Internet world

ネット時代に生きる

櫻井 哲朗

第8回

ビッグデータの時代

膨大なデジタルデータが「予測できる未来」を生む

雨の音が聞こえない空梅雨で終わりそうですが、もう今年も半分が過ぎてしまいました。ああ今年も何もせずに折り返し地点に来てしまったと後悔の念にかられます。ですが、よくよく考えてみれば例年のことなので通常運転であると思うこともでき、さほど悔やむことでもないことに気づきます。

例えるなら、よく遅刻する友達と待ち合わせをするときに、あらかじめ時間に遅れることが予測できるのであまり心を乱されることがないような感覚です。そして、今日の話題は予測です。あいもかわらずの強引な話の展開で申し訳ありません。

2・8ゼタバイト

みなさんは「ビッグデータ」という言葉を聞いたことがありますでしょうか。「ソーシャルメディア」や「クラウド」に続く、いまビジネスやIT業界で最も注目されているキーワードの1つ、それが「ビッグデータ」です。最近TV番組で取り上げられています。例えば、NHKでは2013年3月3日に2時間スペシャルで「〝いのちの記録〟を未来へ～震災ビッグデータ～」[1]と題して震災時の災害情報を使った取り組みを紹介しておりました。またCMでもNECがビッグデータの解析技術を紹介した映像が放送されています。

このようにビッグデータというものが我々の身近な存在となってきています。このビッグデータですが、簡単にいってしまえば巨大なデータのことをさしています。ですが、その巨大さがとてつもなく大きいのです。

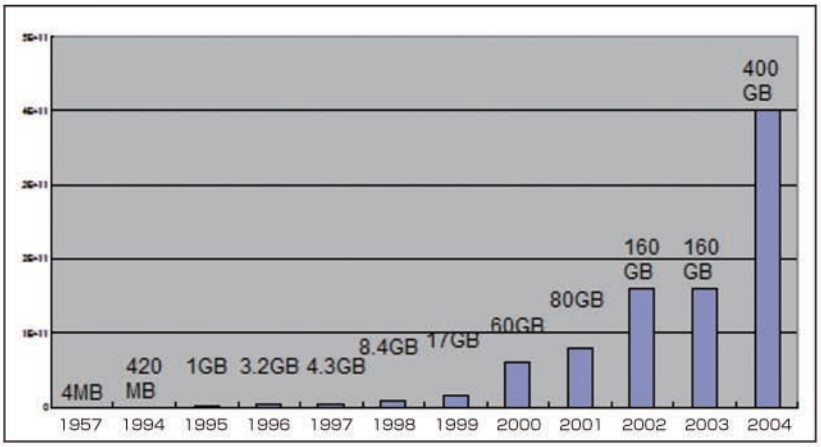
昨年、2012年までに全世界で生成されたデジタルデータの情報量の全体は2・8ゼタバイトになるという結果がアメリカのIT専門の調査会社IDC (International Data Corporation)によって提出された。ここでいうゼタは、メガやギガなどの単位を表すもので10の21乗倍の量である。つまり2・8ゼタバイトは次のような膨大な値です。

マンガ70兆冊分

2.8ゼタバイト＝
2,800,000,000,000,000,000,000

これは、マンガ1冊（40メガバイト）に換算すると70兆冊、ブルーレイ画質の2時間の映画（20ギガバイト）に換算すると1400億本にあたり、1日1冊または1本ずつ見ると仮定すると70兆冊のマンガを読み終わるのに約19億年、

図1：家庭用パソコンのハードディスク容量の推移



１４００億本の映画を見終わるのに約４億年かかる計算になります。そんな膨大な量のデータが存在しており、また調査結果から２０２０年までには40ゼタバイトまで増加すると予想されています。

　このようにいわれても、なかなか実感できないかもしれません。図1は、F社で発売された家庭用パソコンのハードディスク容量の推移を表しております。これから、実際に我々の使っているパソコンのデータ保存領域が増加していることがわかるかと思います。１人１人のデータ量が増加すれば、それにともない全体のデータ量も増加していきます。このように積もり積もったデータ量が２・８ゼタバイトとなります。

　このような巨大なデータの解析は、統計学的にも、とても重要な課題の１つとなっております。実際、米国国立科学財団主催のワークショップ「統計科学：21世紀に対する挑戦と機会」の報告書（２００３年）においてもデータ量が増加した場合の統計解析が重要であると述べています。

