

# 1 ビッグデータ時代を生きる



中野 美由紀 (芝浦工業大学) 豊田 正史 (東京大学)

## ビッグデータとは

### ✦ ビッグデータが起こした現象

ビッグデータ<sup>1)</sup>という言葉から想起される課題はその言葉を問われる立場によってまったく異なる。文字通りの意味にとらえれば「big (大きな) data (データ)」であるが、情報処理に携わる人にとって、もはや「大きな」データが存在することは当たり前となっており、技術的な課題は多くあるが、全件処理は十分に可能となりつつある。しかしながら、そこに「データの内容 (コンテンツ)」「データの利活用」という「社会性」の観点に加わると、とたんにビッグデータは技術の対象にとどまらず、社会における資産へと姿を変え、人が利用し得るデータを生み出すという大きな課題が生じる。経済活動は分かりやすい例であるが、ビッグデータはより幅広い社会活動においても重要な価値を持つと期待され、新たに「データサイエンティスト」と呼ばれる専門家が社会に求められている。

ビッグデータ時代とそれ以前では何が違うのであろうか。1990年代に「データマイニング」が大きく取り上げられた際、データ解析については多くの試みが行われ、コストの高い「全件探索」を効率化するためのマイニング技術が多く提案された。当時、データは収集する側がある意図を持って収集し、その意図に基づいて収集した組織、機関にて解析されていたが、経済的あるいは社会的な有用性から他の蓄積データと併せて利用できることは容易に類推される。たとえば、クレジットカードの信用調査は、返済が可能であるかどうかを知るための「収入」

「雇用」「家族構成」を問うだけであったが、その後、購買履歴、家族構成などに合わせてダイレクトメールが送られるようになった。しかしながら、依然としてデータを収集する側がデータを利用する主体であった。一方、ビッグデータ時代の今、データの収集、解析、利用が多くの場合、異なる組織や機関で行われ、データの収集時に意図されたものとは異なる利用が試みられている。また、利用対象となるデータ自体も人間自体が生成するものに加え、社会活動を維持するためのさまざまなセンシングデータの蓄積が進んでいる。

### ✦ ビッグデータにおける課題

上述のように、ビッグデータを特徴付けるものは日々増大する大量データの「全件処理」である<sup>2)</sup>と同時に「社会活動」における価値を求められていることにある。

コンテンツの「全件処理」を可能とするさまざまな技術として、蓄積のためのストレージ技術、データ解析のための高速かつ大規模分散処理技術 (クラウドコンピューティング等)、大容量の結果を人が扱えるようにするための技術 (機械学習、データマイニング、言語処理等) などが必要であることはいうまでもない。さらに、「ビッグデータ」のもう1つの特徴である「社会活動」に即した成果を得るためには、データの継続性とその変化への追従を可能としなくてはならない。また、データの「社会性」に伴い生じる価値や利用に対する理解、権利などについて、十分に配慮をしなくてはならない。それは、データを利用する側に求められると同時に、データ

を提供する側にも求められる。これらの、ビッグデータ時代におけるリテラシーに関しては、「ビッグデータ」の利用がデファクトとして進んでいる中、急速に進める必要がある。

本稿では以降、本特集の導入としてビッグデータの事例について簡単に紹介する。

## 長期データの収集・蓄積

ビッグデータの大きさについては、具体的な事例を挙げる必要はもはやないかもしれない。全米屈指の小売業 Walmart 社では毎時 100 万人の顧客情報を取り扱い、データは 2.5PB ずつ増加している。Facebook では 1 日にアップロードされる写真が 2 億 5,000 万件、ユーザ同士のやりとりが 8 億件を超える<sup>3), 4)</sup>。米国のオンライン商取引は 2009 年には 30 億ドルに過ぎなかったが、2015 年では 140 億ドルに達すると想定されている。

