

ビッグデータから隠れた知識を 探り出す超高速解析アルゴリズム

情報学プリンシプル研究系 准教授

宇野 毅明



研究背景・目的

ビッグデータの利活用における困難は3つ、どう蓄え、どう使い、どう計算するかです。さらに利用目的・方法も不明確です。「社会をより良く」「生活を便利に」「利益向上」など利活用の目的は明解ですが、データのどこを見て何を調べるか、という具体的な「すべき作業」は見えにくいものです。モデルを立て、データの必要部分のみを取り出し、効率的に計算するという常套手段が使いにくく、解析結果の解釈も非常に難しいのです。本研究では、目的達成のためデータを使って何ができるか、どのような手法で計算できるか、という「利活用のデザイン」をすること、ビッグデータの特徴を活かし、ロングテールの隅々まで網羅的に調べあげる計算手法、その両面の研究をしています。

研究内容

解析の目標がわかりにくい場合、データを見える化してデータがどのようなものかを把握し、適用手法を選ぶことが重要です。それにはデータ項目間の類似性・共通性を網羅的に調べ上げ、ある程度の共通性・均質性を持つグループに分類し、データがどんな種類のもので構成されているかを知る必要があります。今までは、グループ数が少なく均質性が著しく低い分類、高い共通性を持つグループを見つける際、数が非常に大きくなる手法しがなく、こういったデータの根源にせまる方法はありませんでした。本研究では、非常に短い時間でデータの類似性を網羅的に解析する計算手法と、ノイズ混じりのデータを明確化して数の少ない均質な

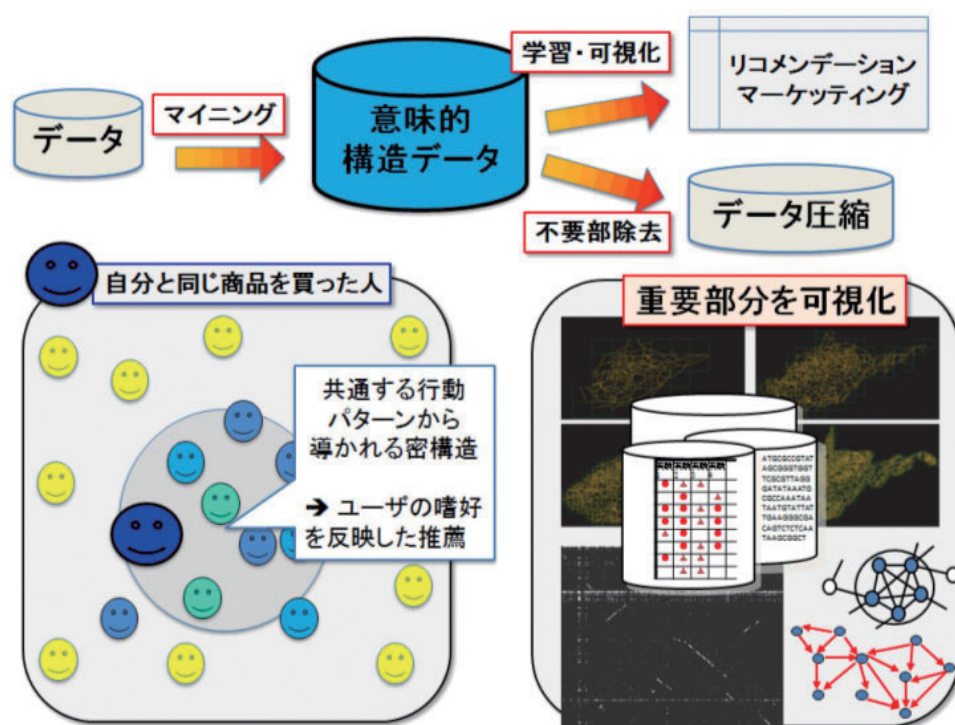
グループを見つけ出す手法を開発しました。これにより、巨大なデータでも1時間程度で1000個程度のグループに分類することが可能となり、各グループに統計／機械学習／認識などの技術を適用することで、非常に高い精度の結果が得られます。

産業応用の可能性

- 人間、動物、企業など、多様な個体の行動データ、購買データ、Twitter・blogなどの解析
- 都市計画・避難計画・交通制御などに用いる、人の分布・移動経路の解析
- ソーシャルグラフ、化学化合物グラフなど、グラフデータの解析
- 文書データ、ゲノムデータなどの文字列データの解析

研究者の発明

- 特許第5267847号：あいまい頻出集合の探索方法及び探索装置



連絡先：宇野 毅明【情報学プリンシプル研究系 准教授】 URL <http://research.nii.ac.jp/uno/index-j.html>