Web検索結果のクラスタ分布の可視化の一手法

松枝　知香†　　伊藤　貴之†

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

あらまし　Web 上の検索結果には非常に多種多様なページが含まれることが多い．特に同音異義語を含む単語をクエリに用いた時には，全く内容の異なる Web ページが混在することが多く，それが検索の工程を煩雑にすることが多い．そこで本報告では，Web 検索結果であるページ群と，そのページ群に含まれるキーワード群を行および列とした表をつくり，そのクラスタリング結果を可視化する一手法を提案する．本手法ではページ群とキーワード群の各々をクラスタリングした結果を 2 つの木構造として表示する．さらに，一方をクリック操作すると他方が反応するという連携機能を提供することで，検索結果の対話的探索のためのユーザインタフェースとして活用できる．

キーワード　可視化，Web 検索，テキストマイニング

Visualization of cluster distribution of Web search results

Chika MATSUEDA† and Takayuki ITOH†

† Faculty of Information Science and Engineering, Ochanomizu University
1529 Otsuka, Bunkyoku, 112–8610 Japan
E-mail: †{coco,itot}@itolab.is.ocha.ac.jp

Abstract　Web search results often contain variety of Web pages. Especially, they may contain complete different fields of Web pages, when particular terms which have homographs are used for the query of search engines. This paper proposes a visualization technique for Web search results. The technique displays the clustering results of lows and colums of matrices, where lows correspond to Web pages, and columns correspond to keywords. The technique also features a linkage mechanizm: when a user operates one of the view by mouse click actions, corresponding parts of the other view reacts. This user interface is useful for interactive exploration of Web search results.

1. はじめに

Web 上の検索結果には非常に多種多様なページが含まれることが多い．例えば意味を一意にとるのが容易な固有名詞などを検索に用いた場合には，検索エンジンから得られたランク付けされた検索結果は有用であることが多い．一方で，広義な単語や同音異義語を含む単語をクエリに用いて幅広く情報を集める場合に，ランク付けされた検索結果には非常に多種多様な内容のページが含まれており，検索結果の全体像をつかむのが難しい場合もある．また，意図に合ったページとそれ以外のページが煩雑に混在されている場合が多く，意図に合ったページだけを読むのが難しい場合もある．

この問題の解決策として，検索結果をクラスタリングして提供する，ということが考えられる．検索結果のクラスタリングサービスとして，日本国内でも(注1)Clusty，(注2)Mooter などが既に商用化されている．これらのシステムは，いくつかの検索エンジンから検索結果を収集し，その内容から検索結果をカテゴリごとに分類している．しかし，検索結果であるウェブページ集合の内容類似性は複雑な構造を有しており，個々のクラスタがどのような類似性によって構成されているのかを表現するのは単純な問題ではなく，検索結果を単純かつ適切にカテゴリ分類できるとは限らない．また，これらのシステムが搭載しているインタフェースの多くはカテゴリごとに検索結果をランキングしているだけであり，依然として検索結果の全体像を表現しているとはいいがたい場合が多い．

本論文では，このような問題を解決するため，Web 検索結果のページ群と，そのページ群に含まれるキーワード群を行および列とした表をつくり、それらにクラスタリングを適用し、可

(注1)：http://clusty.com/
(注2)：http://www.mooter.co.jp/

視化する一手法を提案する．本手法ではページ群とキーワード群の各々をクラスタリングした結果を 2 つの木構造として表示する．さらに，一方をクリック操作すると他方が反応するという連携機能を提供することで，検索結果の対話的探索のためのユーザインタフェースとして活用できる．

2. 関 連 研 究

2.1 ウェブ検索結果の可視化手法

ウェブ検索結果の可視化手法は 1990 年代から多く発表されている．その多くはウェブページ群を構造化して表示するものであり，日本国内でも (注3) Blogopolis などのウェブサイトで搭載されている．一方でウェブ検索結果のキーワード間の関係も可視化するに値する重要な情報であり，その可視化手法もいくつか発表されている [1].

