芸術科学会論文誌 Vol. 7 No. 2 pp. 22 – 33

左京と右京：大規模表形式データの可視化の一手法

橘春帆　伊藤貴之

お茶の水女子大学大学院

Sakyo & Ukyo: A technique for visualization of large-scale table data

Haruho Tachibana　Takayuki Itoh

Graduate School of Humanities and Sciences, Ochanomizu University

{haruho, itot} @ itolab.is.ocha.ac.jp

概要

クラスタリングされた表形式データは，情報科学の多くの場面において重要な役割を果たしている．このような表形式データの内容を効果的に可視化する手法として，表形式データを階層構造やグラフ構造に変換した上で可視化する手法が，近年になって多数報告されている．

本論文では，階層型データ可視化手法「平安京ビュー」を使った表形式データの可視化手法を提案する．提案手法ではまず，表形式データの行および列を構成するデータ要素に，クラスタリングを適用し，各々の結果から階層型データを構築する．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に対しても,「平安京ビュー」を適用して可視化する．この 2 つの可視化結果を相互に操作することで，大規模な表形式データの内容を探索する新しい可視化を実現する．

著者らは実験例として，新聞記事コーパスから作成された表形式データの可視化を試みた．本論文では,「左京と右京」によって新聞記事コーパスから，興味深いいくつかのキーワード群や記事群を発見できた事例を紹介する．

Abstract

Clustered table or matrix data is often very important in various computer science fields. Many of recently reported vsiualization techniques for the table data convert them into hierarchical or graph data.

This paper presents ”Sakyo & Ukyo”, a visualization technique for table or matrix data, applying a hierarchical data visualization technique ”HeiankyoView”. The technique first applies clustering for rows and columns of table data independently, and constructs two hierarchical data from the clustering results. It then independently applies ”HeiankyoView” to the two hierarchical data. It realizes a new visualization schema by providing a mechanism to interact the two hierarchical data each other.

We have applied ”Sakyo & Ukyo” to the visualization of table data constructed from the corpus of newspaper articles. This paper introduces that we discovered intersting clusters of keywords and documents by ”Sakyo & Ukyo”.

1 はじめに

情報技術の普及に伴い，計算機システムのデータベースには非常に多種多様な，かつ膨大な情報が蓄積されている．その情報の中には，表形式データで表現できるものが数多く存在する．例えばテキスト分析の分野では，キーワードとテキスト文書をそれぞれ行と列に配置した表形式データが，しばしば用いられる．例えば顧客管理（CRM）の分野では，顧客と商品をそれぞれ行と列に配置した表形式データが，しばしば用いられる．このような多種多様な表形式データの特徴や傾向を，数理的にというより直感的に理解する一手段として，コンピュータグラフィックスを応用した情報可視化技術の研究が進んでいる．

表形式データの情報可視化の対象として，最近特に注目されているデータに，遺伝子のマイクロアレイデータがあげられる．マイクロアレイデータは，遺伝子の種類を行に，マイクロアレイ（ガラスやシリコン製の小基盤上に DNA 分子を高密度に配置したもの）を列に，それぞれ配置することで作成される表形式データである．この表形式データは，発現率傾向がよく似ている遺伝子同士をまとめるクラスタリング処理を施して分析されることが多い．マイクロアレイデータはその大規模さと乱雑さゆえに，情報可視化手法への期待も高く，すでに多くの手法が発表されている [1]．

表形式データの最も単純な表示手段は，表をそのまま表として表示することである．しかし，上述のような表形式データの中には，情報が非常に大規模かつ疎であるものも多い．そのため，このようなデータをそのまま表として表示することは，画面空間の有効利用の点で必ずしも合理的であるとは限らない．これを改善する一案として，表形式データを木構造データやグラフデータに変換して表示する試みが多く行われている．本論文の提案手法も，表形式データを木構造データに変換して表示する手法の一種であると考えられる．ここで，表形式データを表のまま可視化する手法と，木構造やグラフに変換して可視化する手法には，一長一短の関係がある．そこで近年では，両手法の比較に関する研究も発表されている [2]．

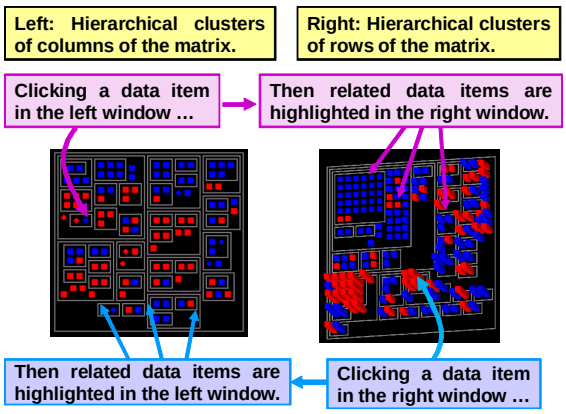

図 1: 「左京と右京」の概観．

本論文では，大規模な表形式データの新しい可視化手法「左京と右京」を提案する．提案手法の概観を図 1 に，そしてフローチャートを図 2 に示す．提案手法ではまず，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適用する．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，階層型データ可視化手法「平安京ビュー」[3, 4] 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に対しても,「平安京ビュー」を適用して可視化する．この 2 つの可視化結果を相互に操作することで，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．

著者らは実験例として，新聞記事コーパスから作成された表形式データの可視化を試みた．本論文では，著者らによる新聞記事コーパスからの表形式データの作成方法，およびその新聞記事から抽出したキーワードと記事群に潜む面白い傾向を紹介する．さらに，いくつかの評価実験結果と考察により，新聞記事コーパスの可視化における提案手法の有用性を検証する．

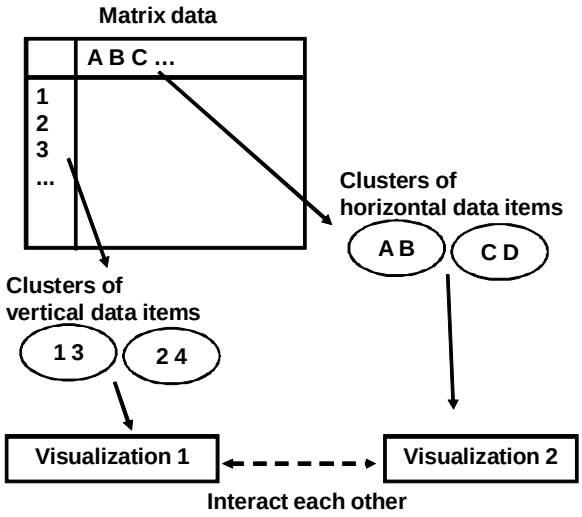

図 2: 「左京と右京」のフローチャート．

なお，提案手法「左京と右京」の基本的概念は，すでに著者自身による口頭講演にて発表されている [5, 6]．本論文はその実装等を詳細化するとともに，新しい実験例として新聞記事コーパスの可視化を紹介するものである．