このように、ビッグデータでは、現在のデータ収集量の大きさに目を奪われがちであるが、実はデジタル化されていない過去のデータにも大きな価値が秘められている。過去のデータのデジタル化は多大なコストを要する一方、過去 100 年近い修理の記録をデジタル化することで多大な経済的効果を得た事例がある。米国ニューヨーク市では、マンホールの爆発事故が年に数回起こり、事前の点検管理では解決しないため、大きな問題となっていた。そこで、記録方式が時代によって異なる修理記録をすべて解析可能な形でデジタル化することで、事故の起きる原因をつきとめ、同様の個所の埋設電線の補修を効率よく行うことにより、マンホール事故を未然に防ぐことが可能となった<sup>2)</sup>。これは高度経済成長期に橋、道路、上下水道など多くの社会基盤が建造された我が国においても、大きな示唆を与える事例である。また、社会におけるデータの蓄積が 100 年にわたる長い時間経過をも耐え得る必要があることを示している。デジタル化されたデータの蓄積においては、記憶メディアの変化、記憶メディアの時間劣化も考慮し、単にメディアに格納するだ

けではなく、必要なときに時間経過にかかわらず利用可能な形で保持する技術が求められており、コールドストレージ等の研究開発が進んでいる。

## リアルタイムな情報要求把握

前章では長期にわたるデータの利用について述べたが、一方で多くの状況において人が求める情報はリアルタイムに変化する。1 日の生活を考えても、朝起きてすぐは当日の天候、交通状況を調べ、昼食には近くにあるレストラン情報が必要となり、アフター 5 には劇場の空席を調べているかもしれない。特に、災害や事故が起こった際には、その対応のための情報が即座に必要となる。

こうしたリアルタイムな実世界状況把握のための情報源として、スマートフォンの普及により爆発的に広まった、Twitter 等のリアルタイムソーシャルメディアが挙げられる。現在では、報道機関等も災害や事故の第一報をいち早く得るために Twitter 等を活用しており、火災や鉄道、航空事故等の映像はこれらのメディアに投稿されたものがテレビでの報道に使われる事例も増えてきている。

2011 年に発生した東日本大震災においては、電話や交通網が麻痺する中、Twitter 等のメディアにおいて、被害状況や救助要請、避難場所等に関するさまざまな情報が、友人間のネットワークを通じて非常に早いスピードで共有された。図-1 は、震災直後から、避難場所に関するつぶやきが、リツイートと呼ばれる情報共有機能で拡散されていった様子を可視化したものである<sup>5)</sup>。図中、各点は個々のつぶやきを示しており、点の間にひかれた線はつぶやきが他のユーザに共有されたことを示す。つぶやきを囲む円は 1 つの話題を示し、その大きさが話題を共有したつぶやきの多さを示している。この図から、ものの数時間のうちに、非常に多くの情報が共有されていったことが見てとれる。

