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○(お茶の水女子大学大学院)
伊藤貴之(お茶の水女子大学大学院)
伊勢昌幸((株)インテリジェントウェイブ開発本部企画部)
宮下光輔((株)インテリジェントウェイブ開発本部企画部)

Visualization in Consideration of Correlation of Large-scale Hierarchical Data
—Application to Visualization of Test Data of Injustice History of Credit Card—

Azusa NAGASAKI, Takayuki ITOH, Masayuki ISE, and Kousuke MIYASHITA

ABSTRACT

Information visualization is a technical field which targets visualization of general information, and its users include people who are not proficient in computer operations. On the other hand, the information on our personal appearance is grown large and complicated quickly, and we cannot discover the numerical feature immediately from them. Therefore, it is one of the important issues of information visualization to establish techniques which semi-automatically represent characteristic phenomena.

In this report, we propose a technique which semi-automatically presents numerical features and overall tendency, as an extended technique of hierarchical data visualization technique "HeiankyoView". The technique chooses three attributes from tabular form data or relational database form, based on their correlativity. It effectively represents the features or tendency, by assigning the three attributes to color, height, and grouping, of "HeiankyoView".

Keywords: Visualization, Hierarchical data, correlativity

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や局所的な事象をすぐに発見できるとは限らない．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，データが大きくなればなるほどデータを解析する際にユーザに負担がかかってしまう．このことから，計算機の操作に十分に熟達していない利用者に対して，複雑な情報の中からユーザが求めるような特徴的な現象や局所的な事象を半自動的に提示する技術を確立することも，情報可視化における重要な課題の一つであるといえる．

我々はクレジットカード不正使用分布の情報可視化に取り組んでいる．クレジットカードの不正使用は深刻な社会問題となっており，その不正使用の数は増え続けている．PIO-NET と呼ばれる全国消費生活情報ネットワーク・システムに寄せられる相談件数は，2000 年には 580 件だったが，2005 年には約 1400 件と，5 年で 1000 件も増加している．さらにスキミングと呼ばれる，クレジットカードを盗み，カード情報を読み取った後にクレジットカードそのものは被害者の元に返し，読み取ったカード情報を用いて不正を行う方法がある．この方法で不正が行われてしまうと，クレジットカードは被害者の手元に残っているために，被害者はカードの明細書をカードの使用履歴と照らし合わせてきちんとチェックしない限り不正使用に気づきにくい．したがって，被害者が気づいていない不正使用も鑑みると，莫大な数のクレジットカードの不正使用がそこには内在している，ということが考えられる．

この不正使用を防ぐための多くのシステムでは，例えば「2000 ドル以上のクレジットカードの使用がガソリンスタンドであれば，それは不正である」といったような「ルール」をクレジットカード会社が定め，それに該当した場合，リアルタイムでクレジットカードの不正使用であるという判断が下される，という仕組みを採用して

いる．しかし，この「ルール」を設ける際には，人間が今までのクレジットカードの不正使用のデータを見て，ヒューリスティックに決めている．そのため，この「ルール」の中には，あまり参照されないものや意味がないものも多く含まれているのが現状である．

したがって本研究では，クレジットカードの不正使用を防ぐために，今までの不正使用履歴のテストデータを分析して不正使用の傾向を掴み，より良い「ルール」を作るための指標を作ることを目指している．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算出し，その結果に基づいて階層型データを構築することにより，生データをそのまま可視化しただけでは見つけるのに時間がかかってしまうような，数値的特徴や全体的な傾向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．

2．関連研究

クレジットカードの不正履歴以外にも，ビジネスの局面においてその大規模データを可視化して分析する，という試みは行われている．VisImpact[1]では統計的な相関分析などデータマイニングのテクニックを使用し，それを基に可視化を行って，データの分析に役立てている．