2 関連研究

2.1 表形式データの可視化手法

表形式データの情報可視化手法として有名なものに，Table Lens [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，旧来から既に知られている [8]．本論文の提案手法は，この一種であると考えられる．

2.2 階層型データの可視化手法

階層型データを可視化する手法は既に多く報告されているが，主に以下のような手法が知られている．

- 画面空間の再帰分割による手法．代表的な手法として Treemaps [9] や，このバリエーションとして Quantum Treemap [10] があげられる．
- 木構造を描画する手法．代表的手法として，Cone Tree [11], Hyperbolic Tree [13], Fractal Tree [12] があげられる．
- 3 次元空間で入れ子状に配置された直方体や球状物体により階層構造を表現し，これを半透明表示する手法． 代表的な手法として Information Cube [14] があげられる．
- 2 次元空間で入れ子状に階層構造を構築する手法.「データ宝石箱」[15] およびその改良手法である「平安京ビュー」[3, 4] が該当する．

本論文の提案手法は，階層型データの親子関係を表示することよりも，階層構造の最下位に属するデータ要素を限られた画面空間に一覧表示することを目的としている．上述の階層型データ可視化手法のうち，Treemaps, Quantum Treemap, データ宝石箱，平安京ビューの各手法は,「最下位階層に属するデータ要素を表現する図形を，できるだけ隙間無く，しかも重なることなく画面に配置する」という点で，この目的に合致した手法である．このような手法は，各データ要素をクリッカブルな状態で表示できることから，一種のユーザインタフェースとしても活用しやすい手法であるといえる．

なお平安京ビューは文献 [4] にて，長方形領域の縦横比，類似データ間における画面配置結果の類似度，などの数値評価結果において Quantum Treemap やデータ宝石箱よりも良好な結果が得られたことが実証されている．

2.3 多変数データの可視化手法

本論文の提案手法は，行と列からなる表形式データを多変数データとして扱い，そのクラスタリング結果を可視化している．よって多変数データのクラスタリング結果の可視化手法も，本論文の提案手法と関連があると考えられる．

多変数データの可視化手法として有名な Parallel coordinates の拡張手法として，多変数データのクラスタリング結果の可視化を試みた手法 [16, 17] がいくつか報告されている．また，主成分分析に基づいて多変数データを平面に投影し，そのクラスタを表現する手法 [18, 19] も報告されている．しかしこれらの手法では，多くのデータ要素が画面空間上でお互い重なり合って表示されてしまうという点から，ユーザが任意のデータ要素をクリックすることが困難だと思われる．著者らは多変数データの任意のデータ要素をクリック可能な形で可視化することで，情報を対話的に探索するユーザインタフェースの一種を実現したいという発想から，これらの手法を採用していない．

多変数データやクラスタデータの探索や視覚的分析のために，上で述べたようなさまざまな手法を統合した，複合的な可視化システムも存在する [18, 20, 21]．本論文の提案手法は，多変数データやクラスタデータではなく，表形式データにおいてこれらと同様な探索や視覚的分析を実現する一手法である，と位置づけられる．

2.4 テキスト文書の可視化手法

テキスト文書は，情報可視化で扱う典型的なデータのひとつである．既に提案されているテキスト文書可視化手法のいくつかは，時系列変化のパターン発見やテキスト文書の動向情報など，テキスト文書の内容上の時間的変化を表現すること目標としている．代表的な可視化手法として Theme River[22] は，特定のキーワードの頻度情報の時系列変化を川の流れのように表現することで知られている．この手法は，主なパターンや動向情報の時間的変化の可視化分析や発見に用いられる．また Wong ら [23] は，テキスト文書の時系列データセットから常習的な連続しているパターンを可視化する手法を提案している．

これらの手法と違って，他の多くのテキスト文書可視化手法は，本論文の提案手法と同様に，大量のテキスト文書の共起関係に着目した可視化を実現している．代表的な可視化手法として Galaxies[24] は，テキスト文書の多次元特徴を 2 次元の散布図として表現している．この拡張手法として InfoSKY[25] は，階層構造化された 2 次元空間にテキスト文書の散布図を表現している．これらの手法は，テキスト文書同士やテキスト文書のクラスタ同士の距離を表現するのに効果的である．しかしこれらの可視化結果では，プロットした結果に粗密が発生してしまい，プロットどうしが重なり合って表示されることが避けられない．著者らは，テキスト文書のメタファが簡単にクリックできるような可視化手法を確立することで，大規模なテキスト文書群を自在に探索するユーザインタフェースを構築したいと考えている．このような発想から著者らは，Galaxies や InfoSKY のような散布図ベースの手法を採用していない．

DualNAVI[26] は，著者らの手法に最も類似しているテキスト文書可視化手法の 1 つである．この手法は画面を 2 つに分割し，左側にテキスト文書の一覧，右側にキーワード間の関連性を表現するためのグラフを表示する．そして，左側と右側が相互に画面上で連動操作できるようになっている．しかし，スクロールなどの操作を使わずに，何百ものキーワードを一画面に見せることは困難である，という欠点がある．

3 平安京ビュー: 大規模階層型データ可視化の一手法

本論文の提案手法では，大規模階層型データ可視化手法「平安京ビュー」を使用する．図 3 は,「平安京ビュー」による階層型データの可視化結果の例である．この可視化結果において，階層型データ中の葉ノードは黒いアイコンで，そして枝ノードは長方形の枠で表示されている.「平安京ビュー」は，その可視化結果における葉ノードの格子状の配列が，まるで平安京の地図のように整然としていることから命名された手法である．

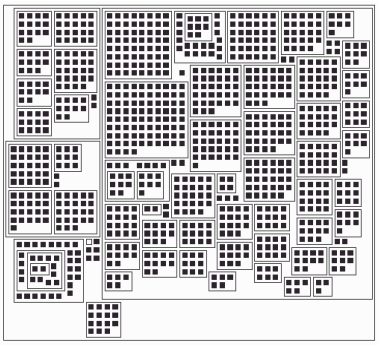

図 3: 「平安京ビュー」による階層型データの可視化の例.

