ビッグデータ時代を迎え注目されるデータ分析技術

Growing Importance of Data Analysis Technologies in Era of Big Data

折原　良平　■ORIHARA Ryohei
西川　武一郎　■NISHIKAWA Takeichiro
佐藤　誠　■SATO Makoto

ビッグデータ時代を迎え，巨大なデータを並列計算機で取り扱うためのデータ処理基盤とともに，データを解析してそこから有用な情報を取り出すためのデータ分析技術への期待とニーズが高まっている。データの量や種類，その生成速度は今後も増加する一方であり，こうした大量のデータを分析できる環境が整えば，これまで困難であった新たな分野にデータ分析を活用し，新たな知見をもたらすことが期待されている。

こうした状況のなかで東芝は，社会インフラ向けからコンシューマー向けまでの幅広い分野に携わっている強みを生かし，これらの事業活動を通して様々なデータを入手して，分析し，活用する技術開発を進めている。また，ビッグデータをより深く活用するために，多種混合データへ対応するための高度な分析方法とデータベースの開発にも取り組んでいる。

Data analysis technologies to extract useful information from large volumes of data, as well as parallel computing platforms to handle such data, have become increasingly important as key technologies in the new era of big data. With the continuous increase in the volumes, categories, and expansion ratios of stored data, data analysis technologies are expected to offer new insights in fields that have remained unanalyzed up to now by making full use of newly constructed databases.

In response to this situation, Toshiba has been engaged in the development of various technologies to collect data from its business activities in a diverse array of fields, ranging from social infrastructure systems to consumer products, and to analyze and utilize such big data efficiently. We are also developing highly sophisticated analysis methods and databases to process combinations of various types of data for the effective utilization of big data.

ビッグデータ時代の到来

ネットワーク技術や，センサ技術，ストレージ技術などの進歩によって，これまでよりもはるかに巨大で複雑なデータにアクセスできる状況が実現した。こうしたデータをビッグデータと呼ぶようになったのは2011年頃から[1]，それ以降，ICT（情報通信技術）分野の流行語として定着している。このようななか，東芝は，ヘルスケアや，保守，エネルギーマネジメント，リテールなどの分野を中心にビッグデータ処理を活用して新たな価値を提供していく方針である。

ビッグデータ関連技術の中では，巨大なデータを並列計算機で取り扱うためのデータ処理基盤[2]が注目されることが多いが，データを解析してそこから有用な情報を取り出すためのデータ分析技術も必要不可欠な要素であることはこれからも変わらない。しかし，ビッグデータが利用できるようになったことで，データ分析技術のトレンドは変化し始めている。

ここでは，データ分析技術やそのアプリケーションをいくつかの観点で分類しながら，データ自体のトレンドを踏まえてデータ分析技術の今後の方向性について述べる。

データ分析の分類

ここでは，与えられたデータに対してなんらかの処理を行い，ユーザーの意思決定に役だつ情報を得ることをデータ分析アプリケーション（以下，データ分析と略記）と定義する。データ分析を，処理に用いるアルゴリズムによって特徴づけることは可能であるが，以下では，より広い立場で，次の観点から述べる。

（1）　分析の出力の特性
（2）　分析の出力と意思決定の隔たり

扱うデータの特性も重要であるが，Volume（量），Velocity（速度），及びVariety（種類）の3Vに基づく分類[3]がビッグデータの定義(注1)と関連して広く受け入れられている。以下では，3Vを，データ分析を分類する観点ではなく，既に到来したトレンドを特徴づける要素と考える。

