左京と右京：2 つの平安京ビューによる表形式データの可視化

橘春帆*　伊藤貴之**
(*) お茶の水女子大学大学院　人間文化研究科
(**) お茶の水女子大学　理学部　情報科学科
{haruho, itot}@itolab.is.ocha.ac.jp

1. 概要

情報可視化は，世の中にある一般的な情報を可視化する研究分野であり，最近になって活発に研究が進められている．我々は大規模階層型データを一画面に可視化する手法「平安京ビュー」[1]を提案しており，これを多くの用途に適用した事例を報告している．

一方で，情報科学の多くの分野では，表形式で格納されているデータが非常に多く存在する．表形式データの可視化手法として我々は,「平安京ビュー」を一画面に二つ用いて可視化する手法「左京と右京」[2]を提案している.「左京と右京」では表形式データに対して，行を構成するデータ要素，列を構成するデータ要素，の各々についてクラスタリングを行う．続いて，行を構成するデータ要素で構成される階層型データに対して,「平安京ビュー」を適用して可視化する．同様に，列を構成するデータ要素で構成される階層型データを,「平安京ビュー」を適用して可視化する．図 1 に示すように「左京と右京」では，この 2 つの可視化画面を相互に操作できるような仕組みを提供することで，大規模な表形式データの内容を探索する新しい可視化手法を実現する．

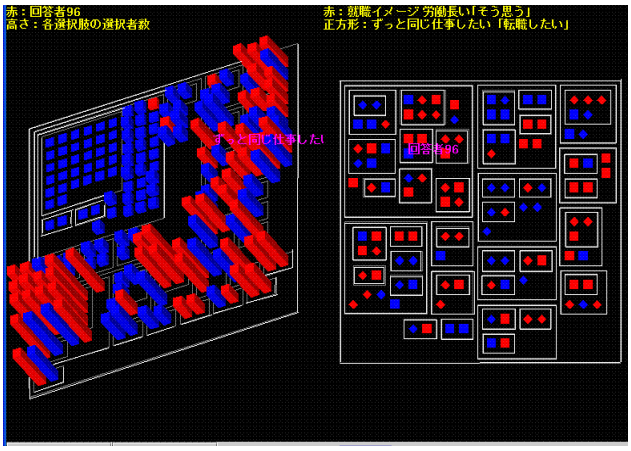

図 1　「平安京ビュー」を 2 つ用いた可視化手法「左京と右京」

情報可視化が有効に活用できると考えられるデータの一つに，アンケートの集計結果があげられる．アンケート集計結果は，回答者数だけで列を持ち，質問に対する選択肢数だけ行を持つ表形式データの一種と考えることができる．我々は「左京と右京」を，アンケート集計結果の可視化に適用している．この適用事例において，右京は回答者数だけのノードを有する階層型データ，左京は選択肢数だけのノードを有する階層型データを表示する．これによって例えば，ある特定の選択肢を選んだ回答者群の分布や，特定の回答者が選んだ選択肢の分布，などを可視化することが可能になる．これにより少数だけど興味深い局所的傾向の発見ができると考えられる．

本報告では，まず 2 章にて「左京と右京」の関連研究を紹介し，3 章にて「左京と右京」について紹介する．続いて 4,5 章にて「左京と右京」をアンケート集計結果の可視化に適用した事例を紹介する．最後に 6 章にて，「左京と右京」を MuST コーパスに適用し，文章とキーワードの共起性などの可視化を行う展望について論じる．

2. 関連研究

「左京と右京」の先行研究である情報可視化手法「平安京ビュー」[1]は，「平安京ビュー」は，大規模な階層型データ全体を一画面に配置する情報可視化手法である．葉ノード，枝ノードの直交配置を意識したアルゴリズムを採用しており，平安京の地図に似て見えることから命名された．

階層型データを可視化する手法には，木構造を表示する手法[3][4]や個々のノードが持つ占有率に比例した面積で画面を分割した手法[5]がある．

これらの手法に比べて，平安京ビューは階層型データを構成するすべての葉ノードを一画面に,同じ大きさで,重なることなく表現できるという特徴がある．
この特徴はアンケート情報を構成するすべての回答者と選択肢を一画面に表示したい，という本研究の目的に向いているといえる．また葉ノードの個数に比例した面積で画面を分割し，葉ノードを平等に一画面に表示する手法[6]もある．この手法は平安京ビューに近い特徴を持つといえる．

