

미래 인터넷 환경의 서비스를 위한 빅데이터 기술 전망

Service WG
이동만 (KAIST)

최근 널리 쓰이고 있는 용어인 빅데이터(Big Data)는 기존의 방식으로는 저장, 관리, 분석하기 어려울 정도로 큰 규모의 데이터를 의미한다. 현재는 대규모의 데이터로부터 저렴한 비용으로 원하는 가치를 추출하기 위해서 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처를 지칭하는 말로 그 의미가 확장되고 있다. 이는 전 세계적으로 일어나는 정보의 폭발적인 증가 추세와 맞물려 있다. IDC에 따르면 2011년에 전 세계적으로 생성된 디지털 정보량은 1.8 Zetabyte(1조 8천억 Gigabyte) 정도로 추정된다. 또한 스마트폰의 확산, 소셜 네트워크 서비스(SNS)의 사용 확대, 기기 간 통신(Machine-to-Machine Communication) 기술을 이용하는 M2M 센서의 확대 보급 등으로 인해 앞으로 정보의 빅데이터 추세는 더욱 가속화될 전망이다.

빅데이터로 간주되는 데이터의 양은 기업의 특성과 규모에 따라 상대적일 수 있지만 공통적으로 기존의 데이터 관리 및 분석 체계로는 감당할 수 없을 정도로 많은 양으로 인식되며, 이로 인해 빅데이터를 부정적인 현상으로만 생각할 수도 있다. 그러나 빅데이터를 잘 활용하면 오히려 여러 분야에서 달성이 불가능하던 일들이 실현 가능한 일들로 바뀔 수 있다. 예를 들어, 과거에 IBM은 캐나다 의회에 있는 수백만 건의 문서를 활용하여 영어-불어 번역 시스템 개발을 시도하였다가 실패하였으나, 구글(Google)은 자체 개발한 빅데이터 처리 시스템을 통해 동일한 방식이지만 훨씬 많은 수억 건의 자료를 활용하여 50개 언어 간의 자동번역 시스템 개발에 성공하였다. 그 외에도 아마존(Amazon)은 대규모 사용자 정보 처리를 통해 제안되는 추천 시스템에서 전체 매출의 30%가 발생하고 있으며, 페이스북(Facebook)은 7.5억 명의 가입자 정보를 실시간 처리하여 광고주가 원하는 광고대상을 즉각 제시할 수 있다. 이와 같이 빅데이터의 효율적인 활용은 산업 부문별로 약 0.5~1% 정도의 생산성을 증가시킬 것으로 기대되며, 맥킨지에 의하면 미국의 의료 부문에서는 연간 3,300억 달러, 유럽의 공공 부문에서는 2,500억 유로를 절감할 수 있는 것으로 파악된다.

현재 빅데이터 기술은 기업의 필요에 의해 경영정보학의 수집, 정리, 분석, 활용의 네 단계에 따라 각 단계별로 연구가 활발하게 수행되고 있다. 데이터 수집 단계에서는 웹, SNS 데이터, 시스템 로그 데이터와 같이 기업의 의사결정에 필요한 데이터를 대량으로 수집하는 기술을 연구한다. 야후(Yahoo)는 Chukwa를 이용하여 분산된 서버로부터 Hadoop 파일 시스템으로 저장, 로깅하고 MapReduce 기반의 중복 제거 기능을 지원한다. 페이스북은 분산 서버 데이터를 중앙 집중 서버로 전송할 수 있고 다양한 프로그래밍 언어를 지원하는 로그 수집 시스템인 Scribe를 제공한다. Cloudera에서는 분산 데이터를 개인화하여 저장할 수 있는 Flume 서비스를 제공한다. 그 외에도 LDspider, DARQ, SQUIN, C-SPARQL과 같이 Linked Data에 대해 질의를 수행하고 데이터를 획득할 수 있는 방법들이 존재한다. 데이터 정리 단계에서는 대량의 데이터를 효율적이고 접근이 용이하도록 저장하기 위한 기술을 연구한다. 수천 대 규모의 단일 클러스터를 구성할 수 있고 특정 서버의 장애를 자동으로 감지/복구할 수 있는 대용량 분산 파일 시스템인 Hadoop File System, 비 관계형 데이터베이스 구조와 분산 scale-out 식의 확장성을 갖는 NoSQL 데이터베이스 시스템 등이 활용된다. 데이터 분석 단계에서는 대량의 데이터에 대해 실시간 분석 수행하기 위해 데이터 수집기에 분석 및 처리 솔루션을 탑재하거나(예: Esper), 별도의 분석 클러스터를 구성한다(예: Gruter, ClouStream, Yahoo S4, Twitter Storm, FacebookPuma). 또한 구글이 고안한 빅데이터 생성 및 처리를 위한 프로그래밍 모델인 MapReduce 방식을 활용할 수 있다. 의사결정 단계에서는 목적과 데이터의 종류에 따라 적합한 뷰를 제공할 수 있도록 데이터를 표현하는 다양한 방식이 연구/개발되고 있다.

빅데이터에서 중요한 세 가지 문제(3V)는 데이터의 양(Volume), 데이터의 입출력 속도(Velocity), 데이터의 다양성(Variety)이며, 이 세 가지 요소의 급격한 증가는 빅데이터를 활용하는 측면에서 도전적인 문제가 되고 있다. 현재까지 연구된 빅데이터 관련 기술은 Hadoop, NoSQL, MapReduce와 같이 대규모 정보의 저장, 검색, 통계 위주로 발전하고 있으며, 데이터의 양과 입출력 속도의 증가 문제를 해결하는 데 주로 활용되고 있다. 그러나 현재까지 연구된 기술은 여전히 빅데이터의 다양성 증가 문제는 해결하지 못하고 있다. 특히 미래

인터넷 환경에서는 다양한 분야의 빅데이터를 융합하여 유용한 서비스를 창출/제공할 것으로 기대되므로, 빅데이터의 다양성 문제를 해결하는 것이 시급하고 필수적이다. 이를 해결하기 위해서는 반정형, 비정형, 실시간 데이터를 처리하고 분석할 수 있는 기술의 개발이 필요하며, 다양한 분야의 빅데이터를 의미적으로 연계하여 지식 자산화할 수 있어야 한다. 이를 통해 상호 연계된 데이터를 실시간으로 활용하여 사용자 맞춤형 서비스를 언제 어디서나 쉽게 생성할 수 있을 것이다.

현재까지 연구된 빅데이터 기술들은 기술의 초기 단계 및 도입 단계에 있기 때문에 미래 인터넷 환경에서의 서비스를 원활하게 제공하는 데에는 한계가 있으며, 빅데이터 기술 영역 전반에 대한 성숙이 필요하다. 빅데이터 자체를 다른 서비스에서 가져다 쓸 수 있는 단위 서비스화하기 위해서는 쿼리 속도가 더 개선되어야 하고, 데이터 자체의 양은 매우 많으나 데이터들 사이의 상호 연관성이나 메타데이터는 충분하지 않으므로 기존 데이터를 Linked Data로 활용할 수 있도록 하는 기술의 연구가 필요하다. 또한 다양한 형태와 요구사항을 갖게 될 미래 인터넷 서비스들이 각자 필요로 하는 데이터의 범위와 연관성이 달라질 것이므로, 빅데이터로부터 즉시적으로 필요한 양과 형태의 데이터를 추출해낼 수 있어야 하며, 비디오와 같은 비정형 데이터와 실시간으로 센서로부터 생성되는 raw data를 최대한 빨리 활용할 수 있도록 하는 서비스 인프라스트럭처 관점의 연구가 진행되어야 할 것이다.