１Ｆ０３

ビッグデータの研究開発推進の注目点

○野村　稔，奥和田久美（科学技術政策研究所）

１．はじめに

現在、ビッグデータをめぐっての研究開発が、産業界・アカデミア・各国政府によって盛んに進められている。筆者らは、ビッグデータへの取り組みについて、米国政府が 2012 年 3 月に発表したビッグデータの利活用を目的とした研究開発イニシアチブの内容とその注目できる諸点について紹介[1)]した。

本稿では、ビッグデータというキーワードが現れてきた背景を概観し、米国の動きなどを参考に、ビッグデータの研究開発を推進する上での注目点を抽出する。

1.　ビッグデータとは何か

1.1.　ビッグデータとは

ビッグデータとは、必ずしも明確な定義はないが、巨大なデジタルデータの総称である。ここでいうデータとは、どこか一箇所に集められたデータだけではなく、ソーシャル・ネットワーキング・サービス（SNS）などの普及に伴って巨大化した Web 情報、インター―ネット上に蓄積される大量の写真や動画、センサーが検出し送出した膨大な「モノ」からの情報、スーパーコンピュータなどで生成される巨大な数値データなど様々な分野の様々な種類のデータが具体例として挙げられる。そのデータは、量的に既存の技術では管理できないほどに増え、そして複雑化している。

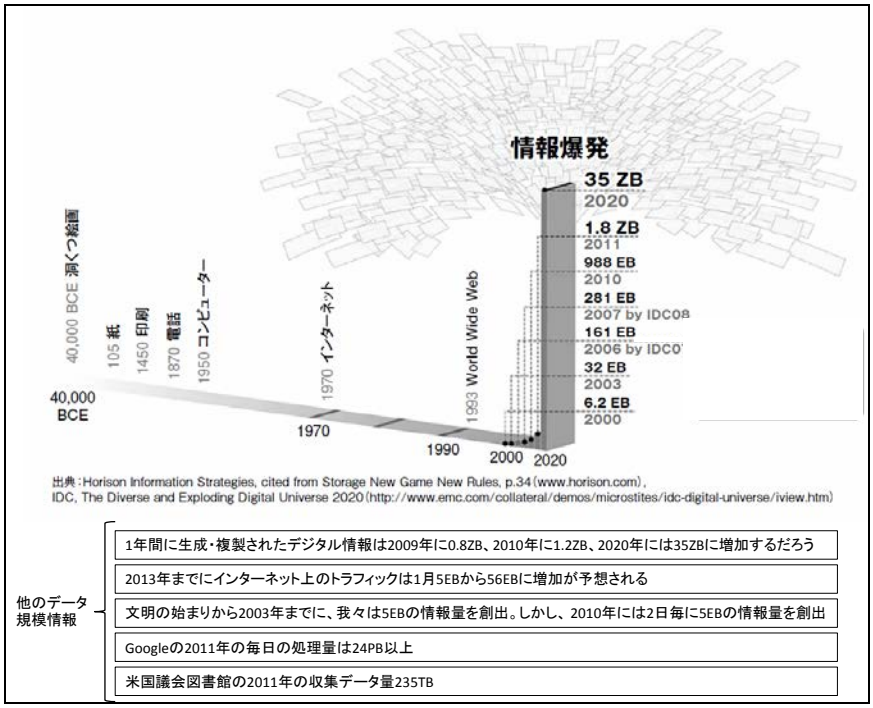
ビッグデータは、文書・画像・センサーデータなどのようなデータが大半を占めている。Facebook や Twitter などの SNS の利用拡大に加え、大容量の映像データのサイトへの投稿が増えており、日々、ネット上で急増しているからである。また、あらゆる「モノ」を Web につなぎネットワーク化するという考え方である「モノのインターネット」（Internet of Things：IOT）の具体化と進展があり、これもデータの急増を招いている。

