大規模表形式データ可視化手法「左京と右京」を用いた文献データの可視化

白鳥佳奈　（指導教員：伊藤貴之）

1. 概要

現代社会には膨大な文献データが存在し，その検索や分析は必ずしも容易ではない．例えば関連文献を探索する際，検索対象とするキーワードが普遍的であればあるほど，多種多様な内容の文献が同時に提示されてしまい，効率の良い探索は難しくなる．そこで，文献データの分布を可視化することによって，目的の文献群を直観的に把握でき，探索をスムーズに行うことが可能となると考えられる．また，可視化結果全体を眺めることにより，その文献データの大局的な動向把握にも役立つと考えられる．

本研究では，大規模表形式データ可視化手法「左京と右京」[1]を用いて，文献と著者の関係について可視化を試みる．本手法は，まず文献中からキーワードを抽出し，そのキーワード，文献，著者の分布を可視化する．これによって，文献の検索作業や動向把握を容易にできると考える．

2. 左京と右京：大規模表形式データ可視化手法

本研究で用いる「左京と右京」は，階層型データ可視化手法「平安京ビュー」[2]の拡張手法の一つで，これを一画面上に左右に 2 つ並べて表示する可視化手法である．「平安京ビュー」は階層型データの葉ノードをアイコンで，枝ノードを長方形の枠で表す入れ子型構造で表示する．

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行い，2 つの階層型データを生成する．続いてこの 2 つの階層型データに対してそれぞれ「平安京ビュー」を適用し，可視化する．このとき，ユーザが対話的に表形式データを探索できるよう，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．このような対話的操作機能により，ユーザは表形式データの探索を容易に行うことが可能である．なお，2 つの「平安京ビュー」のうち右側の「平安京ビュー」を「左京」，左側の「平安京ビュー」を「右京」と呼ぶ．

また文献[1]では，「左京と右京」を新聞記事コーパスの可視化に適用し，キーワードと新聞記事に潜む興味深い関係を発見した事例を報告している．

3. 提案内容

本手法を用いた可視化結果の例を図 1 に示す．

本手法ではまず，各々の文献データから著者・文献・キーワードを抽出する．そして著者と文献を「左京と右京」を用いて可視化し，画面の一番左にボタンの集合として「キーワードパネル」を表示する．

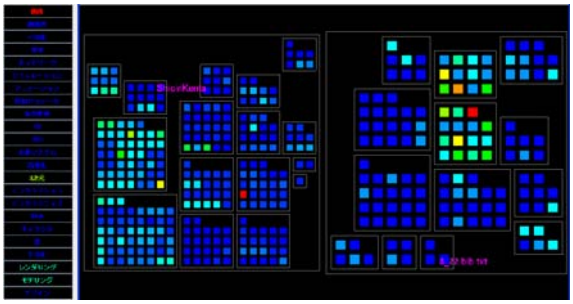

図 1：可視化結果例

3.1 「左京と右京」の拡張

先行研究である文献[1]では新聞記事コーパスの可視化を行っており，新聞記事とキーワードという 2 軸で 1 つの表形式データを可視化していた．

それに対し本研究では，文献データを可視化するにあたり，その 2 軸にさらに著者情報を付加し内部ロジックを 3 次元に拡張し（図 2 参照）2 つの表形式データを可視化する．これによって，著者と文献・キーワードの相関性を表現する．

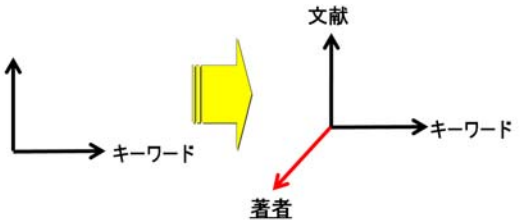

図 2：「左京と右京」の拡張

3.2 文献データ可視化における処理手順

以下，各論文を r_1~r_n(n は論文数)，各著者を s_1~s_l(k は著者数)とし，各キーワードを c_1~c_m(m はキーワード数)と記述する．

本手法ではまず，論文と著者のクラスタリングを行う．ここで a_{ij} は i 番目の論文の j 番目のキーワードの重要度を示し，b_{kj} は k 番目の著者の j 番目のキーワードの重要度を示す．