図 3 からもわかるように「平安京ビュー」は，階層構造を二次元の長方形群の入れ子構造で表現し，その全体を一画面に表示することを目標とした手法である．この手法は，階層型データ中の葉ノードと枝ノードの親子関係よりも，階層型データ全体に分布する葉ノード群を全て一画面に表現することに主眼をおいた視覚化手法である.「平安京ビュー」で技術的に重要な点は，枝ノード群を表現する任意の大きさ・形状の長方形群を，限られた大きさの画面空間に有効に配置できるという点である．言い換えれば「平安京ビュー」は，計算機上の限られた大きさのウィンドウやディスプレイに，できるだけ大規模な情報を詰め込んで表現する技術，ということができる．この要件を実現するために「平安京ビュー」では，以下の条件をできるだけ満たすように，大量の情報を限られた画面空間に配置する．

- 葉ノードや枝ノードが画面上で互いに重ならないように配置する．
- データ全体の画面占有面積ができるだけ小さくなるように配置する．
- 全ての葉ノードが同じ大きさで表示されるように配置する．

なお，これらの条件を満たすような画面配置の処理手順については，文献 [3, 4] に詳しく解説されているので，参照していただきたい．

4 左京と右京: 平安京ビューを応用した表形式データの可視化

本章では，平安京ビューを応用して表形式データを可視化する新しい手法「左京と右京」を提案する．提案手法では，図 2 に示した処理手順によって，表形式データを 2 つの階層型データに変換する．そして図 2 中の「Visualization 1」と「Visualization 2」の部分に「平安京ビュー」を適用し，2 つの階層型データを可視化する．

図 1 にて概観を示した通り,「左京と右京」では,「平安京ビュー」を用いた 2 つの可視化結果を，画面上で左右に並べて表示する．このとき，この 2 つの可視化結果は，相互に操作可能な状態で表示される．左側の可視化結果で特定のデータ要素をクリックすると，右側の可視化結果の対応する部位が色を変えて表示される．同様に，右側の可視化結果で特定のデータ要素をクリックすると，左側の可視化結果の対応する部位が色を変えて表示される．このような対話的操作機能により「左京と右京」は，表形式データを探索するための新しい可視化技術を実現する．

本論文では，図 1 の右側の「平安京ビュー」を「左京」と呼び，図 1 の左側の「平安京ビュー」を「右京」と呼ぶ．このネーミングは，北を上にして描かれた平安京の地図において，右側に「左京」があり，左側に「右京」があることに由来している．

4.1 表形式データのクラスタリング

本論文における表形式データの定義を，以下の通り記述する（図 4 参照）．まず，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，である表形式データを仮定する．また，この表形式データの各欄に格納されている値を，a_{11} 〜 a_{nm} で表す．以下，列を構成する m 個のデータ要素のうち i 番目のデータ要素を，n 次元ベクトル $c_i = (a_{1i}, \cdots, a_{ni})$ で表現する．同様に，行を構成する n 個のデータ要素のうち j 番目のデータ要素を，m 次元ベクトル $r_j = (a_{j1}, \cdots, a_{jm})$ で表現する．

続いて提案手法では，各々のデータ要素ペアについて余弦値を算出し，これをデータ要素ペアの類似度値とする．さらに，類似度値の高いデータ要素どうしが同一のクラスタに属するように，クラスタリングを実行する．提案手法におけるクラスタリングには，階層型クラスタリング法でも，あるいは自己組織マップ法や k-means 法などの非階層型クラスタリング法でも適用可能であるが，著者らは階層型クラスタリング法を用いている．階層型クラスタリング法では，データ要素間，あるいはデータ要素クラスタ間の類似度の大きい順に，これらを併合する処理を反復することで，データ要素を階層的にクラスタ化する．また，類似度に対して複数の閾値を設定し，閾値を基準にしてクラスタを再生成することにより，任意の段階数を有する階層型データを構築することもできる．

図 5 は，階層型クラスタリング，およびその結果の「平安京ビュー」による可視化について説明したもの

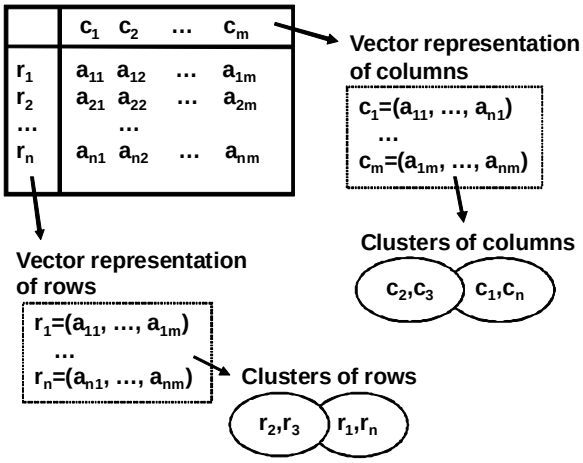

図 4: 表形式データの定義とクラスタリング.

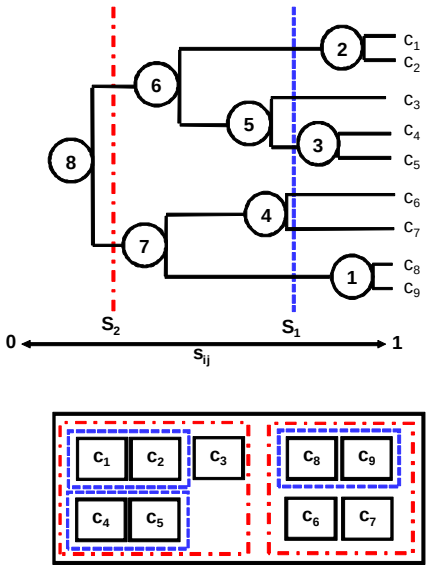

図 5: (上) 階層型クラスタリング. (下) クラスタリング結果から得られる階層型データの可視化結果のイラスト.

