社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

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† 伊藤 貴之†

†お茶の水女子大学理学部 〒112-8610 東京都文京区大塚 2-1-1
E-mail: †{kana, itot}@itolab.is.ocha.ac.jp

あらまし 近年，公開文書の電子化などに伴い，世の中には数多くの文献データが存在する．しかし，その探索や分析は，すべての人に対して必ずしも容易ではない．我々は文献データの例として論文誌に着目し，これらを論文・著者・キーワードという 3 つの面から探索を行うための可視化手法について研究を進めている．ここで論文誌探索の要求として，特定の著者やキーワードに関して局所的に探索したいという要求のほか，論文誌全体を大局的に概観したいという要求も想定される．そこで本報告では、論文誌の局所的探索と大局的概観を実現するために，大規模表形式データ可視化手法「左京と右京」を用いた可視化手法を提案する。本手法では文献データから 2 種類の表形式データを作成し，論文と著者の分布を「左京と右京」で可視化し、キーワードをボタンで配置する．この 3 つを相互に操作することで，論文誌の局所的探索と大局的概観を実現する。

キーワード 可視化，階層型データ，表形式データ，文献データ

Visualization of Literature Data by Sakyo & Ukyo, A Technique of Visualization of Large-scale Table Data

Kana Shiratori† Takayuki Ito†

†Faculty of Science, Ochanomizu University 2-1-1 Otsuka, Bunkyo-ku, Tokyo, 112-8610 Japan
E-mail: †{kana, itot}@itolab.is.ocha.ac.jp

Abstract Recently, there have been enormous document data in the present society, due to digitalization of the published documents. However, search and analysis of such data are not always easy for all people. We focus on academic journals as documents data, and study on visualization of search for the document data from three aspects, articles, authors, and keywords. Here, we assume the following two requirements. One is local search with specific authors or keywords, and the other is overview of overall the journal. This paper presents a visualization technique that satisfies the two requirements, by applying a large-scale matrix data visualization technique “Sakyo&Ukyo”. The technique first generates two matrix data from document data. It then visualizes the distribution of articles and authors by “Sakyo&Ukyo”, and displays keywords as a set of buttons. It realizes the two requirements with the user interface to interact these three results each other.

Keyword Visualization, Hierarchical Data, Matrix Data, Corpus Data

1. はじめに

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. 関連研究

2.1 平安京ビュー：階層型データ可視化手法

本研究で用いる「左京と右京」は，階層型データ可視化手法である「平安京ビュー」[2]の拡張手法の一つである．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．この時，すべての葉ノードは同じ大きさで表示し，それぞれの葉ノードや枝ノードが画面上で重ならないように，かつデータ全体の画面専有面積ができるだけ小さくなるように配置する．こうして，大規模な階層型データの全体を一画面上に表示することができる手法である．

2.2 左京と右京：大規模表形式データ可視化手法

「左京と右京」は 2.1 節で述べた「平安京ビュー」を図 1 のように一画面上に左右に二つ並べて表示する可視化手法である．

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．たとえば，ユーザが左側の可視化結果の角柱をクリックすると，右側の可視化結果の対応する角柱が色や形などを変えて表示される．同様に，右側の可視化結果の角柱をクリックすると，左側の可視化結果の対応する角柱が色や形などを変えて表示される．

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「左京と右京」では、北を上にして描かれた平安京の地図に喩えて，当時東を「左京」，西を「右京」と呼んだことから，図 1 の右側の「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．

