

## ビジネス向けのデータ可視化分析とその応用について A Study of Data Visualization Analysis for Business and its Application

馮鑫† 能上 慎也‡

Xin Feng Shinya Nogami

### 1. 研究背景

21 世紀に入って視覚分析学が生まれるとともに、各学科を統合した新しい理論模型と新しいインタラクティブ手段が出現し、異なる領域の専門家同士が深く結合して、金融、医療、宇宙、地理などの領域における可視化分析方法と技術が形成されてきた。

情報化時代・社会における論理的思考の具象化に対する人々の要求の産物であるこの「データ可視化」は、データが爆発的に増加している現代の重要な課題となっている。

### 2. 研究目的

本研究では、データ可視化分析、特に商用化に向けたデータ可視化分析と関連応用の論理的な考え方を提案することを目的とする。

ネットワークを通して得られた巨大なデータの分析結果とそれをを用いた活動予測が、日常生活、特に商業において活発に活用されている。どのようにして人類活動の視覚化の特徴とデータ分析の結果を結合させて、直観的に展示できるのかが、本研究の中心的テーマである。

特に今回の報告では、収集した第一四半期の企業の製品販売データを用いて分析し、可視化を行う手順を整理する。可視化の効果とは、画像を視覚化することにより抽象的な論理によってデータを理解することができる点である。その可視化結果によって、将来の GMV(Gross Merchandise Volume)と利潤の予測、企業の市場進出戦略の選択などをよりの確に判断できるようになる。

### 3. 研究内容

#### 3.1 研究手法

本研究は、中国での電子商取引会社（「拼多多：ピンドウドウ」）の 2019 年度の 3 つの商品の販売データを用いて、データ補間、データカーブフィッティング、SOM アルゴリズムによるデータを処理した上で、可視化を行う。

†東京理科大学大学院 Tokyo University of Science, Graduate School

‡東京理科大学 Tokyo University of Science

そこから導き出された結論によって企業の生産活動に対して合理的な提案をすることを考えている。

#### 3.2 データソース

本研究では、ピンドウドウにおける売れ筋商品 TOP3 の代表的な商品（表 1 のように示すこと）：ティッシュペーパー、ビデオカメラ、マグカップのデータを選択する。

表 1 売れ筋商品トップ 3 のデータ

年月日	地域	顧客分類	品名	販売量	単価(元)	売上(元)
202001 04	中国		マグカップ	12	6	72
202001 04	マレーシア		マグカップ	14	6	84
202001 05	マレーシア		ビデオカメラ	4	2050	8200

#### 3.3 データ可視化ツールの選択

採用する可視化ツールは、使用方法の視点から見ると、すぐに使えるツールとコード開発ツールの 2 種類に分けることができる。それぞれのデータ可視化開発ツールの長所と短所を組み合わせる上で、本研究ではデータを処理して可視化する必要上、操作性を重視して選択する。

### 4. 研究過程

#### 4.1 データ処理と分析の方法

データ可視化する前にデータを簡単に処理する必要があるが、ここではビジネスデータの特徴を結合して、主にデータ補間、データカーブフィッティング、および SOM アルゴリズム、の 3 種類のデータ処理方法を採用する。

##### 4.1.1 データ補間

データ補間とは、既知の離散点から未知点を決定する計算過程であり、データ解析においてよく使用されるデータ処理方法である。例えば、関数  $f(X)$  が  $[a, b]$  上に定義され、かつ  $a < X_0 < X_1 < X_2 < \dots < X_n \leq b$  の点上の値  $Y_0, Y_1, Y_2, \dots, Y_n$  が既知であるとして、もし単純関数  $p(X)$  が存在するならば、且つ

$$p(X_i) = Y_i \quad i=0,1,2,3,\dots,n \quad \dots (1)$$

が成立すれば、 $p(X)$  を  $f(X)$  の補間関数と呼ぶ。また、点  $X_0, X_1, \dots, X_n$  を補間節点、式(1)を補間条件、 $f(X)$  を被補間関数、 $[a, b]$  を補間区間と呼ぶ。 $p(X)$  を求める方法が補間法であり、 $p(X)$  が次数  $n$  を超えない実係数代数多項式ならば、

