ビッグデータ統合利活用における課題と技術

Subjects and Technologies for Utilizations and Applications of Big Data

中野美由紀

Abstract

ビッグデータで象徴されるように，Web 上のソーシャルネットワークシステム（SNS）を筆頭に日々大量のディジタルデータが生成されている．継続的に生み出されるディジタルデータの収集，蓄積，管理はもとより，従来とは比較にならないデータ量に適した即時解析技術，新たに得られた解析データも含めたデータ共有技術と，データ工学技術の発展が求められている．情報分野のみならず，科学技術分野あるいは社会における多様な問題に対するビッグデータの課題と期待される技術について解説する．

キーワード：ビッグデータ，データ工学，価値創出，データ解析

1. ビッグデータとは

社会が生み出すディジタルデータが爆発的に急増する中，「ビッグデータ」[(1)]というキーワードが近年，脚光を浴びている．Web を例に取るまでもなく，我々の身近においてディジタルデータは常に増大し続けており，必要なデータを単純に探し出すにとどまらず，探し出したデータを有効に利用するために新たな付加価値が求められている．情報爆発，情報大航海，Internet of Things（IOT：もののネット化），Smarter Planet，Central Nervous System for the Earth，Smart＋Connected Communities，Cyber Physical Systems（CPS）など多様なキーワードで，情報社会の未来探索が模索されてきたが，ここ数年は「ビッグデータ」という言葉に集約された感がある．ビッグデータとは，実世界を反映する多種多様なコンテンツ，あるいは，そのコンテンツに加えて社会活動の要求に即時対応し，多角的に処理・解析を行い，新たな社会的価値を生み出す一連の過程を含む全体を指している．

ビッグデータの代表的なコンテンツにソーシャルネットワークシステム（SNS）から生み出されるコンテンツがある．SNS では利用されるコンテンツもテキスト情報から画像，音楽，動画像と多様になっているが，コンテンツの生成，流通する場もメーリングリスト，Web ページ，掲示板，blog と多様な形態を取る．コンテンツ量も増加する一方であり，既に Facebook のアカウントは 10 億を超え，インドの人口に迫る勢いである．日本人の発言率が高いと言われる Twitter では 1 日平均およそ 5,000 万のつぶやきがあり，最大 1 億 4,000 万件を記録したと言われている．また，UGC として動画像の投稿サイトである YouTube では，2012 年統計値として 1 秒に 1 時間の動画像がアップロードされ，1 日の動画像再生回数 40 億回と報告されている．

データの増加は Web 上にはとどまらない．インターネット上の統計を扱う ComScore によれば，2011 年 6 月にはオンラインショッピング利用者は 2 億 8,200 万人であり，2012 年ブラックフライデー（米国感謝祭の金曜日）のオンラインショッピング件数は前年の 26% 増と世界経済の低迷の影響を受けていない．携帯電話，スマートフォンの普及により常時インターネットにアクセス可能となった現代，人の活動は物理的な移動の有無にかかわらず，Web サービスなどの情報社会基盤を通じて営まれている．行政のオンライン化，様々な経済活動情報，社会情報自体がディジタルデータとして蓄積・解析される．人のみならず，ものを媒体とする情報，例えば，都市における交通制御，セキュリティ支援（監視カメラ等），プラント制御，省電力監視，局所的な気象変

中野美由紀　正員：シニア会員　芝浦工業大学教育イノベーション推進センター
E-mail miyuki@shibaura-it.ac.jp
Miyuki NAKANO, Senior Member (Center for Promotion of Educational Innovation, Shibaura Institute of Technology, Tokyo, 135-8548 Japan).
電子情報通信学会誌　Vol.97 No.5 pp.343-347 2014 年 5 月

動情報など枚挙にいとまがない．

これらの情報から新たな情報が創造され，発信源となる場が整い，ビッグデータ時代はデータ双方向性により更に広がると考えられる．

2. ビッグデータの統合利活用における課題

ビッグデータを概観したが，データ工学分野の観点から，ビッグデータの課題は大きくは，①増大するデータの蓄積・管理，②大容量データの効率良い処理，③多様なデータからの価値あるコンテンツの抽出・生成，④プライバシー保護とセキュリティ，が挙げられる．