である．図 5(上) は，階層型クラスタリングが，s_{ij} が最大値を持つデータ要素（またはデータ要素クラスタ）を 1 組ずつ順に併合していることを示している．図 5(上) にて，丸で囲まれた整数値は，データ要素（またはデータ要素クラスタ）を併合する順番を意味する．また，その整数値の水平方向の位置は，s_{ij} の値の大きさを表している．

また，図 5(上) の S_1 と S_2 は，階層型データを構築するための閾値の例である．筆者らの実装ではまず，s_{ij} 値が S_1 より高いデータ要素で構成されるクラスタ (図 5(上) にて 1〜3 で表される併合処理によるクラスタ) を生成する．続いて，s_{ij} 値が S_2 より高いデータ要素で構成されるクラスタ (図 5(上) にて 4〜7 で表される併合処理によるクラスタ) を生成する．図 5(下) のイラストは，2 個の閾値を適用することによって，2 段階の階層型データを生成していることを示している．

なお提案手法に適用するデータは，データ要素間の類似度値を計算できるデータでなければならない．よって現時点の提案手法では，表の中に算術演算ができない情報（文字列など）を含む表形式データは対象としない．また，各欄の値に乱雑性が高い表形式データは，クラスタリングしても意味のある可視化結果が得られないことから，提案手法の適用に向かない．

4.2 平安京ビューによるクラスタリング結果の可視化

提案手法では，前節に示した手法で生成された 2 つの階層型データ各々に対して,「平安京ビュー」を適用して可視化を行う．以下の説明では,「左京」は行を構成する n 個のデータ要素 r_1〜r_n を可視化するものとする．同様に「右京」は列を構成する m 個のデータ要素 c_1〜c_m を可視化するものとする．

「左京」および「右京」は，これらのデータ要素を角柱で表示するものとする．この角柱を 3 次元的に斜め方向から見た画像を生成することで，これらの角柱は 3 次元空間中に立つ棒グラフのように表現される．

「左京と右京」では以下のような設計思想により，角柱を表示するための視覚的特性（色相，高さ，底面形状）を制御する．まず各々の角柱ごとに，0〜1 の範囲で正規化された 3 個の実数値 b_1〜b_3 を用意する．さらに，以下の 3 種類の関数を用意する．

- b_1 を引数として，色相を返す関数 $f_1(b_1)$.
- b_2 を引数として，高さを返す関数 $f_2(b_2)$.
- b_3 を引数として，底面形状を返す関数 $f_3(b_3)$.

「左京と右京」では，この関数の返り値を参照して，角柱を表示するための視覚的特性を決定する．なお b_1〜b_3 の算出式，およびこれらを引数とする関数 f_1〜f_3 は，適用事例ごとにカスタマイズできるものとする．その一例として本論文の 5 章では，新聞記事コーパスの可視化のためのカスタマイズについて論じる．

4.3 2 個の平安京ビュー間の操作

提案手法では，利用者が対話的に表形式データを探索できるように,「左京」と「右京」が相互に操作可能な機能をもつ．例えば利用者が「左京」の角柱をクリックすると，この角柱が表すデータ要素に対応する「右京」中の角柱が，色や形などを変えて表現される．同様に，利用者が「右京」の角柱をクリックすると，この角柱が表すデータ要素に対応する「左京」中の角柱が，色や形などを変えて表現される．

ここで，利用者が「左京」の角柱 r_i をクリックすると仮定する．このとき提案手法は，a_{i1} から a_{im} の値を探索し，値 a_{ij} を用いて「右京」のデータ要素 c_j

に対する実数値 b_1〜b_3 を算出する．これらを関数 f_1〜f_3 に代入することにより，提案手法では「右京」を構成する棒グラフの色，高さ，底面形状を更新する．以上の処理の流れを，図 6 に示す．

なお，a_{ij} から b_1〜b_3 を算出する手段は，適用事例ごとにカスタマイズできるものとする．その一例として本論文の 5 章では，新聞記事コーパスの可視化のためのカスタマイズについて論じる．

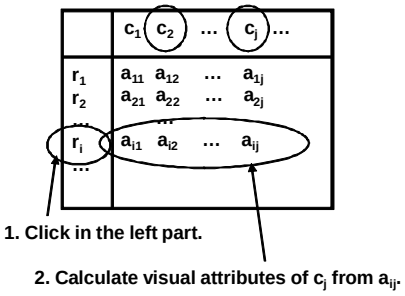

図 6: 左京のクリックにより右京の表示を更新する内部処理手順．

5 新聞記事コーパス可視化のためのカスタマイズ

著者らは新聞記事コーパスの可視化のために，「左京と右京」をカスタマイズしている．本章では，新聞記事の記事文書のデータ要素を r_1〜r_n（n は記事文書数）とする．同様に，新聞記事のキーワードのデータ要素を c_1〜c_m（m はキーワード数）とする．

著者らの実装では，まず 4.1 節で述べた手順に沿って，キーワードと記事文書のクラスタリングを行う．ここで a_{ij} は，i 番目の記事文書の j 番目のキーワードの重要度を示す．著者らの実装では，「右京」（ウィンドウの左側）にクラスタリングされたキーワードを表示し，「左京」（ウィンドウの右側）にクラスタリングされた記事文書を表示する．

図 7 のように新聞記事コーパスデータを可視化するために，著者らはアイコンの属性と左右の可視化結果の連動操作を，以下のようにカスタマイズした．

- 「右京」のアイコンの高さは，j 番目のキーワードの重要度の合計 $\sum_{i=1}^{n} r_{ij}$ を表す．
- 「右京」は特定の条件によってキーワードの重要度を計算し，その重要度に比例した高さでキーワードのアイコンを表示できる．ユーザはその条件を GUI 上で選択できるものとする．
- 「右京」にてキーワードのアイコンにカーソルを合わせると，そのキーワードが文字列で表示される．
- ユーザがキーワードをキーボード入力すれば，入力したキーワードのアイコンが「右京」にてハイライトされる．
- 「右京」のアイコンやクラスタをクリックすると，そのクラスタ内にあるキーワードのリストが表示される．
- 「右京」でユーザがキーワードのアイコンをクリックすると，「左京」の各アイコンの R 値が，キーワードの重要度に比例した値で再算出される．
- 「右京」でユーザが別のキーワードのアイコンをクリックすると，「左京」の各アイコンの G 値が，キーワードの重要度に比例した値で再算出される．
- 「左京」にて，ユーザがアイコンやクラスタをクリックすると，「右京」にて，キーワードのアイコンの色相が再算出される．キーワードのアイコンの色相は，クリックしたクラスタ内の記事文書におけるキーワードの重要度の合計を示す．赤に近いほど重要度が高く，青に近いほど重要度が低い．
- ユーザが「左京」のアイコンをクリックすると，そのアイコンが表す記事文書の本文が表示される．