ビッグデータの特徴

　以上からデータ量が、とんでもなく増加していることがわかったと思います。では次に、そのデータの中身について見ていきたいと思います。データの中身も時代とともに変わっていき、昔は文字データや数値データなどの簡単な形式しか扱うことができませんでしたが、コンピュータ技術の発達とともに画像データ、音声データ、映像データなど多くの種類が扱えるようになりました。また現在、日々生成されるデータは、企業などでは銀行の取引や航空機の予約などの顧客データや病院の通院歴、血圧・血液検査の結果、電子カルテの情報などの医療データなどがある。また個人では、mixiやFacebookなどのSNSに投稿したテキストデータや写真・動画の映像データ、携帯電話のGPS信号などによる地理データなどの様々なデータが大量に生み出されています。

　このような大量で多様なデータをビッグデータと呼んでおり、次のような3つの特徴があります。

爆発的な増加に
容量：Volume

　直感的にイメージされる、ビッグデータの特徴。さきほど述べたように、現在データ量は爆発的に増加しており、そのため各種データ分析で取り扱う容量も増加傾向にある。それに伴い計算量も増大している。計算量の増加もビッグデータ解析を困難にしている原因の1つです。

「非構造」の解析重要
種類：Variety

　これも先ほど述べたようにビッグデータ解析では様々なデータを扱い分析を行う。このとき、データの種類は次の2つに分類することができる。今までのデータ分析で取り上げられてきた顧客データなどの構造化されたデータと文字や音声・映像などの様々な種類の非構造データに分けられる。ビッグデータ解析では、この非構造データの解析が重要となります。

買い物をしても
頻度：Velocity

　データの発生頻度も、ビッグデータの特徴を表すものの1つである。データの量が爆発的に増加しているということは、それに伴ってデータの発生頻度が増加しているとも考えることができる。実際、我々の生活ではコンビニやスーパ

さくらいてつろう

中央大学大学院理工学研究科を卒業し、専攻は統計学。コンピュータなどによって計測される大量のデータをまとめる多変量解析の研究。現在は、諏訪東京理科大学共通教育センター講師。東京都出身、30歳。

ーのレジで買い物をすると、それはPOS（Point Of Sales）データとなり各社に収集されています。

インターネットを閲覧すれば、どのサイトにアクセスしたかというログのデータが発生し、そのデータの中には、どこをクリックしたか、また次のクリックまで何秒かかったかという、とても細かい情報まで蓄積されます。この他にも、SuicaやPASMOによる乗車履歴データと電子マネー履歴データをあわせることで、どのような個人が何を買ったかという詳細なデータを取得することができ、そしてデータは日々または刻一刻と発生しています。

図2：ディサイド・ドットコム（https://www.decide.com/より）

どう活用しているか

では次に、この多様な大量のデータを用いることによって何ができるようになるでしょうか。それは、予測・パターン認識・最適化です。これらをキーワード別に取り上げ、参考文献の[2]および[3]で紹介されている企業のビッグデータ活用方法について簡単に説明させていただきます。

ラプラスの悪魔

まず、その1つが予測になります。突然ではありますが、みなさんはラプラスの悪魔という言葉をご存じでしょうか。これは、フランスの数学者ピエール＝シモン・ラプラスによって提唱された近代の物理学において未来の決定性を論じる時に仮想された超越的存在の概念のことです。ラプラスは自著『確率の解析的理論』（1812年）において次のような主張をしました。

もしもある瞬間における全ての物質の力学的状態と力を知ることができ、かつもしもそれらのデータを解析できるだけの能力の知性が存在するとすれば、この知性にとっては、不確実なことは何もなくなり、その目には未来も（過去同様に）全て見えているであろう。

これは簡単にいうと、現在の全ての状態がわかっているのであれば次の瞬間に何がおこるかが分かるということであり、そのような存在をラプラスの悪魔と呼んでいます。著者のイメージですが、ビッグデータ解析はこのような存在を人工的に作り上げているように思います。コンピュータを使って悪魔を作り出すなんて、まるで、パチンコやパチスロにもなりました「真・女神転生」シリーズや「ペルソナ」シリーズのような話みたいですが。では、実際にビッグデータを使い予測を行っている企業として「ディサイド・ドットコム」や「フライトキャスター」があります。