本研究は，それら 2 種類の可視化手法を合体して，画面上で連携操作することにより，さらに効果的なウェブ検索結果の対話的探索を目指すものである．

2.2 階層型データの可視化手法

従来のウェブ検索結果の可視化手法には，ウェブ検索結果に限らない汎用的な可視化手法を適用しているものが多い．その中でも階層型データの可視化手法はウェブ検索結果にもよく応用されており，その汎用的な手法は既に多く報告されている．階層型データを可視化する最もオーソドックスな手法は，木構造を点と線で表現するノードリンク型の可視化手法であり，Hyperbolic Tree [2] や Cone Tree [3] が知られている．

一方で，空間分割型の階層型データ可視化手法として，入れ子状の帯グラフで階層データを表示する Treemap [4] があげられる．また，そのを改善手法として Quantum Treemap [5] や Voronoi Treemap [6] が知られている．前述の Blogopolis も前提技術として Voronoi Treemap を採用している．また空間分割型手法の一種で，二次元空間で入れ子状に階層構造を構築する手法として，データ宝石箱 [7]，およびその改良手法として平安京ビュー [8] などがあげられる．空間分割型の可視化手法は，各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を主眼においた可視化手法である．本研究では，階層型データの親子関係よりも，階層の最下位に位置するウェブページ群を一覧することを目的としているため，空間分割型の可視化手法が適していると考えられる．

2.3 Web 検索結果のクラスタリング

1 章で紹介した通り，Web 検索結果のクラスタリング技術は Clusty や Mooter などの商用サービスにも利用されているが，学術的にはさらに精度の高いクラスタリング技術が研究発表されている．一例として馬場ら [9] の「キーワード蒸留型クラスタリング」では，数千件単位のウェブ検索結果をクラスタリングし，そこで抽出されたキーワードを蒸留することによって，表記揺れ，同義関係，含有関係などを処理することで，下位に埋もれた話題の発見を促している．

本報告で採用している Web 検索結果のクラスタリングは，現時点では単純に単語と文書のマトリクスに階層型クラスタリングを適用しているだけであるが，上述のような手法を採用することも可能であり，今後の課題として検討したい．

(注3)：http://blogopolice.jp/

3. 本研究が採用する可視化手法

3.1 平安京ビュー:大規模階層型データ可視化の一手法

「平安京ビュー」[8] はその可視化結果において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 1 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．階層型データの葉ノードを黒い長方形で，親ノードをそれらを囲う長方形の枠の入れ子で表現し，データの全体を一画面に表示することを目標としている．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を主眼においた視覚化手法である．

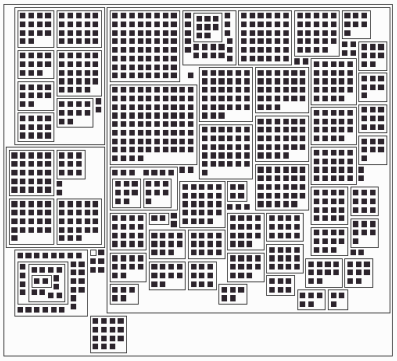

図 1　平安京ビューによる階層型データの可視化の例

「平安京ビュー」において技術的に重要な点として以下が挙げられる．

- 不均一な幅や深さを有する階層型データにも適用できる
- 枝ノードの長方形領域の正方形化
- 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
- 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
- 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

本論文の提案手法では，階層型データの親子関係をすることよりも，データの最下位に属するデータ要素の全貌を一覧することが重要であるため,「平安京ビュー」を用いることは，妥当であると言える．

3.2 左京と右京:大規模表形式データの可視化の一手法

左京と右京 [10] とは，前節で述べた「平安京ビュー」を用いた表形式データの可視化の一手法である．

表形式データの行と列を構成するデータ要素に，クラスタリングを適用し，各々の結果を「平安京ビュー」を用いて可視化する．図 2 は「左京と右京」による表形式データの可視化結果である．２つの可視化結果は，相互に操作可能な状況で表示される．このとき左側の可視化結果を右京，右側の可視化結果を

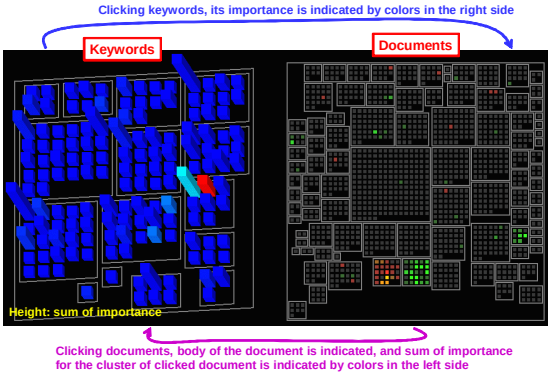

図 2 「左京と右京」による相互操作の例

左京と言う．左京の可視化結果で特定のデータをクリックすると，右京のデータの左京に対応する部分がカラーリングされる．同様に，右京の可視化結果で特定のデータ要素をクリックすると，左京の可視化結果の対応する部分がカラーリングされる．このように 2 つの可視化結果を相互操作することによって「左京と右京」は，大規模な表形式データの内容を探索することのできる新しい可視化技術である．

