1E4-1

グラフ構造に着目した評価文書の極座標可視化法

Polar Coordinate Visualization Method for Review Documents Focusing on Graph Structure

伏見 卓恭 *1 Takayasu FUSHIMI　佐藤 哲司 *1 Tetsuji SATOH　斉藤 和巳 *2 Kazumi SAITO　風間 一洋 *3 Kazuhiro KAZAMA

*1筑波大学　図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba

*2静岡県立大学　経営情報学部
School of Management and Information, University of Shizuoka

*3和歌山大学　システム工学部
Faculty of Systems Engineering, Wakayama University

In this paper, we propose a novel visualization framework for evaluation documents like review text of commodities or services.

1. はじめに

近年，ソーシャルメディアの台頭により，ユーザは商品やサービスに対するレビュー文などの評価文書を容易に投稿することができるようになった．そのため，Web 上には様々な種類の評価文書が蓄積されている．評価文書の中には，同一カテゴリの他社製品，同一メーカーのシリーズ商品など他の商品と比較したコメントや，ユーザの背景知識，苦情，要求など多様な情報が含まれている．

本研究では，こうした評価文書を適切に可視化することで，対象商品の評価の全体像を直感的に把握でき，かつ所望の評価文書へのアクセシビリティを高めることを目標とする．そのために，各商品に対する評価文書を，その文書が表す意図・トピック・評価観点などの項目に従って放射状に，各評価文書と共に投稿された商品に対するスコアに従って同心円状に可視化する手法を提案する（図 1）．

2. 関連研究

Web 上の評価文書に関する研究は，評価表現·意見抽出に関する研究 [Turney 02] や評価観点の抽出に関する研究 [Titov 08] に端を発し，評価極性の定量化 [Scaffidi 07]，評価文書の可視化 [Oelke 09] など多様な研究が展開されている．これらの研究の背景には，サイバー空間での購買行動においてユーザの商品選択の手がかりが，既に商品を購入したユーザによる評価文書であることが社会的な背景としてある．

評価文書の可視化に関する既存研究では，離散的に評価観点を抽出し，タグクラウドや棒グラフ，レーダーチャート形式でプロットする手法が多く見られる [Oelke 09, Wu 10]．抽出した評価観点の数には恣意性があったり，その数が多いほど解釈困難な結果となる．また，隣接する軸の関連性については言及されていない．よって，ユーザの商品選択において有用な可視化結果を出力するには限界があると言える．本研究では著者らの [Fushimi 11] をベースに，類似の要点（評価観点や主張点）の評価文書が近傍にプロットされるような可視化手法を提案する．抽出した離散的な評価観点を軸にする関連研究とは異なり，評価文書群に内在する要点が各文書の配置座標（偏角）を決定する（図 1 参照）．多次元尺度法などの可視化法 [Torgerson 52, 小林 14] と異なり，任意の偏角で放射状に評価軸をとり，スコアにより半径を決定する点で，従来の可視化手法と異なる．

