

社会を変えるビッグデータの使い方

多様なデータの統合で見えてきたCPSの具体像と課題

社会システムの稼働情報や、人々のコミュニティ活動情報は、そのほんの一部を切り出してもペタバイトからエクサバイトにのぼる巨大データだ。しかもその形式は多様で、単一データベースに収斂させることすら難しい。センサーネットワークやSNSなどから得られるビッグデータはどう収集・蓄積・分析すれば、実世界に価値あるフィードバックとなりうるのか。ユビキタスサービスに造詣の深いNIIの今井和雄教授と、データ分析に精通する高須淳宏教授に聞いた。



今井和雄 Kazuo Imai
国立情報学研究所
研究戦略室 特任教授

がソーシャルメディアで発信するテキスト情報も有効になるでしょう。交通状況や除雪進捗の把握には、リアルタイムの情報が必要で、タクシーの移動情報の他、除雪車からも例えばスマートフォンを利用して稼働情報を取得したいと考えています。これらのリアルタイムに発生するデータ以外に、過去の交通渋滞データや除排雪記録、交通事故データ、地図データなども状況の知識化のために必要です。全部の情報を集約し、分析・シミュレーション・可視化などの処理を行った結果、最も効果的な除排雪作業指示（意思決定）に結びつけることができ、バスや緊急車両のスムーズな運行や、渋滞

にわたることです。各種のセンサーによる実世界の物理的な状況を示すデータに加え、映像や、TwitterなどのSNSを通じて人が発するテキスト情報も含まれます。

CPSでは、センサーデータに代表されるように、物理世界から逐次情報が得られますので、従来のようにまずはデータをどこかに貯めてじっくり料理するという方法は適しません。データ1つひとつのサイズは小さいとはいえ、膨大な数量に及ぶため、1箇所への集約はもちろん、分散しての蓄積も容易ではないのです。データ圧縮技術のさらなる改良に加えて、データの取捨選択の仕組みが必要になると考えています。

またデータ表現がそれぞれ異なり、統合が難しいことも問題です。以前、銀行などの企業合併の際のシステム統合にあたって、管理しているデータの「名寄せ」が問題になったことがあります。統合するシステムはお互い同じようなデータを扱っているはずですが、その表現がそれぞれ違うので、精度よく統合しようとする表現の違いをうまく吸収して情報を結びつける技術が必要になります。センサーデータの場合も同様の問題が生じます。CPSでは、この問題をデータストリームという、より厳しい状況下で解くことが求められます。

CPSに関わるデータでは、まず、データ発生の「時間」と「場所」をカギにした統合を行います。例えば地図情報とスマートフォンや自動車のGPS情報、SNSに投稿された地名情報などといったデータに絞りこみ、それらを時間軸で組み合わせれば、複数のデータから得られる情報を相互に補完することにより、交通状況をきめ細かに把握できるわけです。

巨大データを分析するには、いつ、どこで、どういうイベントがあったかという切り口でデータを抽出するのが得策です。抽出データに過去の同種イベントの発生記録などを突き合わせると、例えば「A地点で事故が起こると周辺

CPSは現実世界をどう変えるのか？

—CPS（サイバー・フィジカル・システム）が私たちの暮らす現実世界をどう変えるのか、具体的イメージをお聞かせください。

今井 わかりやすい例として、NIIが北海道大学や大阪大学、九州大学と共同で実施しているCPSのプロジェクトにおいて、北海道大学が中心となって計画している「スマート除排雪実証実験」があります。この実験が目指すのは、積雪による道路渋滞を緩和するために最適な除排雪車オペレーションを行うことです。これを達成するためには、さまざまなデータが必要です。積雪状況の把握には、気象データはもちろんのこと、路面状況の計測器からのデータや歩行者

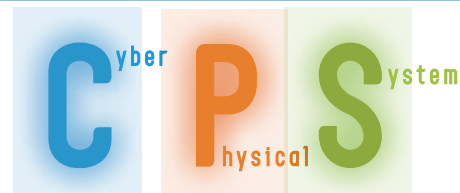
しがちな道路の平均走行速度の向上、また凍結路面の情報発信など事故防止にも役立てられることが期待できます。

膨大で複雑なデータの分析がうまくできれば、例えば公共空間の人の数や位置、動きによって地下街やビル群の空調機器をきめ細かく制御するエネルギー最適利用システムの提供や、災害時には同様の情報を利用して、地下街からの適切な避難経路をそれぞれの個人に提示するなど、さまざまな応用が考えられます。

データ収集・集約・蓄積にまつわる課題は？

—CPSの前提になるデータの収集、集約、蓄積の方法や課題について教えてください。

高須 CPSの課題の1つは扱うデータが多岐



道路状況はこうなる」という予測も可能になるでしょう。NIIの研究プロジェクトでは、自動車のGPSデータから道路の平常時の状況を把握するとともに、平常時と異なる挙動をする自動車を見つけることによって、道路の異常状態をリアルタイムで検知する方法を研究しています。現在は、自動車のGPSデータのみを用いていますが、テレビ放送のニュースやSNS情報の分析を行う研究グループと共同して、情報を多角的に捉える検討も進めたいと思っています。

