ビッグデータの潮流とデータエコシステム

Big data and data eco-system

喜連川 優[1]

KITSUREGAWA Masaru[1]

1 東京大学生産技術研究所（〒153-8505　東京都目黒区駒場4-6-1）
1 Institute of Industrial Science, The University of Tokyo (4-6-1 Komaba Meguro-Ku, Tokyo 153-8505)

原稿受理 （2012-10-16）
情報管理 55(10), 705-711, doi: 10.1241/johokanri.55.705 (http://dx.doi.org/10.1241/johokanri.55.705)

著者抄録

米国が2億ドルの研究開発投資をするとの発表以来，「ビッグデータ」なるITキーワードが急に取り上げられるようになった。本稿では，ビッグデータの背景について述べると同時に，情報爆発，情報大航海プロジェクトを振り返りながら，その本質について考察する。加えて，ITメディアを取り上げ，動画を用いつつ，ビッグデータの有用性について具体的に紹介する。さらに，ビジネスにおけるビッグデータとして，プローブカーのセンサー情報利活用，科学におけるビッグデータの動きとして第4の科学に触れ，最後に，データの利活用を促進するためのエコシステムの必要性について論ずる。

キーワード

情報爆発，ビッグデータ，Twitter，第4の科学，e-science

1. ビッグデータと情報爆発

2012年3月29日にホワイトハウスから発表されたビッグデータへ2億ドルの研究開発投資を行うというメッセージ[1)]は，大きなインパクトがあり，一般誌を含め，にわかにビッグデータなるキーワードが衆目を集めるに至ったと言える。極めて平易な表現であることも功を奏してか，非IT系の人々にも，「大きなデータ」という言葉は，自分のPCに入りきらない膨大なデータとは，どれくらい大きなものを意味するのだろうという素朴な興味を湧き立たせたのかもしれない。学会でも数多くのセッションやパネルがビッグデータを取り上げているのが実情である。

ビッグデータとは，どれくらい以上のデータを指すのか，なぜ今ビッグデータなのか，誰が作った言葉か等，いろいろと矢継ぎ早に疑問が浮かぶのはごく当然ではあるものの，「情報爆発」と呼ばれた文部科学省特定領域研究を2005年から，「情報大航海」と呼ばれた経済産業省のプロジェクトを2007年から推進してきた筆者は，ITの大きな流れの中で非常に自然に生まれた言葉として感じられ，細かい定義への疑問よりも，これまでのわが国のIT分野における研究開発の方向感の妥当性を再確認することができたよい言葉が生まれたと受け止めている次第である。

