修　士　論　文

題　目　評価構造における単語間の関係性可視化に関する研究

指導教員　小山田 耕二 教授

京都大学大学院 工学研究科 電気工学専攻

氏 名　小澤 啓太

平成 28 年 2 月 4 日

目 次

第 1 章 提案手法

本章では, 提案システムで使用する評価構造のテキストデータの説明と, 評価構造のテキストデータから単語の Word Cloud 内座標を計算するまでのプロセスについて述べる.

1.1 使用データ

評価構造データ向けのテキストベース可視化を行うにあたって，評価グリッド法による評価構造データの作成を行った．本研究では, E-Grid を用いて評価構造データの作成を行った．E-Grid とは, 評価グリッド法に基づいた評価構造の抽出や分析をビジュアル的に支援するシステムであり, 評価構造を抽出する機能から評価構造データを作成した. E-Grid でのインタビューは基本的に 1 人ずつで行い，個人の評価構造を作成する．その後，回答者全体の評価構造を把握するため，個人毎の評価構造を統合し全体の評価構造を作成する．E-Grid により作成される評価構造データは評価項目と隣接する評価項目，評価項目の回答者数 (＝重み)，評価項目のテキストラベル, 評価項目の回答者名等が格納されている．提案する可視化手法は評価項目の重み，隣接する評価項目の情報，評価項目内のテキストラベルの情報を使用する．提案手法では文章ではなく単語ごとに可視化する必要があるので，テキストラベルに形態素解析を行った．形態素解析には日本語形態素解析システム MeCab[18] と形態素解析辞書 Unidic を使用した．Unidic は, 語彙素・語形・書字形・発音形の階層構造を持つので，表記の揺れや語形の変異にかかわらず同一の単語と識別することが可能である．提案手法では, 表記の揺れや語形の変異がある単語はそれぞれの単語の語彙素を表示する．形態素解析により抽出した単語の中で品詞が名詞・動詞・形容詞のいずれかである単語を提案システムでは使用する. 以上から得られたデータに以下のプロセスを通し可視化を行う.

1.2　座標計算

提案システムでの可視化プロセスは大きく 3 つに分けられる. はじめに, 評価構造データから評価構造内単語間の距離行列を作成する. 次に, 評価構造内単語間の距離関係を保持したまま単語を Word Cloud 可視化するため, 次元削減法を用いて各単語の x, y 座標を計算する. 最後に, 計算した座標では単語間の重複が発生する場合があるので, 単語の重複を阻止し, しかし単語間の位置関係を崩さず, 指定領域を最大限活かすことのできる単語の座標を再計算する. Word Cloud に可視化する際は再計算された座標を使用する. 以下の項では, 可視化プロセスの詳細を説明する.

1.2.1　単語間距離行列の作成

前章で記述したとおり，提案手法では Word Cloud 内の単語の座標を評価構造内での単語の距離関係を考慮して決定する．評価構造内の単語間の距離関係は単語が使用されている頂点の距離関係から計算され，距離行列 D で表す．単語間距離行列 D は，評価構造から計算される出現行列 F，頂点間距離行列 U，重み行列 A から計算される．Fig?は評価構造と評価構造から取得された出現行列 F，頂点間距離行列 U，重み行列 A の例を表す．以下では頂点間距離行列 U，出現行列 F，重み行列 A の導出方法を述べる．頂点群は $N \in n_1, n_2, ..., n_l$, 単語群は $W \in w_1, w_2, ..., w_m$, で表す。また頂点 n_iを回答した人数を a_i、単語 w_iの全頂点での使用回数を u_iと表す。

頂点間距離行列

頂点間距離行列 U は，各頂点間の距離関係を表す行列である. ネットワーク図の頂点 Vi と Vj の最短距離の経路でノードを n 個経由する場合, 第 i 行第 j 列と第 j 行第 i 列の要素が n+1 となる行列である．また、対角成分の値は全て 0 である。最短経路の計算は幅優先探索アルゴリズムを使用した. このアルゴリズムは横型探索とも言われ, 以下のルールに従って探索を行う.