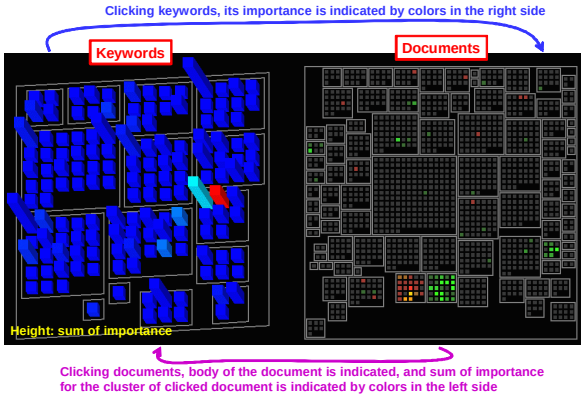

図 7: 新聞記事コーパス可視化のための提案手法のカスタマイズ．

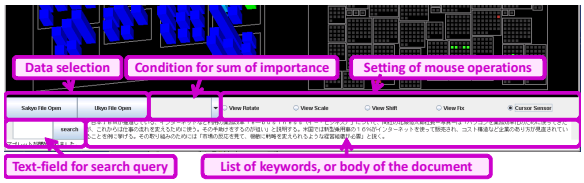

図 8: 新聞記事コーパス可視化のための GUI.

さらに著者らの実装では，図 8 に示すように，新聞記事コーパス可視化のために，ウィンドウの下部に以下のような GUI を設置した．

- 可視化の対象となるデータを選択するボタン．
- 条件にしたがって表示対象となるキーワードや記事を絞り込むための選択メニュー．現在の実装では，記事を掲載月で絞り込めるようになっている．

- マウス操作を設定するボタン.
- 検索のための文字入力欄.
- 選択したアイコンに関連するテキストを表示させる部分．現在の実装では,「右京」でキーワードのクラスタを選択すると，キーワードのリストが表示される．また,「左京」で記事文書を選択すると，記事の本文が表示される.

また著者らの実装では，図 8 に示す GUI 部品とは別に，色相，高さ，枠の線の太さなどの特性を調節するためのメニューもサポートしている.

6 新聞記事コーパス可視化の実験

本章では，著者らが「左京と右京」を新聞記事コーパスの可視化に適用した実験結果を紹介する.

6.1 新聞記事コーパスから表形式データの作成

著者らは新聞記事コーパスとして,「動向情報の要約と可視化に関するワークショップ (MuST)」[1] が提供する毎日新聞全文記事データベース (1998 年,1999 年) を用いた．このコーパスは,「date」「headline」「body」などのタグがつけられた XML 形式テキストファイルの集合で構成されている.

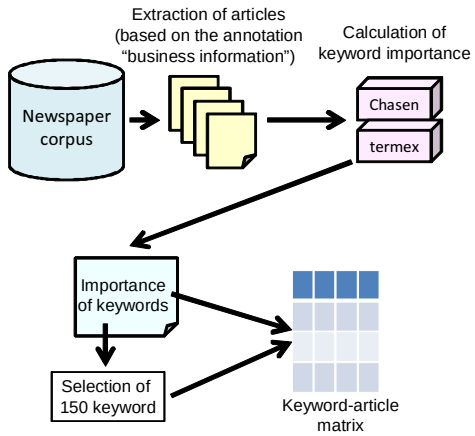

図 9: 新聞記事コーパスから表形式データを作成する処理手順.

本実験による表形式データの作成手順を，図 9 に示す．本実験ではまず,「headline」タグの内部に「ビジネス情報」という単語を含む記事を全て抽出した．その結果として，1998 年の新聞記事から 2178，1999 年の新聞記事から 1400 の,「ビジネス情報」に関する記事が抽出された．続いて，抽出されたそれぞれの記事文書に対して単語の重要度計算を適用し，1998 年と 1999 年でそれぞれ重要度の順位が 200 位までの単語を抽出した．著者らは文書の形態素解析に”chasen” [2] を適用し，単語の重要度計算に”termex” [3] を用いた．続いてこれらの 200 個ずつの単語の中から，1998 年および 1999 年それぞれに対して，手動で 150 個の単語を選んだ．この選択に際して著者らは，ビジネスに関する強い動向情報を表すと思われる単語，具体的には企業名，商品名，技術的，経済的条件を表す単語を優先的に選んだ．そして 1998 年と 1999 年それぞれに対して，キーワードと記事から構成される表形式データを作成し，その各欄に重要度の実数値を埋め，その可視化を行った.

[1] http://must.c.u-tokyo.ac.jp/ 参照.

6.2 実行例

6.2.1 可視化例 (1)

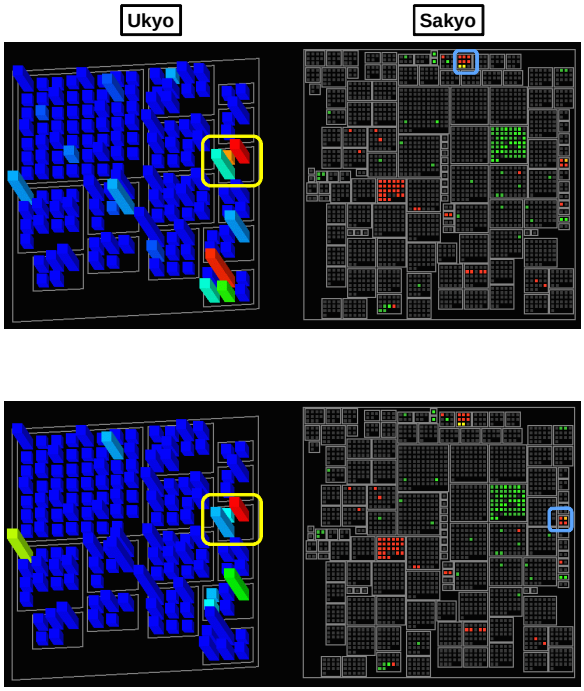

図 10: 可視化例 (1).「米国」「インターネット」の 2 つのキーワードに着目した.

図 10 は，1998 年の新聞記事の可視化結果の例である.「右京」で，あるクラスタ（黄色の丸の部分）をクリックし，そのクラスタに属するキーワードのリストを表示させた．その中から,「米国」「インターネット」の 2 つのキーワードをクリックして,「左京」でそれらのキーワードを含む記事をハイライトさせた.「左京」

[2] http://chasen.naist.jp/hiki/ChaSen/ にて公開されている.

[3] http://gensen.dl.itc.u-tokyo.ac.jp/ にて公開されている.