動画はHTML版または電子付録でご覧ください。

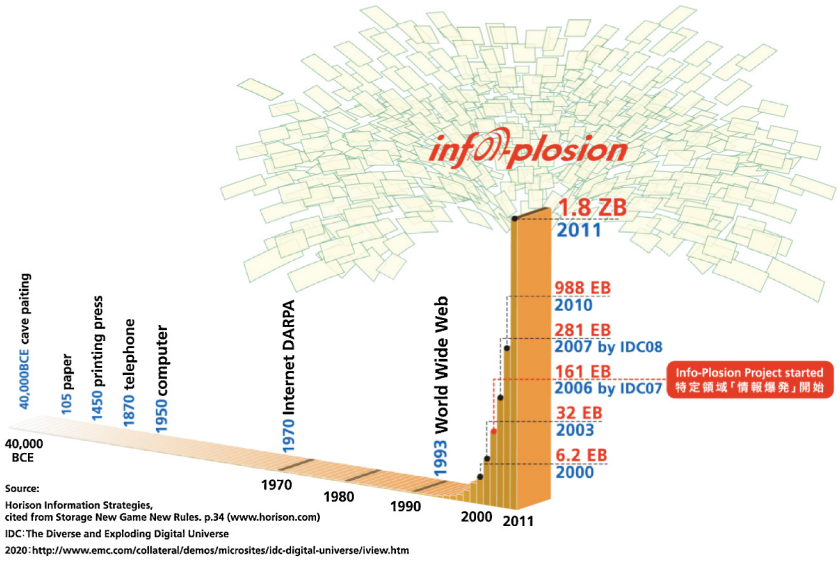


図1　情報爆発

20世紀末から21世紀にかけて，多様な技術の進展により世の中のあらゆる活動が従来に比して非常に精緻に「可観測」となり，結果として，該データの大きさに着目すると，ビッグデータという表現が生まれたと考えられる。ちなみに，われわれは該現象を情報爆発と呼んできた（**図1**）。2004年に情報爆発プロジェクトの申請を計画する際に種々議論を重ねた折には，いわゆるInformation overload（情報の過多による過負荷）が大きな問題として指摘されていた。この現象は現在においても決して解決されているわけではなく，日々，大量の情報が濁流のごとく押し寄せる中で，多くの人々は咀嚼できずにいる，何が本当に正しいのかも判断できずにいるのが実情であろう。もちろん，多様な情報技術の開発が進められているが本質的に困難な課題でもあり，今後も基礎研究の継続が不可欠と考える。一方で，われわれは，これだけ膨大なデータが利用可能となるのは人類史上初めてであり，この機会をチャンスととらえ，新たな価値創出に向けて取り組むべしとの発想に基づく研究も推進してきた。すなわち，情報あるいはデータが過多になることから派生する種々の課題（ネガティブな側面）を自ら同定しつつ解決に向けた研究をすると同時に，今までには考えられなかった大量情報・データを利用した新たなサービス創出に挑戦するという両面作戦を展開してきた。昨今のビッグデータの論点を見ていると，前者ではなく後者を前面に押し出した感が強い。すなわち，大量データの積極的活用とそこからの価値創出に焦点を当てたと言えよう。

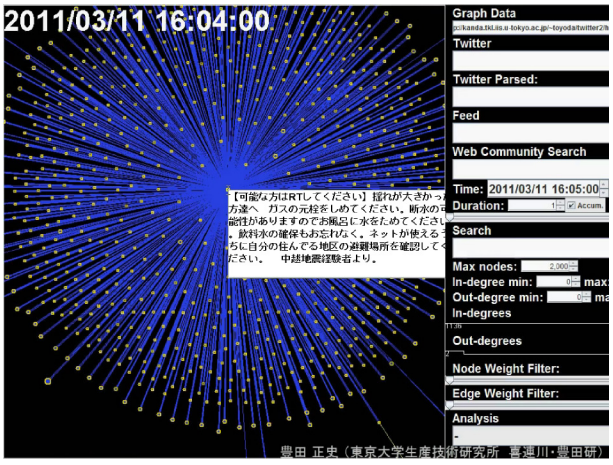
図2　情報爆発に対する2つの視点

2. 膨大なITメディア情報がヒントを与える社会学

20世紀の末よりWebだけではなく，ブログやTwitter，ソーシャルネットワーク等多様なITメディ

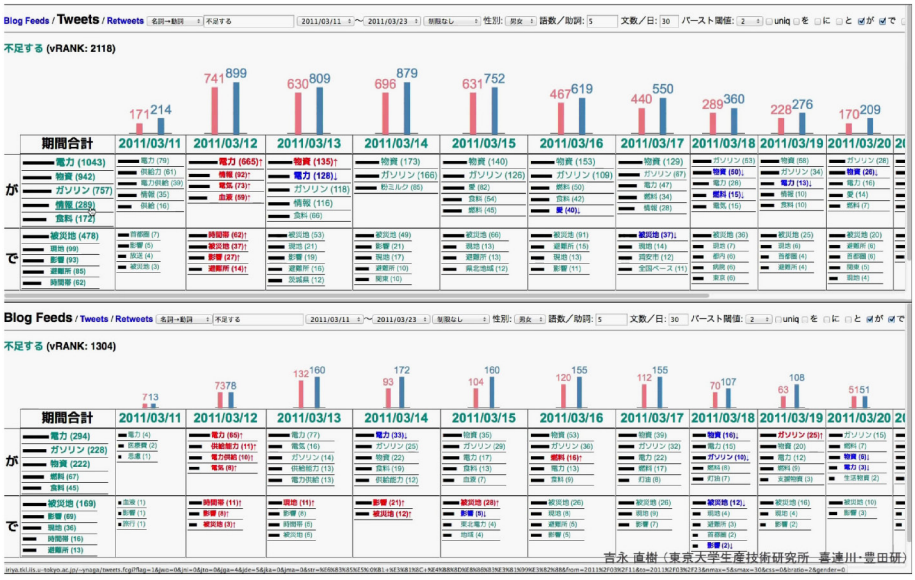
アが多用されるようになった。ブログはクロール可能であり，帯域は制限されているもののTwitterも，部分的には収集可能である。東大の喜連川・豊田研究室では，Webページの収集を1999年より開始し，既に200億ページ以上を収集するに至っている。ブログは2006年から，Twitterは震災時の動きの補足を目的として2011年よりインフルエンサーを中心にクロールを行ってきた。