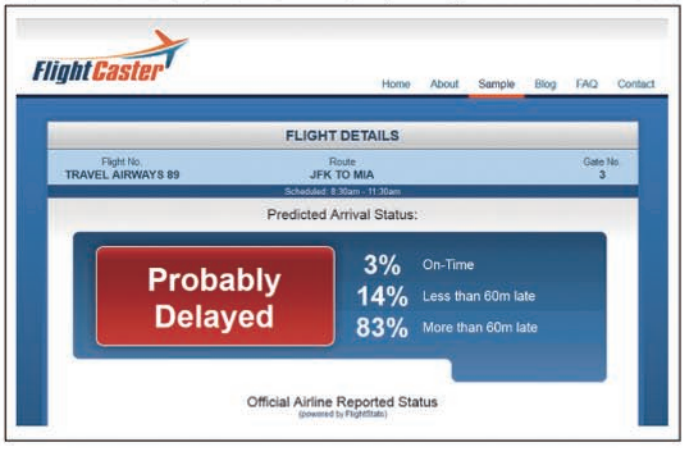


図3：フライトキャスター（http://flightcaster.com/より）

買い時教えます

ディサイド・ドットコム（https://www.decide.com/）

この企業では、TV、PC、デジタルカメラ、スマートフォン、ゲーム機器などの電子機器の買い時を教えてくる会社です。実際には、図2のような価格の予測を行うことによって、その電子機器が値上がりするのか値下がりするのかを教えてくれます。では、なぜ、このような予測が可能になったのでしょう。同社は、インターネット上にあるオンラインショップか

図4：論理方程式

ら毎日各電子機器の価格データを収集するとともに、他にもこれらの製品に書かれているブログやニュース記事を分析し、さらにはニューモデルの発売予定の情報もチェックしています。これらを集計・分析することで価格の予測を可能にしています。残念ながら、まだ日本語版のサイトはないため、我々が使うには少し敷居が高いかもしれません。

飛行機の遅れも

フライトキャスター
（http://flightcaster.com/）

この企業では、天気予報ならぬフライト予報をしてくれるサービスを提供しています。具体的には図3のようにフライトの遅れを予測して教えてくれます。図において、上から定時に着く確率（On-Time）3％、遅れるが60分以内に着く確率（Less than 60m late）14％、60分以上遅れる確率（More than 60m late）83％となっています。まるで天気予報の降水確率のように飛行機が遅れる可能性を教えてくれます。

では、どのようにして、この予測を実現しているのでしょうか。同社は、これらの予測のために交通統計局のデータ、連邦航空局の航空交通管理システム指令センターの警報、飛行機の運航状況を教えてくれるサイトのデータ、米国気象局の天気予報などを使って予測を行っています。そして、これらのデータは先ほどと同様にインターネット上にあり誰でも取得可能なデータであるということがあげられます。

「不確実」も知識に

また、著者の観点からは天気予報の降水確率のように表している点が、とても大きなポイントの1つに考えられます。ほとんどの場合、100％の確率で未来を予測するのは非常に困難です。なぜなら、不確実な要素が多いからです。ですが、これを使える知識にかえることができます。このような仕組みを参考文献[4]では、次の図4のような論理方程式で表しています。

この図の意味について説明いたします。通常、観測されたデータから導き出される可能性はいくつか存在します。たとえば、3日連続で晴れだったからといって次の日は晴れかもしれませんし雨かもしれません。つまり、どの結論も不確実性がともなってきます。どの結論も100％でない場合、そのときどうするか。そこで、各結論に対して、その不確実性の度合いを付け加えます。これが、「不確実性の度合いについての知識」の部分です。

先ほどの例ならば晴れの確率は90％、雨の確率が10％といったことがわかるようになります。これを加えることによって、「不確実な知識」は「利用できる知識」に変わります。みなさんも天気予報で雨の降る確率が80％以上だと傘を持っていくように判断しているかと思います。このように、どれが起こるかわからないが、それらに確率を与えてあげるだけで、人は判断することができます。

このようにビッグデータを用いた予測について説明してきました。いままで、どの企業においてもデータ分析は行ってきたのですが、それは過去の分析や現状の把握にとどまるものでした。しかし、ビッグデータを用いることによって将来の予測が可能となりました。

図5：reCAPTCHA
（http://www.google.com/recaptchaより）

パターン認識する

パターン認識とは色々な情報を含むデータの中から意味のある特徴を見つける方法です。画像データから文字を抽出してテキストデータに変換する技術やデジタルカメラの顔を識別する機能などの技術もパターン認識が使われています。近年では、統計的な考え方を取り入れることによって認識の精度がより高まりました。このときに、使われているのがビッグデータです。では、実際にビッグデータを使い予測を行っている企業として「グーグル」があります。

