「平安京ビュー」によるマトリクス型データの可視化の試み

～「アンケート集計結果」と「遺伝子ネットワーク」の可視化への応用

伊藤 貴之○　橘 春帆　西山 慧子

(お茶の水女子大学 理学部情報科学科)

Visualization of Matrix Data by HeiankyoView
Application to Visualization of Questionnaire and Genome Network

Takayuki ITOH,　Haruho TACHIBANA,　Keiko NISHIYAMA

ABSTRACT Table or matrix data is very popular data structure, and therefore various visualization techniques for such data structure have been presented. However, many of the data in our daily life or work are very sparse, and therefore the usage of display spaces is not always reasonable while using existing table or matrix data visualization techniques. One idea to improve the display space usage is translation of the table or matrix data into hierarchical or graph data. This report discusses the visualization of matrix data by HeiankyoView. We assume that this discussion is useful for visualization of various matrix data including questionnaire or genome network data.

Keywords: Visualization, Hierarchical Data, Matrix Data, Questionnaire, Genome Network.

1. はじめに

表形式データやマトリクス型データは、日常生活や日常業務に非常に多く見られるデータ構造であり、その可視化技術も旧来から幅広く開発されている[Rao94][Tree]。一方、現実に収集されるマトリクス型データは非常に疎である場合が多く、マトリクス型のまま可視化した場合、可視化結果の画面有効利用性に問題がある場合が多い。そこでマトリクス型データそのものを可視化する代わりに、マトリクス型データを構成する各行をノードとする階層型データやグラフデータを可視化する、という試みも見られる。最近では、マトリクス型の可視化手法とノード型の可視化手法の可読性を比較する研究も見られる[Gho04]。

本報告では、表形式データやマトリクス型データを、階層型データやグラフデータに変換し、「平安京ビュー」[Ito03]を用いて可視化する試みについて論じる。まず本報告では、表形式データやマトリクス型データの変換方法について定義する。続いて本報告では、「アンケート集計結果」「遺伝子ネットワーク」という 2 つの題材について、表形式データやマトリクス型データを階層型データに変換して「平安京ビュー」で可視化する方法について検討し、その利点について考察する。

なお本報告の内容は、現時点ではあくまでも構想段階であり、まだ実装や実験は完了していないことをご了承いただきたい。

2. 平安京ビュー

「平安京ビュー」[Ito03]は、大規模階層型データを対象とした情報可視化手法である。図 1 に「平安京ビュー」による可視化の例を示す。「平安京ビュー」では、階層型データを構成する葉ノードを正方形のアイコンで表現し、枝ノードを入れ子状に配置された長方形の枠で表現する。

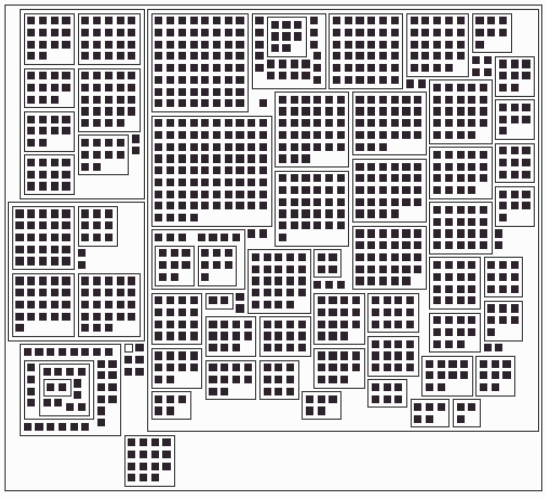

図 1. 「平安京ビュー」による可視化の例。

「平安京ビュー」は数千、数万もの葉ノードを有する階層型データに対し、画面上で互いに重なり合うことなく、できるだけ小さい画面空間に余すことなく、またすべての葉ノードを対等な大きさで表現することができる、

