빅데이터의 중요성 : 미 국세청의 사례

파울로 레오카디오
Phd. Paulo Leocadio

인터넷 시대에 빅데이터(Big Data)란 기존의 데이터베이스 기술의 경계를 넘어 데이터세트에 적용되는 개념이다. 빅데이터는 컴퓨터와 리포지토리(repository)의 힘을 극도로 개별화된 정보의 무더기와 결합하여 현대 생활과 인간 행동의 모든 측면에 대한 새로운 시각을 제공한다. 우리는 매일 약 250경 바이트의 데이터가 생성되는 세상에 살아가고 있다. 전 세계 데이터의 90%는 지난 2년 이내에 만들어졌다. 규제당국에 있어서는 엄청난 과제와 동시에 거대한 기회가 주어지는 세상이기도 하다.

데이터 과학자, 정책 입안자 및 세금 전문가들은 빅데이터 메커니즘, 도구 및 솔루션을 사용하여 조세제도를 더욱 잘 연구하고 개혁할 수 있는 방안을 찾고 있다. 정책, 우선순위 및 전략 구축에 빅데이터 분석을 통한 정보를 활용할 수 있는 접근성이 높아진 가운데, 현재 미국 지도층 사이에서는 포괄적인 조세개혁 계획에 대한 토의가 이뤄지고 있다.

국세청(IRS)과 같은 정부기관이나 공공기관이 수집하는 데이터의 규모 및 상세내용은 유의미한 것이어야 하며, 또한 사용자 기반의 현실을 대표할 수 있는 것이어야 한다. IRS는 미국 및 미국 시민에게 있어 떼려야 뗄 수 없는 존재이다. 생활 속에 함께하며 인생의 결정에, 심지어는 일상의 루틴에도 영향을 준다. IRS는 대부분의 국민 및 모든 종류의 사업에 대한 데이터를 수집하고 조직할 수 있다. 세금 데이터의 경우, 거의 모든 미국 시민 및 법인은 세금을 지불하고 다양한 세금 정보를 보고할 책임이 있다. 이는 자신에 대한 상당한 양의 정보 공개로 이어진다.

이는 곧 수집, 조직, 관리 및 분석할 수 있는 데이터의 양이 상상할 수 없을 정도로 막대하다는 것을 의미한다. 이것이 곧 빅데이터이다.

IRS는 매년 2억 5천만 건 이상의 소득신고를 받고 처리한다. IRS가 연간 4,500억 달러(약 557조 원) 규모 이상으로 추산되는 택스 갭(tax gap)으로 인해 고군분투하는 가운데, 예산 삭감 및 인력 축소는 IRS의 역량에 타격을 주고 있다. 보다 스마트한 업무 방식이야말로 IRS의 범죄수사국(Criminal Investigation Division)에서 규명하는 세금 포탈 및 탈세에 대처하기 위한 도구 및 효율성을 확대할 수 있는 해결책이다.

빅데이터는 그 특징(예: 규모, 속도, 다양성, 정확성)으로 인해 다양한 데이터원으로부터 다양한 포맷으로, 각기 다른 업데이트 주기를 가지고 유입 및 저장되며, 보관을 위한 대규모 저장 공간이 필요하다. 필요한 방법, 기법 및 기술이 알려지면서 빅데이터는 조세 포탈 분석의 판도를 바꾸고 있다. 데이터베이스 속 지식발견 과정에서 분석을 통한 데이터마이닝이 사용되며, 이 과정에서 예측적 태스크(predictive task) 및 서술적 태스크(descriptive task)가 진행된다.

세금 포탈 조사에서는 데이터 마이닝을 통해 다량의 데이터를 분석하여 데이터세트에서 인식되지 않고 숨어 있는 패턴을 찾아낸다. 이 때 패턴을 찾기 위해 통계적 분석 및 데이터베이스 기술이 사용된다.