表形式データの情報可視化手法として有名なものに,TableLens[7]があげられる．この手法は，表形式データを表のまま表示し，利用者が凝視したい部分だけを対話的にズームアップできるようなインタフェースを備えることで，表形式データの対話的な探索ができるツールを提供している．

一方で,表形式データをグラフデータに変換して可視化する手法も，旧来からすでに知られている[8]．このような考え方は，疎な表形式データにおいて有効な手法であると考えられる．

3. 左京と右京

「左京と右京」は，平安京ビューを応用した新しい試

みであり，表形式データから階層的クラスタリングの結果として得られる階層型データを可視化するものである．

「左京と右京」では前処理として，列を構成するデータ要素が m 個，行を構成するデータ要素が n 個である n 行 m 列の表形式データに対して，クラスタリングを行う（図 2 参照）．列を構成する n 次元ベクトルを持つ m 個のデータ要素と考え，ベクトル間距離の近いものをボトムアップ的にクラスタリングし，結果として数段階の階層を持つ階層型データを形成する．同様に，行を構成するm次元ベクトルを持つ n 個のデータ要素に対してもクラスタリングし，階層型データを形成する．このようにして結果として 2 つの階層型データを生成する．

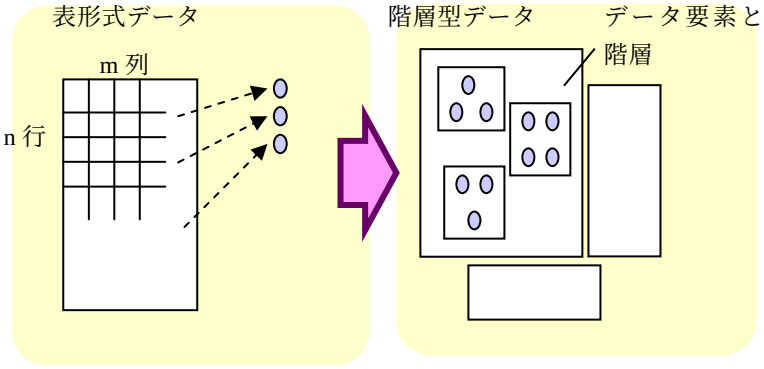

図 2　表形式データに対する階層型クラスタリング

続いて「左京と右京」では，クラスタリングして得られた階層型データについて階層型可視化手法「平安京ビュー」を適用する．n 行に対して 1 つの平安京ビュー，m 列に対しても 1 つの平安京ビューを出し，ひとつのディスプレー上に 2 つの平安京ビューを立ち上げて，その 2 つが相互に操作できることが本研究「左京と右京」の新しい点である．つまり，右京でクリックしたアイコンに反応して，色や高さの変化が左京側のアイコンに現れる．同様に左京のクリックした場合も，右京側のアイコンに変化が現れる．

4．アンケート集計結果データによる実験

この章では，アンケート集計結果を表形式データにして可視化した実験例を示す．私たちの実験は，n 個の選択肢で構成されるアンケートを，m 人の回答者に答えてもらう．図 3 で示すように，私たちのアンケートは，すべての問題が 1 つの項目を選択することによって答えられるものであり，自由回答を認めていない．だたし 1 つの項目に対して複数の選択肢を選ぶことはできる．

本実験では，アンケートの回答の集計結果から表データを作成し，クラスタリングによって「右京」および「左京」で表示する階層型データを作成した．クラスタリングには Cluster3.0[8]というソフトウェアを使用している．

アンケートの内容は「女子大生の就職に関する意識調査」を準備した． 回答者は，お茶の水女子大学理学部情報科学科に属する，1 年生から大学院修士 2 年生の学生 118 名である．アンケートは 31 の質問と 179 の選択肢を含んでいる．したがって，回答の集計結果データから作成された表データは 118 の列データと 179 の行データを含む． この章では，「左京と右京」によるアンケート集計結果データの可視化結果を見せる．アンケート集計結果の実験では，右京は 118 人の回答者を表示する．そして，左京は 179 の選択肢を表示する． 左京では，クラスタリングによって関連性の高い選択肢が組にされる．同様に，右京では似ている回答をする回答者が組になって表示される．

