

研究員 の眼

ビッグデータで何が変わるか

経済研究部 研究員 高山 武士
(03)3512-1824 takayama@nli-research.co.jp

ここ数年で、情報を得るためにウェブサイトを検索する機会が増えた。エコノミストとして電子データを解析する機会も多い。最近「ビッグデータ」という言葉も良く聞くようになり、改めて大量の情報に囲まれていることを認識させられる。

ビッグデータという言葉には明確な定義はないが、「3V」によって特徴付けされることが多い。これは、「量 (Volume)」「多様性 (Variety)」「速度 (Velocity)」の頭文字で、つまり、「高速で多様かつ大量に生み出される」データのことである。グーグルの検索欄には、この瞬間も多くの人々が様々な言葉を入力しているだろうし、ツイッター上には、様々な内容のツイート（つぶやき）が投稿されている。まさに、高速で多様な大量のデータ（ここでは検索ワードやツイート）が生み出されており、3Vのイメージに合致する。大規模なだけでなく、多様かつ高頻度で生み出されていることも大きな特徴だ。スマートフォンから取得される位置情報データや、防犯カメラで記録される人間の顔（表情）や動きのデータもこうした3Vデータの代表と言える。

（今までの統計学とビッグデータ）

ビッグデータと言っても、データをどこからがビッグか明確に区切ることはできない。突如として、画期的な分析手法が生まれた訳でもない。そのため、言葉だけが流行っている感じもするが、今まで非現実的（非常識）と考えられていた分析手法に脚光が当たっているという側面も確かにあるだろう。

ビクター・マイヤー＝ショーンベルガー、ケネス・クキエは、ビッグデータ分析が進んだことで、①できるだけ全データを使う、②精度は重要ではなくなる、③因果関係ではなく相関関係が重視される、といった変化（いわば、パラダイム・シフト）が起きていると指摘する¹。

筆者は、「法則を確かめる道具」としてのデータ分析から「法則を発見する道具」としてのデータ分析に（再び）注目が集まっているように感じる²。上記の③に近いが、因果関係と相関関係の違いというよりは、データを分析する前に、あらかじめ強い仮説（仮定）を設けているか、そうでないかとい

¹ ビクター・マイヤー＝ショーンベルガー、ケネス・クキエ、斎藤栄一郎訳「ビッグデータの正体 情報の産業革命が世界のすべてを変える」（講談社）

² 「再び」と書いたのは、昔からどちらの側面もあるためである（昔も、データ集計や単純化で新しい知見を得ていた）。ただ、ビッグデータが扱えるようになって、違う側面から発見の可能性が広がり、再び「法則を発見するための道具」として注目を浴びているのだろう。

う違いが大きいように思う。

具体的に言えば、前者には、例えば「仮説検定」と呼ばれる伝統的な統計の手法がある。読んで字のごとく「仮説」が正しいか調べるために、データを使って検定（検証）する方法である³。

一方、後者のアプローチは、「データマイニング」と呼ばれる。こちらも読んで字のごとくデータからお宝となる事実を「発掘（マイニング：mining）」する作業である。このアプローチは、とにかく沢山のモノの中からはなんらかの関係があるものを手当たり次第に探す⁴。

感覚的に言えば、同じツイートと株価の関係を調べるにしても、「ポジティブ（楽観的）なツイートが多いときには株価が上がるだろう」と仮説を設定してデータを調査するのは仮説検定、「株価が上がる時に投稿されているツイートの特徴はなんだろうかと内容、言語、単語数、地域など様々なことを調べていく」方法はデータマイニングである。

前者のようにデータを「仮説を検証する」ために使うならば、母集団が大きくてもビッグデータを使う必要はない⁵。ただ、データを「法則を発見する」ために使うならば、ビッグデータはできるだけ大きい方がよい。発掘しがいがあり、思い至らなかった知見が得られる期待も高まる⁶。

こうした「法則の発見」は、これまで熟練労働者の経験などに頼ってきた部分が大きい。しかし、これらの知識や法則の発見というプロセスに、データの力が役立つようになっている。実際、検索ワード、ツイート、位置情報データ、顔や動きのデータなどから、お宝を発掘する動きは進んでいる。

（人間の役割とコンピュータの役割）

もうひとつ注目すべきは、データを解析するコンピュータ（アルゴリズム）の発達である。処理能力の向上だけでなく、機械学習（ベイズ推定）など、3Vのデータと相性が良い分析手法・アルゴリズムが発達し、コンピュータが代替できる領域が広がっている。情報の「蓄積や集計」は機械が最も得意とするところだが、それだけではない。

例えば、検索サイトでは、大量のウェブサイトの中から適切と思われるサイトを上位に表示してくれる。つまり、情報の「取捨選択」もほとんど機械がしてくれる。オンラインショッピングで人の購買履歴を学習し、その人が好みそうな商品をオススメすることなども機械が行うようになった。

また、受け取ったメールを迷惑メールか否かの判断も機械がしてくれる。つまり、情報の「解釈」も、ある程度のルールがあればやってくれる。これらは、多くのデータ（購買履歴やメール）を学習させればさせるほど性能が上がる。道徳的な議論を棚上げすれば、人事評価や就職（採用）などの相性判断はすぐに機械化できるだろう（男女の出会いも…）。

将棋ソフトは、まだ短時間での全数探索が出来ないけれど、大量の局面（良い手・悪い手）を学習することで、プロ棋士に勝利できるほどの実力をつけている。

³ サンプル（標本）から全体（母集団）の性質を推定する方法（統計学的推定）を用いて検証している。仮説があれば、それを検証するためのランダム（無作為）なサンプル（標本）を取り出すことができ、そこから母集団の性質を知ることができることがポイントである。

⁴ もちろん、結果的に因果関係などなんらかの関係が推測できることもある。その推測された関係は、まさに新しい知見となりうる。つまり「言われないと気付かないが、言われてみれば確かにそうだ」という発見をデータに教えてもらったことになる。ただし、あらかじめ確認したい因果関係が明確になっているわけではない。この意味で、因果関係を発見すること自体は目的ではないと言えるだろう。

⁵ 例えば、西内啓「統計学が最強の学問である」（ダイヤモンド社）を参照。

⁶ 仮説検定では、仮説が「間違っている」（棄却される）か「そうとは言い切れないか」のどちらかの結論が得られるのに対し、データマイニングでは必ずしもお宝となる関係が見つかるとは限らない。そのため、サンプルを用いないでできるだけ3Vのデータをつかうデータマイニングは発掘コストが高く、費用対効果が悪くなりやすい。しかし、現在はこの発掘コストの低下が進んでいる。

これまで、コンピュータはデータを「集計や蓄積」することが主流だったけれど、いまは、コンピュータがデータから「判断や評価」を行うようになっている。

今後、ビッグデータの流行を経て、さらに機械の存在感は大きくなるだろう。それだけ人間の役割は少なくなるかもしれない。

ただ、それは人間がより「人間らしい」仕事に注力できる時代なのかもしれない。そういった「人間らしさ」とはなんだろうかと考えつつ、ビッグデータの今後に注目している。