また，階層型データの可視化手法の著名な手法として，階層構造を木構造として表現する手法と，階層構造の末端にあたる葉ノードを 2 次元的に画面空間に配置する手法がある．前者の中で有名な手法には，Hyperbolic Tree[2]や Cone Tree[3]が挙げられる．後者の中で有名な手法には，画面空間の 2 次元的分割により葉ノードを一括表示する TreeMaps[4]が挙げられる．本論文の提案手法が用いる階層型データ可視化手法「平安京ビュー」[5]も，後者に属する手法である．本論文では大規模データを一画面上に展開して一括表示することによりデータの分析に役立てることを目的としているため，後者のような階層型データ可視化手法が適切であると考えられる．

「平安京ビュー」は，大規模階層型データの可視化のための一手法である．本論文の研究手法では，これを用いて階層構造にしたクレジットカードの不正使用履歴データを可視化する．「平安京ビュー」は，階層型データの葉ノードを長方形のアイコンで，枝ノードを長方形の枠で表現し，階層構造を 2 次元の長方形郡の入れ子構造で表現し，これらをできるだけ小さい画面空間に配置することで，階層型データ全体を一画面上に表示する．また，可視化するデータの属性をそれぞれ「平安京ビュー」の色，高さ，グループに割り当てることによって，そのデータの数値的な特徴や全体的な傾向を読み取ることができる．Fig. 1 は「平安京ビュー」を用いた可視化結果の一例である．

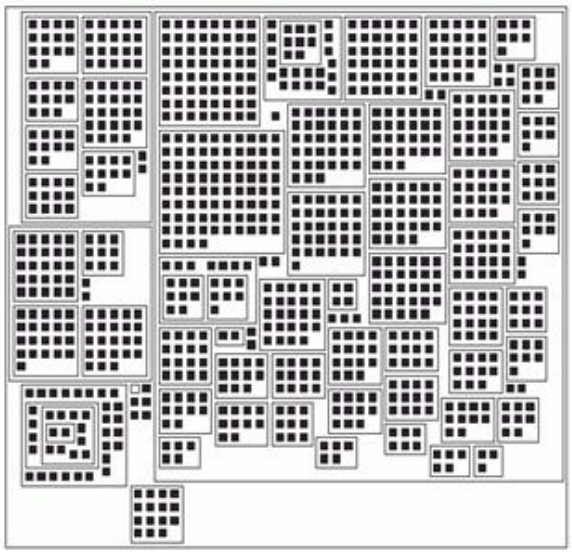

Fig. 1　Sample of the large-scale hierarchical data visualized by “HeiankyoView”

