2変数詳細度制御を用いた大規模データの可視化

長崎あずさ （指導教員：伊藤貴之）

1 概要

私たちの身の回りには多数の大規模データが存在する．コンビニでの買い物や，IC チップに記録される行動の経路など，私たちの暮らしにおいては今や大量のデータを収集されることが日常的となっている．しかしそれらのデータは非常に大規模かつ複雑で，解析が困難なものがあまりにも多い．その理由としては，データ自体の莫大さのみならず，特にデータベース化したときの属性数が多すぎることが挙げられる．たとえばコンビニで買い物をしただけでも，購買者の個人情報や商品情報，コンビニの位置情報，購買日時や天気など，取得される情報は非常に多岐に渡り，かつ膨大である．このような背景からデータの属性数は非常に多くなってしまい，そのためデータ全体の解析への需要は高いにも関わらず解析が困難な事例はまだまだ多い．

本論文では，このようなデータの解析を支援する一手法として，2 つの属性による詳細度制御を用いた大規模データ可視化手法を提案する．ここで大規模データとは，レコード数が多いだけでなく，前述のように属性数が多いデータを指す．このようなデータは，大規模な情報全てを可視化してしまうとデータの特徴や傾向を掴むことが非常に難しい．

本論文では 2 属性による詳細度制御を組み合わせることで，データの特性を保ったままユーザに提示する情報量を減らして可視化することにより前述の問題を解決し，大規模データをユーザにとって解析しやすいものとすることを目標とする．

2 関連研究

提案手法が前提とする可視化手法「平安京ビュー」[1] は，大規模階層型データ可視化手法である．平安京ビューは，階層型データの葉ノードをアイコンで表示し，枝ノードを長方形の枠で表示することで，階層型データの全体を一画面に表現することができる手法である．階層型データの葉ノードが複数の属性を有するとき，平安京ビューでは色・高さ・グループに割り当てる属性を自由に選ぶことができ，これによってデータ中の特定の属性間の相関性などを表現できる．

ただし，前述のような大規模データをそのまま可視化すると図 1 のようになってしまい，ユーザにとって非常に見づらく，解析しづらい可視化結果となってしまう．よって本研究ではこの多すぎる情報量を減らすため，色とグループに着目し，それらに割り当てる 2 属性に基づく詳細度制御を実現した．

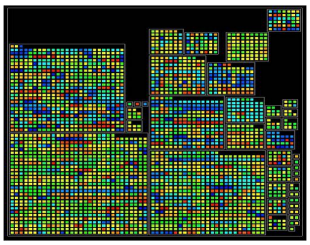

図 1: 平安京ビューによる可視化例

3 提案内容

提案手法ではまず，特定の属性 A_p でレコードをグループ化することにより，図 2 のようにして大規模データから階層型データを構築する．そして，これとは別の属性 A_q を平安京ビューの色に割り当てることにより，図 3 のように平安京ビューにて A_p と A_q の 2 属性を眺めることが可能になる．

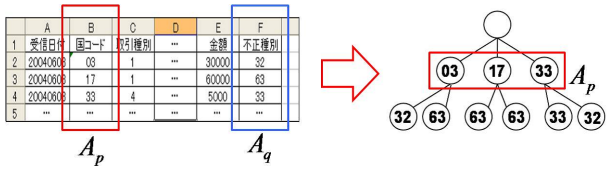

図 2: 大規模データの階層型データへの変換

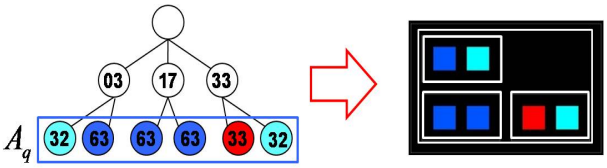

図 3: 階層型データの平安京ビューによる表示

続いて，2 属性による詳細度制御の概要について述べる．図 1 のような画像を特徴や傾向がつかみやすいように表示するため，グループ数を減らすこととアイコン数を減らすことによってユーザに提示する情報量を減らす．前者は階層型データの枝ノードにあたる属性，すなわち平安京ビューのグループに割り当てる属性による詳細度制御を取り入れることによって実現し，後者は色に割り当てる属性による詳細度制御を取り入れることによって実現する．

これは現時点では前者を適用した後，後者を適用している．この 2 属性による詳細度制御の詳細について以下に述べる．

3.1　グループに割り当てる属性に基づく制御

グループに割り当てる属性 A_p に基づく制御では，情報量を減らしつつも全体的な特徴と局所的な特徴の両方を掴むため，ノード数による制御と，グループ間の類似性による制御の 2 つを，以下に述べる手法で行うことにより，階層型データの枝ノード同士を統合する．