式(2)のように書くことができる。

$$p(X) = a_0 + a_1X_1 + \dots + a_nX_n \quad \dots (2)$$

#### ■ 4.1.2 データカーブフィッティング

データカーブフィッティングは、平面上のそれぞれの離散点で表される座標間の関数の関係を連続曲線で近似的に描く処理方法である。連続的な関数を用いて離散的なデータを分析する定型的なデータ処理は、科学研究や実際の生産過程において観測や実験で得られたデータ対照グループ  $(X_i, Y_i)$  ( $i=1, 2, 3, \dots, m$ ) を用いて行うことができる。ここで  $X_i$  の値はそれぞれ異なっており、また  $Y_i$  の値も異なっている。これらの実験によって得られたデータを少なくとも無限に近似して近づける関数が必要となるが、その関数をフィッティング(あてはめ)関数として求める。このフィッティング関数が得られると、データの中での異常値や欠落値を削除したり、補完したりすることができる。このようなデータ処理の方法は、データ可視化の過程で特に重要である。

#### ■ 4.1.3 SOM(Self-organizing maps)アルゴリズム

多次元データの可視化を実現する方法は、1つの高次元のデータ情報の次元を下げて可視化するのが一般的である。SOM アルゴリズムもまた主に次元を低下させて可視化を実現するものである。低次元データの空間は、高次元データの情報を表す複数の仕様メッシュ部分に分割されるが、これを実現するために、低次元空間のメッシュは約  $M$  個のセルから構成されると仮定する。各セル  $j$  は、高次元情報と同じ次元のプロトタイプベクトル (Prototype vector)  $(M_j = (M_{j1}, \dots, M_{jd}))$  を持ち、メッシュ内のセルの位置  $r_j$  は固定されている。SOM のトレーニング過程とは、実際にプロトタイプベクトルのデータを調整する過程であり、トレーニングによってプロトタイプベクトルが低次元空間に対応する位置を得ることができ、これによって低次元空間で高次元の情報を表すことができる。SOM アルゴリズムは、自己組織化ニューラルネットワークに関する手法であり、重みとしきい値をトレーニングする過程で次元削減の対応法則を探索し、データ可視化の次元変化のためのアルゴリズム支援を提供する。

#### ■ 4.2 研究フローチャート

研究を進めるにあたり、全体の手順を図1のフローチ

ャートに示す。

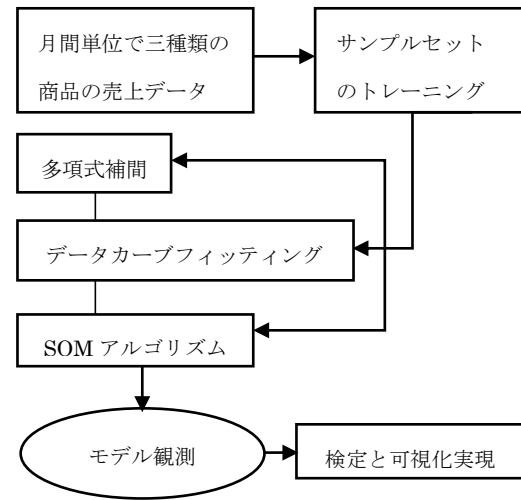


図1 研究全体のフローチャート

#### ■ 5. まとめ (期待される効果)

効率的なデータ可視化技術は、まだ電子商取引のビッグデータの分析管理に広く応用されているとはいえ、データ価値の損失を招いていると思われる。成熟した動的データ可視化技術及びビジネスデータモデルに適合する可視化形式を電子商取引のビッグデータ分野に応用することは、企業の正しい発展方向であると考えている。本研究の目標は、図2のように処理したデータの可視化を行い、それを展示した上でそれぞれの関係を明らかに表示できるようにすることであり、それによって企業の解決策根拠を論理的に提供することを目指している。



図2 可視化結果の表示例

#### 【参考文献】

- [1]松岡大祐, 荒木文明: “大規模シミュレーションデータ可視化における研究の動向”, J-STAGE 2011 (13) : pp.35-63 (2011).
- [2]Bruce H. McCormick, Thomas A. DeFanti and Maxine D. Brown (eds.), Visualization in Scientific Computing. ACM Press (1987).
- [3]Chen, W., L. Ren, M. Zwicker, and H. Pfister, Hardware-accelerated adaptive EWA volume splatting, IEEE Visualization '04, pp.67-74 (2004).