빅 데이터의 핵심은 기계 학습이다.

얼마 전부터 빅 데이터라는 용어가 유행처럼 번지고 있다. 사실 유행처럼 번지고 있다기 보다는, 이미 수많은 IT인들 간에 회자되어 이젠 식상해지는 수준이다. 얼마 전까지만 해도, 빅 데이터가 들어간 연구 제안서는 비교적 용이하게 선정되었는 데, 요즘은 오히려 다소 비판을 받고 있는 실정이다.

이제는 데이터 과학(Data Science)이라는 용어도 심심치 않게 오르내리기 시작한다. 미국의 컬럼비아 대학교에서 처음으로 "데이터 과학 개론"이라는 수업을 열었고, 뉴욕대학교와 노스 캐롤라이나 주립대학교도 데이터 과학자 인증 과정과 학위 과정을 제공하기 시작했고, 시애틀의 워싱턴 대학교는 아예 빅 데이터 박사 과정을 신설하였다. 데이터 과학이라는 용어도 남발되는 경향이 생겨서, 공신력 없는 데이터 과학 자격증도 우후죽순처럼 나오는 상황이다.

과거 인공지능이나 최근의 로봇에 대한 연구도 그랬지만, 특정 용어가 식상해지거나 비판의 대상이 되는 이유는, 만족할만한 결과를 내기 어려운 기술적인 어려움이 가장 큰 이유겠지만, 스스로가 역량이 되지도 않으면서 유행하는 용어를 사용하며 연구를 하겠다고 제안한 사람들도 이러한 경향에 다소 기여했을 거라고 본다.

빅 데이터라는 이름을 단 일부 제안서들을 보면, 클라우드니 하둡이나 NoSQL, 컬럼기반 데이터베이스 등을 언급하고, 통계 또는 수치해석 프로그램인 R이나 매트랩 등을 말하며, 마치 그런 오픈소스 내지 상업용 제품들을 이것저것 가져다 붙이면 바로 뭔가 굉장한 결과가 나올 거처럼 서술해 놓았다. 하지만 결과는 기대보다 안좋았고 그러한 경향이 고객과 평가자들을 실망시켰을 것이다.

그러나, 빅 데이터는 기술이라기 보다는 현실이다. IBM의 통계에 따르면, 하루 250경 바이트의 비정형 정보, 매달 10억 여개의 트윗, 매달 300억 여개의 페이스북 메시지가 생성되고 있다고 한다. 그러나 이 통계는 이 글을 쓰는 지금도 계속 증가하여 과거의 측정값을 무의미하게 만들 것이다.

이러한 상황에서 사람들 또는 기업들이 빅 데이터에 관심을 가지는 이유는 결국 그 활용 사례 때문이다. 빅 데이터의 성공 사례들은 매우 다양해서 체계적으로 정리해서 열거하기 어려울 정도이다. 이러한 성공 사례들에서 빅 데이터는 주로 미래 예측, 상황 분석, 분위기 측정, 이상 감지 등을 통해 품질 개선, 공정 개선, 신상품 개발, 고객 행동 패턴 분석, 부정 행위 판별 등에 활용되고 있다.

이러한 활용 사례를 가만히 들여다 보면, 결국 빅 데이터에서 사람들이 원하는 것은, 몇 가지 특수한 경우를 과감히 무시한다면, 세 가지 과정으로 나눌 수 있다. 첫 번째로 지금 이 시점에도 빠르게 증가하는 대용량이며, 복잡하게 상호연결되어 있는 정형/비정형의 데이터를 효과적으로 분산 저장하고 필요할 때 빠르게 검색할 수 있게 하는 것이다. 두 번째로는, 이렇게 저장된 데이터

에 대해 빠르고 지능화된 알고리즘이 데이터 내부의 변수들 또는 숨어있는 변수(hidden variable)와의 알려지지 않은 연관 관계 더 나아가 인과 관계를 찾아내는 것이다. 마지막으로 세 번째는 이렇게 찾아낸 연관 관계 또는 인과 관계를 토대로 데이터 과학자와 비즈니스 전문가들이 사업적 가치가 있는 것을 찾아내는 것이다.

