빅데이터 이용 확산을 위한 ODI 기반 데이터 액세스 프레임워크

김화종
강원대학교

요약

최근 사회 각 분야에서 빅데이터를 분석하여 새로운 가치를 찾아내려는 시도가 급속히 증가하고 있다. 그러나 빅데이터를 분석하여 소기의 성과를 얻으려면 한 기관이나 기업이 자체로 보유하고 있는 데이터 뿐 아니라 외부에 있는 가치 있는 데이터가 필수적으로 필요한 경우가 대부분이다. 현재 빅데이터 이용에서 가장 어려운 것은 대용량 데이터를 다루는 하드웨어나 분석 소프트웨어 도입이 아니라 핵심적으로 필요한 외부 빅데이터를 어떻게 확보할 것인가이다. 본 고에서는 빅데이터를 효과적으로 공유하고 활용하기 위한 방안으로 오픈 데이터 인터페이스(ODI)를 제안한다. ODI를 사용함으로써 프로그램이 직접 읽을 수 있는(machine readable) 데이터 공유가 확대되고, 데이터 매쉬업이 쉬워지며, 개인의 데이터 가공 능력을 거래할 수 있는 생태계 구현이 가능해질 것이다.

Ⅰ. 서론

인터넷과 스마트폰 이용의 확대로 촉발된 빅데이터는 광고, 마케팅, 건강, 의료, 보험, 금융, 재난관리, 범죄예방 등 거의 모든 사회, 정치, 경제, 문화 분야에서 관심을 끌고 있다[1-3]. 빅데이터의 여러 특성 중 가장 중요한 것은 여러 소스로부터 발생한 데이터를 결합하여 분석하는 것이다. 하나의 데이터 셋만 분석하는 것은 단순한 통계 분석과 다를 바가 없으며, 빅데이터 분석의 매력은 여러 데이터를 매쉬 업(mash up)하여 새로운 가치를 찾는 것이다. 따라서 빅데이터 활용의 성공은 대용량 데이터를 다루는 하드웨어 구축이나 분석 소프트웨어 도입에 있는 것이 아니라 분석에 필요한 가치 있는 여러 데이터를 확보하는 것에 의존한다[4].

그러나 타 기업이나 기관이 보유한 빅데이터를 원시(raw) 데이터 형태로 확보하는 것은 매우 어려우며 은행, 병원, 정부 기관 등이 구축한 데이터는 공유가 거의 불가능하다. 더욱이 향후 사물인터넷(IOT) 등의 확대로 데이터는 더 빠르게 대용량으로 발생할 것이므로 지금과 같은 제한적인 데이터 공유 방식이 아닌 효과적인 데이터 공유 방법이 필요하다.

본 고에서는 빅데이터 이용 확산을 위해 필요한 주요 이슈들을 정리하고 이들을 해결할 수 있는 방안의 하나로 Open Data Interface(ODI) 프레임워크를 소개한다. 2장에서는 최근의 정보통신 서비스 변화 방향과 "데이터 서비스"의 개념을 소개한다. 3장에서는 ODI의 요구사항을 파악하고, 4장에서 ODI의 개념을 소개하겠다.

Ⅱ. 정보통신 서비스의 변화

1970년대 인터넷이 소개된 이후 우리 사회는 정보 사회로 변화하기 시작했으며, 2000년대 스마트폰의 보급으로 정보통신 서비스는 이제 생활의 필수 서비스가 되었다. 정보통신 서비스란 간단히 정의하면 "원하는 데이터를 편리하게 제공하는 것"이라고 할 수 있다.

현재의 정보통신 서비스 구조는 〈그림 1〉과 같이 크게 두 개의 영역으로 나누어진다. 먼저 임의의 장치간에 데이터의 전달을 책임지는 데이터 네트워크(Data Network) 영역이 있고, 이렇게 얻은 데이터를 활용하여 이용자에게 원하는 결과를 제공하는 데이터 응용(Data Application) 영역이 있다.