2.1 増大するデータの蓄積・管理

ビッグデータの特徴は 1. で述べたように保持するデータが日々増大していくことにある．従来のデータベース管理では，例えば過去のデータは磁気テープなどの三次記憶にバックアップとして蓄積されるが，多くの場合，再度頻繁にアクセスを行うことは想定していなかった．しかしながら，ビッグデータ時代においては，蓄積されたデータは全て利用することが前提となる．それは，現時点において，ディジタルメディアとして保持されているデータに加え，日々増えるデータを過去のデータと等しく処理できるように管理することを意味している．更には，社会活動により蓄積されたディジタル化されていないデータの取込みも大きな課題となる．紙などのアナログ媒体に記録されたデータは単位，表記の揺れ，データベース化を想定してない記述形式など，個々のデータごとにディジタル化の問題が異なる．過去のデータの取込みの難しさは，我が国の社会保険のディジタル化において多くの欠損が生じたことからも明らかである．過去のデータのディ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ィジタル化することで，埋設電線の補修を効率良く行えるようになり，マンホール事故を未然に防ぐことが可能となったニューヨーク市の事例のように大きな効果が期待できる[(2)]．

2.2 超大容量データの効率の良い処理

ビッグデータの価値は保持・蓄積しているだけでは得られない．現代社会は「Right Now（今すぐ）」な社会と言われる．我が国では 2011 年度末に携帯電話の台数が総人口を超え，スマートフォンの急速な普及もあいまって，誰しもがいつでもどこからでも必要な情報を得ることが可能な状況となっている．成長著しい SNS を例に取れば，ユーザが増大すると，流れる情報量はユーザ間の相互のやりとりにより爆発的に増加する．この膨大な情報から，各ユーザへの即時情報提示，適切な情報推薦，効果的なオンライン広告の提示など，SNS の典型的な機能を効率化する上では，処理の高速化と結果の高精度化という相反する要素が同時に求められている．すなわち，高精度な結果を得るためには多くのデータを処理する必要があるが，高速化するためには対象となるデータ量は少ないほうがよい．従来は計算機資源の限界により，扱えるデータに制限があった．しかしながら，ビッグデータ時代である現在，保持しているデータを全て利用することが新たな発見を可能にすることであり，全データを対象とした効率の良い解析処理が大きな課題となる．SNS は人によるデータの生成が主となるが，現在，センシング技術の発達により，社会の様々な場所にセンサが設置され，多量の情報がリアルタイムに処理されている．社会的に大きな影響がある事件，事故，災害と直結した事例も多く，地震，津波などの災害検出，大規模プラントの異常検出，インターネット上のサイバーテロ攻撃検出など，即時性が更に重要な課題となる．

2.3 多様なデータからの価値あるコンテンツの抽出

既に述べているように，ビッグデータ時代では，蓄積されている全てのデータが解析の対象と考えられる．したがって，センサが毎秒ごとに送出する気温データから人が書き込む記事，画像，音声・音楽，ビデオ，あるいはシステムの稼動ログまで，多種多様である．

個々のデータに対応する技術はテキストマイニング，Web マイニング，Web グラフ解析，あるいは，マルチメディア認識などの研究分野で大きく進展している．その結果，誰しもが容易に Flickr や YouTube などのサイトに写真，動画像をアップロードできる．これらのサイトでは，ユーザのテキスト情報，アクセスログ情報などから，キーワード検索や推薦ランキングなどの機能を提供している．また，Facebook，Twitter などの SNS でテキストに写真や動画像も紹介されている．Follow や「いいね！」などの推薦機能により，情報の拡散や推薦が可能となる．一方，音声認識では，「しゃべってコンシェル」や Siri などの音声認識ツールが簡単に利用できる．音声認識ツール自体は音声認識技術に加え，辞書，大規模知識データとの連携による適切な語句変換機能など，ビッグデータを利用したアプリケーションの代表例であろう．