1. 根ノードを空のキューに加える.

2. ノードをキューの先頭から取り出し, 以下の処理を行う.

- ノードが探索対象であれば, 探索をやめ結果を返す.
- そうでない場合, ノードの子で未探索のものを全てキューに追加する.

3. もしキューが空ならば, グラフ内の全てのノードに対して処理が行われたので, 探索をやめ”not found”と結果を返す.

4. 2 に戻る.

$$U = \begin{pmatrix} u_{11} & \dots & u_{1l} \\ \vdots & \ddots & \vdots \\ u_{l1} & \dots & u_{ll} \end{pmatrix} \tag{1.1}$$

出現行列

出現行列 F は頂点と単語の関係を表す行列である．出現行列の列は評価構造内の頂点を, 行は全頂点から形態素解析を行い取り出したの全ての単語群を表す．i 行目の単語が j 列目の頂点の文章に出現しない場合は 0 となる．出現する場合、i 行目の単語が n 個のノードで使用されていると第 i 行第 j 列の要素は 1/n となる。要素の値を使用回数の逆数とすることで単語が複数のノードで出現する際に単語間距離を平均の距離となる。この行列から各頂点に出現する単語を知ることができる.

$$F = \begin{pmatrix} f_{11} & \dots & f_{1l} \\ \vdots & \ddots & \vdots \\ f_{m1} & \dots & f_{ml} \end{pmatrix} \tag{1.2}$$

$$f_{ij} = \begin{cases} \frac{1}{u_i} & (n_i \in w_j) \\ 0 & (otherwise) \end{cases}$$

重み行列

各頂点の重み行列 A は，各頂点の回答者数を表す行列である．対角成分以外の値は全て 0 とし、対角成分は回答者数の逆数とする。i 番目のノードの回答者数が n の場合、第 i 行第 i 列の要素は 1/n となる。提案手法では、頻出語の関係性をよ

り詳細に可視化するため、頻出語ほど距離行列の値が小さくなるよう設定した。

$$A = \begin{pmatrix} a_{11} & \dots & a_{1l} \\ \vdots & \ddots & \vdots \\ a_{l1} & \dots & a_{ll} \end{pmatrix} \tag{1.3}$$

$$a_{ij} = \begin{cases} 0 & (i \neq j) \\ \frac{1}{a_i} & (i = j) \end{cases}$$

上記の 3 行列 F，U，A を計算した後，以下の式によって単語間距離行列 D を計算する．ネットワーク図内の単語 w_iと w_jの最短距離の平均値が, 第 i 行第 j 列と第 j 行第 i 列の値となる行列である．

$$D = \frac{1}{2} FAUA^{\mathrm{T}}F^{\mathrm{T}} = \begin{pmatrix} d_{11} & \dots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{m1} & \dots & d_{mm} \end{pmatrix} \tag{1.4}$$

1.2.2 次元削減

次に, 評価構造内単語間の距離関係を保持したまま単語を Word Cloud 可視化するため, 単語の x, y 軸の値を求めるために次元削減を行う．次元削減を行うことで，単語数の行と 2 列の行列式を計算し，この 2 次元の値を単語の x, y 軸の値にする．提案システムでは多次元尺度構成法を適用することで次元削減を行う．多次元尺度構成法とは, 多変量解析の一手法である．分類対象物の関係を低次元空間における点の布置で表現する．多次元尺度法では, 始めにヤング・ハウスホルダー変換を施す．ヤング・ハウスホルダー変換は距離の 2 乗の行列に両側から中心化行列をかける演算であり, 以下の式で表す.

$$P = -\frac{1}{2} JDJ^{\mathrm{T}} \tag{1.5}$$

単語数を n とすると,行列 J は単位行列から全要素に $1/n$ の行列を引いた $n \times n$ 行列を引いた $n \times n$ 行列 行列 D は点 x_iと x_jの間の距離 d_{ij}の 2 乗を要素とする $n \times n$ 行列である．次に, P をスペクトル分解することで固有値, 固有ベクトルを求め, 固有値の大きい方から 2 つ取り, 対応する固有ベクトルを取り出す．各単語の x,y 座標は取り出した 2 つの固有ベクトルの値となる.