■分析の出力の特性

分析の出力の特性については次の2通りを考える。

（1）　簡単に定義できる条件を満たすデータ又はデータ群
（2）　データに基づき構築されるモデルの出力

(1)を選択的分析，(2)を包括的分析と

（注1）　3Vのうち，少なくとも一つが従来と比べて極めて大きいものをビッグデータとする。

呼ぶ。例えば，スーパーマーケットで顧客が同時に購入する商品をPOS（販売時点情報管理）データから求めるバスケット分析は，選択的分析である。一方，過去のデータに基づき天候や気温と電力消費との関係をモデル化したうえで行う電力需要予測は包括的分析である。包括的分析を“深い分析”と呼んでいる文献もある[4]が，バスケット分析のように選択的分析から深い洞察を得られる場合もあるため，ここでは別の語を用いることにする。

■分析の出力と意思決定の隔たり

分析の出力と意思決定の隔たりとは，分析結果から直ちに意思決定ができるのか，あるいは分析結果に基づく別の検討を通して意思決定をするのか，の区別で，前者を直接的分析，後者を間接的分析と呼ぶ。例えば，バスケット分析は，同時に購入される商品がわかっても，それを売上増につなげるには商品配置を検討すべきか，ミックスマッチ販売をすべきかなど，スーパーマーケットにとって最適な施策は明らかでないので，間接的分析である。一方，電力需要予測は，予測結果に基づいて発電量を制御すればよいのは自明であり，直接的分析である。

■データ分析の分類

同じアプリケーションであっても複数の分析種別によるアプローチが可能な場合もある。例えば，クレジットカード使用の不正検出は直接的分析であるが，あらかじめカード会社の設定したルールを使った選択的分析に基づくこともできるし，ユーザーの利用履歴から利用モデルを構築し，それとの一致度を調べるという包括的分析に基づくこともできる。代表的なデータ分析を，これまでに述べた観点で分類して**表1**に示す。包括的分析では，その出力が直ちに意思決定に結びつくようにモデルを設計できれば直接的分析になるが，必ずしもそれが可能とは限らない。

表1．データ分析の分類
Categorization of data analysis applications

	選択的分析	包括的分析
間接的分析	バスケット分析 スポーツ戦略立案支援 SNS人脈分析 生産工程管理	視聴率予測 新聞記事クラスタリング
直接的分析	不正利用検出 経済指標予測	電力需要予測 番組推薦 不正利用検出 スポーツ戦略立案支援 経済指標予測 生産工程管理

SNS : Social Networking Service

データトレンドと分析技術

ビッグデータ時代を迎え，選択的分析が用いられるケースが増えてきた。これには二つの背景がある。第1に，非常にたくさんのデータが利用可能な場合，データ集合自体がモデルとして機能し，包括的分析と事実上同じことを選択的分析で行える場合があるという点である。第2に，選択的分析はアルゴリズムを並列化するのが比較的容易で，並列データ処理基盤との相性が良いという点である。包括的分析は，データ全体の様々な統計的性質を調べなければならない場合が多く，並列化しても通信オーバヘッドが大きくなる傾向があり，注意を要する。

しかし，データの特性は年々変化しているので，これに伴って分析技術も変化していくことが予想される。以下では，データのトレンドと，それに伴う分析技術のトレンドについて，3Vの観点から述べる。

■Velocity

データが高速で生成されるようになり，分析結果の検討に十分な時間を掛けることができないために間接的分析を行えず，直接的分析を用いるしかなくなる可能性がある。例えば，マイクロ秒単位でシステムトレードを行う超高速取引には，間接的分析が利用されることはない。前述のとおり，包括的分析は直接的分析に結びつきやすいので，これは包括的分析の重要性を増大させることにもなる。

■Volume

マルチメディアデータや多様なセンサで監視される生産工程データはデータの空間が極めて高次元であり，原理的にあるいはデータ採取コストの面から十分に大量のデータが得られない場合が増えてくる。このようなときには選択的分析で包括的分析を置き換えることはできないため，包括的分析が必須になる可能性がある。

■Variety

データ種類の増大に伴い，生データから特徴量を抽出するための設計を人が行うのが困難になってくる。特徴量を抽出する処理はマルチメディアデータにおいて特に重要であり，その自動化に向けて近年注目を浴びているのが深層学習[5]である。これは，多段に積層されたニューラルネットワークにより生データからアプリケーションに適した特徴量を自動的に抽出するもので，間接的・包括的分析を前処理として使い，その後で本分析を行うという手法である。