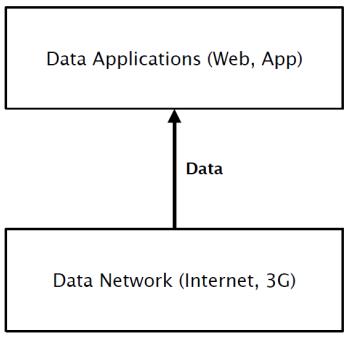

그림 1. 현재의 정보통신 이용 형태 (데이터 네트워크와 데이터 응용 두 영역으로 구성)

데이터 네트워크에는 인터넷과 전화망이 주로 사용되고 있다. 데이터 응용에는 검색, 전자상거래, 원격교육, 게임 등이 포함되며 현재 대부분의 서비스가 웹(web)이나 앱(app) 형태로 제공된다. 데이터 응용은 데이터 네트워크로부터 데이터를 전달받아야 하는데 현재는 응용 프로그램에서 필요한 데이터를 직접 프로그램을 통해 가져오는 방법으로 동작한다.

지금과 같이 응용 프로그램이 필요할 때마다 데이터를 찾아서 가지고 오는 방식은 최초 프로그램 작성에도 많은 시간이 필요하며 기능을 업데이트 하거나, 데이터 소스를 교체하는데 유연성이 부족하다. 예를 들어 날씨 안내 앱을 만들려면 기상 데이터를 찾아서 가공하는 작업을 앱 개발자가 직접 작성해야 한다. 현재 앱을 개발하여 수익을 내기 어려운 근본적인 이유는 데이터 확보와 가공에 많은 개발비가 들기 때문이다. 즉, 앱을 구성하는 사용자 인터페이스나 서비스 개발 자체에 시간을 집중할 수 없고 필요한 데이터를 적시에, 적절한 포맷으로 공급받는데 많은 개발 노력이 필요한 상황이다.

향후 빅데이터 활용 요구가 늘고, IOT, 센서망 등의 확산으로 데이터가 폭발적으로 늘어나게 되면 응용 프로그램이 직접 데이터를 찾는 방식보다, 응용에서 필요한 데이터를 적절하게 공급하는 Data as a Service (DaaS)가 중요한 역할을 하게 될 것이다. 즉, 응용에 필요한 데이터를 공급하는 "데이터 서비스 영역"이 새롭게 부각될 것이며 이러한 구조를 〈그림 2〉에 나타냈다.

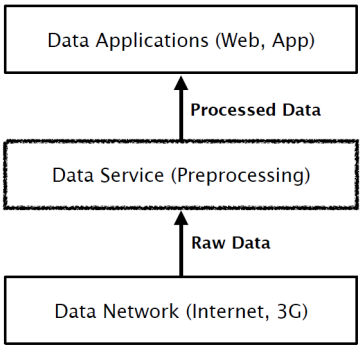

그림 2. 미래의 정보통신 이용 형태(데이터 서비스 영역의 부각)

지금까지는 데이터 서비스 영역에 해당하는 기능을 응용 프로그램에서 직접 구현해야 했다. 즉, 웹이나 앱 개발자가 데이터의 확보, 가공, 분석, 가시화 등을 모두 처리했다. 앞으로는 쉽게 앱을 개발할 수 있는 데이터 공급 체계가 필요하며 이를 다루는 "데이터 서비스" 영역에 대한 체계적인 준비가 필요하다.

데이터 서비스 영역은 웹이나 앱 프로그램이 필요로 하는 데이터를 표준적인 방법으로 제공하는 계층이라고 하겠다. 예를 들어 인터넷 초기에는 ftp로 파일을 주고 받던 것을 http 프로토콜을 도입함으로써 다양한 웹 서비스가 확산될 수 있었던 것과 유사하게, 미래에는 프로그램이 바로 처리할 수 있는 (machine readable) 데이터를 연결해 주는 데이터 서비스가 널리 요구될 것이다[5]. 데이터 서비스 영역은 향후 새로운 비즈니스 영역으로 중요시 될 것이며 기업 뿐 아니라 개인도 데이터 거래를 할 수 있도록 하는 기반 기술이 될 것이다.