3.2.1 表形式データのクラスタリング

「左京と右京」における表形式データの定義を記述する（図 3 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．

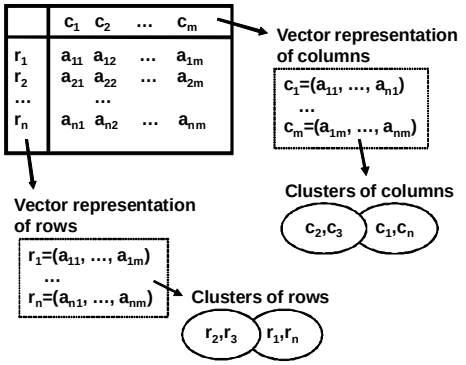

図 3 表形式データのの定義とクラスタリング

「左京と右京」おけるクラスタリングには，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も適用可能であるが，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．

3.2.2 平安京ビューによるクラスタリング結果の可視化

左京と右京では，前節で示した手法で生成された 2 つの階層型データの対して，「平安京ビュー」を適用して可視化を行う．「左京」は行を構成する n 個のデータ要素 r_1〜r_n で，同様に「右京」は列を構成する m 個のデータ要素 c_1〜c_m を可視化される．また，これらのデータ要素を角柱で表示される．

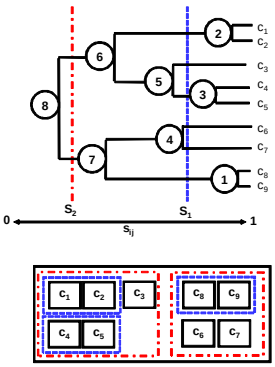

図 4 (上) 階層型クラスタリング (下) クラスタリング結果から得られる階層型データの可視化結果のイラスト

角柱を表示するための視覚的特性（色相，高さ，底面形状）は以下の関数の返り値によって制御される．

- 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 $f_3(b_3)$.

3.2.3 2 つの平安京ビュー間の操作

「左京と右京」では，利用者が対話的に表形式データを探索できるよう，「左京」と「右京」を相互に操作することが可能である．例えば，利用者が「左京」の角柱 r_i をクリックすると仮定すると，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，「左京と右京」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示したものが図 5 である．

4. 提案手法によるウェブ検索結果のクラスタ分布の可視化

本研究では前章で紹介した「平安京ビュー」「左京と右京」を用いてウェブ検索結果のクラスタ分布を可視化する．本章ではその処理手順を示す．

4.1 検索結果からの抽出ワード選出

本手法ではまず，検索結果から上位ページの情報を抽出する．現時点での我々の実装では，(注4)Google で 1 個のキーワードを

(注4)：http://www.google.com/

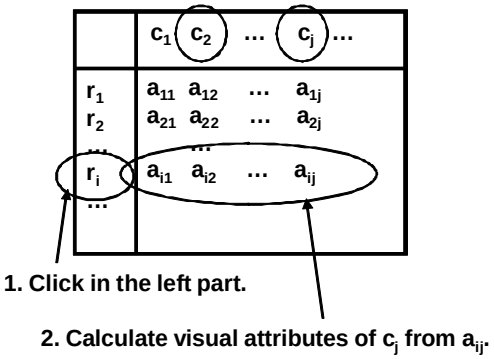

図 5　左京をクリックすることによる右京の表示を更新する内部処理手順

入力して得られた検索結果から，上位 1000 件のページの URL，タイトル，プレーンテキストを抽出している．

次に本手法では，全てのページのタイトルに対して形態素解析を実行し，得られた単語の重要度を算出する．現時点での我々の実装では，形態素解析エンジンに(注5) MeCab を適用し，専門用語の選出とその重要度算出に(注6) TermExtract を適用している．ここで Web ページの本文ではなくタイトルからワードを抽出する理由は，そのページ内容を代表する短い文字列の中からのほうが，ノイズの影響を受けることなく重要な単語を見つけられる可能性が高いと考えられるためである．

そして得られた専門用語の中から，重要度計算結果の高い 50 語程度を抽出し，キーワードとして登録する．この 50 語のキーワードを本論文では「抽出ワード」と呼ぶ．