1.2.3 重複阻止

最後に, 計算した座標では単語間の重複が発生するので, 単語間の相対的位置関係を崩さないように単語の重なりの除去を行う. 提案手法では Erick らが提案した重複阻止手法を参考にした.[17] Erick らの提案手法では指定領域内で格子を生成し, 格子の 1 区画を細胞と呼ぶ. そして座標点が存在する細胞だけを残し, 細胞の位置座標を最適化計算する手法である. この手法では, 細胞間の相対的位置関係保持, 細胞間の重複阻止, 指定領域の最大利用を達成する最適な細胞の座標を計算する. 提案手法では単語内の 1 文字を格子の 1 細胞, 各単語を細胞が隣接する長方形と置き換え, 単語の位置の最適化計算を行う. 提案手法では, 正方形の細胞ではなく長方形の細胞の座標最適化計算なので一部制約条件を修正した. 以下の節では最適化計算で用いる目的関数, 制約条件の説明を行う.

指定領域に配置する単語の集合を $G = g_1, g_2, \cdots, g_n$ とし, 単語の各文字, すなわち各細胞の集合を $H = h_{11}, h_{12}, \cdots, h_{1m}, h_{21}, \cdots, h_{nm}$ とする. 各細胞は正方形であり, $h_{ij} = (x_{ij}, y_{ij}, w_{ij})$ で表される. (x_{ij}, y_{ij}) は正方形 g_{ij} の中心点, $w_{ij} > 0$ は正方形 g_{ij} の辺の長さである. w_{ij} は $w_{ij} = \alpha_i\delta$,α_i は $\alpha_i = e_i(max(e_1, e_2, \cdots, e_n))$ と表し, e_i は単語 g_i の評価構造内での出現頻度を表す. すなわち $w_{i1} = w_{i2} = \cdots = w_{im}$ となる.δ は再編成後の最頻出語の辺の長さを表す. 各細胞は縦, 横の辺の長さが $W \times H$ である指定領域内で細胞間の大きさの比率を維持しつつ, 細胞間の重複阻止と指定領域の最大利用, 細胞間の相対的位置関係の保持を達成するように再編成される.

目的関数

指定領域内で単語を再編成するために, 提案手法では単語の再編成を以下のように最適化問題として定式化した.

$$
\begin{aligned}
minimize\ E(\mathbf{z}) &= E_{comp}(\mathbf{z}) + E_{resize}(\mathbf{z}) \\
subject\,to\ A\mathbf{z} &\leq \mathbf{b}, \mathbf{z} = [\mathbf{x}\,\mathbf{y}\,\mathbf{r}\,\delta]^{\mathbf{T}}, \\
\mathbf{x} &= (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N})^{\mathbf{T}} \in \mathbf{R^N} \\
\mathbf{y} &= (\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_N})^{\mathbf{T}} \in \mathbf{R^N} \\
\mathbf{r} = (\mathbf{r_{12}}, ..., \mathbf{r_{1N}}, \mathbf{r_{23}}, ..., \mathbf{r_{2N}}, ..., \mathbf{r_{N\,-\,1N}})^{\mathbf{T}}&, \mathbf{r_{ij}} \in \mathbf{0}, \mathbf{1} \\
\delta &\leq min(W, H),
\end{aligned}
\tag{1.6}
$$