3.1.1　ノード数に基づく制御

これはデータの全体的な特徴を掴むため，小グループを統合することによって実現する．同階層において，属する葉ノード数が一番多い枝ノードに比べて，属する葉ノード数が無視できるぐらい少ない任意の枝ノードが複数あれば，それらの枝ノード同士を統合する．このようにして少しの情報量しか持たないような小グループを統合することによって，全体的な特徴を見やすくすることができる．これによる可視化例を図 4,5 に示す．

3.1.2　類似性に基づく制御

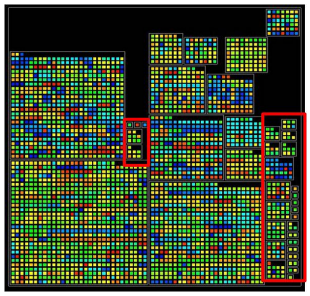

図 4: 平安京ビューによるグループ数が多いデータの可視化例

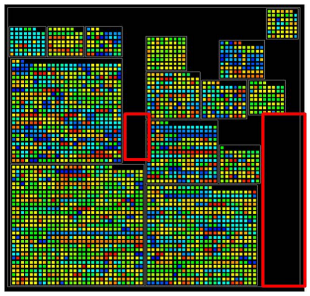

図 5: 図 4 にグループに割り当てる属性による詳細度制御を適用した可視化例

これはデータの局所的な特徴を掴むため，類似グループを統合することによって実現する．それぞれのグループに属する葉ノード群間の属性 A_q の数値分布が類似している場合に，グループ同士を統合して一つのグループとして表す．類似性の判定方法を以下に示す．

1. データ全体を平安京ビューのグループに割り当てる属性でグループ分けをする
2. それぞれのグループにおいて，色に割り当てる属性の値のヒストグラムを作成する
3. 任意のグループ G_i と G_j のヒストグラムの余弦を計算する
4. 3 の処理を全ての 2 つのグループの組み合わせについて適用する
5. 計算結果の値が高い順にグループを統合する
6. 1〜5 の処理を全てのグループに割り当てる属性と色に割り当てる属性の組み合わせについて適用する

これにより統合されなかったグループは局所的な特徴をもつと言える．

3.2　色に割り当てる属性に基づく制御

色に割り当てる属性 A_q の詳細度制御として，特定階層における属性 A_q の数値分布をアイコン数を減らして表示する．特定の枝ノードに属する葉ノードを全て結合して一つのノードとして扱い，結合する前のノードの属性 A_q の値の分布を treemaps[2] のように長方形を分割して描くことにより実現する．

図 7 は図 6 に，「ノード数に基づく詳細度制御」「類似性に基づく詳細度制御」を適用してグループの統合を行ったものに，「色に割り当てる属性に基づく制御」を適用して葉ノードの統合を行ったものである．グループの統合を行った時点で，特定のグループが統合されずに残ることによりユーザは局所的な特徴を発見することができるが，これに葉ノードの統合を行って色の分布を強調することにより更にそれを見つけやすくすることができる．

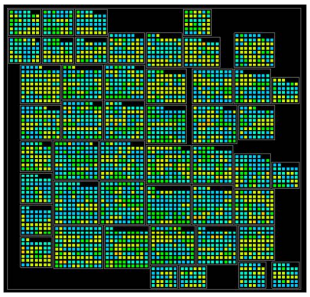

図 6:　平安京ビューによるデータの可視化例 2

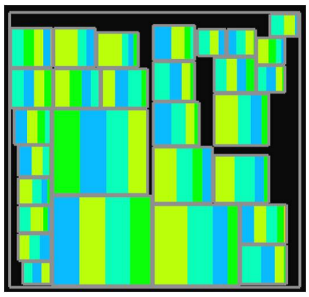

図 7: 図 6 に 2 変数詳細度制御を適用した可視化例

4　まとめ

本論文では，大規模データの可視化の一手法として，2 変数詳細度制御によりデータの情報量を減らして情報を解析しやすくする手法を提案した．

謝辞

貴重なデータとご助言を数多く賜りました株式会社インテリジェントウェイブ様に感謝いたします．

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.

[2] Johnson B., Shneiderman B., Treemaps: a space-filling approach to the visualization of hierarchical information structures. In *Proceedings of IEEE Visualization '91*, pp. 284–291, 1991.