連絡先: 伏見卓恭，筑波大学図書館情報メディア系，茨城県つくば市春日 1-2，029-859-1391

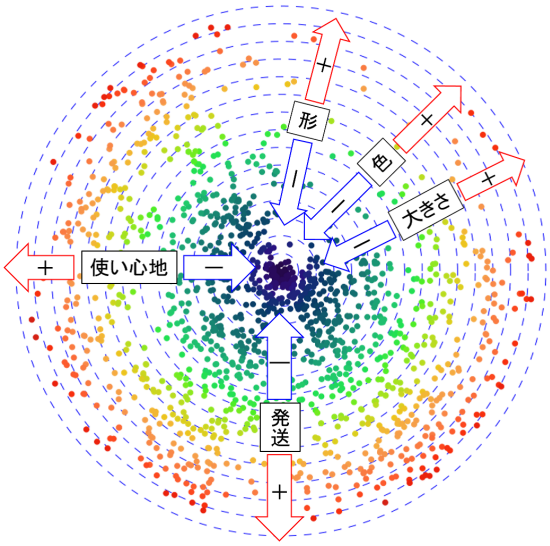


図 1: 提案可視化法のイメージ図

3. 提案手法

提案手法の枠組みでは，評価文書集合を $\mathcal{D} = \{1, i, \ldots, N\}$，単語集合を $\mathcal{W} = \{1, j, \ldots, M\}$，とし，各評価文書を Bag of Words の単語頻度ベクトル $\mathbf{a}_i$ で表現する．$a_{i,j}$ は，$\mathbf{a}_i$ の j 番目の要素であり，評価文書 i に単語 j が出現した回数である．一般に，文書数 $N = |\mathcal{D}|$ と比較して単語数 $M = |\mathcal{W}|$ は非常に大きくなり，次元の呪いや計算量の点で好ましくない．したがって，入力された単語頻度ベクトルを次元圧縮したベクトルを計算する．次元圧縮して得られる各次元を擬似単語と呼ぶことにする．擬似単語は，次元圧縮手法により性質が異なるが，LDA などのトピックモデルを用いることで，トピックの含有率を表現することが可能である．次いで，著者らの手法

[Fushimi 11] をベースとし，評価文書・擬似単語からなる重み付き 2 部グラフを極座標平面上に評価文書を可視化する PCE（Polar Coordinate Embedding）法を提案し，評価文書の布置座標を計算する．最後に，布置された評価文書に対して，どの方向にどんな内容の評価文書が集まっているかを表現するアノテーションベクトルを計算し，可視化結果にアノテートする．

提案法の枠組みを以下にまとめる：

1. 次元圧縮：評価文書と擬似単語の重み付き 2 部グラフを構築する；
2. PCE 法：重み付き 2 部グラフを極座標平面に布置する；
3. アノテーション：布置結果にアノテートベクトルを付与する；

各手順の詳細を次節以降で説明する．

3.1 次元圧縮

本稿では，以下に示す 3 つのトピックモデル [Stevens 12]（次元圧縮手法）を用いる．各手法において，入力された単語頻度ベクトル群 $\mathbf{A} = [a_{i,j}]_{i=1,j=1}^{N,M}$ は，$M \times S$ の行列 $\mathbf{W}$ と $S \times N$ の行列 $\mathbf{H}$ の形で表現され，各評価文書は，S 次元の擬似単語ベクトルに圧縮される．

- SVD（Singular Value Decomposition）は，以下のように，与えられた行列 $\mathbf{A}$ をよりサイズの小さな行列の積に分解する．

  $$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T.$$

  $\mathbf{U}$ と $\mathbf{V}$ の行ベクトルは，正規直交化された特異ベクトルであり，$\mathbf{\Sigma}$ は特異値を対角要素に持つ対角行列である．$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}$ ， $\mathbf{H} = \mathbf{\Sigma}\mathbf{V}^T$ とする．

- NMF（Non-negative Matrix Factorization）は，以下のように，与えられた行列を 2 つの行列の積に分解する．

  $$\mathbf{A} = \mathbf{W}\mathbf{H}.$$

  得られた行列をそのまま $\mathbf{W}, \mathbf{H}$ とする．SVD との顕著な違いは，分解後の行列要素の値が非負値である点である．

- LDA（Latent Dirichlet Allocation）は，各文書が複数のトピックから成ることを仮定した確率モデルであり，各文書 i のトピック分布 $\boldsymbol{\psi}_i$ とトピック s における各単語生成確率 $\boldsymbol{\phi}_s$ を collapsed ギブスサンプリングにより推定する．推定した文書 i におけるトピック s の含有率 $\psi_{s,i}$ を行列 $\mathbf{H}$ の (s,i) 要素, トピック s における単語 j の生成確率 $\phi_{j,s}$ を行列 $\mathbf{W}$ の (j,s) 要素とする．

上述した 3 つのトピックモデルにより得られた行列を $\mathbf{B} = \mathbf{H}$ とする．S 擬似単語と N 文書それぞれをノード，擬似単語 s と文書 i 間の値 $b_{s,i}$ を重み付きリンクとした二部グラフとして扱う．以下，擬似単語ノードを単に単語ノードと呼び，変数は j で表現する．