全国民ではないにしても，また，利用者の年齢には偏りがあるものの，相当数の人々の感情を補足可能となったことは大きな変化と言える。社会を感じ取るシステムを作れるのではないかという考えからsocio-sense[2]と名付けたシステムを2003年より開発してきた。**動画1**に震災直後のTwitterの動きを紹介する。この可視化システムは豊田正史准教授が作成したもので，実際には巨大なタイル型高精細ディスプレイで動作するものであり，より大規模なサイバー空間の挙動解析を可能とするが，ここでは，通常端末の解像度を想定したデモを例示する。3月11日の大震災の際，電話や交通が途絶える中Twitterでは避難場所に関する情報が数多く共有された。震災直後には，以前の震災を経験された複数の方々が，風呂に水をためる，ガス栓を閉める，避難所を確認する，といった極めて的確なアドバイスを発信し，それが多くのフォロワーに伝播している。動画では，点は個々のツイートを示し，線はツイートが他のユーザーにリツイートされたことを示す。このようにして，またたく間に，多くの人々に貴重な経験に基づく示唆が伝播していったことがわかる。もちろん，すべての情報が正しくまた信頼のおける情報であるわけでもなく，愉快犯とも言える情報も少なからずあったことも事実である。今日のITメディア時代において，その利用には配慮が必要であることは事実であるものの，情報の信憑性という問題はITメディアに特有な問題ではないことも事実である。時間を経て多様な避難場所に関する情報が伝播されてゆくことが見て取れるが，動画の最後においては，避難所情報をGoogleマップ上にまとめて公開した人が現れ，同様に広がる姿が見られる。さらに興味深いのは，他のインフルエンサーが，これは素晴らしい試みであると述べ情報提供を呼びかけると，一層広く情報が伝播していくことがわかる。このように，過去には不可能であった速度で情報が拡散する時代に入ったと言える。



動画1　震災時におけるTwitter上での情報伝播の可視化

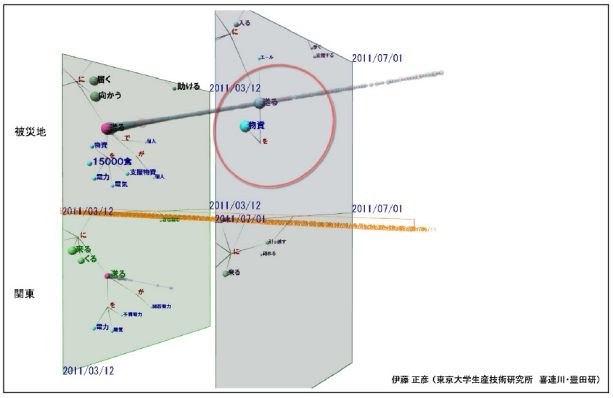
情報の拡散に加え，情報そのものの解析も有益であることは当然である。**動画2**は吉永直樹特任准教授が作成したTwitterとブログの比較解析ツールである。このツールでは，クエリとして行為や感情を指す動詞や形容詞を入力することで，その行為や感情について，人々が各メディアで実際にどのように言及しているかを分析することができる。最初の例は「不足する」という動詞である。収集したすべてのつぶやきとブログ記事を1日単位で構文解析し，「不足する」という動詞の主語・目的語を抽出することで，震災時に人々が実際に「何が」「どこで」不足すると訴えていたかを明らかにすることが可能となる。電力が不足するという結果は，容易に想像可能であるが，震災2日目には情報が不足しているという訴えを多くの人々がしていることは注目に値する。情報を専門としている研究者にとって，災害時の情報提供の重要性をまざまざと提示されたとも言え，真摯に受け止める必要があると感じた次第である。形容詞