図表１に、データ量の増加状況を示す。最近、ゼタバイトレベルのデータ量が現れてきており、それもここ数年、指数関数的に増加している。さらにこの勢いは継続しそうである。

1.2.　ビッグデータが現れてきた背景

データ規模が膨大になってきた背景としては、Web データの収集が以前に比べはるかに容易になったこと、デバイスからのデータ収集(携帯電話などからのデータ収集)やモノからのデータ収集が容易になったこと、そして大量データを扱える蓄積・処理技術の高度化が挙げられる。

まず、データ収集面であるが、Web 情報の収集を例として採り上げる。サーチエンジンンが必要とする Web 情報の収集は、1994 年の後半ごろまでは主に人間による作業に依存していた。しかし、Web の世界的な普及によりその限界が見えてきた。この打開策として登場したのがクローラというプログラムで、これにより、Web 上の文書や画像



図表１．データ量の増加状況
（出典：参考文献[1,2)]を基に科学技術政策研究所にて作成）

などが周期的に取得され、自動的にデータベース(DB)に収集されるようになった。また、デバイスや「モノ」からのデータ収集に必要なセンサーと通信機能の小型化・低価格化の進展もデータ収集の容易化を加速した。また、センサーにより収集等したデータを送信する通信モジュールの低価格化が進展し、契約者数も増加している。

データの蓄積・処理面に関しては、巨大なデータセットが分散処理環境上の様々なコンピュータ上に分散して格納されている状況下でデータ処理を並列的に行う技術として、最近 Hadoop の利用が大きく取り上げられている。Google の MapReduce の仕組みをベースに作られたオープンソースであるが、現実的な利用局面における種々の問題を解消すべく多くの商用版が利用可能となっている。一方でリレーショナルデータベース(RDB)は 20 年以上もの最適化コンパイラ技術の蓄積があり、現時点においては Hadoop と RDB は併存状況にある。

言うまでもなくビッグデータは大量のデータである。その処理のためには、大量のディスクと大量のコンピュータが必要になる。データも並列の入出力が行われないと時間がかかり過ぎてしまうからである。複数のコンピュータにより複数のディスクへの並列アクセスができる環境を提供することで、入出力の短縮ができる。クラウドが普及し、大量のディスクと大量のコンピュータを共に適える手段となったことが、ビッグデータに注目が集まるようになった背景でもある。

大量のデータの収集、蓄積・処理が可能になりつつある現在は、その大量のデータからいかにして価値を生み出し、新産業の創出や社会課題の解決に繋げるかが鍵となって来ている。ビッグデータが注目される最も大きな理由はここにある。

1.3. ビッグデータの特徴

ビッグデータが従来のデータと異なる点には、多量性、リアルタイム性、多種性、データ構造、不正確さ・あいまいさなどが挙げられる。

(1) 多量性

もしデータが大きいことが困るだけなら、サンプリングによって小さくして扱えばいいが、それでは結局、一部しか見ることができない、または重要なものを落とすかもしれないという懸念がある [3)]。大きなデータの集合の中から、特徴的なパターンを発見したり、データの集合をある特徴のグループに分割したりすることでデータから知識を発掘する処理として、データマイニングがある。ビッグデータには、多量であるがゆえに、よくある特徴的なパターンと共に、希なパターンも含まれているはずで、むしろ、この希なパターンを発見することが個性的な価値創出を生むのかもしれない。

また、大量なデータは、別のポテンシャルも内在している。今までの物理では一般的に、まず観察し、内在する法則を「式」に落とし込んで一般化し、この式を使うことで物理現象を再現してきた。例えば飛行機の場合、まず流体の式でシミュレーションを行い、動きを解析した。しかし、さらに高速になったときには、その式が成り立たなくなる。ビッグデータの解析では、「式」に落とし込む方法とは異なる方法で知識を抽出するとも言える。そのためにはより多くのデータがなければならない。同じパターンが見つけられる程度にデータが大きくなければならないからである。

