相関性を考慮した大規模階層型データの可視化

- クレジットカード不正履歴テストデータの可視化への応用 -

長崎あずさ*　伊藤貴之*　伊勢昌幸**　宮下光輔**
*お茶の水女子大学 理学部情報科学科
**株式会社インテリジェントウェイブ開発本部企画部

1. 概要

情報可視化は日常の一般的な情報を可視化する技術であり，その利用者には計算機の操作に熟達していない人，あるいは多忙すぎて計算機を操作する余裕のない人も含まれる．一方で，私達の身の回りの情報は急速に巨大化・複雑化しており，その中から特徴的な現象をすぐに発見できるとは限らない．このことから，複雑な情報の中から特徴的な現象を，計算機を十分に操作しない利用者に対して半自動的に提示する技術を確立することも，情報可視化の重要な課題の一つであるといえる．特に大規模階層型データは数値的特徴や全体的な傾向が掴みにくく，データを解析する際にデータが大きくなればなるほどユーザに負担がかかってしまう．

そこで本報告では，表形式又はリレーショナルデータベース形式のデータを構成する属性間の相関関係を検出し，その結果に基づいて階層型データを構築し，数値的特徴を読み取りやすい可視化結果を半自動的に提示する手法を提案する．

提案手法では階層型データ可視化手法「平安京ビュー」[1]を用いる．「平安京ビュー」は，階層型データの葉ノードを色のついたアイコンで，葉ノードのグループを入れ子のように描かれた長方形群を使って表現し，階層型データの全体像を一画面で見ることができる可視化手法である．棒グラフにカーソルを当てると，詳細情報が表示されることにより，Decision Making などにも活用できる手法であると考えられる．提案手法では，相関性があるとみられる属性を「平安京ビュー」の色，高さ,グループに割り当てることで，半自動的に可視化結果を生成する．

本報告では可視化対象となるデータに，クレジットカードの不正履歴テストデータを用いている．

2. 提案内容

図1は本手法の処理手順を示したものである．本手法では，表形式またはリレーショナルデータベース形式のデータにおいて属性間の相関性を算出し，相関性が高いとみられたものを「平安京ビュー」の色，高さ，グループに割り当てて可視化する．

"Visualization in consideration of correlation of large-scale hierarchical data"
Azusa Nagasaki*, Takayuki Itoh*
Masayuki Ise**, Kousuke Miyashita**
Ochanomizu University　**Intelligent Wave Inc.
{azusa, itot}@itolab.is.ocha.ac.jp

属性同士の相関性の算出には，その属性の特徴に合わせてケンドールの順位相関係数，標準偏差，エントロピーなどを用いる．

標準偏差とエントロピーは，以下の方法で用いる．

- 属性を 1 つ選び(以下属性 A とする)，属性 A の値に従ってデータ全体をグループ分けする
- それぞれのグループについて，別の属性 B の標準偏差あるいはエントロピーを求める

ケンドールの順位相関係数

ケンドールの順位相関係数とは，順位関係を比較することによって相関係数を算出する方法である．具体的には，表形式データから 2 行を抽出し，その 2 行における属性 A の 2 値の大小関係と，別の属性 B の 2 値の大小関係と比較する．この処理を全ての 2 行の組み合わせ M 組について適用する．大小関係が一致する組数を K，不一致の組数を L とするとき，相関係数は以下のように表せる．

$$r = \frac{K - L}{M} \qquad (1)$$

これにより求められる相関係数の絶対値が 1 に近いほど，属性 AB 間の相関性が高いと予想できる．

ケンドールの順位相関係数は，属性値の順序や大小関係に基づいて相関性を計算する．よって，値の間に明確な順序や大小関係がある属性の相関性算出に向いている．

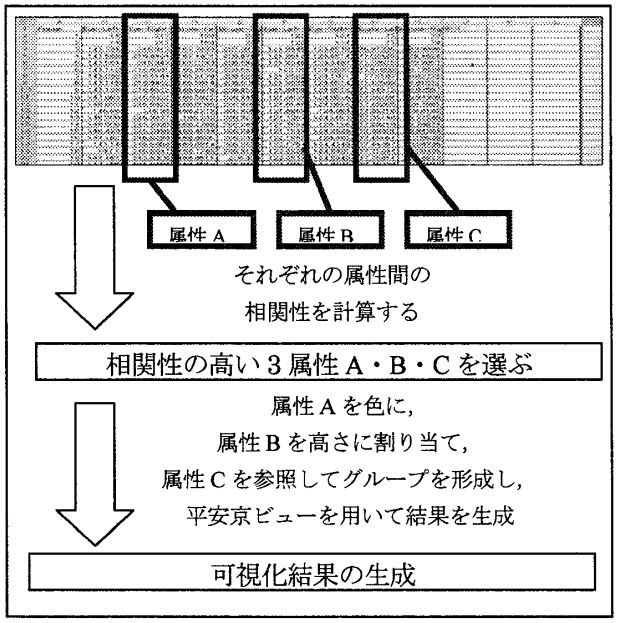

図 1：本手法のフローチャート

標準偏差

標準偏差とは値の散らばり具合を示す指標である．属性Aの値によって属性Bの値をグループ化したとき，属性Bの各グループ内における標準偏差を，それぞれ以下のように求めるとする．

