インターネット計測とビッグデータ

今後、あらゆる分野で重要性が増すであろうデータ解析。
統計やデータ解析を道具として使いこなして、問題を解決する能力が求められます。

3.1 インターネット計測

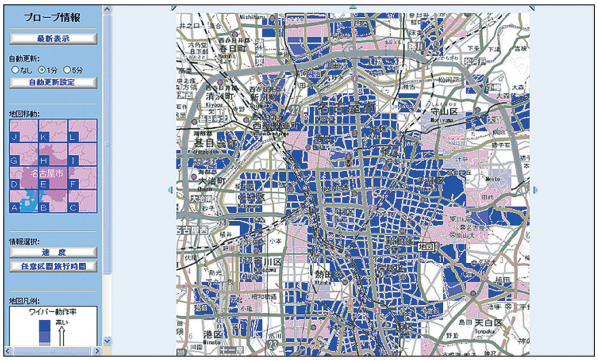
インターネットは常に変化を続けるオープンシステムです。自律分散型のインターネットには、中心もなければ代表点もなく、測る場所や時間によって違う姿が観測されます。このようにインターネットを把握することは難しいのですが、だからこそその実態を把握しようと、インターネット計測と呼ばれる様々な取り組みがされてきています。

インターネット計測としては、トラフィックの量やその内訳の計測、ネットワークの繋がり方を探るトポロジ計測などが代表的です。これには、このレポートで毎回報告している電子メールのSPAMの割合やウィルス感染、セキュリティ攻撃の観測なども含まれます。最近では、ピアツーピア型のシステムの観測やソーシャルネットワークの使われ方、そこでの人と人の繋がり方の観測など、幅広いオンラインサービスの計測があります。ここでは、インターネットとインターネット上のサービス、あるいはその利用に関する計測とその応用を広くインターネット計測と呼びます。利用者に身近なSPAM判定、検索ランキング、オンラインお勧めシステムなども、インターネット計測技術の応用だと言えます。

これらのインターネット計測に共通するのは、大量かつ不完全なデータから有用な情報を見つけ出そうというアプローチです。これは、従来の工学的な計測とは対照的です。従来型の計測では、計測の精度を向上して正確なデータを得ようとしますが、インターネット計測では、正確なデータがないことを前提に、曖昧な情報を突き合わせることで実態を推測せざるを得ません。

例えば、インターネットに繋がっているPCやデバイスの総数の正確な数は計りようがありません。しかし、インターネットのアドレスの使用状況、主要Webサイトへのアクセス、各国のインターネット利用者数調査、PCやモバイルデバイスの出荷台数など、複数のデータを突き合わせることで、おおよその数を推測することは可能です。現在では、「繋がる」という定義にもよりますが、おおよそ30〜50億台ぐらいだと考えられています。

また、自動車の位置とワイパーの稼働状況の情報を収集することができれば、局地的な集中豪雨の様子を細かく知ることができます。個々のワイパー稼働状況は不確かな情報ですが、多数のワイパー情報を集めると、十数km間隔で設置されている気象センサーでは捉えられないきめ細かな状況を把握できるのです(図-1)。



WIDEプロジェクトが2001年に名古屋で行ったインターネット自動車実験では、1,570台のタクシーから位置、速度、ワイパー稼働情報を収集した。図の青い部分がワイパー動作率が高い地域で、細かな降雨状況が分かる。

図-1 自動車のワイパー情報

データに含まれる隠れた情報を見つけ出すためには、多くの場合、複数の要素の関係を分析する多変量解析をはじめとした統計的手法を使います(図-2)。このような手法は、インターネット計測以前から、例えば、心理学や行動科学などの社会科学や、医学や薬学などで応用されています。しかし、インターネットと情報技術によって、データ取得とデータ解析の自動化、システム化が進んで大きく状況が変わったと言えます。それによって、それまで難しかった、膨大なデータへのアクセス、常に更新されるデータを対象にした解析、非線形モデルへの応用などが可能になってきました。今では、あらゆる科学技術分野で、膨大なデータの解析は欠かせない研究手法になってきています。