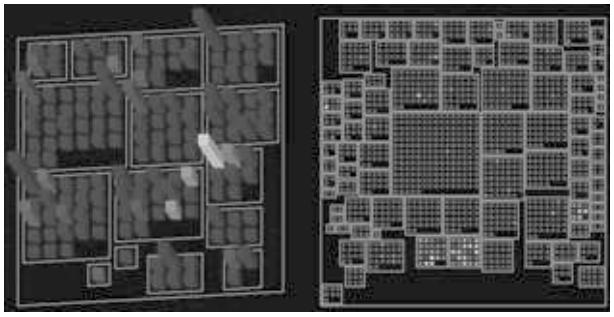

図 1：「左京と右京」による可視化例

3. 提案手法

本手法を用いた可視化結果の例を図 2 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．

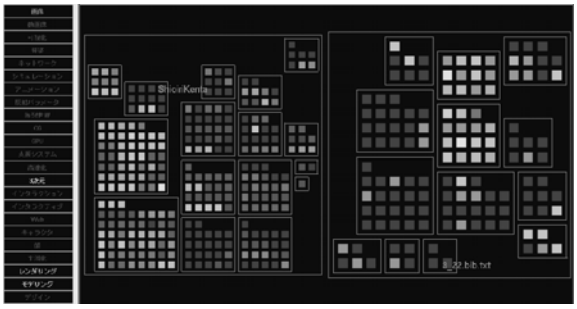

図 2：可視化結果例

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．

それに対し本手法では、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 3 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．

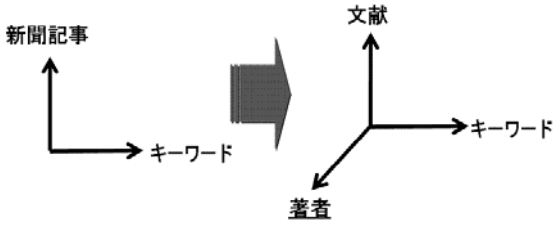

図 3：「左京と右京」の拡張

3.2 文献データ可視化のためのカスタマイズ

本報告では以下，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．まず提案手法では，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．

続いて提案手法では，「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，ユーザが対話的に二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」に相互に操作可能な機能をもたせる．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．

以下，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 つ多いとし，最後の列の要素はすべて 0 で初期化する．

