빅데이터 시대의 노동통계

방형준 (한국노동연구원 부연구위원)

최근 인구에 많이 회자되는 단어 중 하나가 빅데이터이다. 저장장치의 발달과 실시간 무선 통신 기술의 진보로 인해 과거 신문이나 기록에 의존하던 시대와 비교하면 최근 생산되는 데이터의 양은 가공할 만큼 늘어났으며, 컴퓨터 기술의 발전과 각종 기계학습 기법의 발전으로 많은 양의 데이터를 한 번에 처리하거나 원하는 형태로 가공·분석할 수 있게 되었다. 이에 따라 사회 각 분야에서 빅데이터가 주목받고 있으며, 빅데이터를 기업 경영과 관련한 의사 결정이나 정부의 정책 결정, 혹은 각종 홍보나 사회활동에 활용하는 경우가 늘어나고 있다.

이처럼 빅데이터에 대한 관심은 증대되고 있으나 관심의 초점은 빅데이터를 어떻게 활용할 것인가에 있고, 빅데이터가 어떻게 생산되는지 혹은 통계로서 빅데이터가 가지는 특성은 무엇이며 어떻게 공식 통계와 함께 사용할 수 있는지 등에 대해서는 상대적으로 관심이 소홀하다. 그러나 이는 잘못된 현상이 아니며, 오히려 이러한 영역은 통계를 생산하거나 다루는 영역에 종사하는 사람들, 통계를 활용하는 연구자들, 그리고 빅데이터를 다루고 처리하는 기술을 개발하는 개발자들이 관심을 가져야 하는 영역이기에 당연한 일이기도 하다.

빅데이터가 상업적으로 어떻게 활용 가능하며 우리의 실생활을 어떠한 모습으로 바꾸고 있는지에 대해서는 이미 많은 신문 기사와 도서, 그리고 각종 기고가 있기 때문에 이번 호 기획 특집에서는 빅데이터가 기존 통계와 어떠한 관계가 있으며 어떤 방식으로 기존 통계를 대체하거나 보완하는지, 그리고 어떻게 생산되는지 등을 소개하고자 한다. 거시경제의 현황을 보

여주거나 현재 사회의 모습을 파악하게 해 주는 사회경제적 공식 통계를 가공하고 생산하는 것과 관련하여 빅데이터는 역설적인 모습을 띠고 있다. 전 세계 국가를 저소득 국가와 고소득 국가로 나눌 때 빅데이터의 유용함과 활용도는 저소득 국가에서 더 높지만 빅데이터를 생산하고 가공하는 인력과 관련 연구는 고소득 국가에 주로 집중되어 있다.

일반적으로 저소득 국가는 국가가 생산하는 통계의 양이 적을 뿐만 아니라, 국가의 행정력이 사회 전체에 미치지 못하기 때문에 생산된 통계의 정확성도 떨어지는 편이다. 일례로, 시리아처럼 내전을 겪는 국가의 경우, 국가의 행정력이 정부군이 지배하는 일부 지역에만 한정되기 때문에 통계가 정확하지 못할 뿐만 아니라, 국가가 정확한 행정통계를 지속적으로 생산할 여력이 충분치 않아 통계의 양도 많지 않다. 따라서 저소득 국가의 사회를 좀 더 정확히 이해하기 위해서는 공식적인 통계를 추세 파악을 위한 용도로 사용하고 실제값에 대해서는 다양한 빅데이터를 통해 추정하는 것이 더욱 정확할 것이다. 예를 들어, 신문 기사에서 특정한 이슈나 키워드가 얼마나 자주 등장했는지, 혹은 사람들이 사회관계망서비스(SNS)에서 많이 입력하거나 사용한 문구나 단어가 무엇인지를 빅데이터 기법을 활용해서 파악한다면, 실제 그 사회의 관심사와 해당 관심사에 대해 사람들이 직면한 현실을 잘 이해할 수 있다. 실제 저소득 국가의 경제 발전을 연구하는 일부 학자들은 관심 대상국의 사회경제적 현황을 파악하기 위해 각종 문자정보 등을 활용해 공식적인 통계를 보정하거나 같이 사용하는 경우도 있다.