■分析技術のトレンド

このように，データのトレンドである3Vいずれからも，包括的分析が今後もいっそう重要になるという分析技術のトレンドが示される。分析技術のトレンドを，3V及び分析技術の分類と関連させて**図1**に示す。

今後の分析技術のトレンドは，選択的分析により成果を出しながら，包括的分析を効率よく並列化するという困難な目標に向かって技術開発を進めていくことになると思われる。このアプローチはまだ成功例が少なく[4]-[6]，有効な手法を開発できれば，分析技術を大きく前進させることができると期待される。それに加え，分析の高速化に役だつ方向性として，包括的分析における追加学習や，選択的分析による直接的分析の可能性も探られていくと思われる。

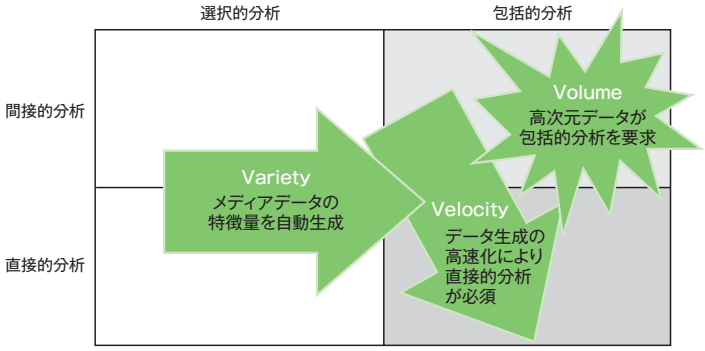

図1. ビッグデータの特性と分析技術のトレンド — データのトレンドである3Vいずれからも，包括的分析が今後もいっそう重要になるという分析技術のトレンドが示される。
Trends in characteristics of big data and data analysis technologies

データマイニングのビジネス活用

これまでのデータマイニング活用の流れ

1993年にデータマイニングということばが生まれて[7]以来，データマイニングの結果を活用してビジネスに生かす取組みは，金融・流通業界で最初に広まった。金融機関にとって，過去のデータを分析して事故率の低い優良顧客を特定し，その集団に属する顧客を優遇するなどの戦略をとることで，データを活用しないライバルに差をつけることができる。

流通業界では，POSデータの分析により店の仕入れや品ぞろえの最適化が最初に図られたが，メンバーカードなどで個々の顧客を把握できるようになると，金融業界と同様に，優良顧客特定のための分析も進んだ。分析結果はダイレクトメールを送付したり顧客に合わせた優待サービスを提供したりするのに活用されている。更に，ダイレクトメールを送付したときに，送付先の顧客から高い確率で反応を得るためのデータ分析も進んでいる。

また，近年では，Web上での閲覧履歴をもとにした顧客分析が大きな成果を上げており，広告の配信や商品の推薦などを通して，企業の収益に大きく貢献している。

このようなデータ分析による成功事例は広く紹介されてきているが，比較的単純な分析によって実現されている事例が多い。しかし今後，使用できるデータの増大と分析技術の高度化により，データ分析を有効に活用できる業界は更に広がると期待される。

例えば，製造業では工程内の品質管理のためにデータ分析を活用しているが，データを収益につなげた事例は建設機器の遠隔監視などに限られている。この理由としては，出荷した製品の稼働データを収集したり蓄積したりしていないこと，及び製品の稼働データを有効活用する分析方法が確立されていないことが考えられる。インフラ機器のデータは収集していても，これを保存するストレージシステムの容量が十分でない場合，一定期間が過ぎるとデータを消去しなければならない。また，コンシューマー向けの製品では，稼働データの収集すらしておらず，市場での使われ方を定量的に把握していないことが多い。