動画2　Twitter とブログに記された震災時の人々の行為・感情の分析

を入れることもできる。「怖い」という形容詞に対してその主語を解析することで，恐怖の対象を抽出することが可能となる。余震が怖い，津波が怖いという表現が多いことはうなずけるものの，（地震警報の）音が怖いと感じている人々が非常に多いということは，1つの発見とも言えよう。繰り返し，大きな音で鳴る警報は，人々に恐怖を与えていたことが明らかになった。加えて，これらの感情は，Twitterでは反射的につぶやかれるのに対して，ブログではほとんど捕捉できていない。ブログは多くの場合，瞬間の感情表現に使われるのではなく，一定の時間を経てまとまった意見を述べるメディアとなっていることがその要因と推察される。このように，比較することにより，メディアの特性を浮き彫りにすることもできる。

動画3は伊藤正彦助教が作成した3次元可視化探索ツールである。このツールでは**動画2**で用いた解析結果を，検索語を中心としたツリーで表現し，3次元空間の時間軸上に可視化する。ここでは，検索語として「被災地」と入力することで震災時における「足りないもの」と「提供されたもの・こと」を俯瞰および詳細に探索するデモを例示する。さらに「関東」と入力することで被災地との状況の違いを比較可能



動画3　災害時における人々の行動・興味に関する時系列推移の3次元可視化・探索

にしている。動画ではまず，3月11日の震災の後，被災地および関東で「足りない」ものを探索する。震災後数日間は，被災地では血液，人手，ガソリン，電力，および物資などさまざまなものが不足している。さらに，長期的に何が不足しているのかを俯瞰し詳細を探索することで，被災地では6月半ばになってもボランティアや人手が足りていないことが容易に発見できる。次に，「被災地に行く」に焦点を当てて見ることで，ボランティアに行くという活動が非常に高頻度かつ長期間にわたって行われていたことが見て取れる。さらに，関東に関するパネルを追加することで，被災地および関東に何が送られたのか

を比較および探索する。関東では関西電力から電力が送られることが話題になっていたことがわかる。「被災地に送る」という活動は非常に長期にわたっているが，6月後半から7月頭にかけて特に活発になっていることが見て取れる。新たなパネルを加えることにより，何を送ることが特に話題になったか，震災直後との違いは何か，それらがどの程度継続的な活動かを詳細に探索し，「物資を送る」が特に強く話題になったことを確認している。最後に，「送る」に関する全期間のノードを展開表示することで，被災地には非常に多様なものが長期にわたって送られてきたことが一望できる。可視化ツールを用いることにより，災害時に，どこで何が足りなくて，何が供給されたのかなど，複数の状況・行動の長期的な変化を俯瞰しさらに詳細を比較することが容易になり，さらに，長期的な需要と供給のズレを浮き彫りにすることも可能になってきている。

これ以外にも，もちろんいろいろな解析が可能であり，実際，多くを試みたが，ここでは誌面の都合上，膨大なITメディアに関する話題を閉じることとする。

3. ものがしゃべる時代に

大量のITメディア解析が拓く新しい世界を紹介したが，人のみならず，ものが発生する大量のセンサーデータがビッグデータの源泉となる。米国においてはスマートグリッドの導入が積極的に推進されているが，スマートメータからのデータ量は年間1エクサ（10の18乗）バイトに達するとの予想もある。米国はセンサーが多用されるようになる現代を「age of observation」と呼んでいる。すなわち，センサー技術の進展と低価格化に伴い，社会活動の多くが飛躍的に精細に捕捉可能となり，従来には考えられなかったサービスが次々と生まれると見込んでいる。