という点において他の階層型データ視覚化手法とは異なる特徴を有する。

「平安京ビュー」に代表される大規模階層型データ視覚化手法は、非常に広い用途への適用が考えられる。実際に伊藤らは、このような階層型データ視覚化手法を、ウェブサイトのアクセス分析[Yam03]、企業組織のコミュニケーション分析[Ito04a]、科学技術計算のパラメータ最適化[Ito04b]、ネットワーク不正侵入の検出[Ito04c]、分散計算環境の負荷分布監視[Ito05]、などの目的に適用している。

3. マトリクス型データから階層型データ・階層型グラフデータへの変換

まず n 個のデータ要素 $\{a_1,...,a_n\}$ を想定し、任意のデータ要素 a_i と a_j の間の距離を d_{ij} とする。このデータ要素を一般的なクラスタリング手法（例えば、階層的クラスタリング、SOM, k-means など）を用いて、d_{ij} の小さいものができるだけ隣接するように並べ替えるとする。

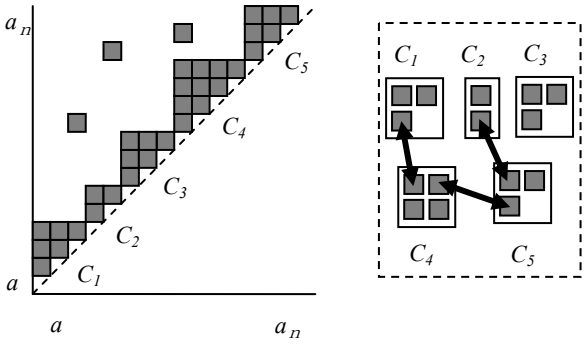

図 2. (左)マトリクス型の可視化の例。(右)マトリクス型データを階層型グラフデータに変換した例。

図 2(左)は、この n 個のデータ要素を、マトリクスの行と列の双方に配置し、マトリクスを構成する各カラムのうち $dij<D$ (D は定数) であるカラムだけを黒く塗った例である。マトリクス型データをそのまま可視化する実用事例の大半では、$dij<D$ であるカラムを発見することを目的としている。

マトリクス型可視化結果が示す画面上で、かなりの面積を占める部分が白いということは、画面有効利用の観点からは好ましい可視化結果であるとは言い切れない。現実問題として、データ要素が数百～数千に及ぶ場合、このようなマトリクス型の可視化を適用すると、一画面にデータの全容を示すことが難しくなる。少なくとも、クリッカブルな GUI としての実用が難しいくらいデータ要素が小さく表示されるか、あるいはズーム操作やスクロールバーを多用しないとデータ全体を可視化できなくなることが多い。しかしそれでも、多くの技術分野において、このようなマトリクス型の可視化手法は普及している。遺伝子ネットワーク分析やテキストマイニングなどの技術分野が、その典型例である。

ここでマトリクス型の可視化結果に、黒く塗られたカラムが、対角線（図 2(左)の点線）に沿って階段状の塊として観察できることに着目する。この階段状のカラムの塊は、クラスタリングの結果として生じるものである。図 2(左)を例にすると、この可視化結果からは C_1～C_5 の合計 5 個のクラスタを観察できる。本報告では、このクラスタリング結果をもとにして、データ要素$\{a_1,...,a_n\}$をグループ化することで、階層型データを構築する。

またマトリクス型の可視化結果に、階段状のカラムの塊とは別に、対角線から遠い位置に黒いカラムが点在することがある。本報告ではクラスタをまたぐカラムを、データ要素 a_i～a_j 間を連結するリンクとして表現する。図 2 の場合、C_1-C_4 間、C_2-C_4 間、C_4-C_5 間にリンクを観察できる。

本報告では、以上の操作によって生成されたリンクつき階層型データを、「平安京ビュー」を用いて可視化することを考える（図 2(右)参照）。

4. 左京と右京：アンケート集計結果の平安京ビューによる可視化

アンケート集計結果は、表形式データとして表現されることの多い典型的なデータである。本報告では以下のようなアンケートを集計した表形式データを想定し、これを「平安京ビュー」を用いて可視化することを考える。