の表示結果で赤色のアイコンは,「インターネット」というキーワードを含む記事を示し，緑色のアイコンは「米国」というキーワードを含む記事を示す．それぞれの明度は，そのキーワードの記事内での重要度を示す．この結果から著者らは，黄色のアイコンを含むクラスタ，つまり「インターネット」「米国」の両方の単語について高い重要度をもつ記事を含む 2 つのクラスタ（図 10 における「左京」の青い丸の部分）に注目した．

まず一つ目のクラスタ（図 10(上) における「左京」の青い丸の部分）をクリックすると,「右京」にてキーワードのアイコンのいくつかが，青以外の色でハイライトされた．ハイライトされたアイコンはそれぞれ,「金融情報」「サービス」「無料」「パソコン」の各キーワードを表すものであった．クリックしたクラスタ内の記事本文を読んでみて，インターネットを利用した金融サービスなどに関する記事でクラスタが構成されていることが確認できた．

同じように他方のクラスタ（図 10(下) における「左京」の青い丸の部分）をクリックして「右京」のアイコンをハイライトさせると,「右京」にてキーワードのアイコンのいくつかが，青以外の色でハイライトされた．ハイライトされたアイコンはそれぞれ,「企業」「ビジネス」「投資」「パソコン」の各キーワードを表すものであった．クリックしたクラスタ内の記事本文を読んでみて，インターネットを利用したビジネスプロセスの革新などに関する記事を含むクラスタであることが確認できた．

以上の結果より著者らは,「米国」と「インターネット」に関連する記事で，意味の違う 2 つのクラスタの存在を発見できた．

6.2.2 可視化例 (2)

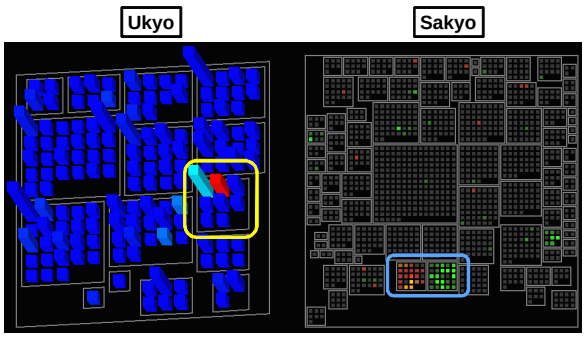

図 11: 可視化例 (2).「パソコン」「デジタルカメラ」の 2 つのキーワードに着目した．

図 11 は，1999 年の新聞記事の可視化結果の例である.「右京」で，あるクラスタ（黄色い丸の部分）をクリックし，そのクラスタに属するキーワードのリストを表示させた．続いてその中から,「パソコン」と「デジタルカメラ」の 2 つのキーワードをクリックした．その結果,「左京」でそれらのキーワードを含む記事がハイライトされた．赤色のアイコンは,「デジタルカメラ」というキーワードを含む記事を示し，緑色のアイコンは,「パソコン」というキーワードを含む記事を示す．それぞれの明度は，そのキーワードの記事内での重要度を示す．

この結果から，2 つのクラスタ（図 11 における「左京」の青い丸の部分）に注目した．このクラスタは，それぞれ赤いアイコンが多いクラスタ，および緑のアイコンが多いクラスタである．このようなハイライト分布より，この 2 つのキーワードはクラスタリング結果に強く貢献する重要なキーワードである，ということがわかる．

また，この可視化結果では，赤いアイコンを含むクラスタの中に，黄色やオレンジのアイコンもいくつか含まれている．この結果により，デジタルカメラに関する記事のいくつかに，パソコンに関する内容が大きく関与していることが発見できた．

6.2.3 可視化例 (3)

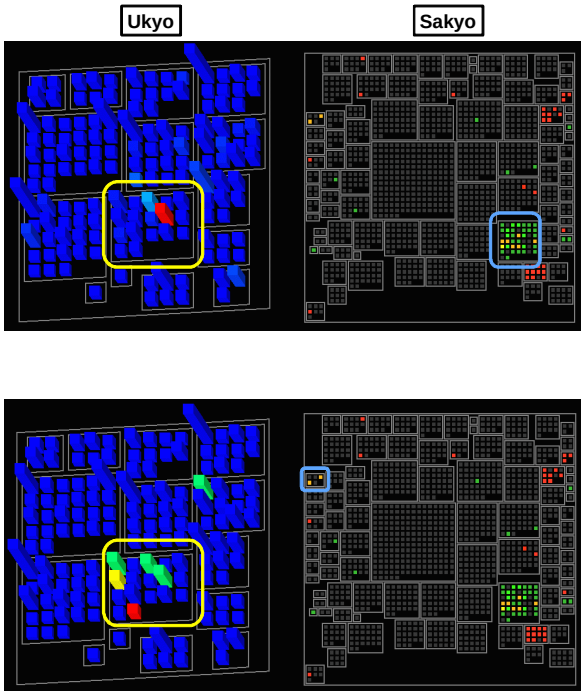

図 12: 可視化例 (3).「デザイン」「トヨタ自動車」の 2 つのキーワードに着目した．

図 12 は，1999 年の新聞記事の可視化結果の例である.「右京」で，あるクラスタ（黄色の丸の部分）で高さが高いアイコンに注目し,「デザイン」と「トヨタ自動車」の 2 つのアイコンをクリックした．すると,「左京」でいくつかのアイコンがハイライトされた．赤色のアイコンは,「デザイン」というキーワードを含む記事を示し，緑色のアイコンは,「トヨタ自動車」というキーワードを含む記事を示す．それぞれの明度は，そのキーワードの記事内での重要度を示す．

この結果から著者らは,「左京」の 2 つのクラスタ (図 12 における「左京」の青い丸の部分) に注目した. それらは, クラスタ内に黄色のアイコンを含んでいることがわかる. まず 1 つ目のクラスタ (図 12(上) における「左京」の青い丸の部分) をクリックしたとき,「右京」にて青以外の色でハイライトされたアイコンは, わずか 2 個であった. ハイライトされたアイコンは「トヨタ自動車」と「デザイン」の各キーワードを表すものであった. 逆に言えば, 他の企業名や商品のキーワードはハイライトされなかったことになる. このことより 1999 年は, トヨタ自動車のデザインに関する記事が, 他の企業や他の商品よりデザインに関する記事と比べて, 突出して多かったことが示唆される.