以上の「1000 ページ」「50 語」という数字は，現在普及しているディスプレイの解像度に対して提案手法が適切に表現可能なデータ規模に相当するものであり，今後のディスプレイ技術の発展によってさらに多くのページ数や単語数に対応できるものと考える．

4.2　抽出ワードと Web ページのクラスタリング

続いて本手法では，抽出ワードと Web ページに対してクラスタリングを実行する．まず本手法では，前節で述べた手法で抽出した 1000 件の Web ページの全ての本文を対象として，各ページにおける各抽出ワードの重要度を算出する．現時点での我々の実装では，ここでも重要度算出に TermExtract を適用している．

続いて本手法では，3.2.1 項で論じた手順によって，抽出ワードと Web ページに対してクラスタリングを実行する．ここで i 番目の抽出ワードを r_i とし，j 番目の Web ページを c_j とし，j 番目の Web ページの本文における i 番目の抽出ワードの重要度を a_{ij} とする．

4.3　クラスタリング結果の可視化

続いて本手法では，3.2.2 項で論じた手順によって，各々のクラスタリング結果を可視化する．ただし我々の実装では，Web ページ側 (図 7 右) の可視化には従来の平安京ビューを適用し，抽出ワード側 (図 7 左) の可視化には平安京ビューの四角いアイコンを文字に置き換えたタグクラウドを適用する．タグクラウドとは、図 6 のように重要なキーワードの文字サイズや色を強調することで、全体像をできる表示方法である．

我々の実装では，抽出ワード r_k の各 Web ページにおける重要度の合計値 $\sum_{j=1}^{1000} a_{kj}$ を求め，この値が大きい抽出ワードを大きく，値が小さい抽出ワードを小さく表示する．

図 6　(注7) WordleWordle によって作成されたタグクラウド

5.　実　行　例

本章では「デフォルト」というキーワードに対する検索結果を例にして，提案手法の実行例を示す．デフォルトという単語を選んだ理由は，「初期値，債務不履行」などの意味の他に，「ブレイブリーデフォルト」といった名前のゲームも存在するため，多種多様な Web ページを収集できると考えられるためである．

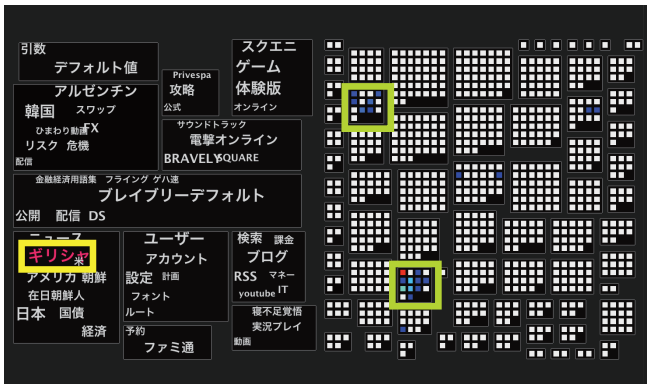

図 7　可視化の実行結果

利用者が左側の「ギリシャ」というワードをクリックしたとする．すると右側の Web ページ一覧の中から，黄色の枠で囲まれた「ギリシャ」というワードに関連するページが色付けされる．このとき右側のノードの色は，そのワードに対する重要度の高さによって変わる．ノードの色が赤に近いほど重要度が高く，青いものほど重要度が低い．

また逆に右側の黄色の枠で囲まれたノードをクリックすると，そのノードの対応する Web ページの中の，左側の抽出ワードの関連するものだけが色付けされる．このとき右側と同様に，右側のページに対する重要度の高いワードが赤く，低いものが

(注5)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
(注6)：http://gensen.dl.itc.u-tokyo.ac.jp/

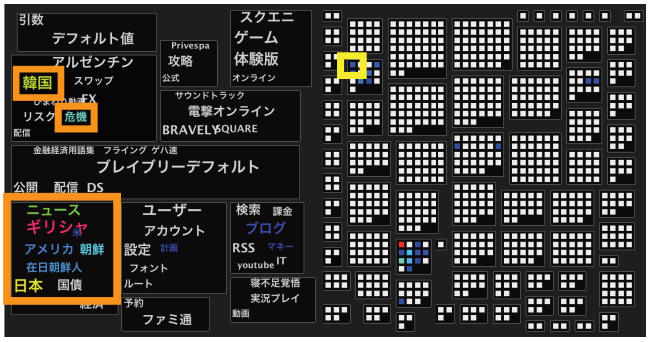

図 8　右側の可視化結果を操作した例

青く表示される．

図 8 の左側の可視化結果を見ると，右側のページの内容に関連するワードは黄緑の枠で囲まれたキーワードグループに集中しており、その中をみてみると、国の名前、「国債」、「経済」といったキーワードが見て取れる．ここから黄色の枠で囲まれたノードの示すページは、様々な国の経済について述べているページだということが推測できる．