3.2 Polar Coordinate Embedding 法

単語ノード，文書ノードはそれぞれ半径 r_1，r_2 の円上に配置し，同心円上に可視化する．まず，円上の最適配置（角度）を計算する．具体的には，重み行列 $\mathbf{B} = [b_{j,i}]_{j=1,i=1}^{S,N}$ に対して多次元尺度法などと同様に中心化を施し，中心化重み行列 $\tilde{\mathbf{B}} = [\tilde{b}_{j,i}]$ を得る．$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_S]^T$ および $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]^T$ を適切に初期化した座標行列を基に反復的に求める．但し，$\mathbf{x}_j$ は，単語ノード j の布置座標ベクトルを表し，$\|\mathbf{x}_j\| = r_1$，$\mathbf{y}_i$ は，文書ノード i の布置座標ベクトルを表し，$\|\mathbf{y}_i\| = r_2$ である．

$$\begin{aligned} J(\mathbf{X},\mathbf{Y}) &= \sum_{j=1}^{S}\sum_{i=1}^{N} \tilde{b}_{j,i} \frac{\mathbf{x}_j^T}{r_1}\frac{\mathbf{y}_i}{r_2} \\ &\quad + \frac{1}{2}\sum_{j=1}^{S} \lambda_j (r_1^2 - \mathbf{x}_j^T\mathbf{x}_j) \\ &\quad + \frac{1}{2}\sum_{i=1}^{N} \mu_i (r_2^2 - \mathbf{y}_i^T\mathbf{y}_i). \end{aligned} \tag{1}$$

ここで，λ_j と μ_i は，各円周上に配置するための制約を表すラグランジュ乗数である．式 (1) において $\frac{\mathbf{x}_j^T}{r_1}\frac{\mathbf{y}_i}{r_2} = \cos\theta_{j,i}$ であり，隣接するノードどうし（文書 i に出現する単語 j）がその重みに応じて原点から見て同じ方向に配置されることによって，$J(\mathbf{X},\mathbf{Y})$ は最大化される．すなわち，同じようなノードと隣接するノードどうし（共通の単語ノードを有する文書ノード）を同一方向に近くに，異なるノードと隣接するノードを遠くに配置する．

また，ベクトル群 $\mathbf{Y}$ を固定すれば，ベクトル $\mathbf{x}_i$ の最適配置は以下のように求まる．

$$\mathbf{x}_j = \frac{r_1}{\|\tilde{\mathbf{x}}_j\|}\tilde{\mathbf{x}}_j, \quad \tilde{\mathbf{x}}_j = \sum_{i=1}^{N} \tilde{b}_{j,i}\mathbf{y}_i \tag{2}$$

$\tilde{\mathbf{x}}_j$ は，単語ノード j が重み $\tilde{b}_{j,i}$ で隣接するノード i の現在の座標ベクトルの合成ベクトルである．そして半径 r_1 上にくるように正規化している．

同様に，ベクトル群 $\mathbf{X}$ を固定すれば，ベクトル $\mathbf{y}_i$ の最適配置は以下のように求まる．

$$\mathbf{y}_i = \frac{r_2}{\|\tilde{\mathbf{y}}_i\|}\tilde{\mathbf{y}}_i, \quad \tilde{\mathbf{y}}_i = \sum_{j=1}^{S} \tilde{b}_{j,i}\mathbf{x}_j \tag{3}$$

PCE 法のアルゴリズムを以下に示す．

1. ベクトル群 $\mathbf{X}$ と $\mathbf{Y}$ を初期化する；
2. ベクトル群 $\mathbf{Y}$ を固定し，ベクトル $\mathbf{x}_j$ を求める；
3. ベクトル群 $\mathbf{X}$ を固定し，ベクトル $\mathbf{y}_i$ を求める；
4. 目的関数 $J(\mathbf{X},\mathbf{Y})$ の変化が十分小さければ終了する；
5. （2）へ戻る；

このアルゴリズムは HITS アルゴリズム [Kleinberg 99] と類似した構造を持つことが分かる．ただし，ベクトル群に対して 2 重の中心化を施す点，および，正規化の施し方の点に特徴を持つ．提案アルゴリズムの 1 反復は，2 部グラフのリンク数に比例した計算量となる．よって，ネットワーク可視化の代表手法の一つバネモデル法 [Kamada 89] などの非線形最適化が必要な可視化法と比較して，高速な方法である．