多種多様なデータの中から「価値のある」コンテンツの抽出，生成を行うためには，①多種なデータを同時に扱い，②多様なデータの中から使える結果を探すという二つの課題が挙げられる．

多種なデータを同時に扱うことの難しさは，前述の推薦の例でも，多種なデータ（テキストと画像）を対象にしてはいるが，画像そのものの情報（色，形状），動画像の情報（音声，画像の変化）を本質的には利用していないことからも分かるだろう．テキスト情報も含め，

Web 上には構造化されていないデータが多数ある．構造化されたデータは数値による大小比較，マッチングなどが容易に行えるが，非構造化データの処理には多くのコストが掛かる．また，非構造化データそれぞれの特性（写真，動画像，音声，音楽，テキスト等）に合わせた加工が必要となる．複数の非構造化データが存在する中で，非構造化データに対する統合的なデータ解析処理はビッグデータ時代における大きな課題である．

また，多様なデータから有用な結果を探すためには，どのようなデータを組み合わせ，得られた結果をいかに評価するか，という更に難しい課題がある．個々の問題設定ごとに「価値のあるコンテンツ」は異なるため，統合的な仕組みによる簡単な数値化などへの置換えが難しい．データサイエンティストという言葉で代表されるデータ解析の専門家が求められる最大の理由である．複数のデータ群の中から抽出された結果が，「価値のある」コンテンツと見定めるためには，最終的には特定の問題に関する専門家による評価が必要となる．

2.4　プライバシー保護とセキュリティ

大規模データの利用に際して問題となるのはそのデータ利用に対する安全性である．現在のインターネット社会ではクリック一つの動作も全て収集されている．Web サービスでは，ユーザの選択，購買履歴に基づいて適切な商品などの情報提示を行う．検索サイトである地名を調べれば，すかさず，その地名に関連した宿泊情報が広告として提示される．多数のユーザから得られる膨大なアクセスログはその情報を提供しているユーザにとっても有用な結果としてフィードバックされる．

このような膨大なアクセスログから個人を特定できる情報がのぞかれ，匿名化されたデータはほかの購買履歴，アクセスログなどとともに解析することで，新たなビジネスの機会，知見を得ることがビッグデータ時代における大規模データ解析処理である．しかし，個人のし好を反映したアクセスログは，匿名化を施されてもなお，現在のデータ解析処理能力により個人が特定できる可能性がある．2006 年に AOL が研究者向けにオープン提供したアクセスログ情報は，地名を含んだキーワードを解析の結果，実在の人物が特定されてしまった[(3)]．また，別の事例では映画の推薦システムのログが公開された．このログは匿名化がなされていたが，別の同種の推薦システムは実名入りであった．二つのデータを組み合わせることで，どちらのシステムでも同じ人は同じ映画を公開時に推薦するという行動から，匿名化のデータでも個人の特定が可能となってしまった[(4)]．

匿名化技術，あるいは，個人が特定できないようなデータ処理により，個人のプライバシーは一見保護されている．しかし，ビッグデータ時代では，個々のデータの匿名化がなされていたとしても，複数のデータソースを組み合わせることで，当初想定していた匿名性が消える可能性は高い．また，複数のデータを組み合わせて得られた結果自体を二次利用する場合，元々のデータ提供者はそれを許諾しているかは，曖昧な状況である．これもまた，ビッグデータの特徴とも言える異なるデータの統合的な解析において，配慮しなくてはならない点と言える．

3. ビッグデータの統合利活用における技術

ビッグデータの統合利用における課題に対し，必要となる技術，あるいは提供されている技術について概観する．

3.1　大規模データの蓄積・管理技術

現状のビッグデータ時代の到来を可能としたのが大規模計算機基盤技術であり，簡単な言葉でまとめれば，データセンターとクラウドコンピューティングと言えるであろう．大規模データセンターによる集約的な計算機資源の管理により爆発的に増えるデータに対応すると同時に，クラウド技術により集約された計算機資源を効率良く様々なアプリケーションにて利用することが可能となっている．