$\mathbf{z}$ は求める変数である．$\mathbf{x}, \mathbf{y}$ は細胞の重心座標, $\mathbf{r}$ は単語の重複を防ぐための調整変数,δは再編成後の最頻出語の辺の長さを表す変数である．$A, \mathbf{b}$ は最適化問題の制約条件を表す行列である．エネルギー項 $E_{comp}(\mathbf{z}), E_{resize}(\mathbf{z})$ はそれぞれ単語と指定領域を操作する関数である．前者は単語間重複と相対的位置関係の保持を考慮する単語集約エネルギー項, 後者は指定領域を可能な限り最大利用するための指定領域最大利用エネルギー項である．以下では 2 つのエネルギー項についての詳細を述べる．

単語集約エネルギー項　一つ目のエネルギー関数 $E_{comp}(\mathbf{z})$ は単語集約エネルギー項であり, この関数の目的は細胞を狭い範囲で簡潔に表すことである. エネルギー関数 $E_{comp}(\mathbf{z})$ は, 細胞の重心座標を変数とした二次関数で以下の式で表す.

$$E_{comp}(\mathbf{z}) = C \sum_{(i,j)} (x_i - x_j)^2 + (y_i - y_j)^2, \tag{1.7}$$

$$C = \frac{1}{(min(W,H) \times \frac{n(n-1)}{2})},$$

$(i,j) = (j,i)$ は単語 $(g_i, g_j), i, j \in 1, 2, \cdots, n$ の組み合わせを表す. x_i, y_iは単語 g_iの重心点の座標であり, $x_i = \frac{\sum_j x_{ij}}{m}, y_i = \frac{\sum_j y_{ij}}{m}$で表す. n は単語数を表し, m は単語 g_i の文字数である. C は標準化のための定数であり, 指定領域最大利用エネルギー項との値の調節を行う. $E_{comp}(\mathbf{z})$ は単語間の距離が短くなるほど小さくなり, 最小となるのは全単語の重心点が重なる場合である.

指定領域最大利用エネルギー項　二つ目のエネルギー関数 $E_{resize}(\mathbf{z})$ は指定領域最大利用エネルギー項であり, この関数の目的は指定領域を細胞で最大限満たすことである. 上記したように全ての細胞の辺の長さは, 最大の細胞の辺の長さに依存している．そこで提案手法では指定領域最大利用エネルギー関数を以下の二次間数で表す.

$$E_{resize}(\mathbf{z}) = (\delta - min(W,H))^2, \tag{1.8}$$

$min(W,H)$ とは指定領域の短辺の長さを表す. $E_{resize}(\mathbf{z})$ は細胞の辺が長くなるほど小さくなり, 指定領域の短辺の長さと最大の細胞の辺の長さが等しい場合に最小になる.

制約条件がない場合だと式 (?) のエネルギー関数は全細胞の重心が重なり, 最大の細胞の辺の長さが指定領域短辺の長さと等しい場合が最小となる．この状況で

は単語重複が達成されていないので 3 つの制約条件を設けた. 次項では 3 つの制約条件の詳細を説明する.

制約条件

式 (?) で表したように, $A, \mathbf{b}$ は制約条件を表す行列であり, 重複阻止, 相対的位置関係, 指定領域からのはみ出しの阻止の保持を行う. 各単語の初期座標は (x_i, y_i) と与えられており, 単語の相対的位置関係を保存するための制約条件は以下で表される.

$$\begin{aligned} x_{p1} \leq x_{p2} \leq ... \leq x_{pn} \Rightarrow x_{pi} - x_{pi+1} \leq 0 \\ y_{p1} \leq y_{p2} \leq ... \leq y_{pn} \Rightarrow y_{pi} - y_{pi+1} \leq 0 \end{aligned} \tag{1.9}$$

$p, q : \{1, 2, \cdots, n\} \rightarrow \{1, 2, \cdots, n\}$ は単語の初期座標 x_i, y_iを大きさ順に並び替えるために使用した.