3．提案内容

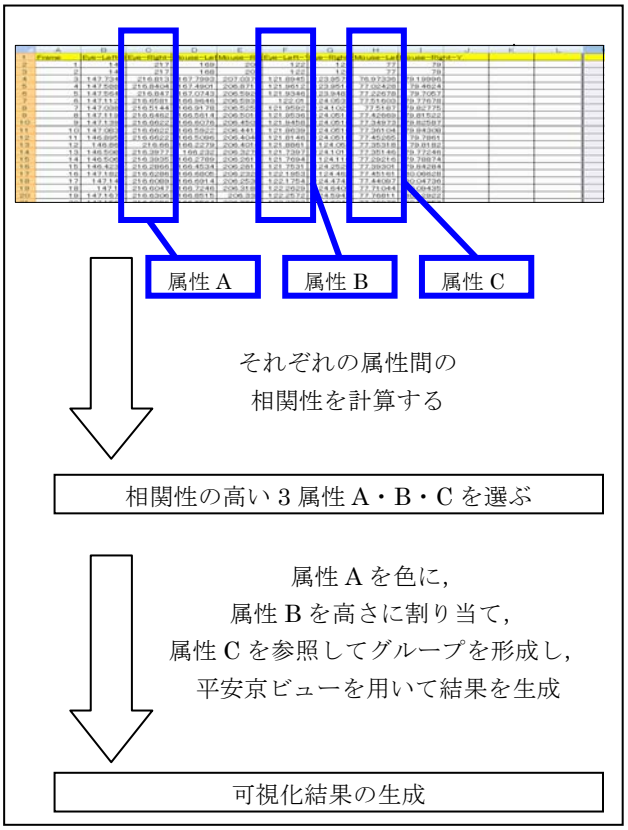

Fig. 2　Flow chart of the technique

Fig. 2 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．

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，ヒストグラムなどを用いる．

標準偏差とエントロピー，ヒストグラムは，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
- 各グループについて，別の属性 B の標準偏差あるいはエントロピー，ヒストグラムを求める
- それぞれのグループにおける最小値やグループ全体の総和を，グループ分けする前の標準偏差やエントロピー，ヒストグラムと比較する．

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，それぞれの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 のある 2 値の大小関係と，別の属性 B の同じ列の 2 値の大小関係とを比較する．この処理を 2 行の全ての列の組み合わせ 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．

$$r = \frac{K - L}{M} \tag{1}$$

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．

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 の値をグループ分けしたとき，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．

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

ここで， は属性 A と属性 B におけるデータの個数の最小値であり，は以下の式で表されるような相加平均を示している．

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

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 の間の相関性は高い，と判断することができる．

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 の値をグループ分けしたとき，属性 B においてそれぞれの値の発生確率 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．

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{4}$$

グループ分けする前とした後とを比較し，グループ分けをした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の値が集中している，と考えられる．このような場合に，属性 A と属性 B の間の相関性は高いと判断できる．

ヒストグラム

ヒストグラムを用いて，属性 A の値によってグループ分けする前とグループ分けした後のそれぞれにおける属性 B の値の発生確率を比較する．値が大きく増減しているものがあれば，それは属性 A でグループ化したことにより属性 B の値の分布に影響が出た，ということが考えられる．よって，そのような場合は属性 A と属性 B の間の相関性は高く，それを基に可視化を行うことにより，特徴的な現象を見つけやすいのではないか，と考えることができる．

比較の際には以下の式を用いる．

$$P = \frac{1 - P_a}{1 - P_b} \tag{5}$$

ここで，P_a はグループ分けした後の発生確率を示し，P_b はグループ分けする前の発生確率を示す．この式はグループ分けした後に，どれほどそのグループに特定の値が集中しているかを示す．これが 0 に近いほど，グループ分けした結果，属性 B の値が同じグループに集中するようになった，ということを示す．

4．実行結果

Fig. 3 に，ケンドールの順位相関係数の計算結果に基づいた可視化例を示す．この例では，支払区分と加盟店コードの 2 属性の間に，ケンドールの順位相関係数が約 -0.685 という強い相関関係があるとみられた．したがって，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を割り当てて可視化を行った．この画像から，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わかる．また，丸で囲ったグループには高さが高いアイコンが集中している．よって，特定の店に高額の不正使用が集中していることも読み取ることができる．

続いて Fig. 4 に，ヒストグラムの計算結果に基づいた可視化例を示す．商品コードと時間帯について，ヒストグラムを用いて計算した結果 P =0 となったので，これらの属性の間には相関があると考えられる．そこで，色に時間帯，グループに商品コードを割り当てて可視化を

行った．なお，時間帯は 24 時間を 30 分ごとに 48 個にわけており，水色が早朝，黄緑が昼の少し前，オレンジや赤が夕方から夜にかけて，を表している．ここで，赤い四角で囲まれたグループは他のグループと比較して非常に水色のアイコンが多く現れている．この水色は 6 時から 9 時の間を示し，このグループは鉄道・バスの商品購入による不正使用を表している．よって，このことから「朝の 6 時から 9 時という混雑している時間に鉄道・バスでのクレジットカードの不正が起きやすい」ということがわかる．これは，クレジットカードの不正使用の防止のための，より効果的な「ルール」作りに役立つだろう．また，同様に白の四角で囲まれたグループには黄緑色のアイコンが頻出していることからこれらのグループでは昼の少し前に不正使用が頻発しており，桃色の四角で囲まれたグループではオレンジ・赤のアイコンが頻出していることから夕方以降の不正使用が頻発していることが読み取れる．

