ビッグデータ活用の未来

The future of big data utilization

京都大学名誉教授

長尾　真

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. はじめに

巨大なデータの解析に興味が持たれたのは 1980 年代の後半から 1990 年代にかけてだった。当時は主として情報科学研究者を中心にデータマインニングという名称で巨大なデータから有用な情報を抽出する研究が行われた。特にテキストを対象とするときはテキストマイニングと呼ばれたし、その中でもウエッブページを対象としたものはウエッブマイニングと言われていた。そしてデータ解析のための多くの手法が開発された。例えば統計的手法を用いた良く現れる相関パターンの検出、クラスター解析、回帰分析などの手法が使われていた。しかしこういった解析の有用性は自然科学分野や情報科学、知識工学などのごく一部の専門家に興味をもたれるだけで一般社会には良く知られていなかった。

21 世紀に入って、取り扱われるデータのサイズがそれまでと比べて格段に大きくなるにつれて、ビジネスにおける巨大データの分析が企業にとって有用であることが認識されはじめ、ビッグデータの解析専門の企業が現れ、これがビジネス分野として認識されるようになって、ビッグデータという名称が広く一般社会に普及したと見ることができる。ただ技術的には、今日テキストマイニングの時代からそれほど発展しているとは言い難い。

2. 仮説検定的性格を越えて

たとえば食べ物と健康との間に関係があるのではないかと考えるときは、それに関係する大量のデータを集めて分析することによって相互関係性が統計的に推定され、その因果関係が学問的に追求される。こういったことは、二つのデータ群の間にある種の因果関係があるのではないかという疑問、つまり仮説を立てて、その関係があるかどうかをデータから検証するという形である。こういった立場の分析はビッグデータ時代といわれる前からいろいろと行われていた。これからは一つのデータ群の解析だけでなく、異なった複数のデータ群の相関解析をすることがますます必要になってゆく。例えば降雨量の季節変化とある特定の農作物の収穫量の関係、その地域的変化、土質との関係などは異種のデータ群の相互関係の典型的なものであろう。

ビッグデータの解析を分類すれば、大きく共時的解析と通時的解析となるだろう。共時的とはある時期に限

定したデータを解析することであり、これはデータを全体的に一つとみて解析する場合と地域的分布や職業別など、種々の特性を配慮した解析となる。通時的とは何年、何十年にわたる経年変化を見るという立場である。この場合に特に注意しなければならないのは長期にわたって同じ条件でデータが取られているか、また社会変動や3.11 のような大事件との関係はどうかといったことに配慮が必要となる。

ビッグデータの分析の最も面白く挑戦的なことは、関係性のありそうなデータ群を仮説検定的に解析するのでなく、仮説のない世界で新しい因果関係を発見することではないだろうか。一見して何の関係もないと思われる幾つかのデータ群をたまたま相互相関解析してみたら、ある種の因果関係が見つかったといったこと、いわば「風が吹けば桶屋がもうかる」という類の発見こそがビッグデータ解析の醍醐味であろう。何らかの仮定を置いて解析することは既にある方向に世界を限定して分析することであって、発見されることはいろいろあっても、それは想定内の期待していたことである。これに対してなんらの仮定を置かずいろんなデータの関連性を調べてみることによる発見こそ創造的で面白く、また有効性が高く、イノベーションに繋がってゆくものである。そのためには膨大な計算量を仮定しなければならないだろうが、これから挑戦してみるべきことと思われる。

別の言い方をすれば次のようなことであろう。人は頭が良いので、ある種の結論を想像し、それに向けた推論や解析をする傾向があるが、コンピュータの場合には一切の先入観を排して客観的に収集した大量のデータを解析するから、意外な事実を明るみに出す可能性があるわけで、この特徴を生かした解析を心がけるべきなのである。ただこのような仮説なしの解析処理で目の覚めるような結論が得られる確率は非常に低いという覚悟を持ってする必要がある。

3.データは膨大でも肝心のデータ群が欠けていることはないか

企業のコールセンター、顧客サービスセンターなどに来る質問や苦情、要求などのデータはどの企業でも保存し解析し、サービスの向上や製品の改良に繋いでいる。これらのデータは種々の観点から分類して利用することが基本であろうが、例えば苦情などを言ってきた人の感情がどのようなものか、どこまで激高しているか、あるいはどの程度の深刻さかといった顧客の感情という観点からのデータは取られているか、あるいは既に保存されているデータから言語分析することによってそういった情報を取り出すことができるかといったことも大切である。

ある種の解析をし、問題解決をしようとすれば、既存のデータだけではできず、新しいデータを入手することが必要ということは多い。それがたまたま他で作られていて、しかもそれが公開されている場合は有難いが、ほとんどの場合そうでない。したがって新しい種類のデータを集めたり、データの精度を一桁二桁上げた計測をしなければならなくなる場合が多い。そのための計測機器の開発を伴うこともあり大変である。国は各種各様のデータを持っているが、それらすべてはどのようなものでどこにあり、どのようなフォーマットで保存されているかということさえ十分に公開されていないが、こういった国の持つ各種データが自由に利用できる環境を整備すればビッグデータ解析が一層進み、国力に反映されてゆくから、国は情報公開によりオープンデータ政策に積極的になってもらいたいものである。