본 고에서는 데이터 서비스 영역을 구현하는 한 방법으로서 ODI 방식을 제안하며 3장에서는 데이터 서비스 영역의 요구사항 즉, ODI가 해결해야 할 주요 이슈를 소개하겠다.

Ⅲ. 데이터 서비스 요구사항

2장에서 소개한 데이터 서비스 영역은 최근 관심을 끌고 있는 다양한 빅데이터를 현실적이며 효과적으로 활용할 수 있는 방안을 제공해야 하는데 이를 위해서 고려해야 할 주요 이슈는 다음과 같다.

3.1 원시 데이터 접근의 문제

최근 빅데이터 처리와 분석에 관하여 하둡(hadoop)과 같은 대용량 데이터 처리 솔루션이 많이 거론된다. 그러나 원시 빅데이터를 다루는 것은 최초의 빅데이터를 확보한 기관에서만 필요한 절차이다. 예를 들어 이동통신사가 보유한 방대한 통화 정보나 카드사가 보유한 거래 정보를 분석하는 것은 기업내 자체적인 빅데이터 분석에 해당한다. 그러나 이러한 자체 빅데이터는 외부로 공유할 수 없다. 2장에서 소개한 데이터 서비스 영역이 필요한 근거는 이러한 빅데이터는 직접 외부로 공유할 수 없다는 데에서 출발한다. 미래 빅데이터의 성공적인 활용은 데이터의 현실적인 공유 방법을 찾아야만 가능하다.

그런데 과연 기업이나 기관이 보유한 빅데이터를 공유할 수 있을까? 현재는 이것이 매우 어렵지만 기관간의 데이터 공유가 가능한 현실적인 방법을 찾아야 한다. 예를 들어 이동통신사 고객 정보 중에서도 공유될 수 있는 데이터가 있다. 어린이가 많이 다니는 길을 시간대별로 공개한다면 교통 안전관리 앱을 만드는데 도움이 될 것이다. 만일 어린이 인구 이동 통계에 대한 더 상세한 정보를 원하면 (예를 들어 분 단위의 정보나 연령대별 정보 등) 이는 유료로 제공할 수 있을 것이다. 이러한 기본적인 공익 데이터 공유에 대한 인식은 창의적인 서비스를 만들어 낼 것이며 이러한 생태계가 만들어질 수 있는 방법이 필요하다.

3.2 표준의 필요

현재까지는 데이터 공유에 관한 표준이 거의 없다. 파일로 데이터를 제공하는 경우 테이블을 정의할 수 있겠지만, 테이블 필

드 정의에 표준화된 것이 거의 없다. 웹 API를 사용하여 데이터를 읽는 경우에도 API 제공 기관이 일방적으로 정한 기준에 따라 프로그래밍을 해야만 한다[6].

현재 발생되는 빅데이터의 대부분은 일정한 포맷이 없는 비정형 (non-structured) 형태를 갖는데 (웹로그, 센서 정보 등) 이러한 비정형 데이터를 분석하려면 정형화되어야 한다. 비정형 데이터를 일단 정형화 하면 방대한 데이터가 생성되는데 이에 대한 표준화를 통해 데이터의 공유와 활용이 쉽도록 해야 한다. 특히 IOT가 성숙되면 센서나 장치들이 끊임 없이 데이터를 양산할 것이며 이의 효과적인 활용에 대비하여 데이터 측정과 관리에 대한 표준화가 필요하다.

향후에 공개되는 데이터가 늘고 이의 변형된 데이터들이 다시 쌓이게 되면 데이터 품질 보장에 문제가 발생할 것이다. 유효기간이 지나 의미가 없는 데이터의 품질은 지금과 같이 파일을 직접 액세스 하는 방식에서는 전적으로 프로그래머가 해결해야 할 문제로 남는다. 데이터 품질에 대한 표준적인 대책이 없으면 잘못된 데이터로 잘못된 분석결과를 얻는 일이 많게 된다.