今井 別の観点からの課題として、プライバシーをどう保護するかという議論も必要です。個人の所在場所、移動状況、SNSへの投稿内容などをどこまで取得してよいのか、また保管や利用の範囲をどうするかについて、社会的な合意が必要になります。

データ管理・分析の方法と課題は？

— データ管理・分析の方法と課題を教えてください。

高須 分析対象のデータには、蓄積された過去データのような「ストレージ系」データ^{※1}と、絶え間なく発生する「ストリーム系」データ^{※2}とがあります。

ストレージ系データは従来からのデータベー

ス管理システム、データウェアハウス、統計解析システムの技術を利用して蓄積・検索・分析されてきましたが、CPSが対象とする膨大なデータを相手に高度な分析を行うには処理性能面でさらに改良が必要です。また、データの背景にある潜在的な情報を抽出するためには、機械学習アルゴリズムの洗練も必要です。私たちの研究グループでは、潜在トピックモデルと呼ばれる統計モデルを用いたCPSデータの分析アルゴリズムの開発に取り組んでいます。

ストリーム系データでは、時間と計算資源の厳しい制約の中で分析を行う必要があります。そのため、いちいちデータベースにデータを格納せずにメモリ上で分析処理し、短時間で結果を出すCEP（複合イベント処理）技術が開発されています。一方で、データストリームの分析では、対象データの範囲を絞ったり広げたりしながら探索的に分

析を行うことも必要になります。現在、データストリームの分析技術と分析のニーズの間にはまだまだ大きな技術的ギャップがあります。このギャップを埋めることは、今後の大きな課題です。

高須淳宏 Atsuhiko Takasu

国立情報学研究所
コンテンツ科学研究系 教授
総合研究大学院大学
複合科学研究科
情報学専攻 専攻長
教授



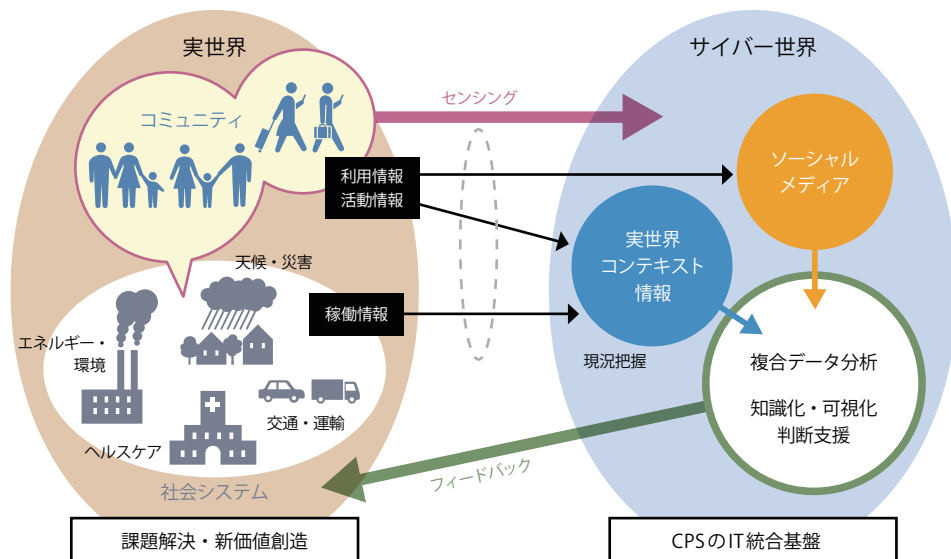
今井 CPSは明快な目的を据えて取り組むことが大切です。その目的に沿って、必要なデータが何かを明らかにし、データ保護を含む管理の条件を洗い出していくというように、総合的に考える必要があります。NIIではここで述べたような課題を解決しつつ、社会・公共・災害対応といった領域に役立つCPS基盤を追究していきます。

(取材・文＝土肥正弘)

※1 一般的な業務システムの処理やIT機器のログ管理、SNSの一括検索などのように、データの発生から分析までの間に時間差があってもかまわない用途に使われるデータのこと。貯めておいた状態で分析処理を行う。

※2 何らかの状態変化に即応した機器制御や、リアルタイムな意思決定支援を行う用途に使われるデータのこと。発生時点で即座に分析処理され、貯められることなく流れすぎていく「フロー型」データ。

(実世界の状況をセンサー等によりリアルタイムに把握し、データ分析に基づくフィードバックにより社会システムの課題を解決し新たな価値を創造する)



図：ソーシャル・サイバー・フィジカル・システム (CPS)