他方のクラスタ (図 12(下) における「左京」の青い丸の部分) をクリックすると,「右京」にて多くのアイコンが, 青以外の色でハイライトされた. これらのアイコンは「アサヒビール」「松下電器産業」「トヨタ自動車」「展開」「デザイン」などのキーワードを表すものであった.「左京」にてクリックしたクラスタ内にある記事は, これらの企業が共同で展開する新ブランドについての記事であった. このような記事群の存在を著者らは想像していなかったが, 以上の操作によってこのような意外な記事群を発見することもできた.

6.3 評価と考察

6.3.1 定量的評価

著者らは Java Development Kit 1.5 を用いて「左京と右京」を実装し, Apple MacBook Pro (CPU 2.2GHz, RAM 2GB, ディスプレイ 1440 × 900 画素) および Windows XP Service Pack 2 の上で実行した.

上記の計算機環境にて, 4.2 節に示した「左京と右京」における可視化処理の計算時間を計測したところ, 図 10 に示すキーワード 150 個, 記事 2178 個のデータにおいて, 右京に 0.04 秒, 左京に 0.06 秒を要した. このことから, 提案手法による可視化処理が非常に高速であることがわかる.

また, 同じデータを用いて, アイコン数の多い「左京」について, マウス操作にあわせてどのくらい縮小表示すると, クリック操作が難しくなるか, という調査を行った. その結果, 平均で拡大率 0.64 倍 (面積比 0.412) を下回るとクリック操作が難しくなる, という結果が出た. この結果から, 上記解像度のディスプレイにおいて, クリック操作が可能な最大表示数は 2178/0.412=約 5300 個, と推測できる. ここで一般的に, 5300 行の表形式データを, スクロール操作せずに可視化するのは困難である. よって提案手法は, 表形式データを表のまま可視化する手法と比べて, 大規模な表形式データの可視化に向いている, ということがいえる.

6.3.2 主観的評価

著者らは, 14 人の被験者に提案手法を操作してもらい, コメントを収集した. 14 人の被験者は, コンピュータグラフィックスを学ぶ女子の大学生であるが, 彼女らはテキスト文書可視化の専門ではない.

本評価では, 以下の 2 種類の設問を被験者に与え, その回答を集計することで,「左京と右京」を用いた新聞記事データベースの可視化の有効性について検証した.

[設問 1:] 6.2 節にて紹介した可視化例 (1)〜(3) における「左京」を提示し, その中から注目すべき記事クラスタに順位をつけさせ, 最大 5 個 (1 位から 5 位まで) を選ばせた. そしてその回答を回収し, 著者らが「このクラスタに注目すべき」と意図したクラスタを適切に選んでくれた被験者の人数を集計した.

[設問 2:] 図 11(上)(下) および図 12(上) の 3 つの可視化例において,「右京」から抽出されたキーワード群を提示し, そこから想像される具体的な記事内容を文章記述させた. そしてその回答を回収し, これらのキーワード群を抽出する際にクリックした「左京」の記事群の内容に近い記事を正しく想像できた被験者の人数を集計した.

設問 1 に対する集計結果を表 1 に, 設問 2 に対する集計結果を表 2 に示す.

表 1: 設問 1 に対する集計結果

	可視化例 (1)	可視化例 (2)	可視化例 (3)
人数	14	12	12

表 2: 設問 2 に対する集計結果

	図 11(上)	図 11(下)	図 12(上)
人数	10	5	11

表 1 より, 設問 1 に対する回答は概ね良好であったことがわかる. ただし可視化例 (2)(3) では, 非常に小さな (しかし重要な) クラスタを見落とす回答が若干見られた.

表 2 より, 設問 2 に対する回答の中で, 図 11(下) への正解率が低かった. この一因として, 抽出されたキーワードが意味の広いものばかりで構成されていたから, ということが考えられる. このことから「左京と右京」を新聞記事データベースの可視化に応用する際には, 可視化技術そのものの機能性だけでなく, 新聞記事からのキーワード選びも重要であると考えられる.

6.3.3 考察

著者らは主観的評価を実施した際に, あわせて被験者に対して, 提案手法について自由にコメントを記述してもらった. 以下, 収集したコメントの概要と, それを反映した提案手法の今度の展望について述べる.

多くの被験者は, すぐに可視化結果の意味を理解し, 操作を習得した. この手法の使いやすい点, わかりやすい点について, 以下のようなコメントがあった.

- 「右京」にて，キーワードの重要度を高さで示すことで，キーワードの新興性やトレンドを発見しやすい．
- また「左京」にて，キーワードの重要度を色相と明度で表現することから，直観的にキーワードと記事の関係を理解しやすい．
- 「左京」でクラスタをクリックしたときに，クラスタ内の記事に含まれるキーワードの重要度が「右京」で表現されるので，ハイライトされたキーワードからどのような記事でクラスタが構成されているのかを想像できるのが興味深い．

一方，この提案手法の持つ問題点について述べた被験者もいた．まず，キーワードのクラスタが直観的に理解しにくいという点を述べる者がいた．一例として本実験のように，ビジネス情報のキーワードを用いる場合には，キーワードがカテゴリごとに，例えば商品名，企業名，一般的なビジネス単語，といったカテゴリごとにグループ化されていたらより直観的だ，というコメントがあった．

別の問題として，この提案手法は,「左京」と「右京」は見かけが似ているが，操作の意味は異なるので混乱する，というコメントがあった．この点について著者らは，より被験者を増やして検討する必要があると考えている．もし提案手法の表現がユーザを本当に混乱させるなら,「左京」か「右京」どちらかを他の階層型可視化手法で差し替える，ということも検討する必要がある．一例として，キーワードの構成を木構造として可視化するために,「右京」に Cone Tree[11] のような木構造可視化手法を適用させる，という案が考えられる．

また複数の被験者から，3 個以上のキーワードの可視化に着手すべき，というコメントがあった．現在の実装では，RGB 値のうち R と G の 2 値を用いて 2 つのキーワードの重要度を表現しているが，これに対する簡単な拡張として，RGB 値の B 値を 3 番目のキーワードの重要度表現に適用することが考えられる．しかしこの方法が適切であるとは言い切れない．なぜなら，もし 3 個のキーワード全てと深く関係している文書があった場合，その文書のアイコンは白やグレーでハイライトされるため，すべてのユーザにとって目だって見えるとは言い切れないからである．また，この拡張が有効なのはキーワードが 3 個までの場合であり，4 個以上のキーワードの相関性を可視化したいというユーザがいる場合には，根本的に新しい方法について議論しなければならないと考えられる．