プローブカーと呼ばれる自動車の例を取り上げよう。どこでブレーキを踏み込んだか，燃費はどうか等，数多くのデータを取得することができる。配送業者にとってこのデータは極めて貴重であり，多様な利用がなされている。例えば，どのドライバーが運転していようとも，ある場所では急ブレーキの頻度が高いとすると，その場所は見通しが悪い等，避けることが望ましい経路と言えよう。少々時間がかかろうとも安全運転の方がはるかに重要であることは言うまでもない。燃費情報は，運転手のエコドライブランキングに利用できる。燃費が悪い運転手は教育をすることにより，その改善が見込まれる。当然のことながら，経路の最適化や配送の効率化にも利用される。物流はありとあらゆる場所において必要不可欠なサービスであり，ASP化が可能な格好の事例と言える。データがクラウド上に蓄積されればされるほど，より精緻な最適化が可能となろう。加えて，将来の方向性としてリアルタイム化が魅力的である。道路工事は頻繁になされており，加えて，天候により，道路の状況も刻々と変化する。自動車がリアルタイムに道路情報をクラウドにアップロードし他の配送車と共有することが可能になると物流は大きく効率化されることは言うまでもない。

4. サイエンスと巨大データ

筆者がお手伝いしている地球環境に関するデータベースは現在，ディスク並びに，テープを合わせると約10ペタバイトのストレージ規模で運用している。従来はいわゆるリモートセンシング画像がその大半を占めていたが，現時点ではスーパーコンピューターによる温暖化予測データが大きな割合を占めている。高エネルギー物理はCERNのLHC（Large Hadron Collider）[3] などの巨大加速器実験装置により膨大なデータが生み出され現在100ペタバイトを超えている。ゲノムは現在数千ドルで1人分のゲノムシーケンスを取得可能なまで，シーケンサー装置の低価格化が急速に進み，データを取得することにもはや障害は無く，そのデータ量は急速に増えておりその量は予測すらできない。1人分のゲノムデータが1テ

ラ（10の12乗）バイトとすると1億人のデータを取得したとしても気の遠くなるデータ量になる。

さて，観測データと計算データの2つが巨大データの源であることは明らかである。巨大科学は巨大データを生み，世界最高水準のスパコン京が大きなデータ生成源になることは間違いなく，同時に，巨大実験装置が膨大なデータを生み出す。スーパーコンピューターの出現によりcomputational ○○なる学問，すなわち，計算物理，計算科学，計算生物学等多様な学問分野が創出された。Jim Grayによれば，これを第3の科学と呼び，そして，観測データとスーパーコンピューターが生み出す2種類のデータを統合解析する新しいデータドリブンサイエンス（あるいはe-science）を第4の科学と位置づけた[4]。すなわち，膨大なデータを目の前にする時代となり，スパコンをチューニングする従来の高度なプログラミングスキルとは別のスキルベースが明らかに必要となっているという見方である。この考えは第4期の科学技術基本計画（2011年8月19日閣議決定）[5]においても参照されるに至っている。データが科学において重要な役割を果たすこと自体はある意味で明白ではあるものの，その量がPCサーバーで扱える限界をはるかに超える時代において，強固なデータ基盤が科学の進歩の屋台骨を支える最も重要なインフラの1つと言えよう。長尾前国会図書館長の提唱する「知識インフラ」[6]も同じ方向感と言える。