上記の手順により，最適配置（角度）が求まった．次いで，各ノードの最適配置（半径）を求める．本稿では，評価文書とともに投稿された評価スコアにより半径を決定する．すなわち，スコアが 1 点の評価文書は半径 1 の円上に，5 点の評価文書は半径 5 の円上にプロットする．単語ノードは最内側の円にプロットする．

3.3 アノテーション

PCE 法により求めた N 個の文書ノードの D 次元 [*1] 座標ベクトル群 $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}^T$ を入力とする．ここで，座標ベクトルは平均が $\mathbf{0}$ に，各座標値の自乗和が 1 となるように正規化されているとする．すなわち，$\sum_{i=1}^{N} \mathbf{y}_i = \mathbf{0}$，任意の整数 d (ただし，$1 \leq d \leq D$) で $\sum_{i=1}^{N} y_{i,d}^2 = 1$ である．一方，アノテーション対象とする文書ノードの属性値を各文書に出現する単語 j に関する単語頻度を用いて $\mathbf{z} = [a_{1,j}, \cdots, a_{N,j}]^T$ の N 次元属性値ベクトルとする．本研究におけるアノテーションでは，属性値ベクトル $\mathbf{z}$ を D 次元空間に埋め込む問題として定式化される．

いま，D 次元の射影ベクトルを $\mathbf{f}$ とする．ここで，$\|\mathbf{f}\| = \sum_{d=1}^{D} f_d^2 = 1$ とする．このとき，ベクトル $\mathbf{f}$ 上への座標ベクトル群の射影値から構成される N 次元縦ベクトルは $\mathbf{Yf}$ となる．よって，属性値ベクトル $\mathbf{z}$ をアノテートする妥当な方向として，次式を最大にする射影ベクトル $\mathbf{f}$ を考える．

$$F(\mathbf{f}) \quad = \quad \mathbf{z}^T \mathbf{Y} \mathbf{f}. \tag{4}$$

PCE 法により得られた座標ベクトル群 $\mathbf{Y}$ を上述のように正規化することで，式 (4) で定義した $F(\mathbf{z})$ はベクトル $\mathbf{Yf}$ と $\mathbf{z}$ の相関係数と等価になる．よって，式 (4) で定義した $F(\mathbf{f})$ を以下では簡単に相関と呼ぶ．

式 (4) の相関を最大化する $\hat{\mathbf{f}}$ はラグランジュ乗数法より以下となる．

$$\hat{\mathbf{f}} \quad = \quad \frac{1}{\|\mathbf{Y}^T \mathbf{z}\|} \mathbf{Y}^T \mathbf{z}. \tag{5}$$

一方，式 (5) を式 (4) に代入すれば以下を得る．

$$F(\hat{\mathbf{f}}) \quad = \quad \|\mathbf{Y}^T \mathbf{z}\|. \tag{6}$$

よって，属性値ベクトル $\mathbf{z}$ を D 次元空間に埋め込むアノテーションとして，その方向と相関を，それぞれ式 (5) と 式 (6) で規定する次式のベクトル（矢印）と定義する．

$$\mathrm{Annot}(\mathbf{z}) \quad = \quad \mathbf{Y}^T \mathbf{z}. \tag{7}$$

明らかに，属性値ベクトル $\mathbf{z}$ に対して，式 (5) の矢印が長ければ相関が高く有意なアノテーションと言えるが，矢印が短ければアノテーションが困難なことを意味する．本稿では，全ての単語 j に関して，属性値ベクトル $\mathbf{z}$ を構築しアノテーションを試みる．そして，相関が上位の単語によりアノテーションを付与する．