画像をテキストに

グーグル

みなさんご存じのグーグルもビックデータの活用が盛んです。みなさんは図5のようなインターネットの画面を見たことがありますでしょうか。アカウントを作るときやパスワードを入力するときなどに時々でてくるかと思います。

これは「reCAPTCHA」というシステムで、ボットと呼ばれる自動的にサイトにアクセスするプログラムからサイトを守る役割があります。でも実は、もう1つ役割があり、それは入力の結果を使い画像データからテキストデータへの変換のために使われています。「reCAPTCHA」は、テキストデータに変換できなかった文字を画像として取り出し、各サイトに送ります。それらは人の目によってテキストデータに変換されます。その入力データをもとに書籍のデジタルデータ化を進めています。日に2億を超える画像データがテキストデータに変換されているという報告もあります。これにより「Googleブックス」による書籍の電子化が進み、最近ではインターネットの検索結果に本ページも出てくるようになりました。

また、グーグルではこれ以外にもビッグデータの活用として翻訳サービスにも力を入れています。グーグルの翻訳のよくある質問を見ますと次のようなことが書かれています。

「自動翻訳」とは―

人手を介さず、最新技術によって自動生成される翻訳です。自動翻訳は「機械翻訳」とも呼ばれます。

Googleで独自の翻訳ソフトを開発したのですか―

はい。Googleのリサーチグループが開発した独自の統計的翻訳システムをGoogle翻訳に使用しています。

統計的機械翻訳とは―

現在市場に出回っている自動翻訳システムのほとんどはルールベースで開発されており、語彙や文法の定義など多くの作業を必要とします。一方で、Googleの翻訳システムの手法では、ターゲットとなる言語で記述された単一言語のテキストと、人間が翻訳した他言語のサンプル翻訳テキストを対にしたものを大量にコンピュータに入力します。そしてこれらのテキストに統計的学習手法を適用して、翻訳モデルを構築しています。Googleのリサーチ評価では、この手法が優れた結果をもたらすことが判明しています。

http://www.google.co.jp/intl/ja/help/faq_translation.html より

ここで注目してもらいたいのが「統計的翻訳システム」という言葉です。グーグルでは、いままでの翻訳システムとは異なり、新たな翻訳システムを作りました。それが「統計的翻訳システム」です。ですが、このシステムを作り上げるためには大量の翻訳データが必要となります。グーグルでは、インターネット上にあるデータを使いながら、このシステムを構築しています。また品質向上のために大量の翻訳データを募集しております。

これ以外にも「アップル」の秘書機能アプリケーションソフトウェアSiriで使われている音声認識技術や対話サービスにもビッグデータによるパターン認識が活用されています。

最適化に使う

最後に最適化について紹介させていただきます。通常、最適化とはあるものを最適な状態に近づけることを指します。ここでいう最適化とはビッグデータを使いシステムなどを最適な状態にすることです。では、実際にビッグデータを使い予測を行っている企業として建設機械大手の「コマツ」があります。

車両の「状況」を把握

コマツ

コマツは建設機械・重機械のメーカーで、国内シェア1位、世界で2位の大手企業です。日本以外にも南北アメリカ、ヨーロッパ、独立国家共同体、中近東、アフリカ、東南アジア、オセアニア、中国にグループ企業を展開するグローバルな企業でもあります。

コマツでは「KOMTRAX」と

参考文献

[1] http://www.nhk.or.jp/special/detail/2013/0303/

[2]城田真琴、「ビッグデータの衝撃―巨大なデータが戦略を決める―」、東洋経済新報社、2012

[3]稲田修一、「ビッグデータがビジネスを変える」、アスキー・メディアワークス、2012

[4]柳井 晴夫、田栗 正章、藤越 康祝、C.R.ラオ、「やさしい統計入門―視聴率調査から多変量解析まで」、講談社、2007

図6：KOMTRAX
(http://www.komatsu-kenki.co.jp
/service/product/komtrax/より)