3.3 데이터 공유의 요구

한 기업이나 기관이 보유한 데이터의 가치는 한계가 있다. 빅데이터의 장점은 여러 종류의 데이터를 융합하는 것이며 기업들은 서로 상대방의 데이터를 활용하고 싶어하지만 데이터를 서로 주고 싶어도 임의로 데이터를 제공할 수 없다. 데이터 관리에 대한 법적인 문제가 발생하기 때문이다. 또한 향후에는 실시간 데이터 분석 요구가 늘어날 것이다. 고객은 점차 빠른 반응을 원하기 때문이다. 그러나 실시간 데이터 공유 요구는 더 복잡한 공유 기준을 필요로 할 것이다.

위와 같은 이유로 기업이나 기관은 데이터 공유에 선뜻 참여하기 어려운 것이 현실이다. 기업 데이터의 어느 범위까지 공유해도 되는지 법적 기준이 없고, 수익배분 문제와 경쟁사와의 관계가 있다. 기업이 데이터 공유에 관심을 갖고 적극 참여하기 위해서는 궁극적으로 비즈니스 모델이 제시되어야 한다. 기업은 무료로 공개할 수 있는 데이터의 범위, 향후 유료화 할 수 있는 범위, 그리고 절대로 공유할 수 없는 데이터 공유 수준을 정해야 한다. 이러한 작업은 단일 기업 혼자서는 해결할 수 없고, 기업들이 연합하여 안을 제시하고 정부의 도움을 받아 데이터 공유 가이드라인을 만들어야 한다.

표준화된 데이터 공유를 통해서 중복적인 데이터 확보 비용을 줄이고, 분석 시간을 줄이고, 데이터 융합을 통한 창의적인 분석과 새로운 서비스를 찾아낼 수 있을 것으며 이는 국가적으로 매우 중요한 일이다.

3.4 데이터 마켓 생태계 요구

데이터 소유권 문제, 데이터 품질관리 문제, 법적 책임 소재 문제 등으로 인해 기업간에 원시 데이터를 직접 거래하는 것은 한계가 있으며 가공된 데이터 거래만 가능할 것이다. 예를 들어 기상정보 데이터가 통째로 거래되는 것이 아니라, 지역별, 시간대별로 제공되어야 하는 것과 같다.

향후 데이터 공유는 데이터를 거래할 수 있는 데이터 마켓을 조성하는 방향으로 발전하게 될 것이다. 데이터 마켓은 단순히 디지털화된 물리적인 데이터를 거래하는 것이 아니라 좀 더 목적에 맞게 가공된 데이터, 의미 있는 분석 결과, 시각적인 처리 (visualization), 또는 개인의 통찰력을 거래하게 될 것이다. 데이터의 거래는 궁극적으로 개인의 소득과 연결되어야 크게 발전할 수 있을 것이다. 빅데이터 활용은 최종적으로 앱이나 웹을 통해서 이루어질 것이며 기업과 개인은 자신의 데이터 가공 아이디어가 수익모델로 이어지는 것을 원할 것이다.

Ⅳ. Open Data Interface

4.1 ODI 개념

4장에서는 앞의 3장에서 논의한 데이터 서비스 영역의 요구사항들을 해결하기 위한 방안으로 오픈 데이터 인터페이스 (Open Data Interface: ODI)를 제안한다. ODI는 기본적으로 API 형태를 사용한다. 그러나 현재 사용되는 일반적인 웹 API와 달리 ODI는 데이터를 읽는 데만 사용되는 것이 아니라, 내가 가공한 데이터를 출판할 수 있는 채널을 제공한다[7]. 예를 들어 구글에서 API로 데이터를 다운로드 받은 후 이를 가공, 분석하여 가치 있는 새로운 데이터를 만들었다면 이를 어떻게 외부에 공개할 수 있을까? 현재 이렇게 새로 만든 데이터를 외부에 공개할 수 있는 방법은 새로운 웹 페이지를 만들거나 특정 앱을 만들어 보

