AI 시대 통계데이터

Column

통계데이터허브국장 **송영선**

챗GPT로 대변되는 생성형 인공지능(AI)은 엄청난 양의 데이터를 사용한다는 특징이 있다. 최근 AI와 관련하여 환경, 저작권, 딥페이크, 윤리 등 다양한 문제가 제기되고 있는데, 그중에서 데이터 처리 과정이나 데이터 자체의 한계로 인해 그 결과가 편향적으로 나오는 문제가 특히 눈에 띈다.

편향성 문제와 관련된 사례는 우리 주변에서 어렵지 않게 찾아볼 수 있다. 예를 들면 최근 영국 내무부는 비자 승인 처리 업무에 AI를 활용하였는데, 백인 인구 비율이 높은 나라의 국민보다 비백인 인구 비율이 높은 나라 국민의 비자 신청 심사가 별다른 이유 없이 더 오래 걸렸다. 또한 페이스북 AI가 흑인 남성을 영장류로 분류하고, 구글 AI가 흑인 남성을 고릴라로 분류하는 것과 같은 문제가 발생하였다.

데이터의 편향성 문제는 AI의 품질과 직결되기 때문에 이 문제를 해결하려는 노력이 다방면으로 이루어지고 있다. 그중에서도 특히 국가통계로 눈을 돌리면 이를 해결할 실마리를 일부라도 발견할 수 있을 것이다.

통계청은 고품질의 국가통계를 생산하기 위해 통계학을 기반으로 삼아 기획에서 공표까지 전 과정을 철저하게 관리하여 모집단의 추정이 가능하도록 하고 있다. 이를 통해 대표성, 타당성과 신뢰성 있는 통계와 통계자료(통계생산에 사용된 원자료)를 생산한다. 유엔통계위원회에서는 과학적 근거에 기초하여 산출된 통계가 “데이터기반행정”의 “데이터”라고 이야기한다.

이런 의미에서 AI가 직면한 문제를 해결하기 위해 국가통계로 눈을 돌리기를 권한다.

우리나라의 국가통계는 2024년 7월 31일 현재 434개 통계작성기관이 고용, 인구 등 다양한 분야에서 1,336종을 생산하고 있으며, 통계생산에 사용된 데이터도 함께 서비스하고 있다. 이렇게 서비스되는 통계자료의 원천은 과거에는 표본을 대상으로 하는 현장조사가 대부분이었으나, 지금은 조사자료 이외에 행정자료, 카드자료, 통신자료 등 다양한 자료를 활용해서 국가통계를 작성하고 있다.

또한 통계청은 통계 및 행정 전수자료를 기반으로 작성된 인구, 가구, 주택, 기업, 취업활동, 아동가구, 청소년 등 분야별 모집단 자료인 통계등록부 7종을 작성하여 제공하고 있다. 이를 기반으로 통계이용자들은 전국 13개의 통계데이터센터를 방문하여 다른 데이터세트와 연계함으로써 부가가치를 높일 수 있다.

앞서 열거한 데이터가 얼마나 가치가 있는지를 확인해 보겠다. 먼저 연금통계는 11종의 각종 연금자료(국민, 직역, 기초, 퇴직 등)를 통계청이 보유한 인구, 가구 및 주택통계등록부 등과 연계하여 작성하였다. 또한 기업통계등록부는 통계청의 전수조사 자료, 국세청 자료, 건강보험 자료, 고용보험 자료 등 15개 기관의 자료를 연계하여 만들었다. 1종으로 표현되어 있지만 이런 통계(등록부) 하나가 엄청난 양의 데이터를 정제해서 만든 고품질 자료인 것이다.

그러나 이렇게 다양한 통계(통계자료) 및 등록부 자료를 제공한다고 하여도 AI 현장에서는 여전히 고품질의 데이터가 부족할 것이다. 이에 국제사회는 그 대안으로 원자료와 자료 형태, 합계나 평균 등 통계적 추론값은 유사하나 응답자의 정보가 노출되지 않도록 합성된 자료인 재현자료(synthetic data)에 주목하고 있다. 통계청도 기업통계등록부, 일자리 행정통계의 재현자료 베타서비스를 2023년과 2024년에 실시하였고 향후 통계청 자료를 재현자료로 생성하여 서비스할 계획이다.

AI가 우리 사회를 완전히 다른 세상으로 이끌 것이라는 막연한 생각이 현실로 다가오고 있다. 이러한 변화에 대비하려면 우선 AI의 품질을 좌우하는 데이터의 품질을 높이려는 노력이 뒤따라야 한다. 아울러 고품질의 데이터를 보유한 국가통계의 중요성에 대해 다시 한번 생각해 보고, 향후 통계청이 주축이 되어 제공할 재현자료에도 관심을 기울였으면 한다.