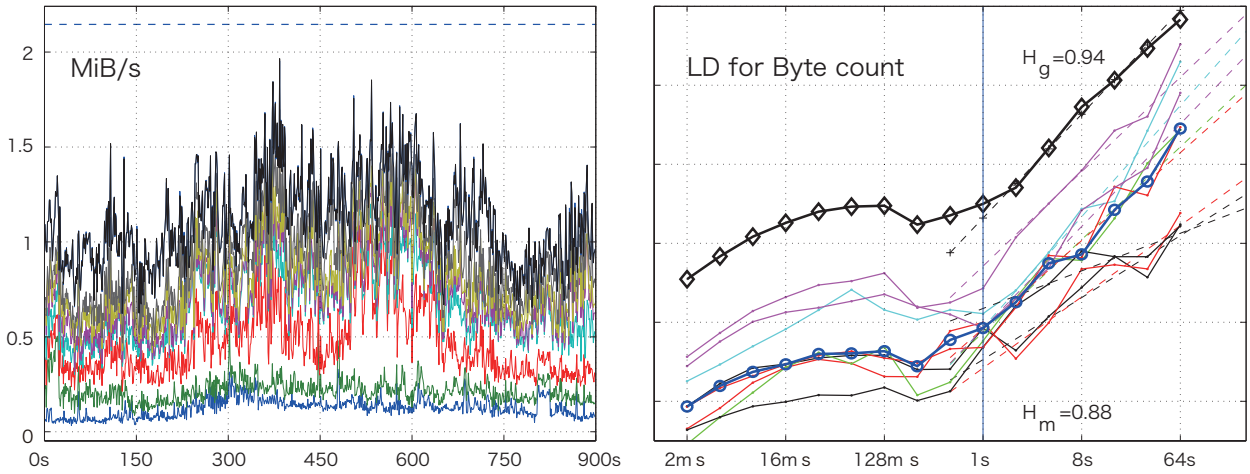
3.2 ビッグデータ

最近「ビッグデータ」という言葉をいろいろなところで見かけるようになりました。ビッグデータは、大量の非定型データから隠れた価値のある情報を引き出す技術の総称として使われています。膨大なデータを収集し分析することで、新たなビジネスモデルの構築や経営革新などのイノベーションに繋げるという考えです。その背景には、この数年、特にクラウドサービスの登場で、ビッグデータを導入するために必要な環境が整い、誰でも使える環境ができてきたことが挙げられます。現状ビッグデータビジネスとして、利用者のオンライン行動履歴のマーケティング利用が注目されていますが、今後は様々な展開が期待されています。

ビッグデータを技術的に見れば、まさにインターネット計測が取り組んできた技術です。オンラインデータ収集システムやデータの保存や共有のためのシステムの構築、膨大で断片的なデータから情報を抽出するための統計処理技術の工学応用などは、インターネットができた時から行われています。インターネット自体は工学的に設計されたコンポーネントから構成されますが、その挙動は無数の要素の相互作用の結果、全体としてみれば個別要素の総和以上の独立な振舞いをみせる複雑系の典型と言えます。また、利用者の行動を反映するので、社会的、経済的、政策的な影響も受けます。インターネットの計測は工学的であると同時に、自然科学や社会科学的な側面も持っています。

データの収集に関しては、インターネットによって状況が劇的に変わりました。インターネット上での情報公開が進んで、誰もが簡単に多様な情報にアクセスできるようになっています。時刻情報や位置情報をはじめとしたセンサー情報が付加されることで、これまで難しかったような関係性についての解析も可能になってきました。また、ソーシャルメディアなどを通して情報が広がるようになり、従来マスメディア中心だった情報伝達と情報共有の在り方にも本質的な変化が生まれてきているだけでなく、例えば、キーワードの拡散を追跡するなど、情報の伝達もデータとして収集できるようになりました。

データの保存に関しては、ストレージの大容量化と価格低下によって、保存可能なデータ量は飛躍的に増えてきてい



ネットワークトラフィック(左)から、統計情報を抽出して比較すること(右)で、異常や故障、またはその兆候を検出することが可能。

図-2 統計手法による異常検出

ます。また、データの処理に関しても、コンピュータの処理能力は飛躍的に上がりました。従来は、ストレージ容量と処理能力の両方の制約から、効率良くデータを保存してアクセスする必要があり、利用形態を想定して構造化されたデータベースが使われてきました。それに対して、文書や画像を含む雑多な情報を保存しておき、後でそこから情報を見つけることができるようになってきたのです。

解析ツールに関しても、データマイニング、機械学習、統計処理などのツールが充実してきて、利用しやすくなったことも挙げられます。MapReduce[*1]などに代表される大規模分散処理も利用可能になっています。

それでも、クラウドサービス以前は、このようなことができるのは、インハウスでデータの収集、管理と解析をできるような組織に限られていました。今では、顧客のオンライン行動履歴を収集して分析するパッケージツールも登場しているので、クラウドサービスとパッケージツールを使えば、僅かな初期投資で誰もが簡単にビッグデータを利用することが可能になっています。

このように、データを基にしたマーケティングやデータを基にした経営判断などのビジネス利用の機会が拡大しています。同時に、あらゆる分野において、データ革命と呼べる技術革新が起こっています。2012年3月には、米国政府がビッグデータの研究開発に巨費を投じる発表を行い、国家としてビッグデータ戦略を推し進める姿勢を示しています。

3.3　データ分析はあくまでも道具

インターネット計測に取り組んできた我々は、これまでデータ収集と分析の必要性や、そのための手間やコストについて理解を得ることに大変苦労してきました。ビッグデータという概念が認知されてきたおかげで、これらの理解が得やすくなってきています。その一方で、最近のビッグデータの話はツールや手法だけが強調されているような印象を受けます。データ解析はあくまでも道具です。目的も持たずに、ただ大量のデータを集めてやたらにCPUを回しても、得られるのは使いようのない数字だけです。