続いて我々の実装では，「右京」にクラスタリングされた著者を表示し，「左京」にクラスタリングされた論文を表示する．また，画面左端にボタンとしてキーワードを表示する．各アイコンの色は重要度によって算出されており，赤色に近いほど重要度が高く，青色に近いほど重要度が低いことを示す．

3.3 二つの「平安京ビュー」およびキーワードパネル間の操作

提案手法では，ユーザが対話的に 2 つの表形式データを探索できるよう，「キーワードパネル」・「左京」・「右京」は相互に操作可能な機能をもつ．例えば，ユーザがキーワードのボタンをクリックすると，このキーワードの表すデータ要素に対応する「左京」及び「右京」の角柱が色や形などを変えて表示される．同様に，「左京」や「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キーワードのボタンや「右京」または「左京」の角柱が色や形などを変えて表示される．

今，文献とキーワードから構成される表を表 T_1，著者とキーワードから構成される表を表 T_2 とする．また，表 T_1，表 T_2 ともに列数はキーワード数より 1 つ多い数とし，最後の列の要素はすべて 0 で初期化しておく．

3.3.1 キーワードパネルクリック時の左京･右京の更新

ここで，ユーザがキーワードボタン c_j をクリックすると仮定する．このとき提案手法は，表 T_1 において a_{1j} から a_{nj} の値を探索し，値 a_{ij} を用いて「右京」のデータ要素 r_i を算出し，「右京」を構成する棒グラフの色，高さ，形などを更新する．その一方，表 T_2 においても同様に b_{1j}～b_{lj} までの値を探索し，値 b_{kj} を用いて「左京」のデータ要素 s_k を算出し，「左京」を

構成する棒グラフの色，高さ，形などを更新する．以上の処理の流れを図 3(赤)に示す．

3.3.2 左京クリック時のキーワードパネルの更新

次に，ユーザが「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，表 T_1 において a_{i1} から a_{im} の値を探索し，値 a_{ij} を用いてキーワードのデータ要素 c_j を算出し，キーワードパネルを構成する文字の色などを更新する．

3.3.3 左京クリック時の右京の更新

2 つの表形式データは列データが共通である．ここで，ユーザが「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，表 T_1 において a_{i1} から a_{im} の値を探索し，値 a_{ij} があらかじめ定めた閾値より大きければ，キーワードのデータ要素 c_j を表 T_2 において参照する．そして b_{1j} から b_{lj} の値を探索し，値 b_{kj} を表 T_1 における値 a_{ij} によって重みづけを行いながら $b_{1\alpha}$ から $b_{l\alpha}$ の値を更新していく．この操作を表 T_1 において a_{i1} から a_{im} まで同様に繰り返し，最終的に値 $b_{k\alpha}$ を用いて「右京」のデータ要素 s_k を算出し，「右京」を構成する棒グラフの色，高さ，形などを更新する．以上の処理の流れを図 3(青)に示す．