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (2)$$

ここで，nは各グループ内におけるデータの個数であり，$\bar{x}$は以下の式で表される相加平均である．

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (3)$$

このようにして算出される標準偏差が小さいほど，同じグループ内の属性Bの値の散らばりが小さい．このような場合に，属性Aと属性Bの間に相関性が高いと判断できる．

エントロピー

エントロピーとは不確かさを示す指標である．属性Aの値によって属性Bの値をグループ化したとき，属性Bにおいてそれぞれの値の生起確率p_iを計算し，それを用いて以下の式でエントロピーを計算する．

$$H = -\sum_{i=1}^{n} p_i \log p_i \qquad (4)$$

グループ化の前後を比較し，グループ化後の方がエントロピーの合計が減っていれば，それは同一グループに同一値が集中していると考えられる．このような場合に，属性Aと属性Bの間に相関性が高いと判断できる．

3. 実行結果

本報告では，ケンドールの順位相関係数を用いた実行結果を示す．

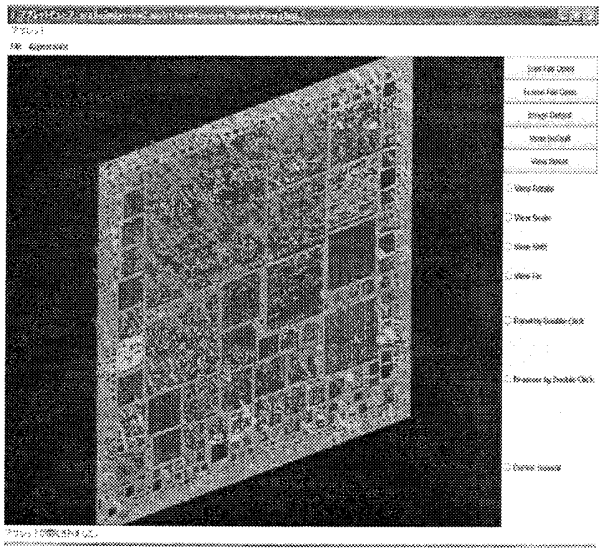

図2：実行結果1

図2は，ケンドールの順位相関係数が約-0.685と，強い相関関係があるとみられる支払区分と加盟店コードについて生成した画像である．ここでは，色には支払い区分，グループには加盟店コード，高さには不正使用金額が割り当てられている．ここから，同一グループ内に同じ色が集中して現れていることが読み取れる．よって，店によって支払区分の傾向に差があることが読み取れる．

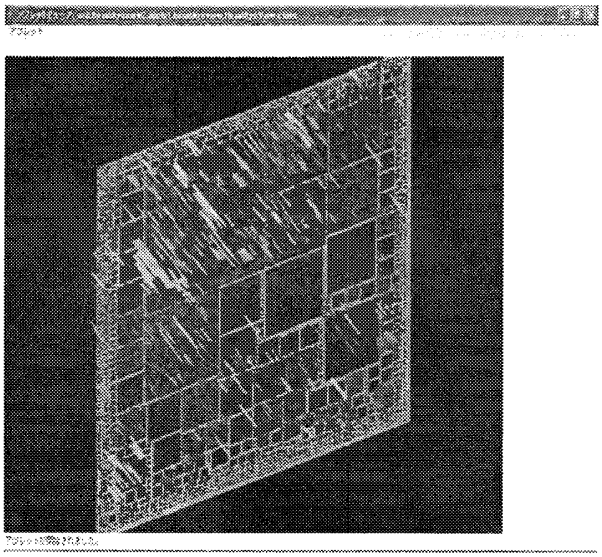

図3：実行結果2

図3は，図1において高さを強調表示するために色と高さに不正使用金額，グループに加盟店コードを割り当てて生成した画像である．この画像では，赤>橙>黄緑>青の順に高額の不正使用となっている．よって，特定の店に高額の不正使用が集中していることがわかる．

4. まとめと今後の課題

本報告では，データの傾向や全体像をより見やすい形で可視化するために，自動的に属性同士の相関性を検出し，その検出結果に基づいて階層型データを構築し，「平安京ビュー」を用いて可視化する手法を提案した．

今後の課題として，まずクレジットカードの不正履歴テストデータを用いた検証を進めるために，以下に着手したい．

- 属性情報を追加する
- 時間帯，金額などの属性の数値を直接用いずに，ヒストグラム化した形で相関性を算出する

また本手法の有用性を高めるために，以下を考察したい．

- 相関性の算出手法の考察
- 用いたデータにおけるそれぞれの属性の数値的特徴とそれに相応しい相関性の算出手法についての考察

参考文献

[1]伊藤，山口，小山田，「長方形の入れ子構造による階層型データ視覚化手法の計算時間および画面占有面積の改善」，可視化情報学会論文集，Vol. 26, No. 6, pp. 51-61,2006.