逆に、データから何を読み取りたいかがはっきりすれば、やるべきことは見えてきます。どのようなことが分かれば何にどのように役立つかを常に考え、問題を設定したり、結果に疑問を持つことが重要で、データ解析は手段にすぎません。データ解析は、あらかじめ仮説を立てて、それをデータで検証する作業の繰返しです。もし結果が予想と違っていたら、そこから新たな問いを見つけ出すことができます。このプロセスの繰返しから、役立つ情報や興味深い事実が見つかるのです。

情報技術によって、データに基づいて考え、考えをデータで検証するという思考プロセスの本質的な変化が起こっているのです。もちろん以前からもデータを基に考えることは重要でした。しかし、扱えるデータの質と量やその表現方法が桁違いに変わって、データをイメージ化しながら、文字通りデータと対話しながら考えることができるようになってきたのです。

3.4　データの時代の課題

これからは、あらゆる分野でデータ解析の重要性が増えていきます。それぞれの分野で、その分野の知識を持った上で、データ解析ができるプロ、データサイエンティストと呼ばれる人材が必要となっています。統計やデータ解析ができるだけでは問題設定はできないので、その分野の専門知識を持った上で、既存の考えや解釈に疑問を持つことができ、問題を明確に設定し、統計やデータ解析を道具として使いこなして問題解決をする能力が求められます。このような能力を持つ人材は圧倒的に不足しているので、人材の育成が大きな課題です。

データの時代には、データの収集と蓄積が財産になります。特に、過去に遡った解析を可能にする長期間のデータは貴重です。また、大量のあいまいなデータを扱う場合でも、

＊1　Googleが開発した分散データ処理技術。ビッグデータ解析に広く使われている。

データの質は重要です。もし誰もが同じデータを基にデータ解析をするなら、データから有益な情報を見つけ出す能力が優劣を決めることになります。しかし、データの質にばらつきがあれば、より良質のデータを持つ方が有利です。実際、インターネットのトラフィックの詳細や、オンラインサービスの利用者の行動履歴など、外部には公開されないデータがほとんどです。したがって、現実によく利用されているサービスの情報にアクセスできると圧倒的に有利になります。つまり、他社が持っていないような実データを持つ会社が強いのです。

一方で、データの共有が進むことは社会全体に有益です。そして、データの共有とプライバシーへの配慮が今後の大きな課題です。これからは、複数のデータを突き合わせることや、多様なデータを関連付けて解析することの重要性が増します。そのためにはできるだけ多くの関連データが、できるだけ広く共有されることが大切です。科学の基本は第三者が検証できることです。データを共有することで、第三者による検証が可能になり、科学として技術が発展する礎になります。

また、データの共有はオンラインのプライバシーとのバランスの問題です。ソーシャルメディアは、友人や知り合いと個人的な情報を共有することで、幅広い人間関係が作られます。また、オンラインでの買いものは、使い込むに従って自分の指向に合うように自動的にカスタマイズされてきて、大変便利です。それと同時に、データを関連付ける技術が発達すると、予想もしないような推測が可能になります。利用者のちょっとした行動の変化からも、プライバシーに関わるようなことを推測できる可能性があります。現状、オンラインプライバシーに関しては、情報技術の専門家でも、過敏な反応をする人から楽観的な人までいます。ましてや、一般の人にとっては潜在的リスクの評価は難しく、社会的な合意形成に至るにはまだまだ時間がかかりそうです。結局は、情報を公開や共有することによるメリットとプライバシー漏えいのリスクとのバランスの問題です。

企業が営利目的で、あるいは公共機関が非営利でどこまで個人を追跡することが許されるかとか、個人に関する医療情報などをどのように共有して社会に役立てるかなど、今後のオンラインプライバシーに関する合意形成は社会的な課題です。

3.5　受け取り側のリテラシ

データを理解する、あるいはデータに疑問を持つということは、情報を受け取る側にも大切です。そもそも、同じデータを見ても異なる解釈は可能ですし、複数のデータから関連性を考えれば、多様な解釈が存在して当然です。更に、「統計のうそ」というテーマで多くの書籍があるように、データが重視されてくるにつれて、疑わしいデータやデータを基にした怪しい議論も増えてきます。実際、発信者のバイアスによる作為的な統計データや情報操作の氾濫が目につきます。

これからは情報を受け取る側にも統計データを理解し、疑う力が必要です。我々はともすると白黒の判定を求めがちですが、そもそもほとんどの物事はグレーであり、白黒はあくまで便宜的にグレーに線を引いただけのことです。情報の受け取り側が白か黒かを求めるのは、自ら判断することを避けて、発信者に判断の責任を求める行為です。しかし、多様な情報が入ってくる現代社会では、受け取り側がグレーをグレーとして受け取った上で、必要ならば自分で判断し白黒の線を引く必要があります。オンラインプライバシーに関しても同様で、ある程度の社会的合意は必要だと思いますが、最終的には自分が判断して自分の行動には自分で責任を持つことが必要な社会になってきているのです。

執筆者：
長 健二朗（ちょう けんじろう）
株式会社IIJイノベーションインスティテュート 技術研究所 所長。トラフィック計測やデータ解析などのインターネット研究に従事。慶應義塾大学 環境情報学部 特別招聘教授。北陸先端科学技術大学院大学 情報科学研究科 客員教授。