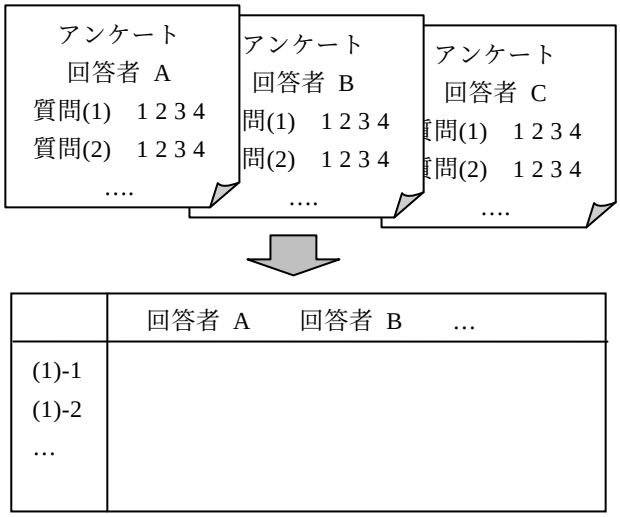

図 3　アンケートと表データ

「左京と右京」では，注目している特定の回答者，特定の選択肢に対する対話的な詳細表示をする．右京の回答者ノードをクリックすると，その人が回答した選択肢ノードが左京にて反応する．逆に，左京の選択肢ノードをクリックすると，その選択肢を回答した回答者ノードが右京にて反応する．

この可視化方法では，クラスタリングした結果として得られるグループ内の相関性の真偽，強弱などを知るのに有効と考えられる．またアンケート集計結果をクラスタリングした結果の大人数のグループだけでなく少人数のグループについてもその回答の詳細を可視化するのが容易になる．

我々はアンケート集計結果の表示のために，以下のように「左京と右京」をカスタマイズした．

- 左京のアイコンの高さは，対応する選択肢を選択した回答者の人数を示す．
- カーソルが選択肢のアイコンの 1 つを指すとき,左京では対応する選択肢の意味を示す．
- ユーザが左京である選択肢のアイコンをクリックすると，右京では選択された選択肢を回答した回答者のアイコンは赤で表示され，他のアイコンが青で表示される．
- ユーザが左京で別の選択肢アイコンをクリックする

と，右京ではその選択肢に回答した回答者のアイコンをひし形として表示する．

・ カーソルが回答者のアイコンの 1 つを指すとき，右京は対応する回答者の名前を示す．

・ ユーザが右京で回答者のアイコンをクリックすると，左京は回答者が回答した選択肢を赤で，他の選択肢を青で表示する．

5. 実験結果

図 4 は右京の実行結果を示す．図 4 では，アイコンの形状は「あなたは，就職活動のイメージで精神的にきついと感じますか？」という質問に対する答えを示す．(ここでは，正方形のアイコンは「はい」と答えた回答者，ひし形のアイコンは「いいえ」と答えた回答者を示す．) 図 4（上側）の結果では，赤いアイコンは学年が一年生の回答者を指示す．そして，青いアイコンは他の学年の回答者を示す．ここでは，一年生には 5 人「いいえ」を選択した学生がいることがわかる．この 5 人の分布状態をみてみると，バラけて表示されていることがわかる．この結果は，「いいえ」と答えた回答者が同じクラスター内にかたまって表示されてない，あるいは，近い位置に密集して表示されていないので似ている性質の 5 人ではないことが分かる． 図 4 (下側) の結果では，赤いアイコンは 3 年生の回答者を示す．この結果をみてみると，3 年生で「いいえ」と答えた学生が全くいないことが分かる．このことから 3 年生のほうが，実際に，自分が就職活動をしている，あるいは，まわりの友人が就職活動をしている姿をみているので 1 年生より就職活動を厳しいものだというイメージが強いことが予想される．