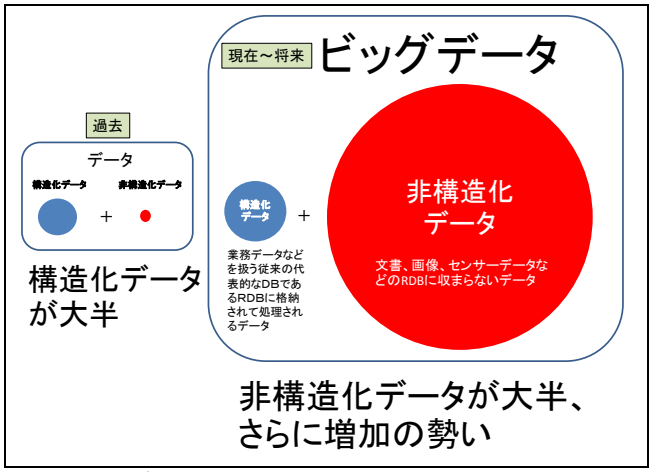
(2) リアルタイム性と多種性

デバイスや「モノ」からのデータ収集が可能になることで、データがリアルタイムに入力され、収集されることになる。

また、様々なデバイスや様々な「モノ」からの多種類のデータ収集が可能になることは、その収集されるデータの粒度が、今まで以上に細かくなり、要素ごとの深い情報の把握ができる。

(3) データ構造

データは、その構造でみると構造化データと非構造化データに大別できる。構造化データは、業務データなどを扱う従来の代表的なデータベース（DB）であるリレーショナルデータベース(RDBとする)に格納されて処理されるデータである。一方、非構造化データは、文書、画像、センサーデータなどの RDB に収まらないデータであり、コンピュータの利用範囲が広がるにつれて、この非構造化データが増えてきている。ビッグデータは、これら両データ構造からなるが、現在から将来は非構造化データが大半を占めるということになり、この点が従来のデータ構造と根本的に異なる注目点である。（図表 2）



図表 2　データの構造上の変化

(4) 不正確さ・あいまいさ

データ量の増大とともにデータ中の不明確さやあいまいさが大きく増加している。データ解析では、様々な不明確なデータを斟酌する必要がある。この点は、ビッグデータ解析のひとつの要点となりうる。

2. 米国政府のビッグデータへの取り組み [1)]

米国政府はいち早くビッグデータに注目する政策を打ち出している。オバマ政権の科学技術政策には具体的に推進するイニシアティブが５つあり、ビッグデータはその一つとして位置づけられている。「ビッグデータ（Big Data）」イニシアティブの内容は以下である。

2.1. ビッグデータ（Big Data）イニシアティブ

ビッグデータの利活用を目的とした研究開発イニシアティブであり、このために新規に２億ドル以上を投じるとしている。ここでは、大規模で複雑なデジタルデータから知識や洞察を引き出す能力を高めることで、国家の喫緊の課題解決に役立てることを目標としている。当初、6 つの政府機関（NSF、NIH、DOD、DARPA、DOE、USGS）が、ビッグデータを取り扱うためのツールや技術の向上に向けた研究投資を行う。次の諸点が目的として挙げられている。

- 大量なデータの収集、蓄積、保存、管理、分析、そして共有のために必要となる最先端の革新的技術を前進させる
- それらの技術を、科学工学における発見の速さの加速・国家安全保障の強化・教育と学習の変容のために利用する
- ビッグデータ技術の開発とその使用に必要とされる労働力を増強する

2.2. 各組織各組織の主な研究概要と研究対象

主な内容を参考資料 [1)]から抜粋して示す。

NSF と NIH では、ビッグデータの科学工学の進展に向けた中核技術の研究開発が行なわれる。具体的には、大規模・多種類のデータセットの管理・分析・可視化・有用な情報抽出の手段となる中核の科学技術の進展を共同でサポートする。NIH は、この中で、分子・細胞・電気生理学・化学・動作・疫学・臨床・健康や病気に関係するデータセットのイメージングに関心を抱いている。