여기서 재미있는 점은 최근 빅 데이터에 대해 크게 떠드는 사람들 중 다수가 첫 번째 과정에 있는 사람들이라는 점이다. 그들의 주장에 따르면, 궁극적으로 데이터는 당연히 클라우드에 집어넣어야 하고, 맵리듀스를 하기 위해 하둡을 설치해야 하고, 궁극적으로는 NoSQL을 따라야 한다. 그리고 나서, 데이터가 채워지기 시작하면 그 다음은 어떻게 해야 하냐라는 질문에는 R을 얘기하고 MatLab을 얘기하고, 또는 Python을 언급하지만 그 이상 구체적인 것은 말하지 못한다. 물론 이 첫 번째 과정은 매우 중요하다. 그러나 어떻게 보면 데이터의 용량에 따라, 당연히 준비되어야 할 기본적인 과정이고, 시작일 뿐인 것이다. 그것이 빅 데이터 성공을 위한 셀링 포인트가 되서는 안되는 것이다.

두 번째 과정에 전문적인 분들은 사실 과거 기계 학습, 통계, 데이터 마이닝 쪽의 전문가들이다. 재미있는 건 이러한 사람들의 대부분은 빅 데이터에 대해 크게 떠들지 않거나, 오히려 다소 비판적으로 보고 있다는 점이다. 그 이유는 이제서야 빅 데이터라고 떠들지만 사실 그를 위한 학습 및 추론 알고리즘은 옛날부터 연구되어 왔기 때문이다. 예를 들면 맵리듀스는 데이터를 집합(set)이나 중복집합(multiset, bag) 또는 다른 집적(aggregation) 함수로 요약하되, 여전히 데이터에 대해 수행하고자 하는 추론 작업은 가능하도록 충분통계량(sufficient statistics)을 유지하는 것으로 볼 수 있다. 점증적 학습(incremental learning)이나 분산 데이터 마이닝 분야에서는 그리 새로운 얘기는 아닌 것이다.

사실 현실적으로 가장 중요한 부분은 세 번째 과정일 것이다. 그런데 이 세 번째 과정은 과거의 사례들에 대한 체계적인 연구가 존재한다면 비교적 용이한 부분일 수 있다. 실제로 국내의 많은 업체들이 세계적인 기업들에 의해 검증된 길을 따라가는 미투(me too) 전략을 구사하는 것도 이러한 경우이다.

만일 그러한 사례가 없다면 결국 비즈니스적인 결정이 필요할 것이다. 그러나 그러한 결정을 내리는 것은 쉽지 않다. 때로는 범죄자의 수가 줄어든 이유는 낙태 때문이라는 식의, 이른바 통념을 거스르는 결정이 내려질 때도 있을 것이다. 영국의 경우, 통계를 오용한 어느 소아과 의사 때문에, 유아 돌연사로 자식 둘을 잃은 어느 여성이 뮌하우젠 증후군 환자로 몰려서 억울한 감옥살이를 하다가 결국 알코올중독으로 죽기도 했다. 그 외에도 주어진 데이터와 알려진 연관 관계에서 올바른 판단을 내리는 것은 매우 중요함을 보여주는 사례들은 허다하다.

그러나, 궁극적으로 이 세 번째 과정은 기술이라고 보다는 결국 여전히 사업적인 감각이나 운의 영역일 수도 있다. 감각이나 운이란 표현의 의미는 결국 여전히 기업의 의사결정자나 관련 실무자는 자신이 처한 불확실한 상황에 대해 도전해야 한다는 것이다. 이러한 의사 결정을 했을 때, 과연 효과가 있을 것인가 하는 질문에 대해 그들은 소위 빅 데이터가 자동으로 해결해 주었으면

하고 바랄 것이다. 그러나, 세상 일이 그러하듯이 여전히 미래는 불확실성(uncertainty)으로 남아 있고, 앞으로도 여전할 것 같다.

그럼에도 불구하고 두 번째 과정은 여전히 중요할 것이다. 두 번째 과정은 제대로 수행된다면, 불확실성을 아예 없앨 수는 없을지 몰라도, 상당히 줄여줄 것이기 때문이다. 그리하여, 의사 결정을 하는 기업 관계자나 이해 당사자가 좀 덜 위험한 결정을 할 수 있도록 지원해 줄 것이다. 그런 의미에서 앞으로의 빅 데이터 논의는 무엇을 설치하고 구성할지 보다, 데이터에서 진정으로 무엇을 어떻게 찾아야 할지에 대한 보다 근본적인 논의가 되었으면 한다.