大規模表形式データ可視化手法「左京と右京」による新聞記事コーパスの可視化

橘春帆　　(指導教員：伊藤貴之)

1. 研究背景と目的

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．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用いられている．本研究では，大規模な表形式データの可視化手法「左京と右京」[1]を用いて，1998年，1999年の毎日新聞の記事の傾向や特徴を可視化する．本研究は，記事中のキーワードの重要度，出現期間，共起関係などに着目し，記事同士の関連性を可視化し，その結果を利用してユーザが分析や意思決定に役立てることを目的としている．

2. 大規模表形式データの可視化手法「左京と右京」

2.1 「左京と右京」の概要

「左京と右京」のフローチャートを図1に示す．「左京と右京」は，大規模階層型データ可視化手法である「平安京ビュー」[2]を一画面に2つ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個の階層型データを生成する．続いて，図1中の「可視化1」「可視化2」の部分に「平安京ビュー」を適用することで，2個の階層型データを可視化する．

「左京と右京」の概要を図2に示す．図2の右側の「平安京ビュー」を「左京」と呼び，左側の「平安京ビュー」を「右京」と呼ぶ．

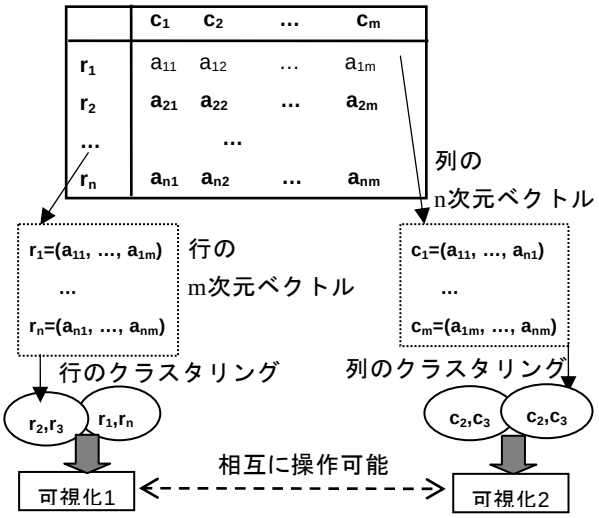

図1.「左京と右京」のフローチャート

2.2 2個の「平安京ビュー」間の操作

提案手法では，ユーザ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．例えばユーザが「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，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，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を算出し，「右京」を構成する棒グラフの色，高さなどを更新する．なお，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．

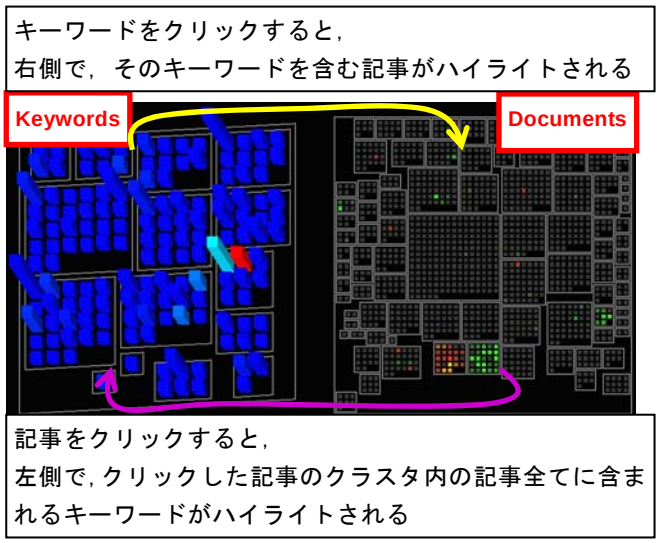

図2.「左京と右京」の概要．

3.「左京と右京」による新聞記事コーパスの可視化

3.1 新聞記事コーパス可視化のためのカスタマイズ

本研究では，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．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．まず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．本論文の実装では，「右京」にクラスタリングされたキーワードを表示し，「左京」にクラスタリングされた記事文書を表示する．新聞記事コーパスデータを可視化するために，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．

1) 「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．2)「右京」は月別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，表示する月を GUI 上で選択できるものとする．3)「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．4)「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．5)「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．6)「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．7)「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最算出される．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．8)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．

3.2　新聞記事コーパスから表形式データの作成

本研究では，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)」[3]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，XML形式テキストファイルの集合で構成されている. このコーパスから,「ビジネス情報」という単語を含む記事を全て抽出すると，1998 年の新聞記事から2178，1999年の新聞記事から1400 の記事が抽出された．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. 本研究では文書の形態素解析に”chasen”[4]を適用し，単語の重要度計算に”termex” [5]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，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．

4 実行結果

図 3 は 1999 年の新聞記事の可視化結果である．「右京」（キーワード側）で「パソコン」というキーワードに注目して月毎の変動を調べていたところ，8 月でアイコンの背が高くなった. 同じ 8 月に背の高いキーワードを見てみるとそのひとつに「デジタルカメラ」があった．この 2 つは同じクラスタにはないが, 相関性があるキーワード 2 つとして考えられる．この 2 つのキーワードをクリックして「左京」（記事側）のハイライト分布を見てみた．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.「左京」の丸で囲ったクラスタには，黄色やオレンジのアイコンがいくつか含まれている. この結果より, デジタルカメラに関する記事にパソコンというキーワードがかなり関与していることがわかる. また,「右京」側のアイコンの色は，「左京」側の丸で囲ったクラスタをクリックしたときの結果である．「デジタルカメラ」と「パソコン」がハイライトされていることから、この記事のクラスタには, この 2 つのキーワードを含む記事で構成されていることが分かる. このクラスタの記事を調べてみると，8 月の記事に,『デジタルカメラ購買層が，専門家だけでなくパソコン利用者の間で広がっている』という内容が確認できた． 1999 年は実際に，デジタルカメラの購買数が大きく増加した年であった．よって，この可視化結果は, このような後日の傾向を示す記事の発見につながると考えられる．

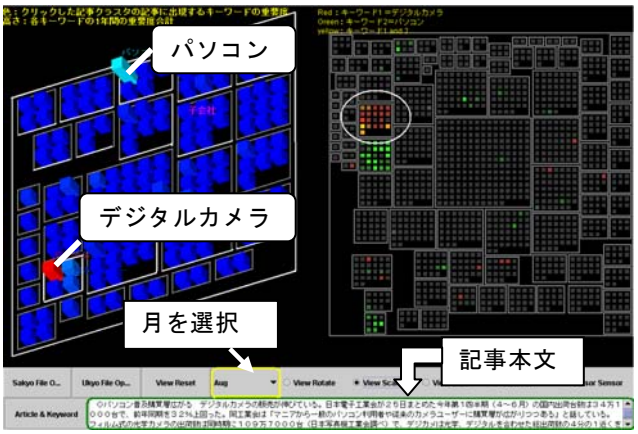

図 3.「パソコン」「デジタルカメラ」の 2 つのキーワードに着目した結果.

5 まとめ

本研究では, 新聞記事コーパスを表形式データに変換させたものを大規模表形式データ可視化手法「左京と右京」に適用させ，可視化実験を行った．キーワードの共起関係に着目して記事同士を可視化することで, いくつか興味深い結果が得られた．

参考文献

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

[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] 加藤，松下，平尾，動向情報の要約と可視化に関するワークショップの提案,信学技報 NLC2004-25, pp. 13-18, 2004.

[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/