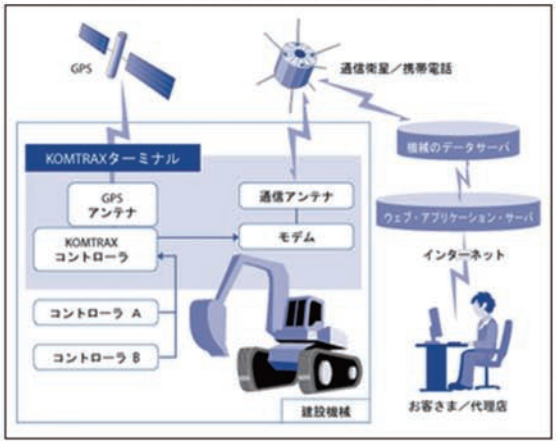
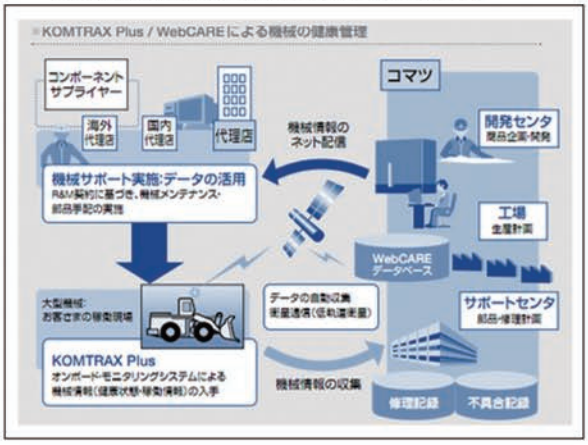


図7：KOMTRAX Plus
(http://www.komatsu-kenki.co.jp
/service/product/komtrax_plus/より)



いうシステムを開発し建設機械の稼働状況を遠隔で管理しています。2011年4月時点で、約6万2000台の建設機械にKOMTRAXが装備されています。

「KOMTRAX」では、図6のようにして建設機械に取り付けられたセンサー情報やGPSによる位置情報が通信システムよって送信され、それらのデータがインターネットを通じて顧客やコマツ販売代理店に提供されます。

また、鉱山向け大型機械の遠隔管理システムとして「KOMTRAX Plus」があり、図7のようにして車両の「健康状態」や「稼働状態」をリアルタイムで把握しています。これらのシステムを使うことによって、各種車両の最適化を行うことができます。各種センサーによる稼働状況から車両ごとの各部品の摩耗率が分かるようになりました。この摩耗率のデータを使うことで、劣化による故障を未然に防ぐための最適なオーバーホール時期を予測することができるようになりました。これより、車両を管理している顧客側では修理コストを削減することにつながります。またGPSによる位置情報から盗難がなくなるメリットもあります。

「モノ」がネットする

このようにして近年、センサー情報の活用が盛んに行われています。また最近では「モノのインターネット(Internet of Things)」というキーワードが注目を集めています。これは、その名の通りモノがインターネットをする時代が来ようとしています。モノに取り付けた各種センサーを使いインターネットを介してモニターしたり、インターネット上からモノをリモートコントロールしたりすることができるようになってきています。

実際、2013年4月にNECは、工場や発電所などの大規模施設(プラント)における故障の予兆を各種センサー情報から分析し、故障に至る前に設備の不健全な状況が把握できる「大規模プラント故障予兆監視システム」の開発を発表しました。また、我々の生活においては最近流行りの、インターネットを介して家電を操作する、スマート家電も、これにあたります。

ビッグデータのこれから

以上のように、ビッグデータを活用することで我々の生活は、より便利になってきます。しかし、ビッグデータの活用には問題点もあります。その1つがプライバシーの問題です。ビッグデータの活用上、どうしても大量のデータを必要とします。極端な話、各個人の全てのデータが必要とされる場合もあります。そうなると、個人の行動が筒抜けになってしまう危険性があります。

たとえば、スマートフォンなどに搭載されているGPS装置を使って、自分自身の行動履歴を全て取得され、つねに誰かに監視されているような状態です。このような観点からビッグデータの活用におけるプライバシーの保護に関する法律的な整備などが各国で行われています。

また冒頭で紹介したようにデータの総量は増えておりますが、この全てがビッグデータとして活用されているわけではなく眠っているデータも多く存在します。たとえば、監視カメラのデータは監視が目的だったのですが、顧客の動きなどを解析しマーケティングに生かしている企業もあります。このように各企業やインターネット上には金脈となるデータが多く眠っているのかもしれません。そんな金脈を掘り当てることができたらなぁと日々夢想するばかりです。