



データを読み解く技術

——ビッグデータ，e-サイエンス，潜在的ダイナミクス——

特集編集にあたって

編集チームリーダー 麻生英樹

近年，インターネットやセンサ技術の進展に伴い，サイバーと実世界の両面で，膨大，かつ，多種多様な情報が大量に流通し，利用可能になっている。これに伴い，高度な電子情報通信技術を駆使してこうしたデータを有効に活用し，科学・工学的な課題や，社会的な課題を解決することが強く期待されている。

こうした状況を背景として，本特集では，大量データの処理やそこからの学習的な情報処理に関連する「データ工学研究専門委員会」，「コンピューテーション研究専門委員会」，及び「情報論的学習理論と機械学習研究専門委員会」の御協力を頂き，大量のデータを読み解くための最新技術の展開について，それぞれの分野で日本を代表するような気鋭の研究者の方々に，多様な側面から解説をして頂いた。

特集の全体は大きく三つの部分から構成されている。まず第1部では，「ビッグデータを読み解く機械学習技術」として，最新の機械学習技術について，それぞれ応用を含めた形で解説して頂いた。ビッグデータを扱う技術として，現時点ではHadoopなどに代表される分散データ処理技術があり，その上に構築された集計やクラスタリング，オンライン学習処理といった軽い情報処理が主流である。しかしながら，今後，より一層データを活用してゆくためには，高度な機械学習技術が適用されるようになると期待されている。

ここでは，まず冒頭に「ビッグデータ統合利活用における課題と技術」として，ビッグデータの現状とその利活用に関する俯瞰的な解説を頂いた。次に，「ベイズモデルに基づく関係データ解析技術」として，最近話題になることの多い「関係データ」すなわち，SNSから得られるような人と人の関係，取引状況などから得られる企業と企業の関係，あるいは購買データなどから得られる人との関係，などに関するデータを読み解き，対象のグループ化などを行うための汎用的な技術について解説頂いた。次の「密度比推定によるビッグデータ解

析」では，確率密度関数の推定をせずに，密度の比を直接推定することによって，様々な種類のデータの解析が，効率良く，簡便に，しかも頑健に行えることが示されている。4番目の「生命科学データからの組合せ発見問題」では，生命現象のような非常に多数の変数に関わる複雑な現象において，重要な変数の組合せを発見するとともに，その重要性を統計的に検定することで有意性を保証するための最新技術が紹介されている。最後に「ビッグデータに挑むクラウドソーシング」においては，ビッグデータの解析処理に大量の人間の力（＝クラウド：計算機のクラウドとは異なる）を利用することで，計算機だけでは解決できない課題の解決を可能にする「クラウドソーシング」について，その可能性や最新の研究が紹介されている。

ビッグデータの応用領域は，大きく分けて，ビジネス領域，公的サービス領域，そして，e-サイエンス領域に分けることができるが，第2部の「e-サイエンス時代のアルゴリズム研究」では，実験装置とデータベースやデータ分析計算機を高速ネットワークを介して接続し，それらの設備を利用することで大規模データを活用しつつ行う，新しい科学の方法論「e-サイエンス」のために必要となる最新の技術を，主に情報処理アルゴリズムの観点，特に，グラフ解析や検索などの観点から御紹介頂いた。また，その社会的な応用事例として，大災害時の通勤困難問題や避難行動計画問題についても解説を頂いている。

冒頭の「e-サイエンス時代に向けたアルゴリズムの最新潮流」では，多項式時間での処理を前提とした従来の<高速>アルゴリズムでは対応することはできないペタスケールやそれ以上のデータを扱うための新しいパラダイムを，具体的な社会的課題に即して構築してゆくための総合的な計画について御紹介頂いた。次に「次世代スーパーコンピュータ技術を用いた超大規模グラフ解析と実社会への応用」では，ビッグデータ解析の最重要課題の一つである「超大規模グラフ解析」について現時点での技術と今後の技術的目標を解説して頂いた。グラフ解析は，第1部の「関係データ解析」とも密接に関わっており，機械学習的なアプローチとデータマイニング的なア