예측적 태스크는 머신 러닝 및 관련 기술을 사용하여 데이터 마이닝을 통해 얻어낸 관측 값 각각에 대한 예측을 하기 위해 사용된다. 예측에는 독립변수와 종속변수 사이의 관계를 검증하기 위해 회귀분석이 사용된다. 금융의 복잡성으로 인해 빅데이터가 제공하는 변수의 규모가 충분해야만 보다 정확한 예측이 가능하다. 이러한 분석에 사용되는 통계적 기법으로는 선형 회귀(include linear regression), 다변량 선형 회귀(multivariate linear regression), 비선형 회귀(non-linear regression), 다변량 비선형 회귀(multivariate nonlinear regression) 및 더욱 복잡한 로지스틱 회귀분석, 의사결정트리(decision trees), 신경망(neural networks) 등이 있다. 조세 포탈의 적발 또는 예방에 적절한 데이터 마이닝의 더욱 복잡한 예측적 기술로는 규칙베이스 퍼지추론, 유전 알고리즘, 베이지안 신뢰 네트워크(Bayesian belief networks), 퍼지 신경망(fuzzy neural networks) 등을 예로 들수 있다.

연관규칙 및 클러스터 분석 등을 포함하는 서술적 태스크는 분석을 통해 데이터를 서술한다. 이러한 태스크는 의심스러운 것으로 분류될 수 있는 행동 모델(또는 거래 모델)을 정립하는 데 사용될 수 있다. 서술적 태스크의 종류로는 다계층간 연관규칙, 다중차원 연관규칙, 정량적 연관규칙 등 연관규칙 분석이 있다. 연관규칙 알고리즘은 포탈의 가능성이 있는 상황을 묘사하는 법칙을 생성한다. 클러스터 분석은 데이터를 수집하여 관련 하위 패턴으로 분류한다. 이는 금융 사기 등을 적발 또는 예방하는 데 사용될 수 있는 패턴들을 발견하기 위해 사용된다.

IRS의 사기 적발 활동 등에 사용되는 복잡한 대규모 분석은 빅데이터 혹은 다양한 데이터원의 사용을 필요로 한다. 사기 행위 적발을 위해 시행되는 감사에는 대규모 내·외부 데이터세트(인구통계자료, 납세자 또는 법인 프로필, 이전 기록물, 콜센터 데이터, 감사 내역 등)가 모두 사용된다. 분석되는 데이터 종류로는 수 년에 걸친 과거 데이터 및 외부 데이터 등을 들 수 있다. 데이터의 규모와 다양성이 매우 높기 때문에 빅데이터에 사용되는 분석도구 및 데이터 과학자의 도움 없이는 분석이 어려울 수 있다.

IRS가 스파이더(spider, 자동화된 컴퓨터 프로그램)를 사용하여 SNS 사이트를 검토한다는 풍문이 있다. 또한 몇몇 보도에서는 IRS가 전화추적기술[예: 이동통신기지국 시뮬레이터인 스팅레이(stingray)]을 채택할 가능성을 시사하기도 했다. 이에 더해 IRS는 비교적 전통적인 기술[예: NRP 및 개별마스터파일(Individual Master File) 데이터베이스]을 사용하여 상당한 규모의 데이터를 보관하기도 한다. 이러한 소식통 및 보도의 정확성과 별개로 한 가지 확실한 결론은 IRS가 수많은 데이터세트에 대한 접근권을 가지고 있다는 것이다.

IRS는 이러한 데이터세트를 교차참조 및 마이닝하여 패턴 인식 알고리즘을 실행하고, 이를 통해 추이를 파악하여 데이터 내의 관계를 이해하고자 한다. 이러한 노력의 일환으로 IRS는 여러 선진 기술 및 도구(이상감지, 어드밴스드 클러스터링, 신경망 등)를 사용하고 있으며, 그 목표는 IRS 산하의 부서(division)간 사건 선정 및 협력의 수준을 증진하는 것이다. 데이터 분석 및 예측적 감시 활동은 IRS가 세금 신고에 있어서의 이상을 감지하고, 더 넓은 시야를 가지고 탈세를 파악할 수 있도록 도움을 준다.

회계 및 세법 직군에서 빅데이터와 분석은 자동화와 연관되어 있다. 데이터 관리 및 처리 능력을 컴퓨터에 맡기는 것은 곧 숫자 분석, 모델 구축, 개별 분석 시행 등에 필요한 수작업이 줄어든다는 것을 의미한다. 이는 세금 전문가의 호기가 끝났음을 의미하는 것이 아니다. 반대로 새로운 시작, 신선한 기회, 새로운 지식, 그리고 이런 기계를 활용하여 일할 수 있는 전문가의 중요성 확대를 의미한다.