- アンケート中には n 個の質問があり、すべての質問は $s_i(i=1..n)$個の選択肢の中から選択形式で回答するものとする。
- 回答者は m 人で、すべての回答者はすべての質問に回答しているものとする。
- 選択肢の総数 Σs_i を行数、 回答者の総数 m を列数とするマトリクス型データを想定し、各カラムには各々の回答者が対応する選択肢を選択した場合は「真」、さもなければ「偽」が記録される。
- 選択肢および回答者は、一般的なクラスタリング手法を用いて各々クラスタリングされる。よって回答パターンの似ている回答者群が隣接するように列が並べ替えられ、隣接する回答者どうしが選ぶことの多い選択肢群が隣接するように行が並べ替えられる。

以上の前提において著者らは、

- Σs_i 個の選択肢をクラスタリング結果に基づいてグループ化して形成される階層型データ
- m 人の回答者をクラスタリング結果に基づいてグループ化して形成される階層型データ

の 2 データを生成し、2 つの「平安京ビュー」を同時に用いて可視化する技術を開発中である。以下に、その概要を論じる（図 3 参照）。

- 本報告では、この 2 つの平安京ビューを「左京」「右京」と呼ぶ。「左京」は回答者をノードとする階層型データを表示する。「右京」は選択肢をノードとする階層型データを表示する。
- 左京の回答者ノードをクリックすると、その回答者が回答した選択肢ノードが、右京上でハイライトされる。
- 右京の選択肢ノードをクリックすると、その選択肢を選択した回答者ノードが、左京上でハイライトされる。
- 左京のノードの高さと色は、各回答者の回答した選択肢に関わる項目値を表現できるものとする。右京のノードの高さと色は、各選択肢の回答者数を表現できるものとする。

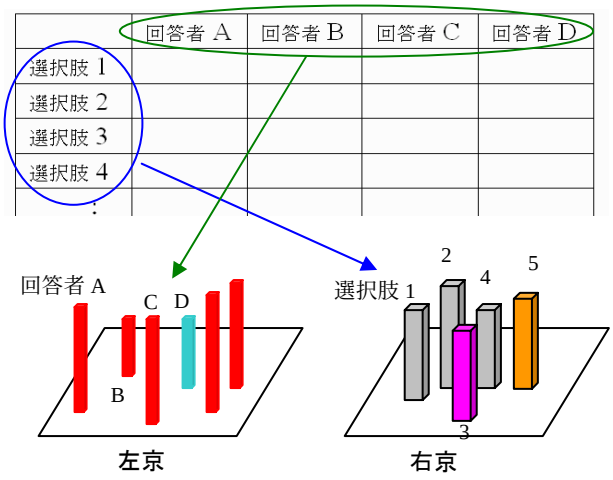

図 3. 表形式アンケート集計結果から「左京と右京」への変換

アンケート集計結果は一般的に、帯グラフや円グラフなどを用いて概略的な統計値を表現するか、あるいは表形式のまま扱うことが多い。しかし帯グラフや円グラフだけでは局所的な傾向を見ることが難しいし、回答者を単位とした表形式のままでは統計傾向を見ることが難しい。本章で示す可視化手法は、「左京」で回答者を単位とした局所的特徴の可視化を実現しながら、同時に「右京」で各選択肢に対する統計傾向を見ることができる、という点に特徴があるといえる。

またクラスタリング結果を反映して生成した階層型データを可視化することで、回答者や選択肢の相関性の高さを視覚的に分析すること、また多数意見だけでなく少数意見にあえて着目した可視化ができること、なども特徴であると考えられる。

