巻頭

ビッグデータ

国立情報学研究所　佐藤　一郎

1．ビッグデータとは

今IT業界ではビッグデータという言葉がもてはやされている。新聞やテレビにも頻繁に登場しており，IT業界以外の方でも聞いた方は多いであろう。ただ，ビッグデータに明確な定義があるわけでないが，仮にひと言でいうならば，大量のデータや多様なデータのデータ分析やその技術全般をさすことになる。むしろビッグデータという言葉が注目される背景の方が重要である。

その背景のひとつはデータの増加である。海外シンクタンクの予測では，世界のデータはインターネット普及以降，データ量が増加しており，毎年2倍に増え，2020年には世界のデータ量は40ゼタバイトに達するという予測がある。ちなみにゼタとは10の21乗であり，喩えると2500億枚のDVDに相当する。実感のない喩えで恐縮であるが，実感できないぐらい大量ということである。

さて，どうしてデータ量が急速に増えているのだろうか。これまでデータというと人間が書いた文章や何らかの入力データ，そして人間が自ら撮った画像が主体であったが，2020年になると半分ぐらいのデータは機械から生まれると予想されている（図1）。例えばスマートフォンには，人間が操作するマイクやカメラだけでなく，方角，磁場，直線加速，輝度，ジャイロ，重力，加速度，気圧，温度，位置（GPS）などのセンサーが組み込まれており，測定した結果をすべて記録・保存しているわけではないが，1台1台のスマートフォンが大量のデータを生成しており，そうしたスマートフォンは世界中に何億台とある。

また，ソーシャルネットワークサービス（SNS）に代表される情報サービスでは，ユーザが入力したテキストや画像によるデータよりも，各ユーザのこれまでのSNS利用状況や友達やフォローというようなユーザ同士の関係に関するデータの方が，データサイズ的に多くなっているといわれる。例えば3人のユーザがSNS上で互いに友達になれば，それぞれの関係は3つの階乗（3！），つまり6つとなる。10人が互いに友達になれば関係の数は10の階乗（10！），362880となり，データ量は大きく増えることがおわかりいただけるはずである。

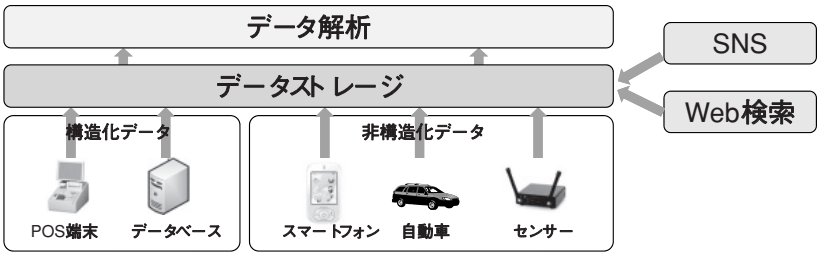

図1　ビッグデータのデータ多様性

2．ビッグデータは何に役立つのか

ビッグデータの事例はすでに多い。東日本大震災の直後，カーナビから自動車の位置情報を集め，震災地

域のうち，どの道路が通行可能または不能かを調べることで，支援物資の移動に役立てた。ちなみにカーナビによる位置情報は誤差やエラーもあり，質が高いデータとはいえないが，そうしたデータでも大量にあれば誤差やエラーを相殺し，質の高いデータに変えることができる。また，ビッグデータは高度なデータ分析にも役に立つ。スーパーマーケットなどではデータ分析を行うことで，個々の顧客の購買品を分析すれば牛乳や食パンなどは顧客ごとに購入品が決まっていることがみえてくる。逆にある商品が欠品している場合はその商品を買っている顧客は来店しなくなることもみえてくる。また，データ分析を通じて購入商品の相関を調べることで，ある商品Aそのものがあまり売れていなくても，商品Aを買う顧客が商品BやCを大量買いしていることがわかった場合，商品Aの販売打ち切りは商品BやCの販売数にも影響がでることも予測できる。

また今後は医療や食糧問題の解決にも役に立つであろう。全国で，ある病気となった患者の症状や治療データを集めることで，個々の患者の診療では見落とされていた症状や治療方法を発見することも期待できるし，農業も気象から土の状況まで各種データを広く集めることで，効率的な栽培方法や収穫時期がみえてくるはずである。また，日本では笹子トンネルの事故に代表されるように，トンネルや橋，道路などの社会インフラの老朽化が深刻な問題となっているが，社会インフラに各種センサーを付けることで，老朽化状況を把握することができる。また，昨今の財政状況を考えると，センサーによる情報から，今後の老朽化状況を予測して，安全を担保しながら点検・修繕時期を遅らせることで，社会インフラの維持費用を最小化することもできるが，その老朽化状況の把握・予測で中心的な役割を果たすのもビッグデータである。