NSF は、前記の中核技術開発に加え、データから知識を引き出す新しい方法、データを管理し、キュレートし、コミュニティへ提供するインフラストラクチャ、教育や人材開発へ、新アプローチを含めた総合的で長期的な戦略を表明している。具体的には、データを情報に変える３つの強力なアプローチである「機械学習」「クラウド（Cloud）コンピューティング」「クラウド（Crowd）ソーシング」を統合する研究に対しカルフォルニア大学バークレイ校を拠点とするプロジェクトにファンディングする。

DoD は、Data to Decisions イニシアティブと名付け、各プログラムを開始している。具体的には、新しい方法で大量のデータを利用し、自ら操作して意思決定ができる完全な自律的システムを作るため、センシング・知覚・意思決定支援などの要素を結びつける、戦闘員や分析者を支援し、オペレーションを高度にサポートできるように状況認識機能を改善するを挙げている。例えば、分析者が任意の言語のテキストから情報を引出すための能力を 100 倍改善することを目指す。また、分析者が観察可能な、オブジェクト数・活動数・イベント数を同様の規模で改善するとある。

DARPA では、半構造化データ（例えば、表・リレーショナル・カテゴリ・メタデータなど）や非構造化データ（例えば、テキスト文書・通信文のトラフィックなど）の両方から成る大量のデータを解析するための、計算手法やソフトウェアツールを開発する。分散データストア内の不完全なデータを処理するスケーラブルアルゴリズムの開発、多様なミッションに応じて迅速にカスタマイズ可能なビジュアルリーズニングを容易にする人間とコンピュータ間の効果的なインタラクションツールの作成などがある。

NIH は、前記の NSF と共同の中核技術開発の他に、クラウド上で利用可能な 1,000 ゲノムプロジェクトの推進を行う。

DoE は、SDAV 研究所を設立する。この研究所は、ローレンスバークレイ国立研究所がリードする形で、6 つの国立研究所と 7 つの大学の専門知識をとりまとめる。そのゴールは、科学者が、データ管理や可視化を容易に行えるような新しく改良されたツールを開発することで、データの管理・解析・可視化の 3 領域における技術的なソリューションを開発・配備し、その使用を通して各分野の科学者を支援する。

2.3. その他の注目すべき点

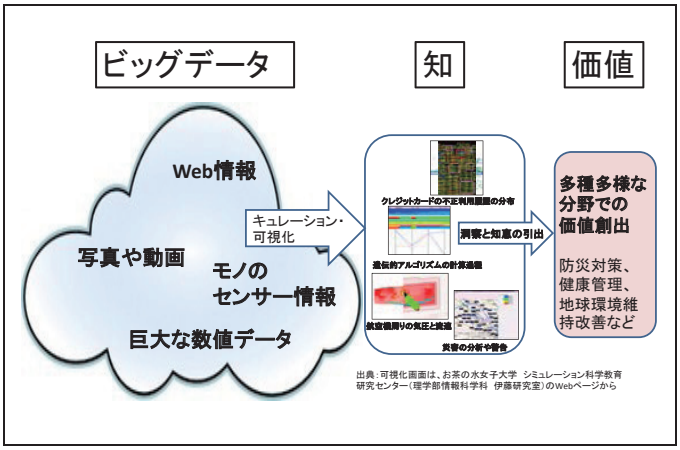
ビッグデータを、インターネット同様、新たなパラダイム創出に寄与しうる科学技術とみなし、様々な領域に非常に大きい影響を与えるものととらえている。また、「可視化技術」「クラウドコンピューティングとの関係」「人材育成への配慮」「産業界および大学の積極的参画」「データ共用」なども重視しており、共同作業を促進するための施設や計算パワーの提供などの推進施策への配慮も見える。