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 4 に示す．

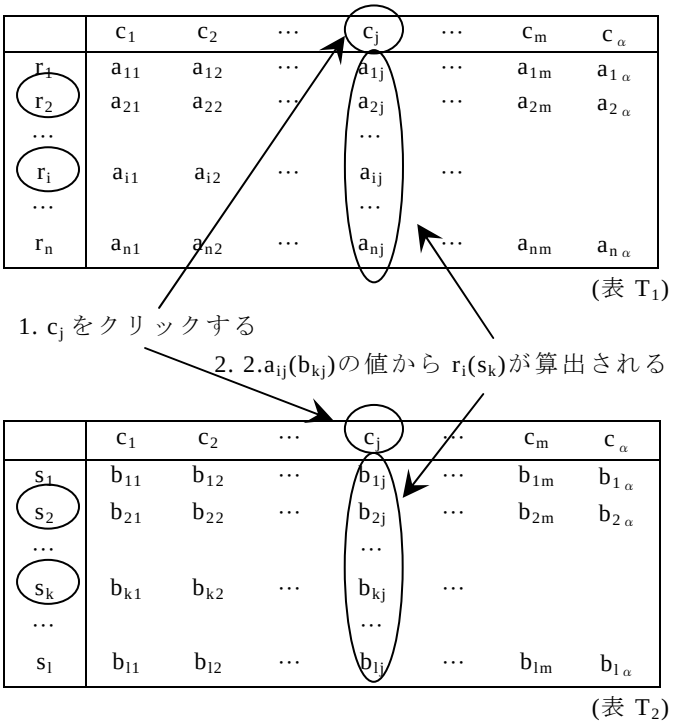

図 4：キーワードパネルのクリックにより左京・右京の表示を更新する内部処理手順

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

二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 5 に示す．

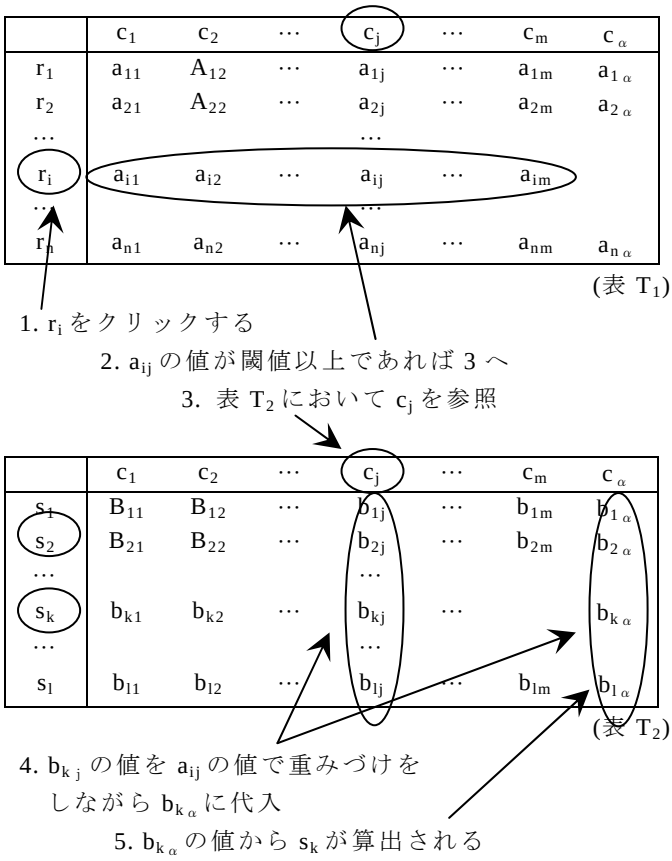

図 5：左京のクリックによりキーワードパネル・右京の表示を更新する内部処理手順

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ている．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．

本適用事例による表形式データの生成手順を図 6 に示す．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る．我々の実装では文書の形態素解析に”茶筅”[4]を用い，単語の重要度計算に”termex”[5]を用いる．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，キーワードとする．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二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，可視化を行う．現在，その可視化結果について検討中である．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．

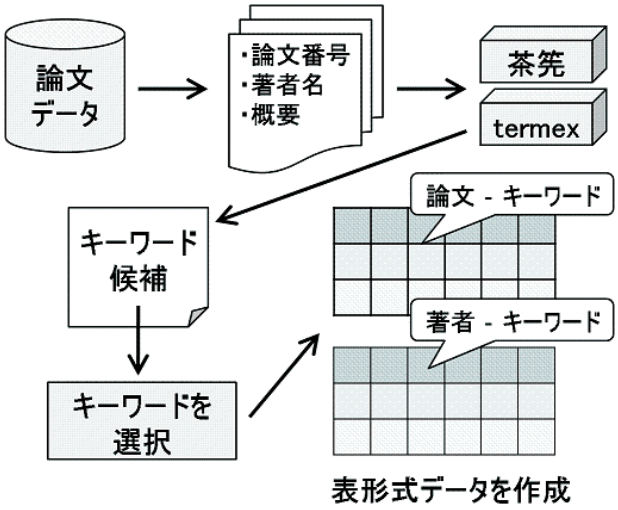

図 6：論文から表形式データを作成する処理手順

図 7 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，中心付近に最も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最も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，上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，下のクラスタでは「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一目でわかる．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．

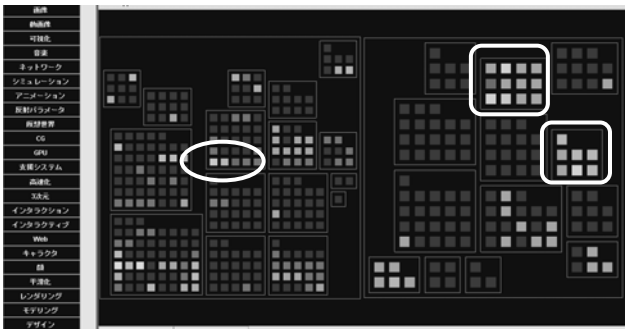

図 7：キーワード「CG」を選択した際の可視化結果

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を提案した．

今後の課題として，より大きな文献データの可視化を試みるとともに,以下の機能の実装に取り組みたい．

- 執筆論文数などの情報に基づき，右京のアイコンに高さを与える
- 選択中の項目を画面上にテキスト表示する
- 右京のアイコンをクリックすると，共著経験のある著者を左京でハイライトする
- 左京のアイコンをクリックすると，そのアイコンが表す論文の概要を表示する

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ある．

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] 奈良先端科学技術大学大学院松本研究室，形態素解析システム「茶筅」,
http://chasen.naist.jp/hiki/Chasen/

[5] 東京大学情報基盤センター中川研究室，横浜国立大学環境情報研究院森研究室共同開発，専門用語抽出自動システム「termex」,
http://gensen.dl.itc.u-tokyo.ac.jp/