데이터 과학자, 프로그래머 및 세법 석사를 갖춘 법조계 전문가(LLM)는 빅데이터를 통해 찾은 막대한 정보에 올바른 질문을 던지고, 알고리즘과 데이터 쿼리를 통해 얻은 피드백을 해석하고, 미래 정책 개발을 위한 지침을 제공하기 위해 필요하다.

회계 직군의 경우, 이 분야를 더욱 깊이 탐구하고 싶어하는 전문가는 조세법 석사를 고려해볼 수 있다. 앞서 언급한 사례와 마찬가지로 이 분야의 전문성 또한 수집·관리·조직되는 데이터를 보관하고 가용한 기술적 도구를 통한 분석 잠재력을 최대한으로 활용하기 위해 중요하다. 정책 입안자들이 빅데이터를 바탕으로 조세 개혁에 있어 보다 과학적인 방법을 사용할 수 있게 된 만큼, 새로운 정보의 접근성, 의미 그리고 유용성을 확보하기 위해 세법 및 회계의 전문가는 앞으로도 계속 필요할 것이다.

빅데이터 및 이를 바탕으로 얻을 수 있는 분석적 능력이 다른 업계에서도 큰 영향을 미치고 있음은 이미 드러나고 있다. 인터넷의 시대에, 빅데이터의 영향으로부터 자유로운 21세기의 주체(사람, 기업, 정부)는 거의 없다. 세금 처리, 정책 및 관행 또한 예외가 아니다. 납세자들은 세무전략 및 세법 준수를 보다 원활히, 그리고 자동적으로 할 수 있는 솔루션과 기회를 찾고 있다. 정부 또한 마찬가지로 미국 세법의 집행 및 개혁을 위해 빅데이터를 수집, 조직 및 사용할 수 있는 새로운 방안들에 투자하고 있다. 현대 기술의 분석력과 인간 창의성의 결합은 조세 부문에 있어서 새로운 시대가 열렸음을 의미한다.

지난 10년 동안 빅데이터 분석에 들인 IRS의 투자는 국제적으로 이루어지는 세금 추징 관련 프로그램(국가의 자금지원 프로그램, 문서 유출 등)과의 협력을 통해 국제적 세금 추징 등의 부문에서 좋은 결실로 이어질 것이다. 정보 보고 및 공유협약 등은 세금 관련 정보 측면의 글로벌 협력에 있어 중요한 구조적 변화를 이끌어냈다. 그렇기 때문에 새로운 이니셔티브 또한 더욱 강화되어 나갈 것이다("J5"로 불리는 글로벌조세동맹(Joint Chiefs of Global Tax Enforcement)이 좋은 예다).

이러한 투자로 인해 만들어진 긍정적인 결과물은 발전의 기회가 무르익은 다른 부문에 집중할 수 있는 새로운 노력의 길을 열어 줄 것이다. 누구나 쉽게 떠올릴 수 있는 좋은 예는 바로 암호화폐다. 암호화폐 관련 과세정책은 아직 잘 알려져 있지 않으나 그 범위가 광대할 가능성이 높다. IRS의 보고서에 따르면 2013-2015년의 기간 동안 비트코인 등의 암호화폐로 발생한 소득을 신고한 납세자의 수는 1천 명이 되지 않는다. IRS는 미국에서 서비스를 제공하는 다양한 거래소에서 나타난 활동을 통해 새롭게 취득한 데이터를 대상으로 적극적인 마이닝을 실시 중이다.

IRS와 조세포탈과의 싸움에는 새로운 미래가 열렸다. 여러 가지 과정, 도구 및 노력 중에서 IRS는 빅데이터 분석을 선택하여 수용했고, 현재 표면에 드러나 확인한 것은 빙산의 일각에 불과하다. 매년 조세포탈 적발률이 400% 증가하고 다른 금융범죄로 인한 소득 적발률이 1,000% 이상 증가하고 있는 것으로 밝혀진 가운데, IRS는 빅데이터, 빅데이터 기술 및 그 도구에 더욱 많은 것을 맡길 것으로 보인다.