3．ビッグデータの歴史

ところで，ビッグデータはいつから始まったのだろうか。もし最古のビッグデータ事例をあげるとすると，その代表例は19世紀末の米国の国勢調査となる。米国は現代的国勢調査の先駆けであり，下院議会の議員数を決めるデータとして10年ごとに実施される。しかし，1880年の国勢調査では，集めた調査票の集計に７年以上かかったとされ（集計期間は文献により違いがある），次に実施される1890年の国勢調査では，移民の増加により，集計に10年以上かかる，つまり次の国勢調査までに集計が終わらないことが予想された。当時の米国の人口は7000万人以下であり，仮に１人の国勢調査票のデータが100文字分100バイトとしても，全データ容量は７ギガバイト。今でいうとDVD２枚程度であり，大量データとはいえないが，当時は手に余るデータ量であり，まさにビッグデータだったと推測される。

米国政府は集計を高速化する新しい技術を公募し，採用されたのが米国の発明家ハーマン・ホレリス（Herman Hollerith）によるパンチカード（紙に空けた穴で情報を表すカード）とその集計装置であり，これにより集計は１年強で終わったとされる。ところで，ホレリス氏が，タビュレーティングマシンの製造販売のために設立した会社は，その後のIBMの母体となる。ご存知のようにIBMはITを主導してきた企業で，その意味ではITがビッグデータを生み出したのではなく，ビッグデータがITを生み出したといえる。

4．今ビッグデータが脚光を集めるわけ

前述のように，手持ちの情報システムでは手に余るデータ量やデータ種類を扱うという点では，ビッグデータは古くて新しい問題といえる。ここ数年，ビッグデータが話題になるのも，前述のようにデータ量が増えていることに加えて，世の中が大量データや多様なデータを扱うことを求めているからともいえる。その背景の１つはマーケティングの変化である。大量生産・大量消費時代であれば，不特定多数の消費者からなるマス市場向けに商品を開発して，そのマスを対象にした広告，例えばテレビCMや新聞広告などをうって，販売促進をしていればよかった。しかし，今の消費者はどうか。そもそも企業の宣伝は信じない。むしろ影響を受けるのは他の人が何を買っているかであることが多く，企業の広告よりも，自分が興味を持っている商品をすでに買っている人のレ

ビューや口コミの方を信じる傾向がある。例えばAmazonに代表されるようにネットショップでは，カタログ的な商品の説明は少なくして，むしろ他のユーザが何を買っているのか，自分が欲しいと思った商品を買った他のユーザのレビューを掲載している。

ただ，そうなると企業は購入者１人１人の行動を調べ，その情報を活かして他の顧客への販売促進をすることになる。小売業を例にとると，今までは商品アイテムごとに，例えば月単位，週単位に販売数を調べるだけでよく，そのデータ量はアイテム数×販売個数で済んだ。しかし，顧客ごとに何をいくつ買っているのかを調べることになると，データ量は顧客数×各顧客が購入したアイテムの名×その顧客のそのアイテムの購入数を表す整数×これまでの購入アイテム数となり，データ量は桁違いに増えることになる。その結果，大量データを処理するためには新しい技術が必要になってくる。例えば，Amazonは協調フィルタリングと呼ばれるアルゴリズムを導入しており，図２のように全商品と全ユーザの表をつくって，各ユーザが買った場合には数字の重みづけを増やしていき，ユーザとユーザの重みづけを比べて，重みづけの似ているユーザは，似た購買履歴を持っているので，その情報を使って商品をリコメンデーションするという方法を使っている。当然，そのデータ量は大量になるので，まさしくビッグデータの世界となる。

５．ビッグデータにおけるデータ分析

当たり前だが，ビッグデータという言葉が流行る前から，データ分析は広く行われている。例えば企業ではBI（Business Intelligence）といった言葉で，高度な統計手法を使った企業データ分析が行われてきた。それでは，ビッグデータにおけるデータ分析と従来のデータ分析では違いがあるのだろうか。両者は多くの部分は同じであるが，決定的な違いもある。今までのデータ分析は，手元のデータ量は少ないことが前提となり，データをとことん調べ尽くすことに注力してきた。それに対し，ビッグデータではデータ量が多すぎて分析しきれないことが前提となる。

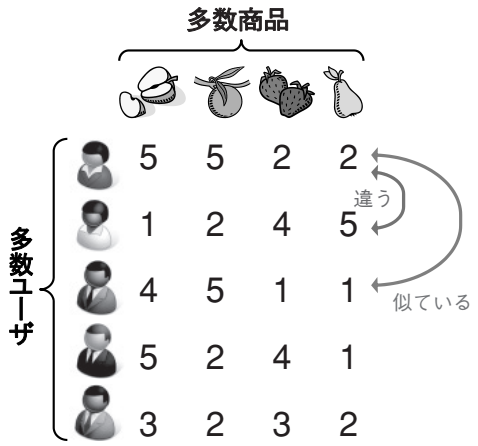

図２　協調フィルタリング

これはレストランのコース料理とビュッフェ形式の違いを思い浮かべるとわかりやすい。今までのデータ分析はコース料理に近い。出てきたものをとことん，美味しく食べる。データ分析でいうと，手持ちのデータを最大限活用することになる。ビッグデータはビュッフェ形式に近い。たくさんの種類の料理が並んでいて，種類ごとの量も多い。当たり前だが，全部を食べることも，全種類を１個ずつ食べることもできない。そうなると，どれを食べるのかを選ばなければいけない。ビッグデータも同じで，データ量が多いので，全部のデータを分析することはできない。したがって，まずどのデータを分析対象にするのかを選ぶことが第一歩になる。