4. 評価実験

実レビューデータに対する提案手法の処理結果が，1）全方向に満遍なく拡がる，2）特定の方向に評価が固まるという観点から定性的に評価する．

4.1 データセット

提案手法の評価に際して，価格.com の商品レビューデータを用いる．紙面の都合上，ランダムに選択した，商品 1：Xperia SO-01B docomo（レビュー数：474，単語数：3856），商品 2：iPod touch 第 4 世代 [32GB]（レビュー数：284，単語数：2595）に対する処理結果を示す．それぞれ，次元圧縮後の次元数（トピック数）は $S = 10$ とした．次元圧縮せず，TF の値 $a_{i,j}$ を直接リンク重みとした手法と比較する．

*1 本稿では 2 次元極座標平面に可視化しているため，$D = 2$ である．

4.2 可視化結果

図 2 に商品 1，図 3 に商品 2 に対する処理結果を示す．ノードの色は，最も近いアノテーションの色を割り当てている．他の可視化との色の関連はないことに注意されたい．どちらの商品に対する可視化結果でも，TF・PCE 法では極端な方向にノードが分離してプロットされており，評価文書へのアクセシビリティの観点では好ましいとは言えない．これらの結果は，TF の値の範囲が極端に大きな値から小さい値までとるため，大きな重みを付してベクトルを合成することが原因と考えられる．また，相関を表すアノテーションベクトルの長さに関しては，次元圧縮後に PCE 法で可視化した結果より大きくなっている．ある方向に文書ノードが集まっているため，それらの文書ノードに含まれる一部の単語に関するアノテーションベクトルが非常に高い相関を持ったと考えれる．

次元圧縮後の値を用いた結果（SVD，NMF，LDA）では，TF・PCE 法と異なり，文書ノードが全体的に万遍なくプロットされている．3 手法の顕著な違いは，算出される重みの値が SVD では実数値，NMF では非負値，LDA では確率値という点である．いずれの手法による重みを用いても，PCE 法による可視化結果に大きな違いはなく，適用可能であることを確認した．さらに，隣接するアノテーションベクトルは，“イヤホン，スピーカー，プレイヤー，音質”などの関連する単語であった．他のアノテーションベクトルとしては，“デザイン”，“問題”，“拡張機能”，“アップデート”などが採用された．一方で，全体に万遍なくプロットされているため，アノテーションベクトルの長さが表す相関に関しては，TF より低い値となった．

5. おわりに

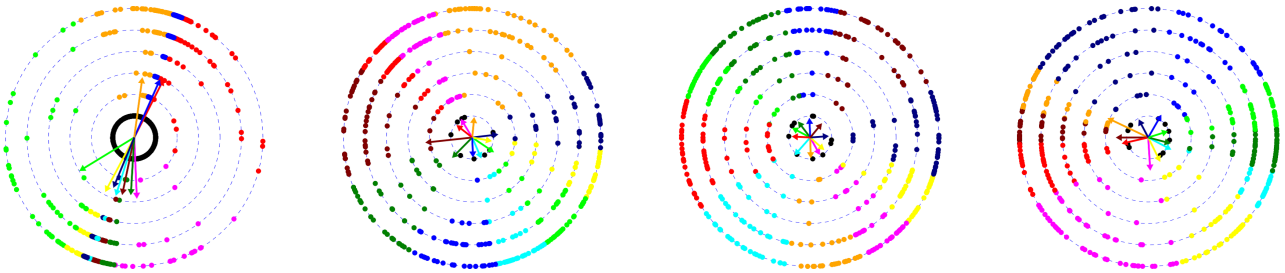
本研究では，レビューサイトにおけるレビューコメントなどのような評価文書を対象に，評価文書へのアクセシビリティ向上を目的とした評価文書可視化法を提案した．具体的には，ユーザが投稿された評価文書を参照する際に考慮する評価項目と評点スコアを軸とした極座標可視化法を確立した．価格.com レビューデータを用いた評価実験では，TF を直接用いる場合より次元圧縮したベクトル群を使用する方がアクセシビリティの高い可視化結果が得られることを確認した．本研究の成果は，大量の評価文書を効果的かつ適切に可視化することでその全体像を把握し，さらに各評価文書へのアクセシビリティを向上させるための方法論である．この成果により，レビューサイトユーザの意思決定支援だけでなく，商品開発側の企業にも有用な情報発見が容易となることが期待される．さらに，これらの技術を一般化させ，インターネット上にある大量の文書データの整理，アクセシビリティ向上への貢献も十分期待できる．

