Big Dataにどう取り組むか

京都大学名誉教授 **長尾　真**

PROFILE　1997～2003年京都大学総長、2007～2012年国立国会図書館長、2005年日本国際賞、レジョンドヌール勲章、2008年文化功労者

1 Big Data の時代

以前から、スーパーマーケットなどで出口のキャッシュレジスタの手前にガムや剃刀の刃などを置いておくと、買い物の直接の目的ではないが買ってゆく人がかなりいるとか、ある品物の隣にそれに付随して買うであろう品物を置いておくとかいった配列の工夫がいろいろと為されていた。こういったことは店舗の売り上げのデータを分析することによって得られていたノウハウである。また各店舗の売り上げの品物を比較することによって、どの地域ではどのようなものがよく売れるかということも分かってきて、各店舗に置く品物の種類を区別することも当然のこととして行われてきている。

近年はインターネットを通じた販売が行われるようになり、商品に対する利用者の意見、特に商品の使いずらさ、故障に対する苦情、その他種々の意見がどんどんと寄せられるようになってきた。年に何百万台と販売する自動車産業の場合などでは、そういった顧客の意見を徹底的に分析することによって製品の改良に繋いで行ったり、新製品の開発のためのヒントをそこから得るということが行われている。

それが最近になってにわかに注目を浴び、Big Data の分析が大切だと叫ばれだしたのはなぜだろうか。その理由は必ずしも明確ではないが、一つにはあらゆることがデジタルで記録されるようになったにもかかわらず、これらの巨大なデータは多く捨てられるか、放置されていて、積極的に利用されていなかったこと、それに対してグーグルやアマゾンなど世界の最先端の企業がこれをうまく利用して新しいビジネスを創造し成功するといった事例が出てきたこと、それにもかかわらず、こういった各種の巨大データを記憶し、実時間的に解析して有用な結果を出すということは必ずしも簡単なことではなく、巨大な記憶装置、超高速の処理機構、そして強力な解析のための種々のソフトウエアを開発しなければならないことなどから、多くの企業が関心を示しながらも、どうしたらよいか分からないといった状況があって、これに突破口を開けようという動きが米国を中心として出てきたことによるのではないだろうか。事実、米国ではオバマ大統領がその重要性を認め、新しい創造につながるものとして、２億ドルをその研究開発に投入することを決定した。

2 科学技術分野における Big Data

Big Data の持つ力、有用性は企業の分野に限られない。科学技術分野においてもそういった事例はすでにいろいろと出てきている。たとえば医療分野では、個々の患者の病状はすべて異なっていて、教科書に書いてあるような形でこの症状に対してはこう処置すればよいというわけにはゆかない。そこで最近は類似の体質の類似の病状の患者にどう治療したらどういう結果になったかといったカルテを集めて分析し、そこからより安心できる治療法を見つけてゆくという努力がなされている。以前は経験豊かな医者が持っていたノウハウを巨大な患者のカルテ情報から抽出し、これを参考にしてベテランと同じように治療ができるようにするという考え方である。

私が研究してきた機械翻訳の分野でも Big Data が最近ますます重要性を高めてきている。機械翻訳は最初文法規則を頼りにして文を解析する方法がとられていたが、言語の文法は文の構造についてのかなり荒い近似的な説明の枠組みであって、それでうまく説明のできる言語の部分は非常に少ない。それに気が付いて１９８２年にアナロジーに基づく機械翻訳という考え方を提唱した。後にこれは例文翻訳、あるいは用例翻訳方式と呼ばれ、広く用いられるようになってきている。

この方式ではまず句を単位とした対訳辞書を作り、この辞書を頼りに与えられた表現の各句を翻訳し、翻訳されたそれらの句をうまくつなぎ合わせることで訳文を作るという方式である。いろんな表現に対処するためには対訳辞書に入れるべき句の数は何十万、何百万となるだろう。これは巨大な量の対訳文章を分析することによって作り上げるのである。今日では数十億文という数の文がコンピュータに記憶され分析され、対訳辞書の増強、改良がおこなわれる時代になってきており、対訳文章の量が多ければ多いほど翻訳の品質が良くなるわけである。

一方では統計的機械翻訳という新しい概念も出てきている。これは巨大な量の対訳文章を統計的に処理することによって与えられた表現の翻訳に利用しようとするもので、文法という概念にとらわれずにまったく機械的に翻訳を行おうとするものである。これも巨大なデータの存在と処理ができる時代なればこその方法といえるだろう。

科学技術分野では計測技術の高度化、精密化に伴って観測されるデータ量が爆発的に増えている。こういったことがあらゆる分野で起きているので、これをどのように保存し、利用するかが大きな問題である。同様のことは科学技術分野だけでなく、社会科学分野などにおいても起こりつつある。

今日科学技術分野の研究は国の主導のレベルで大規模な研究が行われるようになってきており、その研究過程で作り出される各種の研究データは他の研究者が観たときにまったく想像もしなかった形で利用される可能性があるということで、米国の NSF や NIH では、研究によって生じるデータを保存し他でも利用できるようにするために、研究費の申請時に研究データ管理計画書をださせ、そこには情報図書館学の素養を持った人を当てることを奨励するようになってきている。

3 Big Data からの知識の抽出

このように巨大なデータ群には目に見えぬ貴重な情報や知識が潜んでいることが最近になって認識されるようになってきたが、そこには大きな技術的課題が存在することが分かってきて、これを克服するにはどうしたらよいかに関心が集まり、ここに「Big Data 研究」という課題が急速に浮かび上がってきた。