重複阻止は以下の制約条件を用いて行われる.

$$\begin{aligned} |x_j - x_i| \geq \frac{(\alpha_i n_i + \alpha_j n_j)}{2}\delta, \\ or \\ |y_j - y_i| \geq \frac{(\alpha_i + \alpha_j)}{2}\delta, \end{aligned} \tag{1.10}$$

式 (?) から $x_j \geq x_i, y_j \geq y_i$が成り立つ場合, $|x_j - x_i| = x_j - x_i, |y_j - y_i| = y_j - y_i$となり, 式 (?) は以下のように表すことができる.

$$\begin{aligned} x_j - x_i \leq -\frac{(\alpha_i n_i + \alpha_j n_j)}{2}\delta, \\ or \\ y_j - y_i \leq -\frac{(\alpha_i + \alpha_j)}{2}\delta, \end{aligned} \tag{1.11}$$

この数式の or 条件はバイナリ値 $r_{ij} \in \{0, 1\}$ を用いて次式のように表すことができる.

$$x_j - x_i \leq \alpha_{ij}\delta + M r_{ij} \Leftrightarrow y_j - y_i \leq \alpha_{ij}\delta + M(1 - r_{ij}), \tag{1.12}$$

$$\alpha_{ij} = -\frac{(\alpha_i n_i + \alpha_j n_j)}{2} \tag{1.13}$$

M は値のとても大きい定数, n_iは単語 i の文字数を表す. r_{ij}は式 (?) の不等式の片方が成立する場合, もう片方の式を成立させる. 例えば, $x_j - x_i \leq \alpha_{ij}\delta$が成り立つ

場合, r_{ij}を 0 とする. その場合, $y_j - y_i \leq \alpha_{ij}\delta + M(1 - r_{ij})$ は y がどんな値であろうと成り立つ

最後に 単語が指定領域外に出ることを防ぐために以下の制約条件を設定した.

$$\begin{aligned} 0 \leq x_i - \frac{a_i n_i}{2}\delta \, and \, x_i + \frac{a_i n_i}{2}\delta \leq W, \\ 0 \leq y_i - \frac{a_i}{2}\delta \, and \, y_i + \frac{a_i}{2}\delta \leq H, \\ i = 1, 2, ..., N, \end{aligned} \tag{1.14}$$

この 3 つの制約条件により, 重複阻止と指定領域の最大利用, 細胞間の相対的位置関係の保持を満たした単語の再編成が達成される.

1.3 計算方法

提案手法では, 単語間距離行列を計算したが, この計算は python を用いた. 単語間距離行列を求めるために頂点間距離行列を求めたが, その際の幅優先探索は MATLAB を用いて行った.

また, 提案手法では単語の重複阻止のために最適化計算を行っており, この計算は混合整数二次計画問題である. 最適化計算には, Gurobi Optimization Package* により提供されるソルバー Gurobi Optimizer を用いて行った. Gurobi Optimizer では, 最適化計算は混合整数二次計画問題は分枝限定法を使って解く. 分枝限定法とは最適化問題の最適解を求めるための汎用アルゴリズムであり, アルゴリズムは以下となる. P_0 を最適化問題 (最小値を求める問題), z を最適解の最小値, f(P) を目的関数の値, g(P) を目的関数の緩和問題の値とする.

1. 初期設定： $A = P_0, z = \infty$
2. $A = \phi$ ならば終了.
   - $z = \infty$: 問題 P_0は解を持たない
   - $z < \infty$: $f(P_0) = z$
3. 集合 A の中から，部分問題 P_iを選択
4. もし，緩和問題 P_i'が解を持たなければ,

* http://www.gurobi.com/

$A = A - P_i$として，2 へ戻る

そうでなけでば，

もし，$g(P_i') = f(P_i)$ として，P_iの最適値が得られたら，

$z = min\{z, f(P_i)\}$，$A = A - P_i$として，2 へ戻る．

そうでなければ

もし，$g(P_i') \geqq z$ ならば

$A = A - P_i$として，2 へ戻る．

そうでなければ

問題P_iを，部分問題$P_{i1}, ..., P_{ik}$に分解し，$A = (A - P_i) \cup \{P_{i1}, ..., P_{ik}\}$として，2 へ戻る．

参 考 文 献

1) 奥西智哉, 炊飯米を生地に添加したパンの官能評価. 日本食品科学工学会誌, 56, 424-428, (2009).