図示しきることはできないが、詳細につぶやきを見ていくと、震災直後には、過去の大震災を経験した人々から、風呂桶に水をためたり、火災を防ぐた

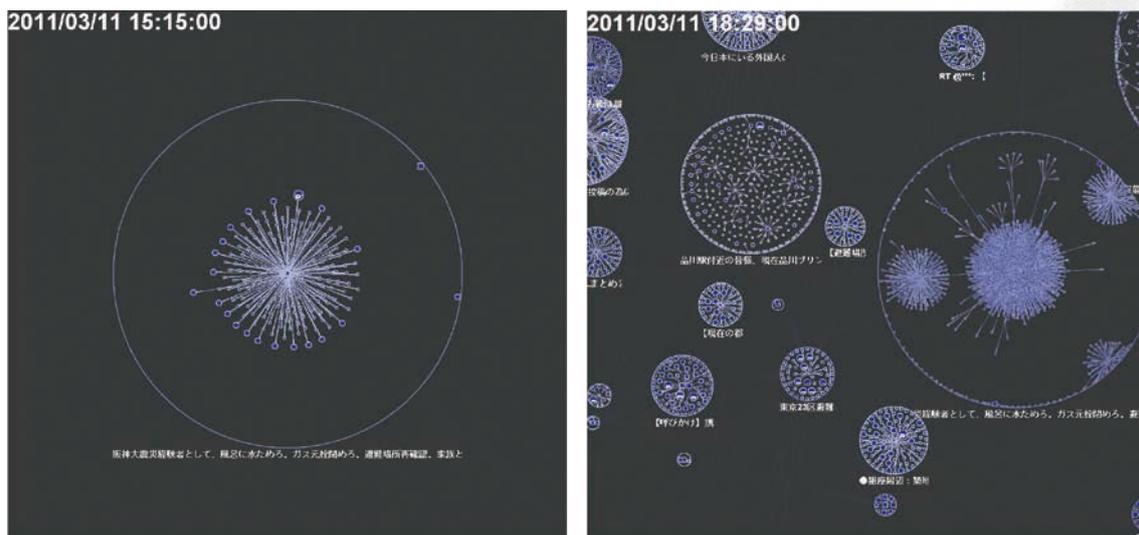


図-1 東日本大震災時の避難所に関するつぶやきの拡散

めにガス栓を閉めたりするべき等の経験を活かしたアドバイスが発信され、それが多くのフォロワーに共有され伝播された。その後、多くの人々が徒歩で帰宅を開始し、大量の帰宅困難者が発生したため、各地でどこが避難場所や休憩所になっているかの情報が共有され始め、最終的には散逸した避難所の情報を集約して Google マップ上にまとめて公開する人が現れるようになるという、非常にダイナミックな情報の変遷を見ることができる。

東日本大震災にかかわるビッグデータ分析については、ソーシャルメディアデータのみでなく、カーナビから得られる走行実績データや、携帯電話の位置情報を用いた混雑統計データ等の多様なデータをマッシュアップした試みが多数行われている<sup>6)</sup>。被災地における人や自動車の避難行動など、今後起こり得る震災に備えるために重要な知見が得られている。

## IoT の活用

米国 Cisco 社の調査によると、インターネットに接続されたデバイスの数は 2010 年以前に世界人口を超えており、デジタルデータの増加は、IoT (Internet of Things) に拠るところが大きいと見込まれている。スマートフォン等のモバイルデバイス

のみならず、自動車に設置されたカーナビやドライブレコーダ、工場や発電所等の機器に設置されたセンサなどを活用したさまざまなソリューションが模索されている。

自動車の走行データを例にとると、米国 INRIX 社は北米・欧州の自動車から大量の位置情報を収集し、渋滞予測等の交通量分析を行い、官公庁、商用車、投資ファンド等に提供している。欧米の自動車保険業界においては、自動車に GPS や加速度計を設置し、運転時間や急ブレーキ・急発進といった運転記録により、保険の料率を変化させるテレマティクス保険のサービスが広まっており、日本においても同様のサービスが始まりつつある。

社会的貢献的な取り組みとしては、HONDA 社においてカーナビから収集した急ブレーキデータをもとに安全運転地図を作る、SAFETY MAP という取り組みが進められている。

東京大学生産技術研究所では伊藤らにより多数のドライブレコーダから得られた大規模な運転操作データを 3 次元空間に可視化し、ヒヤリハット地図の作成を支援する試みが進められている (図-2)。高さ方向に時間軸を表現することで、時間帯ごとの各場所の運転操作回数を可視化しており、操作時の速度や加速度によってフィルタすることで、注意が必要な道とその時間帯などを炙り出すことができ、よ