もう１つ考えなければならないことは，選び方である。ビュッフェ形式で取る料理にも食べ合わせがある。例えば刺身とミートソースのスパゲティを一緒に食べても美味しくないだろう。逆に料理によっては単品で食べるより，違う料理を組み合わせることによって美味しくなる場合もある。ビッグデータの場合も，いろいろな種類のデータがあったときに，相違な種類のデータを組み合わせて分析することで，一種類のデータの分析ではわからなかったこともわかるようになる。例えば家庭の水道使用量とガス使用量を別々に調べたとき，それぞれの使用量から住民が何をしているかはわからない。しかし，水道使用量とガス使用量の両方を見比べることで，例えばお風呂を焚いて

いるなどがみえてくる。これがビッグデータによる分析の利点となると同時に，問題も引き起こす。前述の水道使用量とガス使用量は単独では住民の行動分析は不可能なデータでも，組み合わせによって行動分析ができてしまう。ただ，事前に組み合わせ方はわからないことから，どのデータがプライバシー情報に関わるかもわからない。

さて，ビッグデータのデータ分析では，k－近傍分類，k－平均法，EMアルゴリズムといった最新かつ高度なデータ分析技術を使われているが，一方で大量または多様なデータは間違いや誤差，ノイズが含まれることが多い。このため，高度なデータ分析手法が効果を発揮するとは限らず，高校レベルの統計手法でも十分なケースも少なくない。

6．構造化データと非構造化データ

ビッグデータを知る上で，構造化データと非構造化データという言葉も知っておきたい。従来のデータ分析では，小売ならばアイテム名，販売数，納品日など，ある種類のデータをある定まった形式でデータベースに格納し，そのデータを分析していた。しかし，今はメールのテキストデータや動画像，さらに各種センサーの測定値のように，従来は取り扱わなかったデータも対象となる。こうしたデータは構造化されていないことも多いわけだが，基本的に非構造化データを直接データ分析することはできないので，一旦，構造化する必要がある。例えばメール中に「９月１日の始業式は10時に始まります」という文章があったとする。人間が読めば内容は一目瞭然だが，データとして利用するには，図３のように構造化されたデータに変換する必要がある。例えば，Google は Web 検索サービスを提供しており，そのサービスのために世界中のWebサーバからコンテンツを集めている。ただし，Webコンテンツをそのまま検索することは難しい。そこで検索しやすいデータに変換して，その変換済みのコンテンツを利用して，検索サービスを提供している。

非構造化データ

“9月1日の始業式は
10時に始まります”

変換

項目	データ
日付	09-01
時間	10:00AM
イベント名	“始業式”
アクション	開始

構造化データ

図３　非構造化データから構造化データに変換

7．ビッグデータと情報システム

データ量が大きくなると，メモリに全データを載せることができないだけでなく，１台のコンピュータではデータを処理しきれない場合もあるし，そのコンピュータのハードディスクに全データを格納できない場合もある。そこで考えられたのが，複数のコンピュータを協調させて動作させることで，１台のコンピュータではできなかったような大量データ処理を実現している。このため，GoogleやAmazonなどの事業者はビッグデータに相当する大量かつ多様なデータを処理するために，数万というオーダーのコンピュータをもつ大規模データセンターを構築・運用している。一般の企業でも数十台のコンピュータを使ってデータ分析をすることは珍しくなくなっている。この場合，個々のコンピュータのデータ処理能力とともに，コンピュータ同士の連携をいかに効率よく行うかが重要となる。

8．おわりに

情報処理というと，情報よりも，コンピュータを含む処理そのものを中心に考えがちである。しかし，今はビッグデータに代表されるように大量かつ多様なデータを活かすことが求められており，データに応じて処理方法を選ぶことが求められる。また，ビッグデータに限らず，データ分析では，データにある特性があると仮説を立てて，その特性を抽出する分析方法を選ぶ必要があり，そのためにはデータから仮説を思いつく，つまりデータを読み，そこに特性があることを見抜く力が必要であり，そのために今まで以上にデータを読む能力が必要となる。ビッグデータのブームはいつまで続くかは不明であるが，データを読む力，データを適切に分析する力は，いつの時代でも求められる情報教育の基礎となるはずである。