アプローチとを対比しながら読んで頂くのも興味深いのではないかと思う。3番目の「ビッグデータのための簡潔データ構造」では、ビッグデータを高速処理するための必須技術の一つとなっている「簡潔データ構造」について、基礎からビッグデータへの適用までを含めて解説して頂いた。次の「統計モデルを活用したビッグデータ検索超高速化」では、たん白質の立体構造データベースの検索を例題として、対象となるデータの統計的構造を利用した超高速検索アルゴリズムの設計について解説して頂いた。この技術は、第1部の統計的機械学習と第2部のデータマイニングとをつなぐ架け橋とも考えられるものであり、元々異なる出自を持つ両者が歩み寄っている様子も伺えて興味深い。簡潔データ構造も含めて、現在普通に使われているような超高速の検索などの影では、こうした技術が高度に発展していることを感じて頂ければと思う。5番目の記事である「首都圏における大地震発生後の通勤困難問題」と6番目の「避難計画問題への離散アルゴリズムの適用」では、大規模な交通ネットワークの解析事例として、大災害時の通勤困難率の地理的な分布の推定、津波に対する避難計画の作成といった問題へのアルゴリズム適用結果を御紹介頂いた。抽象的なデータやアルゴリズムが、身近な具体的問題にどのように利用される可能性があるのかを感じて頂ければと思う。

第3部「潜在的ダイナミクス——深い変化を読み解く——」では、データを読み解く技術の新しい研究方向の一つである「潜在的ダイナミクス」の研究の広がりについて御紹介を頂いた。データを読み解くということは、観測可能、あるいは観測可能なデータから、観測不可能、あるいは観測困難な潜在的な構造を読み解くということでもある。多くの情報処理の課題はこうした問題として捉えることができるが、ここで紹介している潜在的ダイナミクスでは、更に新たな方向として、データから、潜在的な構造の「変化」を検出し、読み解くことを目指している。

最初の記事「潜在空間モデリングによる時系列からの再構成」では、多次元の時系列データの処理に潜在的ダイナミクスの考え方を導入した事例について御紹介頂いた。具体的には、脳磁図から脳内の信号源を推定する問題、オンオフする信号源からの混合信号をブラインドで分離する問題、ネットワークの観測データから潜在的な

ネットワークの構造変化を推定する問題を取り上げて、新しい汎用的な枠組みが適用できることを示している。次の「潜在トピックモデルを用いたデータマイニング」では、主にビッグデータの一つである文書データを対象として、文書集合から話題である「トピック」を抽出する技術、更に、そのトピックの時間的変化を抽出する技術について解説して頂いた。3番目の「潜在的ダイナミクスと異常検知」では、潜在的なグラフ構造の変化を「異常」として検知するための最新手法について御紹介頂いた。この記事は、第1部の密度比推定に基づく異常検出手法とも関連しており、実世界の深い構造を知ることの有効性について考えるヒントにもなると思われる。4番目の「「都合」のつながり——イノベーションの潜在ダイナミクスとして——」では、一転して潜在的ダイナミクスとイノベーションとの関係を解説頂いている。ステークホルダー間の会話のログデータを分析するにあたって、それぞれの行動の背景にある状況、意図、制約を、それぞれの「都合」としてまとめて扱い、そのダイナミックな変化を考察することで、イノベティブな提案を生み出すような会話への道筋が示されている。最後の「潜在的ダイナミクスの学習理論」では、潜在的ダイナミクスの推測を支える基盤的理論について紹介して頂いた。この記事でも書かれているように、潜在構造変化の検知は、変化し続ける現象の本質を読み解く鍵であり、今後多くの分野で重要になる技術として期待されている。

データを読み解く技術の代表である機械学習やデータマイニングの技術は、近年になって大変多くの研究者が参画し、まさに日進月歩の状態である。そうした中で、ディープラーニングなど今回取り上げることができなかった技術も多く発展しつつある。今回の特集の個々の解説記事は、そうした中で最新の技術、特に、今後重要となるであろう技術に関するものも多く含まれており、やや取りつきにくい部分もあるかもしれない。しかし、興味のある技術や応用について読んで頂くことで、データを読み解く技術の最前線での高速な展開を少しでも感じ取って頂き、少しでも会員各位の研究・開発の御参考にして頂ければと願っている。

本特集に記事を書いて頂いた執筆者の方々に深く感謝するとともに、編集について御協力頂いた各研究会の関係各位及び編集チームメンバーに深く謝意を表する。

特集編集チーム

麻生 英樹	伊藤 大雄	今井 桂子	上田 修功	中野美由紀	山西 健司
今井 篤	河本 満	青木 啓史	荒木 健治	石田 明	植松 美幸
勝山 裕	川村 春美	蔵田 武志	甲田 泰照	佐藤 一誠	柴田 智行
鈴木 雅実	椿 泰範	中沢 実	峯 恒憲	弓場 竜	吉川 大弘
和泉 勇治					