고소득 국가의 경우 앞서 언급한 통계의 숫자와 정확성 측면에서 저소득 국가보다 질 높고 정확한 통계가 풍부하게 생산되고 있기 때문에 다양한 빅데이터를 활용하여 통계를 보정할 필요성이 높지는 않다. 그러나 반대로 고소득 국가이기에 생산되는 빅데이터의 양이 많을 뿐만 아니라, 분석 과정에서 개발된 기법이나 도구를 활용하여 상업적으로 이용할 경우 높은 수익을 기대할 수 있기 때문에 오히려 빅데이터 종사 인력이 풍부할 뿐만 아니라 다양한 연구도 진행 중이다. 일례로 미국에서는 이미 20세기 중반부터 '경기침체지수(recession index)'라 하여 신문 기사에서 '경기침체(recession)'라는 단어가 얼마나 많이 등장했는지를 가지고 실제 경기 상황을 파악해보려는 노력이 있었다. 간단하지만 '경기침체지수' 역시 빅데이터를 활용하여 경기 상황을 짐작하는, 통계를 활용한 한 가지 간단한 예이다.

고소득 국가에서 빅데이터를 이용해 기존의 통계를 보완하거나 새로운 통계를 생산하는 것은 빅데이터가 가지는 시의성에 기인하는 바가 크다. 전통적인 방식으로 수집하는 노동시장 관련 통계나 거시경제 관련 통계들은 조사 및 수집에 시간이 오래 걸리기 때문에 실제 통계가 생산되는 시기와 발표 시기 간의 시차가 큰 편이다. 예를 들어, 분기별 경제성장률의 경우 추정치만 하여도 최소 한 달여가 지나야 집계되며 추정치가 아닌 실제값은 그보다 더 오랜 시간이 걸려 발표된다. 비교적 통계 생산 시차가 짧은 수출입 관련 통계나 고용률 및 실업률 관련 통계 역시 약 보름 정도가 지나야 지난달의 통계값을 얻을 수 있다. 반면, 빅데이터를 이용하면 정확성은 조금 떨어질 수 있어도 즉각적으로 추정치를 얻어낼 수 있다는 장점이 있다. 예를 들어, 신문 기사에서 실업에 관한 기사의 수가 늘어나면 실업률이 올라가는 추세라는 것을 확인할 수 있고, 인력 부족에 관한 기사나 SNS 언급이 많다면 경기가 호황 국면에 접어들고 노동시장이 완전 고용에 가까울 가능성이 높다고 판단할 수 있다.

이번 호 기획특집에서는 빅데이터를 노동시장에서 어떻게 활용할 수 있으며, 기존의 통계들을 어떻게 보완하거나 대체할 수 있는지에 대해서 고소득 국가와 저소득 국가의 사례로 나누어 살펴보고자 한다. 고소득 국가의 경우, 수치화가 어려워 기존 통계에서는 찾을 수 없는 통계를 어떻게 생산할지 혹은 시의성 높은 추정치를 어떻게 얻어낼 수 있는지에 대해 좀 더 깊이 살펴볼 것이다. 반면 저소득 국가에 대해서는 기존 통계를 보완하고 사회 현실을 잘 반영할 수 있는 활용 사례에 대해 집중하여 살펴볼 것이다.

아울러 우리가 사용하는 통계의 다수는 국가 간 비교를 위하여 국제기구에서 정리·가공한 형태이다. 그러나 국가별로 통계를 수집하는 방식과 대상, 기준이 상이하기 때문에 국제기구에서 이를 정리하여 일관된 통계를 생산하는 것이 결코 쉬운 것은 아니다. 따라서 빅데이터와는 다소 동떨어져 있지만, 국제기구에서 각 국가의 통계를 어떻게 취합하여 시계열 자료를 생산해 내는지 알아보는 것은 향후 국제 통계를 사용할 때 도움이 될 것이다. 더하여, 국제기구에서 생산하는 다양한 유형의 통계 중 노동시장 관련 통계들은 다른 유형의 통계, 예를 들어 산업 통계나 금융 통계 등과 어떠한 차이가 있으며 어떤 다른 과정을 거쳐 생산되는지도 소개할 것이다.

빅데이터에 대한 두 편의 원고와 국제기구에서 통계를 생산하는 방식에 대한 한 편의 원고를 통해 우리는 빅데이터의 활용과 상업적 이용에서 잠시 시선을 돌려 빅데이터 자체에 대한 이해를 높임으로써, 빅데이터를 전보다 더 넓은 시각에서 다룰 수 있을 것이다. 아울러, 빅데이터 시대에 기존의 통계를 어떻게 이해하고 바라보며 향후 어떠한 방법으로 개선하거나 보완할지에 대한 논의가 더욱 생산적으로 진행되는 데 일조할 수 있기를 바란다. KLI