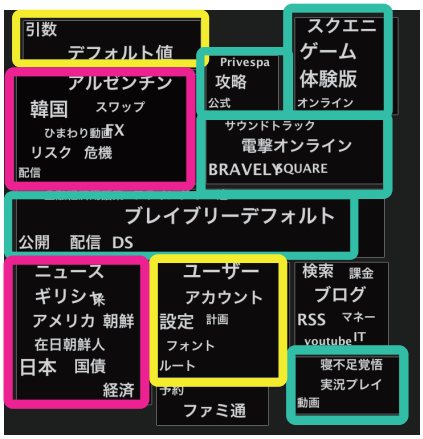

図 9　クラスタリング結果を単語の意味ごとに色分けした図

図 9 の可視化結果を見ると，キーワードがゲームに関するグループ，「初期値」という意味を持ったデフォルトに関するグループ，経済に関するグループに分かれているのがみて取れる．

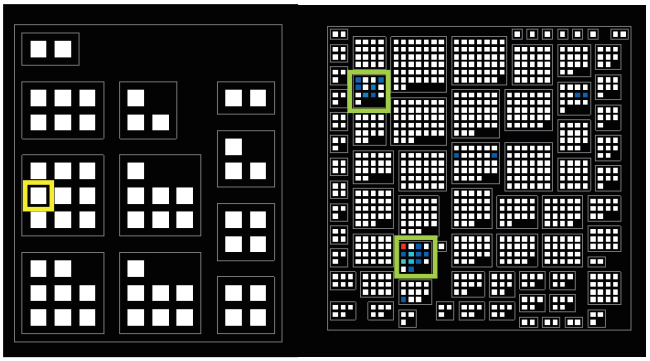

図 10　「左京と右京」で「デフォルト」を可視化した結果

図 8 と同じ結果を、「左京と右京」で可視化した結果が図 10 である．図 7 同様、図 10 は「ギリシャ」というキーワードをクリックした結果であるが、図 10 の左側の可視化結果に比べて、図 7 ではひと目で検索キーワードの全体像が見て取れるようになったことがわかる．

図 11　Clusty で「デフォルト」を検索した結果

また、図 11 は、2.3 項で紹介した「Clusty」で「デフォルト」を検索した結果である．左側のグループをみると、「債務不履行」、「初期値」という意味の項目しかあがっていないことがわかる．これに対して提案手法では、辞書に存在しないようなキーワード、この場合「ブレイブリーデフォルト」などといったキーワードも除外されずクラスタリングされていることがわかる．

6. ま と め

本報告では，Web 検索結果であるページ群と，そのページ群に含まれるキーワード群を行および列とした表をつくり，そのクラスタリング結果を可視化する一手法を提案した．本手法ではページ群とキーワード群の各々をクラスタリングした結果を 2 つの木構造として可視化し，その連携操作によってページ群とキーワード群との関連性を対話的に探索できる仕組みを提供する．本報告では「デフォルト」という単語に対する検索結果を例にして，その実行結果を示した．

今後の課題として，クラスタリング手法の改善や，タグクラウドの配置アルゴリズムの改善などに着手したい．またユーザテストによって本手法の有効性を検証したい．

文　献

[1] 吉田, 小山, 中村, 田中, Web 検索結果におけるキーワード出現相関の可視化と対話的な質問変換, 第 18 回データ工学ワークショップ (DEWS2007), C7-2, 2007.

[2] 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.

[3] Carriere J., et al., Research Paper: Interacting with Huge

Hierarchies beyond Cone Trees, IEEE Information Visualization 95, pp. 74-81, 1995.

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

[5] 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.

[6] Balzer M., Deussen O., Voronoi Treemaps, IEEE Symposium on Information Visualization, pp. 49-56, 2005.

[7] 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.

[8] 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.

[9] 馬場，新里，柴田，黒橋, キーワート蒸留型クラスタリングによる大規模ウェブ情報の俯瞰，情報処理学会論文誌, Vol. 50, No. 4, 1399-1409, 2009.

[10] 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, 2008.