5. データエコシステム

サイエンスではその分野にもよるが，観測データを取得することに膨大な努力が必要とされる場合がある。もちろん，該データを分析，モデル化し，新たな知見を得るプロセスもきわめて重要であるが，データそのものの重要性を再認識すべきであるという考えも生まれつつある。論文だけが引用数によって評価される今日，データは論文の一部に過ぎないことから，データのオープン化はややもすると遅れ気味である。データそのものの重要性を評価し，引用の対象とする試みも始まりつつある。筆者はこのような意識の変化はきわめて重要であると考えている。データが誰によって取得され，そのデータが誰によってどのように加工され，その加工されたデータがさらに，別の研究者によって利用されてゆくという一連のプロセスをすべてきっちりと管理するフレームワークの構築が不可欠である。このエコサイクルが回ることによりデータを取得した研究者にデータを提供する大きなインセンティブが生まれ研究のスピードが大きく加速されるものと考える。データベースの研究分野ではこれらの仕組みをプロビナンスと呼び，誤った結果が得られた場合，どの段階でエラーが混入したかを素早く捕捉することを可能とすることの有用性が明らかにされつつある。NSFでは2009年より，GEO分野でデータの共有が強く推奨されているものの，いまだに必ずしもうまく機能していないようである。Data Cite，ODATAなど多様な試みがなされつつある。3章で述べたようなビジネス分野においては，データを第三者が利用する場合，その制度やガイドラインの整備が未だ不十分で一層の努力が望まれるものの，インセンティブは明快であり，データがイネーブラとなるデータドリブンスタートアップが生まれつつある。一方，サイエンスにおいては，エコシステムのデザインは必ずしも容易ではない。わが国だけの問題ではなく，世界的に見て，エレガントかつ実効的なデザインがサイエンスの進展上強く求められるところと言えよう。

6. おわりに

動画を記事に含める最初の試みとして，執筆のご依頼を頂戴した次第であるが，動画の説明は図の説明以上に難しく，動画の作り方も含めて，今後の大きな挑戦になろうかと感ずる次第である。

ビッグデータの動きについて概観した。わが国においても，遅れること無く，JSTをはじめ，ビッグデー

タの研究が推進されることを切に祈願する次第である。大きなデータを扱い，そこから新たな発見，価値創出を行うためには，大きなデータを自在に操ることを可能とする基盤が不可欠である。これまでとは異なるファンディングスキームが必要となろう。データのエコシステムのデザインをはじめ，新しいルール作りも必須であり，研究支援機関が果たす役割はビッグデータ時代において益々重要になると考える。

参考文献

1) Office of Science and Technology Policy. “OBAMA ADMINISTRATION UNVEILS “BIG DATA” INITIATIVE: ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS”. White House. 2012-03-29. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf, (accessed 2012-11-15).
2) Kitsuregawa, Masaru; Tamura, Takayuki; Toyoda, Masashi; Kaji, Nobuhiro. “Socio-Sense: A System for Analysing the Societal Behavior from Long Term Web Archive”. Progress in WWW Research and Development: 10th Asia-Pacific Web Conference, APWeb 2008, Shenyang, China, April 26-28, 2008. Proceedings. Springer, 2008, p. 1-8.
3) LHC アトラス実験．http://atlas.kek.jp/, (accessed 2012-11-15).
4) Hey, Tony; Tansley, Stewart; Tolle, Kristin. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009, 252p.
5) 文部科学省. “第4期科学技術基本計画”. 科学技術政策. http://www8.cao.go.jp/cstp/kihonkeikaku/kihon4.html, (accessed 2012-11-15).
6) 長尾真. “知識インフラの構築”. 総合科学技術会議．http://www8.cao.go.jp/cstp/tyousakai/seisaku/haihu05/nagao.pdf, (accessed 2012-11-15).

Author Abstract

“Big data” has increasingly attracted our attention as an IT keyword since the U.S. announcement of 200 million dollar investment in R&D. This article reviews how the concept evolved and what its essence is. In Japan, Information Explosion (Info-plosion) project (2005-2010) by MEXT and Information Grand Voyage project(2007-2009) by METI were conducted, which had been launched based on quite similar motivation. An actual example video on twitter and blog diffusion and its content analytics shows the new value created by big data. Business applications which utilizes probe information generated by car are also introduced. Big data also are giving an impact of style of research in science:fourth paradigm. Finally the importance of data ecosystem is addressed so that various types of data driven innovation be enabled.

Key words

Info-plosion, big data, Twitter, data centric science, e-science