今後は，自然言語処理の技術を用いて評価表現，評価極性を可視化法に導入し，さらに，どの方向にどんな評価項目が布置されているかを自動的にアノテートする枠組みを組み込むつもりである．

謝辞 本研究は，JSPS 特別研究員奨励費 15J00735 の助成を受けたものである．

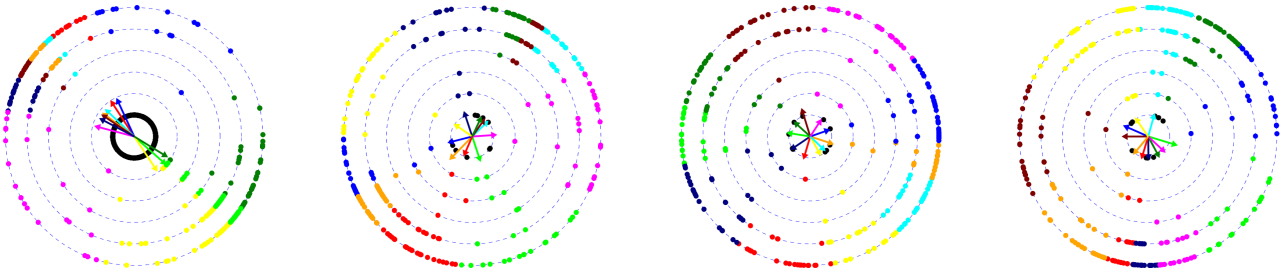
参考文献

[Fushimi 11] Fushimi, T., Kubota, Y., Saito, K., Kimura, M., Ohara, K., and Motoda, H.: *AI 2011: Advances in Artificial Intelligence: 24th Australasian Joint Conference, Perth, Australia, December 5-8, 2011.*



(a) TF・PCE 法　(b) SVD・PCE 法　(c) NMF・PCE 法　(d) LDA・PCE 法

図 2: 商品 1 に対する可視化結果



(a) TF・PCE 法　(b) SVD・PCE 法　(c) NMF・PCE 法　(d) LDA・PCE 法

図 3: 商品 2 に対する可視化結果

Proceedings, chapter Speeding Up Bipartite Graph Visualization Method, pp. 697–706, Springer Berlin Heidelberg (2011)

[Kamada 89] Kamada, T. and Kawai, S.: An algorithm for drawing general undirected graphs, *Inf. Process. Lett.*, Vol. 31, pp. 7–15 (1989)

[Kleinberg 99] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol. 46, pp. 604–632 (1999)

[Oelke 09] Oelke, D., Hao, M., Rohrdantz, C., Keim, D., Dayal, U., Haug, L.-E., and Janetzko, H.: Visual opinion analysis of customer feedback data., in *IEEE VAST*, pp. 187–194, IEEE Computer Society (2009)

[Scaffidi 07] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C.: Red Opal: Product-feature Scoring from Reviews, in *Proceedings of the 8th ACM Conference on Electronic Commerce*, EC '07, pp. 182–191, ACM (2007)

[Stevens 12] Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D.: Exploring Topic Coherence over Many Models and Many Topics, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pp. 952–961, Association for Computational Linguistics (2012)

[Titov 08] Titov, I. and McDonald, R.: Modeling Online Reviews with Multi-grain Topic Models, in *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pp. 111–120, ACM (2008)

[Torgerson 52] Torgerson, W.: Multidimensional scaling: I. Theory and method, *Psychometrika*, Vol. 17, pp. 401–419 (1952)

[Turney 02] Turney, P. D.: Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 417–424, Association for Computational Linguistics (2002)

[Wu 10] Wu, Y., Wei, F., Liu, S., Au, N., Cui, W., Zhou, H., and Qu, H.: OpinionSeer: Interactive Visualization of Hotel Customer Feedback, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 1109–1118 (2010)

[小林 14] 小林 えり, 斉藤 和巳, 池田 哲夫, 大久保 誠也：L1 埋め込みによるアノテーション付き可視化法, 第 7 回 Web とデータベースに関するフォーラム (WebDB Forum2014) (2014)