개인정보가 포함되어 있다. 개인정보가 포함된 내용을 꼭 필요하지 않은 다른 부처에 제공할 경우 개인정보의 누출 가능성을 높일 수 있기에 데이터 제공에 있어서는 개인정보의 익명처리(anonymisation)나 그룹처리(aggregation)가 필요하다.

공공 데이터는 가급적 민간에게 개방되어야 하는 것이 원칙이다. 그러나 민간 개방에 따른 문제가 해결되지 않아 링크드 데이터로의 전환이 늦어진다면 최소한 공공기관간의 데이터 연계만이라도 우선적으로 시행하여 대국민 서비스를 개선하는 것이 바람직하다. 빅 데이터에 해당되는 대규모의 공공 데이터 처리는 우선적으로 비교적 처리가 쉬운 정형 데이터 처리로부터 시작할 수 있다. 그러나 이미 민간에서 사용되고 있는 SNS(Social Network Service)를 통해 얻어지는 데이터를 활용한다면 일부 사람이 감지한 건축물 붕괴, 복지 사각지대에 있는 빈민층 보호와 같은 서비스에 활용할 수 있다.

V. 결 론

지금까지 정보의 급격한 증가문제는 정보처리 분야에서만 발생하는 것이 아니라 과학기술 분야의 제4 패러다임, 국가 안보 분야의 비주얼 애널리틱스와 같이 새로운 접근방법을 필요로 하는 패러다임의 변화를 요구하고 있다는 것을 살펴보았다. 기업과 공공기관의 경우 이는 빅 데이터라는 이슈로 귀결되며 새로운 도전분야 또는 기회라고

볼 수 있다. 이를 제대로 활용할 경우 국가나 기관의 경쟁력을 확보할 수 있는 좋은 수단임을 확인할 수 있었다. 빅 데이터 중 비정형 데이터의 경우에는 이를 해결하기 위한 많은 기술 개발이 필요하다. 그러나 정부가 생산하는 빅 데이터는 대부분 비교적 처리가 용이한 정형 데이터이며 기관간의 벽을 허물 수 있다면 바로 활용이 가능한 데이터이다. 팀 버너스리가 제안한 링크드 데이터는 이러한 문제에 대한 해답이 될 수 있다. 국가 차원에서 최소한 공공기관 간에 보유한 데이터를 링크드 데이터 또는 연계가 가능한 적절한 방식으로 제공하도록 하고 이를 기반으로 데이터를 분석 내지 시각화하는 도구를 제공하여 대국민 서비스를 한 차원 높인다면 현재 우리가 자랑하는 전자정부를 빅 데이터 기능을 제대로 활용하는 새로운 형태의 전자정부로 격상시킬 수 있을 것이다.

▶ 참고문헌

국가정보화전략위원회 (2011). 『빅 데이터를 활용한 스마트 정부 구현』. 서울: 국가정보화전략위원회.

최현수 (2009. 12. 22). 사회복지통합관리망의 구축과 통합조사관리. 『사회복지통합관리망 운영에 따른 기대효과와 과제 정책토론회 자료집』, 42-73.

Berners-Lee, T. (2009). Linked data. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. *International Journal on Semantic Web and Information Systems,* 5(3), 1-22.

Borja, E. C. (2008). *Brief documentary history of the department of homeland security: 2001-2008.* Washington, DC: Department of Homeland Security.

Dohrmann, T., & Pinshaw, G. (2009). *The road to improved compliance.* New York: McKinsey & Company.

Gantz, J., Boyd, A., & Dowling, S. (2009). *Cutting the clutter: Tackling information overload at the source* (IDC White Paper). Framingham, MA: International Data Corporation.

Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., & Toncheva, A. (2008). *The diverse and exploding digital universe* (IDC White Paper). Framingham, MA: International Data Corporation.

Gartner (2011, June 27). *Gartner says solving 'big data' challenge involves more than just managing volumes of data.* Gartner Press Release. Retrieved from http://www.gartner.com/it/page.jsp?id=1731916

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space.* San Rafael, CA: Morgan & Claypool.

Hey, T., Tansley, S., & Tolle, K. (2009). *Jim Gray on eScience: A transformed scientific method, the fourth paradigm.* Redmond, WA: Microsoft Research.

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

Kitsuregawa, M. (2007). Challenge for info-plosion. *Proceedings of the 10th International Conference on Discovery Science,* 1-8.

Kitsuregawa, M., & Nishida, T. (2010). Preface: Special issue on information explosion. *New Generation Computing, 28*(3), 207-215.

Lavigne, V., & Goulin, D. (2011). Applicability of

visual analytics to defense and security operations. *Proceedings of the 16th International Command and Control Research and Technology Symposium.*

Lesk, M. (2007). How much information is there in the world? Retrieved from http://www.lesk.com/mlesk/ksg97/ksg.html

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition and productivity.* New York: McKinsey & Company.

Robinson, D. G., Yu, H., Zeller, W. P., & Felten, E. W. (2009). Government data and the invisible hand. *Yale Journal of Law & Technology, 11,* 160-175.

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: The research and development agenda for visual analytics.* IEEE CS Press. Retrieved from http://nvac.pnl.gov/agenda.stm