今後のビッグデータのビジネス活用に向けて

米国の市場調査会社であるインターナショナルデータコーポレーション（IDC）社は，全世界で生成され，複製され，消費されるデータは，2013年には4.4 Z（ゼタ：10^{21}）バイトであったのに対し，2020年には44 Zバイトになると予想している[8]。今後，ストレージシステムのビット単価が低下して，蓄積されるデータの量は増加し続け，また，これまで以上に多くのデータがインターネットを介して共有されると予想されている。

このような大量のデータを安価に活用できる環境が整えば，時系列センサデータ，メディアデータ，あるいはこれらを組み合わせた多種混合データを用いた，高度な分析が可能になると思われる。この結果，これまで困難であった難しい課題にもデータ分析を活用できるようになると期待される。

一方，収集した大量のデータを様々な形で組み合わせても，そこから望んでいるような情報を抽出できるかはわからない。有益な分析結果を得るためには，集めたデータに有益な情報が含まれている必要があり，そのようなデータを集めることが重要な課題である。いったん適切にデータを集めることができれば，それらを意味のある特徴量に変換したり，製品に適した構造の統計モデルを構築したりするなど，高度な包括的分析により成果を上げられると期待される。また今後，分析対象が多くなると，一つひとつの分析対象について，人が検討して結論を出すのは困難になるため，直接的分析が重要になると考えられる。玉石混交の大量データの中から，意思決定に役だつモデルを構築するのには課題が多いが，これを実現できれば，新しいデータマイニングのビジネスでの成功事例を増やせるものと期待される。

東芝のデータ分析技術

以下では，当社のデータ分析，及び分析を高度化させるための基盤技術について述べる。当社の強みは，社会インフラ事業では対象設備に備え付けられたセンサから，コンシューマー事業では製品に搭載されたセンサから収集したデータを利用できることである。

当社のデータ分析技術と，それを活用して実現できる効果を**図2**に示す。

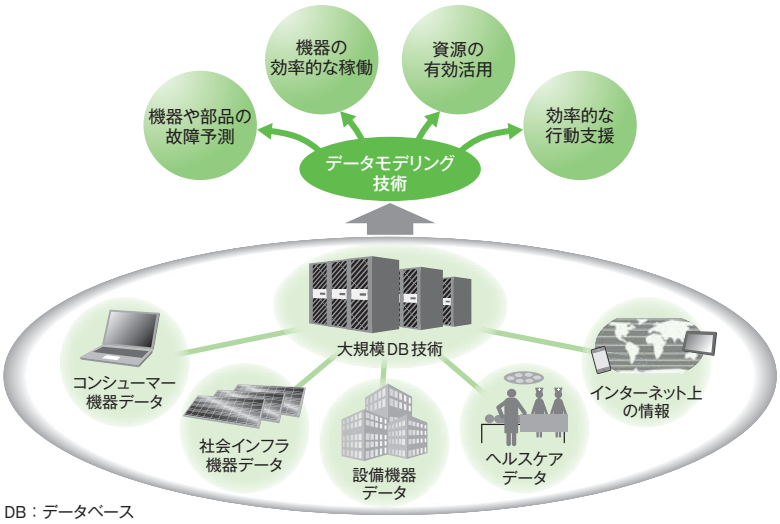

図2．東芝のデータ分析技術 ― 当社は，社会インフラ向けからコンシューマー向けまでの幅広い事業活動を通して様々なデータを入手して，分析し，活用する技術開発を進めている。
Toshiba data analysis technologies

■社会インフラ向けのソーシャルメディア分析

コンシューマーによって作成され，共有されるデータの例としてソーシャルメディアへの投稿データがある。特に近年，外出中にスマートフォンを使って投稿する例が増えていることから，投稿者の位置情報と投稿データをセットで得ることができる。そして，これを社会インフラを監視するためのセンサ情報とみなして活用できる。当社は，Twitter(†)のテキスト内容を分析して単語の出現頻度や時系列変化を調べることにより，鉄道の運転見合わせや遅延などの運行情報を自動抽出する技術を開発した（この特集のp.19－22参照）。

