

# Web 検索結果のクラスタ分布の可視化の一手法

松枝知香 (指導教員：伊藤貴之)

## 1. はじめに

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 Clusty [1], Mooter [2] 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。

本研究では、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、可視化する一手法を提案する。

## 2. 本研究が採用する可視化手法

### 2.1 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 [3] は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、データの全体を一面面に表示することを目標とした階層型データ可視化手法である。

「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。

- 幅や深さの不均一な階層型データにも適用できる
- データ全体の画面占有面積が小さくなるように配置
- 葉ノードが同じ大きさで表示されるように配置

### 2.2 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 [4] 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。

「左京と右京」の行と列の可視化結果は、相互に操作可能な状況で表示される。この手法は、2つの可視化結果を相互操作することによって、大規模な表形式データの内容を探索することのできる可視化技術である。

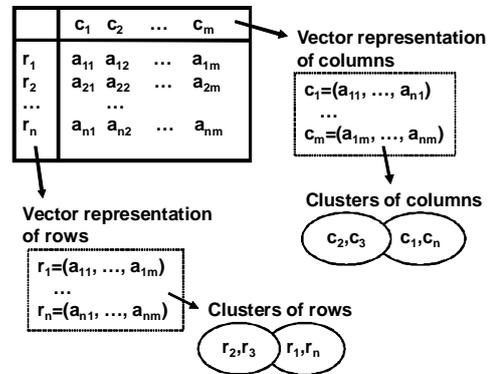


図1 表形式の定義とクラスタリング ([4] より転載)

### 2.2.1 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1参照)。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、 $a_{11} \sim a_{nm}$  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  $c_i = (a_{i1}, \dots, a_{in})$  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  $r_j = (a_{j1}, \dots, a_{jm})$  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。

### 2.2.2 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  $r_1 \sim r_n$  で、同様に「右京」は列を構成する  $m$  個のデータ要素  $c_1 \sim c_m$  を可視化する。また、これらのデータ要素を角柱で表示する。

### 2.2.3 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  $c_j$  に対する情報を抽出し、その情報を元に構成する棒グラフの色、高さ、底面形状を計算し、更新する。

## 3. 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。

### 3.1 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、Google [5] で1個のキーワードを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 MeCab [6] を適用し、専門用語の選出とその重要度算出に TermExtract [7] 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からである。

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研究では「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る。

### 3.2 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、2.2.1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。

### 3.3 クラスタリング結果の可視化

続いて本手法では、2.2.2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2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2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、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。

## 4. 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。

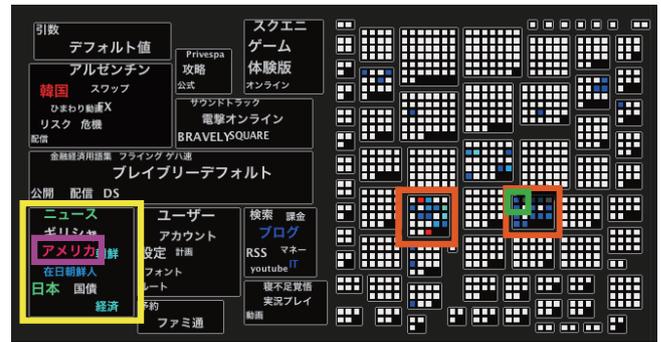


図2 右側の可視化結果を操作した例

図2は、提案手法の実行例である。利用者が左側の紫の枠で囲まれた「アメリカ」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、「アメリカ」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。

また逆に右側の緑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、左側の抽出ワードに関連するものだけが色付けされる。

図2の左側の可視化結果を見ると、緑の枠で囲まれたページの内容に関連するワードは赤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。ここから緑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。また、図2の左側の可視化結果は、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。

## 5. まとめ

本研究では、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。本研究では「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] <http://clusty.com/>
- [2] <http://www.mooter.co.jp/>
- [3] 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可視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.
- [4] 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.
- [5] <https://www.google.co.jp/>
- [6] <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>
- [7] <http://gensen.dl.itc.u-tokyo.ac.jp/>