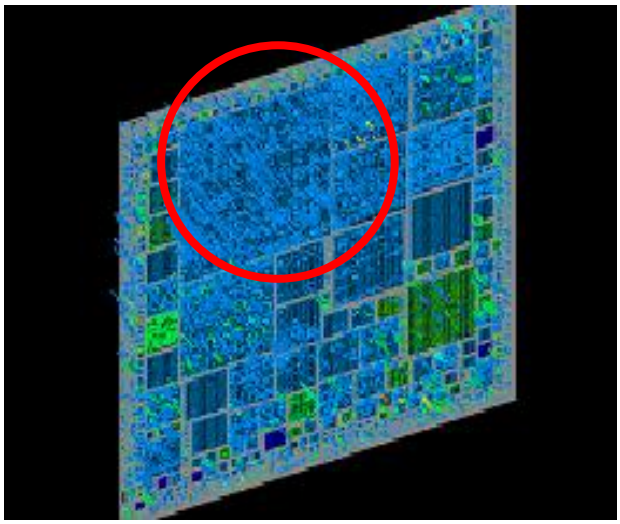

Fig. 3　Result(1)

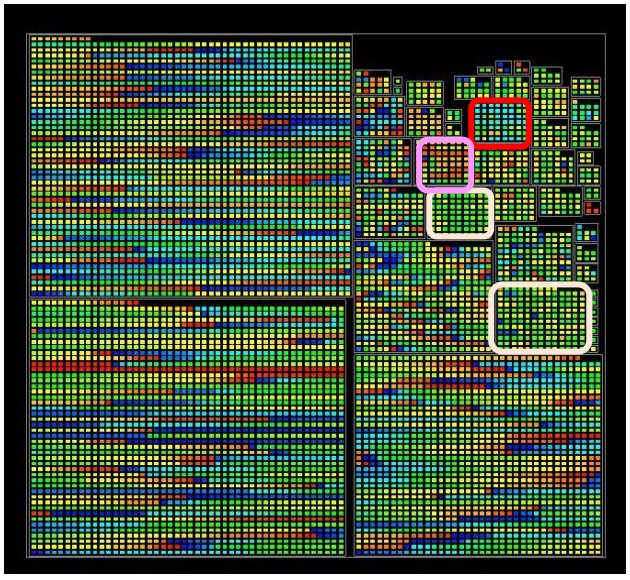

Fig. 4　Result(2)

5．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，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．

- 相関性の算出手法の考察
- 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
- グループ内の値の内訳の詳細表示
- 3 つ以上の属性間における相関性の算出
- データの有効範囲の設定

参 考 文 献

[1] Ming C. Hao and Daniel A. Keim and Umeshwar Dayal and Jo¨rn Schneidewind, “Business Process Impact Visualization and Anomaly Detection”, Information Visualization, pp. 15- 27, 2006.

[2] Lamping, J. and Rao, R., ″The Hyperbolic Browser: A Focus + Context Technique for Visualizing Large Hierarchies,″ Journal of Visual Languages and Computing, Vol. 7, No. 1, pp. 33-55, 1996.

[3] Carrire J. and Kazman R., ″Research Report: Interacting with Huge Hierarchies: Beyond Cone Trees,″ Proceedings of the IEEE Conference on Information Visualization ’95, IEEE CS Press, pp. 74-81, 1995.

[4] Saraiya P., North C., Duca K., An Evaluation of Microarray Visualization Tools for Biological Insight, IEEE Information Visualization 2004, pp. 1-8, 2004.

[5] 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.