データマイニングとは、大量のデータの中から、有用な (価値のある) 情報を抽出する技術のことです。

従来、データの量が少なかった頃は、低次元のグラフにマップするなどして、えいやっと相関をみつけだすことが一般的なデータ解析手順でした。たとえば、化学反応などは、1 〜 2 次元に射影して議論するのが一般的でした。なにをグラフの軸に選べばいいかが自明な問題だけを扱っていたとも言えます。

しかし、データの量がある程度以上多くなり、さらにその中に埋もれた本質的な次元が多くなると、何をグラフの軸に選ぶか、何本の軸があれば十分かは自明ではなくなります。人間は 3 次元空間の生物なので、4 次元以上の情報を把握するのは非常に困難です。かといって、本質的に 10 次元の問題を 3 次元にむりやり射影しては本質を見失います。自然は人間にあわせて作られているわけではないので、どんな問題でも 2 〜 3 次元に落とせるとは限りません。

計算機シミュレーションは莫大な情報を生産します。分子シミュレーションであれば、すべての分子のフェムト秒単位の動きを完全に追跡することが可能です。しかし、人間は、もちろんすべての分子の動きを理解することはできません。アラビア語を知らなければ、コーランのありがたい言葉も文字どころか文様にしか見えないのと同じように、シミュレーションが生みだす大量のデータの背景にある文法構造を知らなければ、何もわからないのです。

シミュレーションデータは現在は、平均化され、低次元な相関関数にされたあとで実験結果と照合されているケースがほとんどです。これではまるで、英語版の聖書の中のアルファベットの個数を数えて、「これは英語のアルファベットの出現比率と一致するので、英語で書かれていると言える。思ったとおりだ。」と言っているようなものです。アルファベットの隣接関係を調べ、単語の単位を調べ、単語の並び順を調べ、他の英語の文章と照合すれば、聖書の内容やその背後の思想にまで迫ることができるかもしれませんが、そこまでの解析はほとんどおこなわれていません。このような解析が伴わなければ、いかに高速なコンピュータで巨大なシミュレーションを行えるようになっても、虚しいだけです。

データマイニングは、そのような大量のデータの中に隠された文法構造を、データ自身に語らせる技術と言えます。計算機シミュレーションが生みだした膨大なデータを、計算機自身が解析することによって、人間の直感の及ばないような複雑な情報をひきだすことができる場合があります。そこに人間の直感を援用すればさらに情報の次元を適切に縮約できるでしょうし、人間の理解しやすく、かつ本質を失わない情報に還元することができると期待されます。また、未解決で重要な問題そのものを見付ける手段としてもデータマイニングが使われて実際に成果を上げています。

こういうデータマイニングをやればいい、という処方箋があるわけではないので、そこにはやはり人間の操作や発想力が必要になります。人間のすべきことが一段メタなレベルになる、と言えばいいでしょうか。

[4 日前 (今月 13 日)12 時]