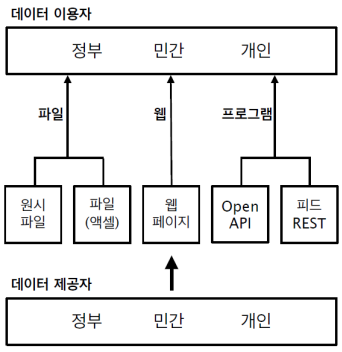

그림 3. 인터넷을 통한 현재의 데이터 액세스 방식

급하는 방법 밖에 없다. 그러나 ODI를 사용하면 웹 사이트나 앱을 만들지 않고도 "데이터 자체"를 공개할 수 있게 된다.

현재 인터넷을 통한 데이터 액세스는 원시 데이터, 정리된 파일, 웹 데이터, API, 피드(feed) 등의 방식으로 전달되며 이를 〈그림 3〉에 나타냈다.

현재의 데이터 액세스 방식을 빅데이터 시대의 데이터 공유 방법으로 확대하는데는 다음과 같은 문제가 있다. 먼저, 파일로 데이터를 받으면 이를 처리하는 작업을 모두 응용 프로그램 개발자가 수행해야 한다. 예를 들어 학교 식당 메뉴를 보여주는 앱을 개발하려면 식단 파일을 받은 후 이를 파싱하여 앱으로 보여주는 프로그램을 작성해야 한다. API로 데이터를 제공하는 경우도 결국 데이터를 읽고 처리하는 프로그램을 응용 프로그래머가 일일이 개발해야 한다. 지금과 같이 단품 서비스 중심의 앱에서 발전하여 향후에는 프로그램과 프로그램이 직접 데이터를 주고 받는 서비스가 늘어날 것이다. 예를 들어 검색, 추천, 예약 그리고 구매가 서로 연결되어 한 번에 이루어지는 앱을 만들려면 여러 사이트에서 제공되는 데이터 융합이 편리하게 처리되어야 한다.

〈그림 4〉에 ODI를 이용한 데이터 액세스 방식을 소개했다. 데이터 이용자는 ODI를 통해 데이터를 찾고 이용할 수 있다. ODI를 사용하면 쉽게 앱을 개발할 수 있으며 ODI의 일차적인 주요 고객은 바로 앱 개발자하고 하겠다.

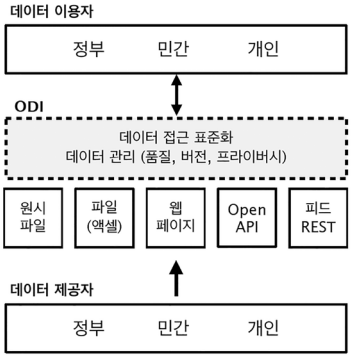

그림 4. ODI 기반의 데이터 액세스 방식

ODI를 사용하면 3장에서 소개한 데이터 서비스 영역의 여러 요구사항들을 해결하기가 용이해진다. 원시 데이터의 직접적인 접근 대신 가공된 데이터를 공유하기 쉬우며, 데이터 생성과 공유에 필요한 표준화된 방법을 만들 수 있고, 개인이 가공한 데이터의 거래가 가능해진다.

4.2 ODI 구성 예

〈그림 5〉에 ODI를 구현하는 예를 보였다. ODI는 기술적으로는 API를 확대한 형태로 구현되는데 이러한 방식은 프로그램이 직접 데이터를 읽을 수 있어야 한다는 조건 (machine readable)을 만족시키기 가장 좋은 방법이기 때문이다.