5. 遺伝子ネットワークの平安京ビューによる可視化

遺伝子解析結果は、マトリクス型データとして扱われる機会の多い典型的なデータである。特に最近では、マイクロアレイ法を用いた遺伝子解析技術が普及し、多数の遺伝子に関するデータをマトリクス形式で可視化する研究事例やソフトウェアが増えている。

マイクロアレイ法では一般的に、スライドガラスに多数の DNA スポットを生成し、核酸のハイブリッド化という反応を誘引するために種々の操作を行い、その操作ごとに遺伝子の発現率を測定する。このとき数種類の遺伝子に対して、同じ操作に対して発現しやすいという相関性が発見される。また逆に数種類の操作に対して、同じ遺伝子が発現しやすいという相関性が発見される。そこで遺伝子解析では、一般的なクラスタリング手法を用いて、発現性の相関性の高い遺伝子が近接するように遺伝子を並べ替え、その遺伝子群に対して図 2 に示したようなクラスタやリンクを形成することで、遺伝子の特徴を発見する、というような方法をとることが多い。

このようにして形成されたクラスタやリンクの中から、遺伝子解析において意味のある現象として、例えば以下のような現象を観察することができる。

[パラログ] 生物学的に類似した働きを担う遺伝子群。クラスタリングによって同一クラスタ中に発見される。

[オーソログ] 複数の生物の遺伝子を同時に解析した際に、複数の生物にまたがって存在する遺伝子群。クラスタリングによって同一クラスタ中に発見される。

[マルチドメイン] 2 種類のクラスタの発現傾向を同時にもつ遺伝子群。

[中間ノード] 重要な 2 種類の遺伝子の間にリンクを介して存在する遺伝子群。遺伝子発現の複雑な相関性を分析する過程において、中間ノードの追跡は重要であると考えられる。

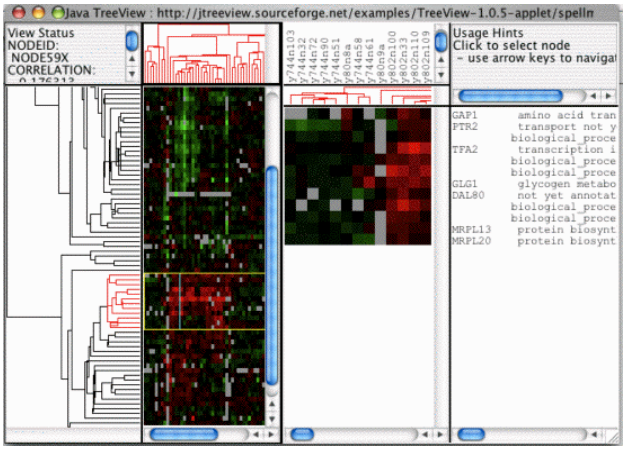

図 4. TreeView による遺伝子解析結果の可視化。

このような遺伝子解析結果の可視化技術に関する代表的な事例について、文献[Sar04]にて詳細な比較がなされている。代表例の一つである TreeView[Tree]の画面表示例を図 4 に示す。画面の左半分は、遺伝子のクラスタリング結果の全体像を、マトリクス形式で可視化する。画面の右半分は、そのマトリクスの一部をズーム表示し、

その該当部分に関する詳細情報を表示する。このマトリクス形式の表示結果の中で、ユーザーが興味をもつ部位は、一般的にはごく一部である。3 章でも述べた通り、このような可視化手法は、必ずしも画面領域を有効に活用しているとは言い切れない。

以上の点を背景として著者らは、

- 遺伝子をクラスタ単位で可視化する、あるいは逆にクラスタを最小単位とした（＝クラスタ内部の詳細構成を隠蔽した）可視化を実現する
- クラスタをまたぐ例外的なリンクや、重要な遺伝子間を介在するリンクなどを強調する可視化を実現する
- クラスタやリンクとみなされない、可視化されなくてもよい情報を、画面表示から排除することで、画面空間を有効活用する。