図 5 は，「右京」のもうひとつの実行結果である．この結果では，「現在，親と同居していますか？」という質問に対して，「はい」と答えた人は赤，「いいえ」と答えた人は青で示されている．この結果の色の分布状況をみてみると比較的，赤いアイコンは赤いアイコン同士，青いアイコンは青いアイコン同士でクラスターを形成していることがわかる．よってこの表示結果が，クラスタリング結果に強く寄与していることが分かる．またアイコンの形状は，「就職する際，親の意見を重要視しますか？」という質問に対して，「はい」と答えた人は正方形，「いいえ」と答えた人は，ひし形で示されている．この結果もひし形と正方形の分布状況をみてみると，表示結果が比較的クラスタリング結果に強く寄与していることが分かる．また意外だったのが，この 2 つの質問内容からして，高い相関性を考えていたが，丸で囲った部分を見ると，赤と青のひし形が混在してひとつのクラスターを形成している．このことからこの 2 つの質問の相関性はあまり高くないことが発見できた．

図 6 は「左京」の実行結果を示す．この図ではひとつひとつのアイコンは選択肢で，高さは各選択肢の回答者の人数である．よって高さの低いアイコンは少数意見だと分かる．ある回答者が答えた選択肢だけが赤色で示されている．この結果の丸く囲った部分を見ると，2 つの赤いノードが 1 つのクラスターを構成している．この 2 つの選択肢の内容は，

- 就職を見通したアルバイトをしている
- 就職活動のイメージが楽しい

であった．

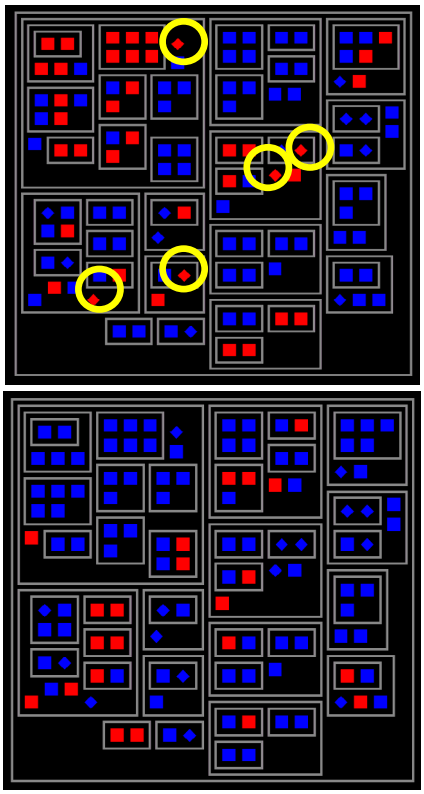

図 4 「右京」による可視化結果(1)． 回答者の学年と就職活動のイメージ間の関係． (上) 赤は 1 年生の回答者を示す．(下) 赤は 3 年生の回答者を示す．