国際的に市場を展開している企業などでは世界の地域ごとの顧客の特徴を把握しておく必要があるから、その国の言葉の特徴を知り、感情分析もする必要が出て来るので、言語の機械翻訳のほかに種々の言語処理についても先進的な技術を開発しなければならない。

いじめ問題やテロなど事件性のある情報がネットに現れた時に素早くキャッチして対処することはこれからますます重要となる。これはビッグデータの中から特異的な情報や、それまでに現れていなかった情報を検出するという技術であり、いわばロングテイルの極端な先に存在するものを調べるといったことに通じる。こういったことは非常に困難であるが、今後ますます重要になってゆくだろう。

このようにビッグデータのテキスト解析は単なる字づら処理でなく、これからはコンテンツ解析、意味解析に向かわざるを得ないが、これを巨大データに対して実時間的に行うには巨大な計算能力が要求される。

4.ビッグデータの解析結果は適切な形で表示して人の理解を促進することが大切である

テキストを解析して得られた結果は文章で表示することが多いかもしれないが、場合によっては表にして出す、あるいは日本地図、世界地図の上にプロットして示すといったこともあるだろう。数値データの場合にはグラフ表示がありうるし、パラメータの軸を設けてグラフを２次元、３次元表示にするなどの工夫もありうる。時間軸によってゆっくりと変化させる動的表示も考えられる。このように表示の仕方の工夫をすることによって、見る人は結果を概念的によりよく把握できるし、俯瞰的に見ることによって結果にある種の不十分さ、疑問などを感じてより深い検討をするようになるキーを得る事にもなるだろう。人間の一覧性の力に頼ることが大切である。解析した結果を人に強制するのでなく、理解させる努力が必要であり、なぜそのような結論になったかという理由も提示できるよう工夫することが大切である。

5.ビッグデータ解析の結果はどこまで信用できるか

スーパーでおにぎりを買った人でキャッシャーのそばに置いてあるチューインガムを買う人が多いという相関が発見されたから、おにぎりのそばにチューインガムを置くのが良いといった戦略が言われたことがあった。しかしおにぎり以外の商品では相関が低かったかどうかをチェックして言っているのかどうか、またキャッシャーで列を作って待っているときの手持無沙汰のためについチューインガムを買ってしまうというのが本当であって、相関が高く出るからと言ってそれをうのみにするのは危険であるということもあろう。

だからビッグデータの解析結果はいろんな立場から吟味することが必要である。ビッグデータが収集された時の条件は何だったか、それに関係する他のデータは取っていなかったか、といったことをチェックし、測定されたパラメータ以外の、データの置かれている環境条件についても十分配慮する必要がある。

関係のありそうなデータ群の相関的な解析をしてある種の結果を得たとしよう。しかしもう一つ別のデータ群を加えて解析をし直したら、先に得られた結論が否定される結果が出て来ることがありうる。例えば毎朝起床したときに冷たい水をコップ一杯飲むと健康に良いということがデータから得られたとしよう。しかしこれに胃腸の強い人か弱い人かというデータを付け加えたら、胃腸の強い人には正しくても、弱い人には却って良くないという結果が出る可能性がある。したがってデータ群によって結論が違ってくる可能性を考えておく必要性がある。

さらに厄介なのは、キャッシャーで行列を作っている間にふと手を出して買うとか、買う予定になかった商品を見るとこれは家ではもう切れていたから買っておこうといった人間の心理的な行動をどのように推定するかということも重要であり、こういった人間の行動傾向はどうすれば把握できるかも考えねばならないだろう。単純なビッグデータの解析からは出てこないことである。

6.データ量の問題が深刻化してくることにどう対処するか

いろんなことを調べようとする場合、より精度の高いより大量のデータがあるに越したことはない。自然科学の観測では観測の精度を上げるとともに時間的にも、分、秒というように間隔をどんどんと狭くしていっている。人工衛星からの地球表面の観測も数メートルの精度から数十センチまで精度を上げつつある。社会のデータにおいても同様な方向にある。

データ量は今後とも増える一方であり、これを保存しいつでも自由に取り出して解析に使うためには、巨大な記憶システムを必要とする。グーグルはネット上のすべての情報を常時収集保持しているが、そのための記憶システムの維持のために発電所を付けねばならないところまで来ているという。十年先を考えた時、記憶システムとそれを動かし続ける電力のことを考えると、これまでのようにどんなデータでも無差別に集めればよいという考え方でやってゆけるのかどうか、という問題が出て来る。データ解析から得られる結果が有用で費用対効果が高いということには今でも疑問であるが、これが続いてゆくのだろうかという心配がある。グーグルのように半

ば独占的な地位を確立している場合にはメリットはあるだろうが、ビッグデータ解析はあらゆるところで行われ、企業間競争の激しい中での費用対効果について検討し、他の手法、他の道を見つける方向に転換してゆかざるを得なくなる企業も出てくるかもしれない。

何年か先にはビッグデータ解析はいわば日常的なこととして行われているほかに、ビッグデータから特定の個人や特定の事件、案件に関する情報を抽出しトレースしたりすることも行われるようになるだろうが、プライバシー問題や微妙な問題を誘起しかねないので注意が必要である。科学技術分野は別として、企業などではビッグデータ解析がいつまでも宝の山という訳にはゆかない時代になっているかもしれない。