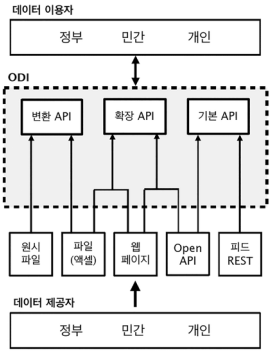

그림 5. ODI의 구현 예

〈그림 5〉에서 기본(primary) API란 현재 제공되는 웹 API를 그대로 사용하는 것을 말한다. 즉, ODI는 기존의 웹 API 프로그램 환경을 수용한다. 변환(conversion) API는 액셀 파일이나 표를 API화 하여 프로그램이 직접 테이블의 (행, 열)을 읽는 등, 파일을 API로 접근하게 해주는 기능을 말한다. 변환 API는 하둡 등을 이용하여 기초적인 데이터 처리와 분석을 한 결과를 공유하는 것을 포함한다. 확장(extended) API는 다른 소스로부터 얻은 데이터를 매쉬업하여 새로운 데이터를 만드는 것을 말한다. 확장 API를 이용함으로써 데이터 분석 결과를 재공개 하거나 그래프를 공유하는 것이 가능하다.

4.3 ODI 특징

ODI의 가장 중요한 특징은 누구나 자신의 아이디어를 구현하여 가치있는 데이터를 만들고 이를 남들이 사용할 수 있도록 한다는 것이다. 개인의 아이디어는 데이터 형태로 제공될 수도 있고 알고리즘 형태로 새로운 API에 포함시킬 수도 있다. 즉, 개인이 처리한 데이터를 출판(publishing) 할 수 있으며 요약 데이터나 그래픽 데이터로도 판매할 수 있게 된다. 궁극적으로 데이터의 읽기와 쓰기가 가능해진 양방향(bi-directional) API 풀을 제공하게 된다. ODI를 통해 개인의 소규모 거래(micro selling)가 가능해진다.

또한 ODI를 이용하면 다른 기업이나 개인 등이 공개 또는 가공한 데이터를 공유하여 사용할 수 있기 때문에 데이터의 중복 생산을 막고 또 이로 인해 발생할 수 있는 트래픽의 양을 줄일 수 있다[8].

현재 빅데이터를 보유하고 있는 곳은 정부와 대기업, 통신사, 포털, 인터넷 관련 기업 등이다. 빅데이터 중에는 공공성을 갖

는 데이터가 존재하며 이들은 기업의 전유물로만 볼 것이 아니라 공익의 목적에 사용될 수 있어야 한다. 또한 빅데이터를 확보할 수 없는 중소기업과의 협력도 가능할 것이다.

V. 결론

본 고에서는 빅데이터 시대에 필요한 효과적인 데이터 공유 방안으로 ODI 프레임워크를 제안하였다. ODI 프레임워크를 이용하면 기관이나 기업이 소유한 데이터를 원시 데이터 형태가 아니라 제한적인 형태로 공유할 수 있으며 개인의 아이디어를 판매할 수 있는 데이터 마켓을 구축할 수 있을 것으로 기대한다. ODI 프레임워크가 정착하려면 데이터 액세스에 대한 기술적인 표준 제정과 함께, 데이터 공유와 이를 통한 다양한 앱 개발이 가능한 것을 보여줄 수 있는 시범 서비스 운영이 필요할 것이다.

참고문헌

[1] Thomas Davenport and Jeanne Harris, Analytics at Work, 2010.

[2] Thomas Davenport, big data@work, Havard Business Review, 2014.

[3] 김화종, 데이터 사이언스 개론, 홍릉과학출판사, 2014.

[4] Bill Franks, Taming the Big Data Tidal Wave, Wiley, 2012.

[5] White House, Executive Order: Making Open and Machine Readable the New Default for Government Information, 2013. 5.

[6] http://www.programmableweb.com

[7] Hwa-Jong Kim, Seung-Teak Lee, Yi-Chul Kang, "The Open Data Interface(ODI) Framework for Public Utilization of Big Data", DATA ANALYTICS 2012 (Barcelona).

[8] 정보화진흥원, 빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼, 2014. 5.

약 력

김 화 종

1982년 서울대학교 공학사
1984년 KAIST 공학석사
1988년 KAIST 공학박사
1988년~현재 강원대학교 IT대학 컴퓨터정보통신공학과 교수
관심분야: Computer Network, Data Science