その第一の課題は毎日テラバイト単位で増加するデータをどのように記憶すれば後の処理に好都合であるかという課題である。大企業や大きな研究機関でなければこれらのデータを独自に保存することが困難なため、クラウドに任せるという道が開かれてきた。もう一つの課題はこの巨大なデータをどのように分析すればどのようなことが分かって来るかという問題とともに、その処理のための理論とそれを実現するソフトウエアもよく整備されていないという問題である。特にこの第二の課題には種々のアプローチや方法が考えられるが、今後の大きな研究開発分野となっているといってよい。

第一の課題で特に大きな問題は、収集の対象となるデータには各種センサーなどから得られる数値データのほかに、文字・テキストデータ、画像・映像データ、音声・音楽等の音のデータ、３次元ビジュアルデータなど種々のものがあり、これらは常に時系列的に時々刻々と増えてゆく。またこれらのデータの中には、例えば音と映像のように、異なるデータ間に存在する時間的な同期を捉えておかなければならないものがある。そこでこれらの様々なデータを統合的にうまく処理するために、それぞれのメディアやデータの種類によってある種の標準形式を作れないかという問題も出てくる。

これらの巨大データからの情報や知識の抽出には大きく分けて３つの場合が考えられる。

（1）巨大な量のデータから特異なデータを検出すること：これが情報となる。

たとえば装置の異常を検出したり、ネット上の Blog や Twitter の中からそれまでになかったような異常な内容の文章や発言、警告などを発見する場合がこれにあたり、貴重な情報となる。

（2）巨大な量のデータに内在する共通の性質を発見すること：これが知識となる。

たとえばスーパーでの売り上げを分析することによって、春にはどの年齢層がどのような商品を買うかといったことを知る。東日本大震災の直後、多くの車が被災地に入ったが、これらの車の GPS データを分析することによって、この道は主要道路なのに車が通っていないからその道路が破壊されているに違いないといったことが推定できたという。

（3）異なった複数分野の巨大な量のデータ群相互間にある種の因果関係を発見すること：これが新しい知識、法則の発見となり、また新しい学問分野を作り出してゆく。

たとえば天候と病気の相関から、湿度の高いときはある種の病気が出やすいといった経験的知識が得られる。こういった病気にまつわる経験則はいろいろと存在するが、膨大な異種データ間の相関を調べることによってその因果関係がより信頼のおけるものとなってゆく。また異なる巨大データの因果関係の分析によって新しい経験的な法則が発見され、医学的にその妥当性が検証される可能性もある。これからの健康科学の一つの新しい分野といってよいだろう。

4 Big Data 解析の手法

Big Data の解析をする手法としては次のようなものがすでに存在する。

（ⅰ）クラスタリングと決定木
（ⅱ）データマイニング
（ⅲ）テキストマイニング
（ⅳ）相関分析
（ⅴ）回帰分析
（ⅵ）言語解析、意味検索等の自然言語処理
（ⅶ）その他

これらすべての手法においてこれから研究開発すべきことは多い。いずれの場合もデータが刻々と増えてゆく中でより良い結論が得られるようにするためには学習の機能を導入することが必須である。

膨大な量のデータを取り扱うためには効率の良い分散ファイルシステムを設計しなければならないし、分散処理ができるよう処理のアルゴリズムに工夫をしなければならない。また相関計算や回帰分析においては膨大な次元数を対象にして実時間的に分析をしなければならないといった難しさがある。

社会における Big Data の多くは文字データであるので、言語解析をし、重要な単語や概念を抽出し、また意味的な検索をするといったことが必要となるので、自然言語処理の各種技術が非常に重要となる。これからのグローバル社会の時代においては国内のデータだけを対象にするのでなく、世界のいろんなところで発生するデータを取り込み、分析の対象としなければならない。したがって多くの言語を対象とした機械翻訳システムの開発が急務となる。特に急速に発展してきている中国の科学技術情報、学術全般の情報、さらには中国の中を飛び交うネット上の Twitter 情報などの分析のために中日機械翻訳システムが大切となる。ところが中国語の解析は世界の主要言語の中では最も困難なものであり、この課題の解決のためには国として相当な努力を傾注する必要があろう。

筆者は2011年の Japio Year Book に「東日本大震災アーカイブの構築」という小論を寄稿した。そこではあらゆる種類の震災に関するデータや震災の影響を直接、間接に受けている関連データ等を集め種々の観点からそれらの間に存在するであろう相互関係を明らかにすることの重要性を述べた。それをさらに学術一般に広げて、学術分野の「知識インフラ」を構築することの努力の重要性についても述べたが、こういったことも Big Data の一つの活動とみることができるだろう。

5 終わりに

特許情報とそれぞれの特許に関連した研究論文は今日Big Data を形成している。このデータの利用を考えるとき、単に必要な特許情報の検索だけに留めず、種々の手法で分析することによって、どのような新しい技術が創造されつつあるか、どの国のどの企業がどのような分野、課題に力を入れているかといったことが分かってくる。

こういった分析は各企業がそれぞれの立場と観点から行うことであるが、国としてもどのような分野を積極的に育成してゆくべきかを検討する際に、特許関連データの分析を欠くことができないだろう。これまでもJapioにおいては技術動向調査を含み各種の統計的分析などを行ってきているが、これからは世界各国の特許関連データについて特許文章の内容分析を含んだもっと詳細かつ高精度の分析を行う必要があろう。発展途上国における特許や技術の詳細などの情報は個々の企業で収集することは難しいから、Japio などがこれを行い、分析をし、有用な情報を提供することも大切であろう。Japio がこういったことに力を入れてゆくことを期待したい。