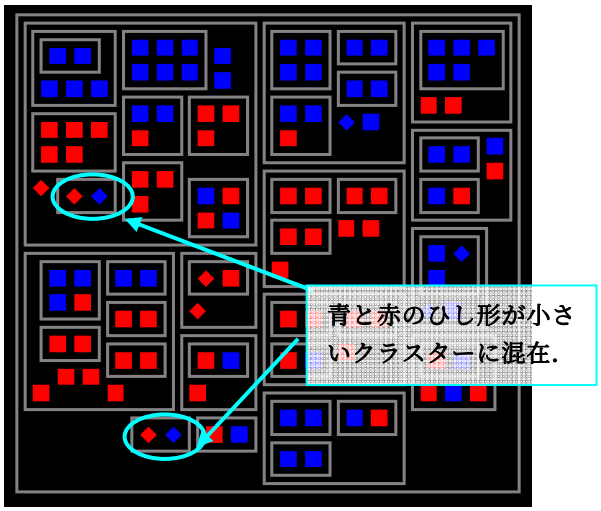

図 5「右京」による可視化結果(2)．回答者の就職活動と親の関係を示す．

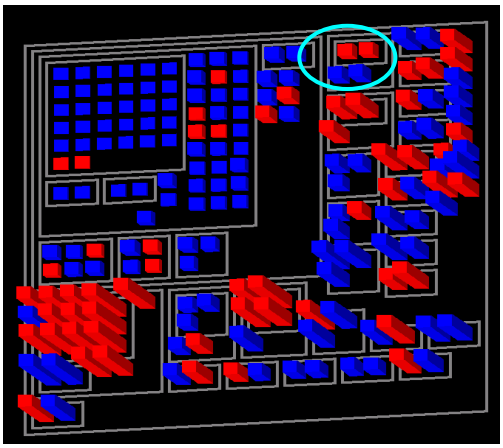

図 6「左京」による可視化結果．回答者のアルバイトと就職動活動に関する彼らの印象との関係を示す．

このクラスター構成から「実際に就職を見通したバイトをしている人は，目的意識が高く，就職活動を楽しんでいる」ということが考えられる．このようにして，少数かつ興味深いクラスターに属する選択肢を 2 つとも選択している回答者をすぐに発見することができた．

6. MuST コーパスの新聞記事を「左京と右京」に適用させる構想例

我々は今後の課題として，MuST コーパスのタグ付き記事を「左京と右京」に適用して，文章とキーワードの共起性などに関する可視化を試みたいと考えている．

現在では単純な実験として，MuST コーパスの中のピックアップした一年間のすべての新聞記事から見出しタグ(headline)の部分に出てくるキーワードを全て抽出し，「左京と右京」の片方に抽出したキーワードを表示し，もう片方に見出しを表示する，という実験の準備をしている．その結果，いかにも関係がありそうでありながら，共起性の低い単語と想像されるキーワードなど，意外な共起関係の発見ができるようにしたいと考えている．

他の例としては，同じトピックの新聞記事（例えば，地震情報を扱っている記事群，ガソリン価格を扱う記事群）の単語を抽出し，その単語を 2 種類の基準でクラスタリングして，それぞれ 2 つの結果に「左京と右京」を適用させるというものを考えている．この方法によって，同じ内容の文章群からクラスタリングの手段を変えることで，記事中の単語の位置づけについて興味深い発見できないかと考えている．

その他にも多くの実験方法が考えられるが，例えば以下の点で我々の専門性から離れた問題点があり，文書要約の専門家の意見を聞きながら実験を進めたいと考えている．

- 「左京と右京」による可視化に適切なサイズのデータを構築するために，MuST コーパスからどのように文章をサンプル抽出するか．また，膨大な新聞記事に潜む興味深い共起関係などの特徴を失わずに，一部の文章を抽出するには，どうすればよいか．
- どのようなクラスタリング方法を採用すれば，文書や単語に関する興味深い特徴が見えてくるか．

謝辞

生活科学系アンケートの採取方法に関する情報提供をしてくださった本大学生活科学部の御船美智子教授，永瀬伸子助教授，畑江敬子教授，香西みどり助教授,Cluster3.0 を開発された Michael de Hoon 氏，テキスト分析について情報提供してくださった本大学情報科学科の小林一郎助教授，アンケート作成時にアドバイスしていただきました本大学人間社会科学科小園絢子氏，アンケートの回答にご協力していただいた本大学情報科学科の皆様に感謝の意を表します．なお本研究の一部は，日本学術振興会科学研究費補助金の助成に関するものです．

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] 橘，伊藤，左京と右京：大規模表形式データの可視化の一手法，情報処理学会データベースと Web 情報システムに関するシンポジウム(DBWeb2006), 2006.

[3] Carriere J., et al., Research Paper: Interacting with Huge Hierarchies beyond Cone Trees, IEEE Information Visualization 95, pp.74-81, 1995.

[4] Lamping J., 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.

[5] Johnson B et al., Tree-Maps: A Space Filling Approach to the Visualization of Hierarchical Information Space, IEEE Visualization '91, pp. 275-282, 1991.

[6] Bederson B., Schneiderman B.,Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies, ACM Transactions on Graphics, Vol. 21, No. 4, pp. 833-854, 2002.

[7] Rao R., 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.

[8] Becker R. A. Eick S. G., Wilks A. R., Visualizing Network Data, IEEE Transactions on Visualization and Computer Graphics, Vol. 1, No.1, pp. 16-28, 1995.

[9] Cluster 3.0, http://bonsai.ims.u-tokyo.ac.jp/^mdehoon/software/cluster/software.htm#ctv