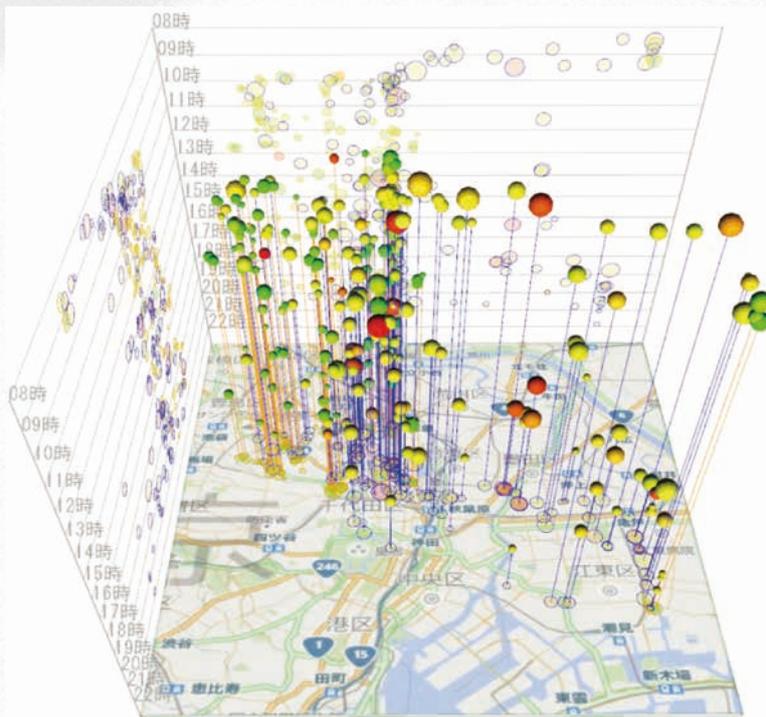


図-2 ドライブレコーダデータの3次元可視化

り詳細なヒヤリハット地図の生成を目指している。

## ビッグデータ時代を生きる

これまで述べたように、ビッグデータという単純な言葉の中にはさまざまな意味が込められている。多様なデータが多様な形で蓄積され、多様な目的で利用されている。つまり、ビッグデータ時代において多角的な視点における考察のもとに、目的に沿ったデータ群、処理技術、さらには得られた結果の精査を行う必要がある。データ解析技術としてさまざまな技術が提案されているが、最終的に得られた結果を活かすのは人であり、実社会である。

### 参考文献

- 1) McKinsey Report, Big Data : The Next Frontier for Innovation, Competition and Productivity (2011).
- 2) Mayer-Schonberger, V. and Cukier, K. : Big Data, A Revolution That Will Transform How We Live, Work and

Think, Eamon Dolan/Houghton Mifflin Harcourt, ISBN-13: 978-0544002692 (2013).

- 3) Big Data Meets Big Data Analytics - Three Key Technologies for Extracting Real-Time Business Value from the Big Data That Threatens to Overwhelm Traditional Computing Architectures (2013). <http://www.sas.com/resources/whitepaper/wp46345.pdf>
- 4) Russom, P. : TDWI Best Practice Report : MANAGING BIGDATA, Technical Report, TDWI (2013).
- 5) 喜連川優 : ビッグデータの潮流とデータエコシステム, 情報管理, 55 (10), pp.705-711 (2012).
- 6) 阿部博史 (編) : 震災ビッグデータ, NHK 出版 (2014). (2015年8月7日受付)

### 中野 美由紀 (正会員) miyuki@shibaura-it.ac.jp

東京大学理学部情報科学科卒業。博士 (情報理工学)。富士通 (株) 勤務後、1985年東京大学生産技術研究所助手 (2004年助教)。2008年特任准教授。2013年芝浦工業大学特任教授。データベースシステム、ストレージシステム、データ工学の研究に従事。IEEE、電子情報通信学会、ACM、日本データベース学会各会員。

### 豊田 正史 (正会員) toyoda@tkl.iis.u-tokyo.ac.jp

東京工業大学情報科学科卒業。博士 (理学)。1999年科学技術振興事業団計算科学技術研究員。2001年東京大学生産技術研究所学術研究支援員、同大学同研究所産学連携研究員、2006年同大学生産技術研究所特任助教授。現在同大学准教授。Webマイニング、ユーザインタフェース、情報可視化の研究に従事。日本ソフトウェア科学会、ACM、IEEE CS各会員。