2) 入江正和, 豚肉質の評価法. 日本養豚学会誌, 39, 221-254, (2002).

3) 来田宣幸, 赤井聡文. 野球における球速と球速感の関係. 日本認知心理学会発表論文集, 42-42, (2009).

4) 中前光弘, 順位法を用いた視覚評価の信頼性について: 順序尺度の解析と正規化順位法による尺度構成法. 日放技学誌, 56, 725-730, (2000).

5) 大山正, 瀧本誓, 岩澤秀紀. 順位法を用いた視覚評価の信頼性について: 順序尺度の解析と正規化順位法による尺度構成法. 行動計量学, 20, 55-64, (1993).

6) J. Sanui, Visualization of users ' requirements: Introduction of Evaluation Grid Method, Proceedings of the 3rd Design and Decision Support System in Architecture and Urban Planning Conference, 365-374, (1996).

7) 讃井純一郎, 乾正雄. レパートリー・グリッド発展手法による住環境評価構造の抽出：認知心理学に基づく住環境評価に関する研究 (1). 日本建築学会計画系論文報告集, 367, 15-22, (1986).

8) 尾上洋介, 久木元伸如, 小山田耕二. 可視化情報学会における会員満足度の因果関係分析. 可視化情報学会論文集, 34, 43-51, (2014).

9) 本村陽一, 金出武雄. ヒトの認知・評価構造の定量化モデリングと確率推論. 電子情報通信学会技術研究報告, 104, 25-30, (2005).

10) G. A. Kelly, The Psychology of Personal Constructs, 1 and 2, (1955).

11) Y. Onoue, N. Kukimoto, N. Sakamoto, K. Koyamada, Network Coarse-Graining for Evaluation Structures, In Proc. of International Conference on Simulation Technology, 34, 447-450, (2015).

12) 樋口耕一, テキスト型データの計量的分析. 理論と方法, 19, 101-115, (2004).

13) Riehmann. P, Gruendl. H, Potthast. M, Trenkmann. M, Stein. B, Froehlich. B, WORDGRAPH: Keyword-in-Context Visualization for NETSPEAK's Wildcard Search. Visualization and Computer Graphics, IEEE Transactions on 18.9, 1411-1423, (2012).

14) Strobelt. H, Spicker. M, Stoffel. A, Keim. D, Deussen. O, Rolled- out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives, Computer Graphics Forum, 31, 1135-1144, (2012).

15) Huang. X, Lai. W, Force-transfer: a new approach to removing overlapping nodes in graph layout, Proceedings of the 26th Australasian computer science conference, 16, 349-358, (2003).

16) Gomez-Nieto. E, San Roman. F, Pagliosa. P, Casaca. W, Helou. E. S, Oliveira. M. C. F, Nonato. L. G, Similarity Preserving Snippet-Based Visualization of Web Search Results, Visualization and Computer Graphics, IEEE Transactions on, 20, 457-470, (2014).

17) Gomez-Nieto. E, Casaca. W, Motta. D, Hartmann. I, Taubin. G, Nonato. L, Dealing with Multiple Requirements in Geometric Arrangements, Visualization and Computer Graphics, IEEE Transactions on, 1, (2015).

18) T. Kuo, K. Yamamoto, Y. Matsumoto, Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 conference on empirical methods in natural language processing, 230-237, (2004).

19) 尾上洋介, 評価構造のビジュアル分析に関する研究, 博士論文, 2016.