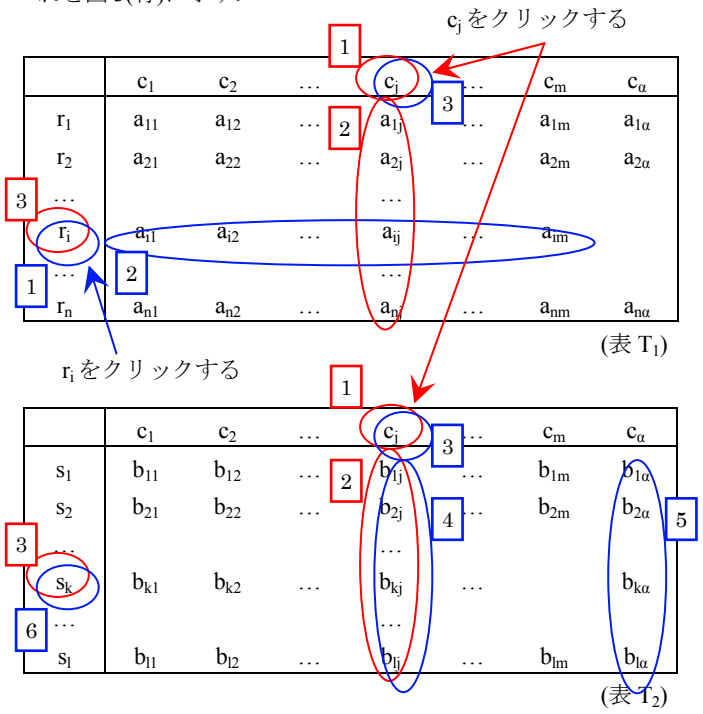

図 3：クリック操作時における各表示更新の内部処理手順

4. 適用事例

我々は適用事例として，芸術科学会論文誌[3]の掲載論文を題材とした．この論文誌に掲載された全ての論文は，キーワードや概要をまとめたカバーシートと PDF ファイルを有する．

本適用事例では，まず論文データ中の各論文から論文番号・概要・著者情報を抽出し，さらに抽出された概要に対し文章の形態素解析，および重要度計算を適用し，単語ごとにそれぞれの論文に対する重要度を求めた．我々の実装では文書の形態素解析に”茶筌”[4]を用い，単語の重要度計算に”termex”[5]を用いた．そして，重要度が上位の単語の中から，各論文のキーワードとして意味をもつ単語を手動で選択し，キーワードとした．続いて，論文とキーワード，著者とキーワードをそれぞれ行と列として構成される 2 つの表形式データを作成し，その各欄に各キーワードにおける重要度を埋め，可視化を行った． なお，当論文誌の現時点での著者数は 283，論文数は 134，キーワード数は 23 である．図 4 はキーワードとして「CG」を選択した際の可視化結果である．

まず「右京」のハイライト分布を見ると，中心付近に特に重要度の高いアイコンが 2 つ存在することが見てとれる．このアイコンはそれぞれ「千葉則茂先生」と「藤本忠博先生」であった．この結果より，上の 2 人の人物が当論文誌において「CG」というキーワードに深く関係していることがわかる．

次に「左京」のハイライト分布を見ると，特に右上の 2 つのクラスタに重要度の高いアイコンが集中していることが見てとれる．このそれぞれのクラスタに属するアイコンをクリックし，キーワードパネルのハイライトを見ていくと，赤色のクラスタに含まれる論文は「CG」「3 次元」「モデリング」「画像」といったキーワードを含み，一方，黄色のクラスタでは「CG」「レンダリング」「支援システム」「デザイン」といったキーワードを含むことがわかった．よって，「CG」という共通のキーワードを含む論文でも，CG のアルゴリズムについての論文群と，CG を用いた応用システムについての論文群に分類されることがわかる．この結果より，ユーザは自分の興味のある方のクラスタを重点的に探索することができ，効率の良い論文探索につながると考えられる．

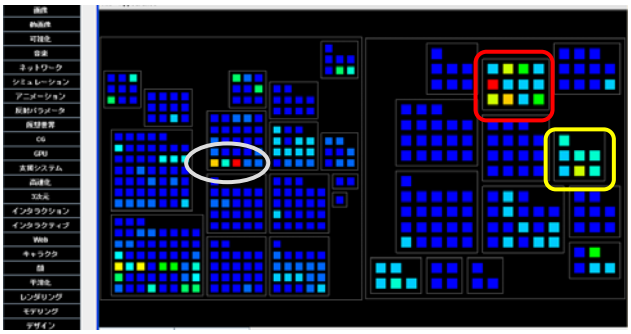

図 4：キーワード「CG」を選択した際の可視化結果

5. まとめ

本報告では，大規模表形式データ可視化手法「左京と右京」を用いた文献データの可視化について提案した．

今後は，他の論文誌を適用した結果からの考察を行うとともに，「左京」「右京」に高さ情報を付与させる他，操作性の向上のための機能の実装や GUI の構築に取り組みたい．

参 考 文 献

[1] 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”，芸術科学会論文誌，Vol. 7, No. 2, pp. 22-33, 2006.

[2] 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.

[3] 芸術科学会論文誌 http://www.art-science.org/journal/index.html

[4] 形態素解析システム「茶筌」，http://chasen.naist.jp/hiki/Chasen/

[5] 専門用語抽出自動システム「termex」，http://gensen.dl.itc.u-tokyo.ac.jp/