■昇降機部品の寿命分析

長期間にわたって稼働しているB2B（Business to Business）向け製品では，大量の保守履歴が蓄積されていることが多い。当社では昇降機の十数年にわたる長期間の保守履歴を電子データとして保管している。このデータを活用して部品の交換基準の最適化を実現した。ビッグデータ解析では，データ分析を目的としてデータベースを設計し，データを蓄積している例は少ない。この事例で活用したデータでは，故障による交換と予備的な部品交換とを区別していない。そこで，部品の寿命を評価するため，テキストマイニングによりテキスト文章の内容から両者を区別するモデルを作成し，データにラベル付けを行っている。この結果，過去の蓄積データから寿命モデルを作成できるようになり，昇降機の属性に応じて寿命が有意に異なることを発見した（設備保守向けデータ分析について**囲み記事参照**）。更に，この結果を活用して属性に応じた最適なメンテナンス計画を立てることで，故障率，保守コストをともに低減できることをシミュレーションにより確認した（同p.15－18参照）。

■太陽光発電所の異常検知モデル

新しい製品への取組みとして，太陽光発電所の発電性能モデリング技術を開発し，異常検知の実証を行っている。これは，太陽光発電所から収集するデータだけでなく，気象データも活用して発電量を推定し，これと実際の発電量の差によって異常を検知するものである。発電量を推定する発電性能モデル構築にあたっては，実際の太陽光発電所の構造を加味した2段階の多変量多項式回帰モデルを活用して性能改善を行い，その有効性を評価した（同p.11－14参照）。

■HDD故障予兆の検知

B2C（Business to Customer）製品でデータ分析を活用した例として，パソコン

設備保守向けデータ分析の３種類の性能指標

設備保守向けの分析を行うデータサイエンティストは３種類の性能指標を利用できる。蓄積されたデータを使って対象機器の異常プロセスを表現する確率モデルを決める際には，モデルの性能指標を参考に試行錯誤をする。適切な異常モデルが得られたら，異常診断の性能指標を用いて最適な機器の異常診断ロジックを導出する。最適な異常診断ロジックが得られたら，実現コストや導入効果を顧客価値に関する設備管理の性能指標に変換しながら異常診断システムを設計し提案していく。

AIC　：Akaike Information Criterion
BIC　：Bayesian Information Criterion
AUC：Area Under the Curve
FIR　：Field Incident Rate
MTBF：Mean Time between Failures
RONA：Return on Net Assets

設備保守向けデータ分析の３種類の性能指標

(PC) に搭載されたHDD (ハードディスクドライブ) の故障予兆検知エンジンがある。当社のPCにはPCヘルスモニタが搭載されており，ユーザーの承諾を得たうえでPCの稼働データをサーバに送信する仕組みができている。既に収集した220万台以上のノートPCの時系列データから，HDDの故障予兆を検知するエンジンを開発した。この事例では，時系列データからHDDの故障予兆を捉えるための特徴量を，故障メカニズムを考慮して756個作成している。そのうえで，ブースティングアルゴリズムにより故障と相関が高くなるような特徴量の組合せを決定している (同p.7－10参照)。

■スケールアウト型データベース GridStore™/NoSQL

今後センサデータを活用するうえで，センサデータを保管し管理するための適切なデータベースの構築も重要になる。東芝ソリューション (株) は，大量かつ多様なデータを高速に蓄積して活用できるスケールアウト型データベースGridStore™/NoSQLを開発した。これは，将来，データ量が増加しても，それに応じて容易に拡張できるほか，センサ数が増加してもほとんどスループットが低下しないという特長がある (同p.23－27参照)。

データ収集の課題

データ分析を行うためには，データの収集，管理，及び分析を，特定の目標に向けて一貫して実施することが理想であるが，現状のビッグデータの収集ではこのように進められないことが多い。