ビッグデータでは異なるデータを解析することにより一層価値の高いコンテンツを得ようとする．そのためには，様々な媒体から異なる手法で蓄積されるデータを利用しなくてはならない．そこで，データクレンジング技術，データ統合技術などが今まで以上に重要となる．また，ビッグデータの特徴の一つはデータ全件を解析対象としていることであろう．蓄積されてきたデータの全てを利用することを意味している．つまり，従来は破棄されていたデータも保持することとなる．一方でハードウェア資源は有限であるから，全件処理するまでもない不要なデータは積極的に廃棄しなくてはならない．

また，プライバシー保護・セキュリティを担保するために，あらかじめデータ加工する工夫も必要となる．例えば，個人情報の消去や特定できる情報の削除，置換えなどの匿名化，あるいは，データを共有して処理するために，データに対するセキュリティを組み込むなどの工夫も必要となる．

3.2　超大容量データ解析処理

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 2.5 PByte ずつ増えている．このような超大容量データの解析処理を 1 日のうちに行い，数千台の配送トラック，数億個の在庫，数千店の店舗に対し，当日の配品，在庫の発注，各店舗への配送ルートなどを決定しなくてはならない．また，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．日々更新される超大規模な情報をやはり10億を超えるユーザへ，常に最新の情報を提示しなくてはならない[5),(6)]．

上記のような大規模データ処理の高速化には，月並みではあるが高速な二次記憶としてのSSDの利用，数千台に及ぶ並列処理化が図られている．また，Walmart社のように単一のデータベースによるデータ解析処理を行う例もあるが，Google，Facebookなど，リポートされているように[7),(8)]データは分散格納され，並列処理される事例が多い．クラウドコンピューティングにおける典型的な並列処理ツールとしてMapReduce，Hadoopなどのオープンソースが提供され，広く利用されている．また，OpenStackなどのクラウド基盤技術の開発も行われ，異なる企業，プラットホーム上のデータを共用しやすい環境が作られている．

3.3 価値あるコンテンツの抽出

データの中から価値ある内容を見いだそうとする試みは社会の中では常に行われている．一定期間で行われる国勢調査，各省庁が出す白書などの伝統的な統計情報から企業の売上げ集計まで社会活動の一環として組み込まれていると言ってよい．それが今，ビッグデータという言葉が大きく取り上げられているのは，解析の対象が非常に大きくなっているというだけではなく，そのデータを「全て」処理することが可能な解析処理基盤が実現したからである．

文献(2)では，「因果関係から相関性」という言葉で表現している．文献(2)では，典型的な事例として，Amazonの書評販売が紹介されている．Amazonは書籍のオンライン販売から始まった．当初は書評委員による推薦という形で顧客に次の書籍の推薦を行っていた．これは何がしがある本を購入したら，同じ著者の本を推薦する，あるいは，同じ分野の本が推薦される．つまり，赤ちゃんの本を購入した人には妊娠に関連した本が何冊も推薦されることになる．そこで，書評委員の推薦とは別に，売上げの記録から本の購入に関する相関性を導き出し，とある本を購入した人が次に購入した本を薦めることにした．その本を次に購入したかの理由は突き詰めず，次に購入したという事実のみに着目したのである．結果，相関性の高い本を薦める方式の売上げが書評委員の推薦に勝つに至った．

ビッグデータという言葉が現れる15年前に，データ工学の分野では「データマイニング」が一躍脚光を浴びた．動機はまさにビッグデータと同じく，データの山から価値あるデータを掘り起こすことである．それまでは，計算機パワーの制限で全件比較するコストが高かった相関関係処理が，効率良いアルゴリズムの提唱[9)]とともに急速に普及したのである．全件検索による相関解析はいまやデータ解析処理システムには通常の機能として提供されている．相関解析は有用なツールであるが，一方，多量のデータに適用すれば多量の結果が得られ，ややもすると人による判断が難しくなる．