また，新聞以外のコーパス，例えば論文や特許データなどの文献，料理レシピなどにこの可視化手法を適用させると面白いのではないかという提案もあった．この点についても，今後の課題として実験してみたい．

7 まとめ

本論文では，階層型データ可視化手法「平安京ビュー」を 2 個使うことで表形式データを可視化する新しい手法「左京と右京」を提案した．また，新聞記事コーパスから得られる表形式データを提案手法に適用し，その結果について検討した．

今後の課題として，前章の考察で述べた点に加えて，以下の点を検討中である．

- より多くの被験者を使ってユーザビリティのテストを行う．
- 提案手法に向いていると思われるクラスタリング手法（双クラスタリングなど）の適用を検討する．

また 1 章にて前述したとおり，表形式データの可視化手法の中でも，遺伝子のマイクロアレイデータを対象とした可視化手法は特に多く発表されている．その手法の多くは，遺伝子をクラスタリングして並べ替えた結果を，表形式のまま表示する手法である [27, 28]．これらの関連手法と比較して提案手法が，どのように効果的な可視化結果を提示できるか，今後の課題として検討したい．すでに「平安京ビュー」はマイクロアレイデータの可視化に適用されている [29] ので，その単純な拡張により，提案手法を用いてマイクロアレイデータの可視化を試みることができると考えられる．

謝辞

テキスト文書データに関する情報提供をしてくださったお茶の水女子大学小林一郎准教授，東京大学加藤恒昭准教授，実験時にアドバイスをいただいた被験者の皆様に，感謝の意を表します．

本論文で用いた毎日新聞全文記事データベース (1998 年,1999 年) は,「動向情報の要約と可視化に関するワークショップ (MuST)」によって提供されました．

また本研究の一部は，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] Saraiya P., North C., Duca K., An Evaluation of Microarray Visualization Tools for Biological Insight, *IEEE Information Visualization 2004*, pp. 1-8, 2004.

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

[3] Itoh T., Takakura H., Sawada A., Koyamada K., Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space, *IEEE Computer Graphics and Applications*, Vol. 26, No. 2, pp. 40-47, 2006.

[4] 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.

[5] 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), pp. 127-134, 2006.

[6] Tachibana H., Itoh T., Sakyo & Ukyo: Visualization of Clustered Matrix Data Applying Dual Hierarchical Data Visualization Technique, *NICOGRAPH International 2007*, 2007.

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

[10] 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.

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

[12] Koike H., Fractal Views: A Fractal-Based Method for Controlling Information Display, *ACM Transactions on Information Systems,* Vol. 13, No. 3, pp. 305-323, 1995.

[13] 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.

[14] Rekimoto J., The Information Cube: Using Transparency in 3D Information Visualization, *Third Annual Workshop on Information Technologies & Systems*, pp. 125-132, 1993.

[15] Itoh T., Yamaguchi Y., Ikehata Y., Kajinaga Y., Hierarchical Data Visualization Using a Fast Rectangle-Packing Algorithm, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 10, No. 3, pp. 302-313, 2004.

[16] Artero A. O., de Oliveira M. C. F., Levkowitz H., Uncovering Clusters in Crowded Parallel Coordinates Visualizations, *IEEE Information Visualization 2004*, pp. 81-88, 2004.

[17] Johansson J., Ljung P., Jern M., Cooper M., Revealing Structure within Clustered Parallel Coordinates Displays, *IEEE Information Visualization 2005*, pp. 125-132, 2005.

[18] Hibbs M. A., Dirksen N. C., Li K., Troyanskaya O. G., Visualization Methods for Stastical Analysis of Microarray Clusters, *BMC Bioinformatics 2005*, Vol. 6, pp. 115-124, 2005.

[19] Marks J., et al., Design Galleries: A General Approach to Setting Parameters for Computer Graphics and Animation, *ACM SIGGRAPH '97*, pp. 389-400, 1997.

[20] Seo J., Shneiderman B., A Knowledge Integration Framework for Information Visualization, *Integrated Publication and Information Systems to Virtual Information and Knowledge Environments 2005*, pp. 207-220, 2005.

[21] Ward M. O., XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data, *IEEE Visualization '94*, pp. 326-333, 1994.

[22] Havre S., Hetzler E., Whitney P., Nowell L., ThemeRiver: Visualizing Thematic Changes in Large Document Collections, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, No. 1, pp. 9-20, 2002.

[23] Wong P. C., Cowley W., Foote H., Jurrus E., Thomas J., Visualizing Sequential Patterns for Text Mining, *IEEE Information Visualization 2000*, pp. 105-111, 2000.

[24] Wise J. A., et al., Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents, *Reading in Information Visualization: Using Vision to Think*, pp. 442-450, 1999.

[25] Andrews K., et al., The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities, *Information Visualization*, Vol. 1, No. 3, pp. 166-181, 2002.

[26] Takano A., et al., Associative Information Access Using DualNAVI, *Kyoto International Conference on Digital Libraries (ICDL'00)*, pp.285-289, 2000.

[27] Eisen M., Spellman P., Brown P., Boststein D., Cluster Analysis and Display of Genome-wide Expression Patterns, *PNAS* Vol. 95, Issue 25, pp. 14963-14968, 1998.

[28] Seo J., Shneiderman B., Interactively Exploring Hierarchical Clustering Results, *IEEE Computer*, Vol. 35, pp. 80-86, 2002.

[29] 西山, 伊藤, 「平安京ビュー」を用いた階層型遺伝子ネットワークの可視化, 芸術科学会論文誌, Vol. 6, No. 3, pp. 106-116, 2007.

橘 春帆

2006 年お茶の水女子大学理学部情報科学科卒業．2008 年お茶の水女子大学大学院人間文化研究科数理・情報科学専攻博士前期課程修了．情報処理学会会員．

伊藤 貴之

1990 年早稲田大学理工学部電子通信学科卒業．1992 年早稲田大学大学院理工学研究科電気工学専攻修士課程修了．同年日本アイ・ビー・エム (株) 入社．1997 年博士 (工学)．2000 年米国カーネギーメロン大学客員研究員．2003 年から 2005 年まで京都大学大学院情報学研究科 COE 研究員 (客員助教授相当)．2005 年日本アイ・ビー・エム (株) 退職，2005 年よりお茶の水女子大学理学部情報科学科助教授．ACM, IEEE Computer Society, 情報処理学会，芸術科学会，画像電子学会，可視化情報学会，他会員．