という観点から「平安京ビュー」を用いた可視化手法を開発中である。以下に、その概要を論じる（図 5 参照）。

- 遺伝子をクラスタリング結果に基づいてグループ化することで形成された階層型データを、「平安京ビュー」を用いて可視化する。遺伝子を表現する各ノードの色や高さを用いて、特定の操作に対する発現率を表現することができる。この可視化は、パラログ、オーソログなどの現象の分析に向いていると考えられる。
- それとは別に、クラスタをまたぐリンクを 3 次元的に表現する。この可視化手法は、特定ノードに関するリンクを 3 次元的に引き上げて表示するという意味で、納豆ビュー[Shi97]に類似した手法と考えられる。この可視化は、マルチドメイン、中間ノードなどの現象の分析に向いていると考えられる。

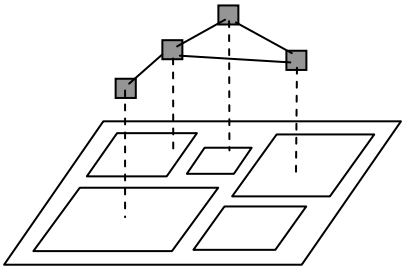

図 5. 遺伝子クラスタの「平安京ビュー」による可視化と、特定ノード周辺のリンクの 3 次元的可視化。

6. まとめ

本報告では、表形式データやマトリクス型データを、階層型データに変換して「平安京ビュー」で表現する試み、およびその具体的な応用例として「アンケート集計結果」と「遺伝子ネットワーク」の可視化を構想していることを述べた。今後は本構想の実装と実験を行い、その有用性について検証したい。

謝辞

遺伝子ネットワークについて貴重なご教示を賜りました東京大学大学院新領域創成科学研究科中谷明弘助教授に感謝の意を表します。

参 考 文 献

[Rao94] Rao R., and Card S. K.., The Table Lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. Computing Systems (CHI'94), pp. 318-322, 1994.

[Tree] TreeView, http://genetics.stanford.edu/~alok/TreeView/

[Gho04] Ghoniem M., Fekete J., Castagiloia P., A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations, *IEEE Information Visualizatoin 2004*, pp. 17-24, 2004.

[Ito03] 伊藤，小山田，平安京ビュー ～ 階層型データを碁盤状に配置する視覚化手法，可視化情報学会第 9 回ビジュアリゼーションカンファレンス, 2003.

[Yam03] 山口，伊藤，池端，梶永,階層型データ視覚化手法「データ宝石箱」とウェブサイトの視覚化,画像電子学会論文誌, Vol. 32, No. 4, pp. 407-417, 2003.

[Ito04a] 伊藤，山口，水田，中村，情報視覚化手法「データ宝石箱」による企業組織コミュニケーションの視覚化，画像電子学会ビジュアルコンピューティングワークショップ, 2004.

[Ito04b] 伊藤，比戸，小山田，酒井，皿井，平安京ビューを用いた細胞シミュレーションのパラメータ最適化 GUI，日本バイオインフォマティクス学会第 4 回システムバイオロジー研究会, 2004.

[Ito04c] 伊藤，高倉，沢田，小山田，ネットワーク不正侵入監視のための視覚化の一手法，情報処理学会第 9 回分散システムインターネット運用技術シンポジウム, pp. 63-68, 2004.

[Ito05] 伊藤，山口，情報視覚化手法「データ宝石箱」のハイパフォーマンス計算技術への応用，日本計算工学会論文誌, Vol. 10, No. 1, pp. 1075-1078, 2005.

[Sar04] An Evaluation of Microarray Visualization Tools for Biological Insight, IEEE Information Visualization 2004, pp. 1-8, 2004.

[Shi97] 塩澤他，「納豆ビュー」の対話的な情報視覚化における位置付け，情報処理学会論文誌, Vol. 38, No. 11, pp. 2331-2342, 1997.