人が見る範ちゅうを超えた多量のデータをいかに扱うかが現在のビッグデータの課題である．データの分類，カテゴリー化するためのクラスタリング，クラシフィケーションなどの技術に加え，機械学習が大きな注目を浴びている．データ探索，解析のための機械学習アルゴリズム及び具体例に関しては，本特集にて専門家からの紹介があるのでここでは触れない．ゼタバイトのデータの全件解析を行うには，有用なデータの関係性，あるいは，単純な統計解析では見つけることは到底できない量の壁が存在する．機械学習アルゴリズムを用いることで「因果関係は分からないが相関性のある」データを抽出し，人が理解可能な形まで結果を集約できることが期待されている．

3.4 データを守る技術

プライバシー保護技術，データアクセスセキュリティ技術，暗号化技術など，データを守る技術は様々である．しかしながら，課題で挙げたように，ビッグデータにおけるデータセキュリティの問題は，そこに保持しているデータセットの匿名性が複数のデータ群を扱うことにより，予想をしない形でその匿名性が破られる可能性が存在することにある．また，一度流通したデータ，あるいは，二次利用されたデータを回収することは，ディジタルの世界ではほぼ不可能である．

そこで，データ品質の保証あるいはその経歴を何らかの形で示すlineage/provenanceなどの機能や物流では当たり前になったトレーサビリティなどの機能をデータ自体に付与するようなシステムが必要となろう．

一方で，ビッグデータ時代においては，データを保持すること自体に価値がある．これは，従来の社会にはない財産が出現したと考えられ，今までの法規則，商慣習にはなじまない事態が起きると考えられる．ビッグデータ時代におけるデータ利活用では「データ利用のガナバンス」について早急に検討すべき段階に来ている．

4. ま と め

ビッグデータとは何か，ビッグデータ時代におけるデータ工学の課題とそれを解決するための技術について概観した．なかでも，本特集に紹介される機械学習アルゴリズムはあふれるデータを人が理解するための大きな手段として期待されている．「ビッグデータ」は超巨大なデータを現象として観察するだけではなく，そのデータを用いて人類に有用な結果を導き出すことに本質がある．本稿以降で紹介される様々なアルゴリズム，処理技法とその結果を見ることで，巨大なデータから抽出され

る「ファクト」の面白さを楽しんで頂きたい．

文　　献

（1） McKinsey Report, “Big data : The next frontier for innovation, competition and productivity,” May 2011. http://www.mckinsey.com/mgi/publications/big_data/pdfs/MGI_big_data_full_report.pdf

（2） V. M.-Schonberger and K. Cukier, Big Data, A Revolution That Will Transform How We Live, Work and Think, Eamon Dolan/Houghton Mifflin Harcourt, 2013.

（3） M. Barbo and T. Zeller Jr., “A face is exposed for AOL searcher No. 4417749,” New York Times, Aug. 21, 2006.

（4） R. Singel, “Netflix spilled your bokeback mountain secret, lawsuit claims,” Wired, Dec. 17, 2009.

（5） 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, Sept. 2013, http://www.sas.com/resources/whitepaper/wp46345.pdf

（6） P. Russom, “TDWI best practice report : MANAGINGBIGDATA. technicalreport,” TDWI, 2013, http://tdwi.org/research/2013/10/tdwi-best-practices-report-managing-big-data/asset.aspx?tc=assetpg

（7） T. Claburn, Google plans to use Intel SSD storage in servers, 2008, http://www.informationweek.com/infrastructure/storage/google-plans-to-use-intel-ssd-storage-in-servers/d/d-id/1067741

（8） D. Mituzas. “Flashcache at Facebook : From 2010 to 2013 and beyond,” 2013, https://www.facebook.com/notes/facebook-engineering/ashcache-at-facebook-from-2010-to-2013-and-beyond/10151725297413920

（9） R. Agrawal, T. Imielinski, and A.N. Swami, “Mining association rules between sets of items in large databases,” Proc. of ACM SIGMOD93, pp. 207-216, 1993.

（平成 26 年 2 月 21 日受付）

中野　美由紀（正員：シニア会員）

東大・理・情報科学卒．博士（情報理工学）．富士通株式会社勤務．1985-07 東大生産技術研究所助手（2004 助教）．2008-07 特任准教授．2013-11 芝浦工大教育イノベーション推進センター教授．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，情報処理学会，ACM，日本データベース学会各会員．