3.　ビッグデータの研究開発推進の注目点

ビッグデータの特徴や米国政府の取り組みを参考に、ビッグデータに関わる研究開発を推進する上で考慮すべき注目点を抽出する。

3.1.　可視化して「価値」を創出する

ビッグデータから価値を創出するためには、多くの課題を総合的に解決していくことが望まれるが、その中でも最大の技術ポイントを挙げるとしたら可視化ではないかと考えられる。解析と可視化が密接な関係になっており、可視化に結びつかない処理は価値の創出につながりにくい。可視化して「知を抽出」し、その後の「価値の創出」への行動に結びつけることが重要であり、この点がイノベーションの源泉であると考えられる（図表 3)。米国のイニシアティブでは、 NSF と NIH、DARPA、NIH、DoE の研究テーマには可視化が盛り込まれており、その重要性を十分に認識している。



図表 3　ビッグデータからの価値の創出

3.2.　データがすでにあるという視点

ビッグデータとその解析について、「集めたもので、多くの人に、（再）利用される」「手に入るデータで何か有意なことをしようというアプローチがビッグデータの基本」[3]という指摘がある。裏返せば、従来は、目的に合ったデータを収集し、それを解析していたと言えるだろう。膨大なデータがすでにあり、これを解析して価値を創出するというアプローチは、今までにはない新しい注目すべき点と言えよう。

3.3.　データの構造変化に追随する処理

図表 2 に示したようにビッグデータは非構造データが大半を占め、今後も指数関数的に増加することは確実である。非構造化データは、企業で言えば、社内の内部情報よりも社外から得られるデータであろう。社内での構造化データと社外の非構造化データとをマッシュアップしてソリューションを作ることにより、顧客ニーズをより取り入れたソリューションが開発される。これは、企業に限らず、研究機関、政府におけるビッグデータ対応にも通じる。いかにして社会の情報を内部活用するかが、今後の研究開発の大きな注目点と言える。前記したが、DARPA も非構造化データの解析を挙げている。

3.4.　データインテンシブ・データドリブン対応

最近、データインテンシブ、データドリブンという用語がよく見られる。共にデータに照準を合わせた処理・対処であり、計算インテンシブ処理に加え正に必要とされてきている。米国のイニシアティブ推進上で、NSF の Subra Suresh 長官は、「米国の科学者は、この新しい「データドリブン革命」によって生じた機会をしっかりととらえて欲しい。現在行っている研究は、新しい事業のための地ならしとなり、数 10 年先の米国の競争力の基盤強化につながるだろう」と述べている。

3.5.　その他の注目点

時間的観点から言えば、ビッグデータの特徴であるデータ入力のリアルタイム化に伴い、即時処理によって、出力やフィードバックをすることが重視されることになる。また、得られるデータの粒度が詳細化できることを利用すると、より細かいセグメントに向けた製品やサービスも提供できるようになる。

3.6.　留意点

データの共用に向けた、コミュニティでのデータの記述や交換のための標準化には意味がある。データのインポート、エクスポート、結合、理解をより容易にし、データの再利用を促す。分野融合の研究もより容易になる [4]。

ビッグデータの議論においては、個人情報やセキュリティなどの解決すべき課題も多い。収集されたデータから個々人の識別や行動が特定されない配慮や注意深い活用が重要である。「データの利活用に関して、法を整備し企業がビッグデータの利活用に萎縮することのない状況を作れば画期的なアイディアが創出する」との意見もある。

(参考文献)

1)「米国政府のビッグデータへの取り組み」（科学技術動向 2012 年 9・10 月号）
2) 喜連川優、「情報爆発のこれまでとこれから」、電子情報通信学会誌、Vol. 94, No. 8, 2011
3)「ビッグデータ高速処理に向けた計算理論的アプローチ」（宇野毅明、国立情報学研究所）
4) “Big data:How Do Your Data Grow? Nature 455, 2008 (Lynch, Clifford)