単にデータのサンプル数や，項目数，種類などを増やしても，収集に失敗すれば，有用なデータを得ることはできない。

これを解決するための手段として，Web上で画面デザインやダイレクトメールの内容をユーザーごとにランダムに変えるなど，アクティブに働きかけることで，ユーザーの有益な反応を収集する取組みがなされている。また一般に，センサデータの収集が容易になっても，ラベル付けは人手に頼るしかないことが多く，大量のデータへのラベル付けが難しいという問題がある。例えば，故障／非故障のラベル付けや，時系列データに正しい意味を対応させる作業は人手に頼るしかない。この問題を解決してデータ分析を効率化するためには，アクティブラーニングと呼ばれる技術によりラベル付けするデータを特定したり，クラウドソーシングという手段を利用して人手で短時間かつ低コストでデータマイニングに必要な教師データ(注2)を整備したりするなどの試みがなされている。

東芝の今後の取組み

ビッグデータの特性を十分に生かして有効に活用するためには，多種混合データから高度な分析によってデータに隠された情報を抽出する技術が重要である。当社は，前述したように，多種混合データへの対応に向けた高度な分析方法とデータベースを開発した。

今後は，ビッグデータの蓄積や単純処理に加えて，多種混合データの活用に向けた分析方法のいっそうの高度化，及び意味のあるデータを収集するための技術が重要になる。当社は，こうした動向を見据えて，データ分析技術の開発と高度化を推進する。

文　献

(1) Manyika, J. et al. "Big data: The next frontier for innovation, competition, and productivity". McKinsey Global Institute. <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>, (accessed 2014-06-19).
(2) Apache. "hadoop". Apache Hadoop Homepage. <http://hadoop.apache.org/>, (accessed 2014-06-19).
(3) Laney, D. "3D Data Management: Controlling Data Volume, Velocity, and Variety". META Group Application Delivery Strategies. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>, (accessed 2014-06-19).
(4) 岡野原大輔．大規模データ分析基盤Jubatusによるリアルタイム機械学習．人工知能学会誌．28，1，2013，p.98－103.
(5) Le, Q. V. et al. "Building High-level Features Using Large Scale Unsupervised Learning". Proc. 29th International Conference on Machine Learning (ICML). Edinburgh, Scotland, UK, 2012-06, International Machine Learning Society (IMLS). 2012, p.81－88.
(6) Panda, B. et al. "PLANET: Massively Parallel Learning of Tree Ensembles with MapReduce". Proc. 35th International Conference on Very Large Data Bases (VLDB), 2. Lyon, France, 2009-08, VLDB Endowment. 2009, p.1426－1437.
(7) 山端　博．ビジネス・インテリジェンスとCRM―データマイニング・ビジネスの実際―．オペレーションズ・リサーチ．43，12，2009，p.653－657.
(8) IDC. "The Digital Universe of Opportunities". EMC homepage. <http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf>, (accessed 2014-06-19).

(注2)　入力データの対となる正しい出力結果。

折原　良平
ORIHARA Ryohei, D.Eng.
研究開発センター 知識メディアラボラトリー研究主幹，博士 (工学)。発想支援技術，類推，機械学習，データ・テキストマイニングの研究に従事。人工知能学会，情報処理学会，日本ソフトウェア科学会会員。
Knowledge Media Lab.

西川　武一郎
NISHIKAWA Takeichiro, Ph.D.
研究開発センター システム技術ラボラトリー研究主幹，博士 (理学)。データマイニング，リスクマネジメント分野の研究・開発に従事。日本オペレーションズ・リサーチ学会，日本品質管理学会会員。
System Engineering Lab.

佐藤　　誠
SATO Makoto, D.Eng.
研究開発センター システム技術ラボラトリー主任研究員，博士 (工学)。データマイニング及び応用統計分野の研究・開発に従事。情報処理学会，電気学会，米国統計学会 (ASA) 会員。
System Engineering Lab.