1Q-09

埋め込み手法の違いによるアノテーション付き可視化の特性評価

大畑　圭佑‡　　斉藤　和巳†

†静岡県立大学 経営情報学部　　‡静岡県立大学大学院 経営情報イノベーション研究科

1　はじめに

大規模なデータ間の相互関係や特徴，その上での現象を分析しようとする動きが様々な分野でみられる．データが有する特徴や関係を理解するための有効な手段のひとつとして「可視化」があり，これまでに様々な可視化法が提案されている [Lee07, Sammon69, Tenenbaum00, Torgerson58]．しかしながら，これらの手法を用いて可視化した場合，どのような特徴を持つデータがどこにプロットされているかを把握することは難しい場合がある．

そこで，オブジェクト集合に対する属性情報から，可視化結果（プロット図）のどの辺りにどのような共通の特徴，属性を持つオブジェクトが布置されているかを自動的に示す方法としてアノテーションを用いた可視化法に着目する．既存の研究 [Kobayashi14, Oohata 16] では，MST 法によるアノテーション付き可視化に絞って研究してきたが, 本研究では PCA,RNG,KNN という違う手法も用いて比較研究を行い, 各手法において, どのような特徴が導かれるのかを検証する．

2　アノテーション付き可視化法

アノテーション付き可視化法 [Kobayashi14] は，オブジェクトの特徴ベクトル，属性情報，カット数 K が与えられたとき，以下の手順により可視化結果を生成する．

step1　特徴ベクトルに基づき, 各種可視化法によりオブジェクトを埋め込む.

step2　step1 の埋め込み座標での距離で，最小全域木を生成.

step3　その最小全域木と属性情報から特徴的部分集合を抽出.

step4　Z-スコアでその部分集合に対しアノテーションを付与.

step5　そのアノテーションとともに K+1 個の部分集合を彩色した最小全域木の可視化結果を出力.

以下では各手順を詳しく説明する．N 個のオブジェクト集合 $V = \{1, \ldots, N\}$ をとし，ここでは各オブジェクトを整数で表す．オブジェクト集合 V に対し，特徴ベクトル群 $Z = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$ と，属性ベクトル群 $Y = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ が与えられるとする．なお本稿では，オブジェクト集合はレビューアイテムとし，特徴ベクトル $\mathbf{z}_n$ は，アイテム n をレビューしたかユーザ次元ベクトルとする．また，ユーザが多種多様な視点で付与するタグを属性と見なし，総属性（異なるタグ）数は L とし，オブジェクト n に対し，第 l 属性に対応するタグが付与されていれば $y_{n,l} = 1$，さもなければ 0 とする ($y_{n,l} \in \{0, 1\}$).

step1 では,MST,RNG, または KNN を作成し, バネモデルで可視化する手法と PCA で可視化する合計 4 手法を対象とする.

step2 では，step1 の埋め込み座標での距離を用いて，オブジェクト集合をノード集合とする最小全域木 $G = (V, E)$ を作成する．E はリンク集合を表す.

step3 では，最小全域木 $G = (V, E)$ の複数リンクを切断し，ある特徴を持った $K+1$ 個の部分集合にオブジェクト集合を分割する．詳細には，K 個の要素からなる切断リンク集合を $E_K = \{e_1, \cdots, e_K\} \subset E$ とする．各要素 $e_k \in E_K$ での切断により，ある連結成分は 2 つの連結成分に分割されるため，切断リンク集合 E_K はオブジェクト部分集合群 $\{V_1, \cdots, V_{K+1}\}$ を決定する．なお，オブジェクト n は一つの部分集合にのみ属し，複数の部分集合に属することはないため，$V = \bigcup_{k=1}^{K+1} V_k$，$k \neq h$ において $V_k \cap V_h = \phi$ が成立する．いま，オブジェクト集合 V_k に属するオブジェクトのうち，属性 l に該当するオブジェクト数を $a_{k,l} = \sum_{n \in V_k} y_{n,l}$ とし，集合 V_k 内の全オブジェクトの該当属性数の総和 $A_k = \sum_{l=1}^{L} a_{k,l}$ をとする．ここで $a_{k,l}$ と，A_k を用いて以下の尤度関数 $F(E_k)$ を定義する.

$$F(E_K) = \sum_{k=1}^{K+1} \sum_{l=1}^{L} a_{k,l} \log \frac{a_{k,l}}{A_k} \tag{1}$$

尤度関数 $F(E_K)$ の最大化は，オブジェクト集合の属性分布が大きく変化する箇所で分割で実現されるので，故に特徴的な属性を有する部分集合の抽出が期待できる.

step4 では，分割された部分集合 V_k の特徴属性を Z-スコアを用いて抽出する．ここで，全オブジェクトが属性 l を有している確率を $p_l = \frac{\sum_{n=1}^{N} y_{n,\,l}}{N}$ としたとき，Z-スコアは以下の式で定義される.

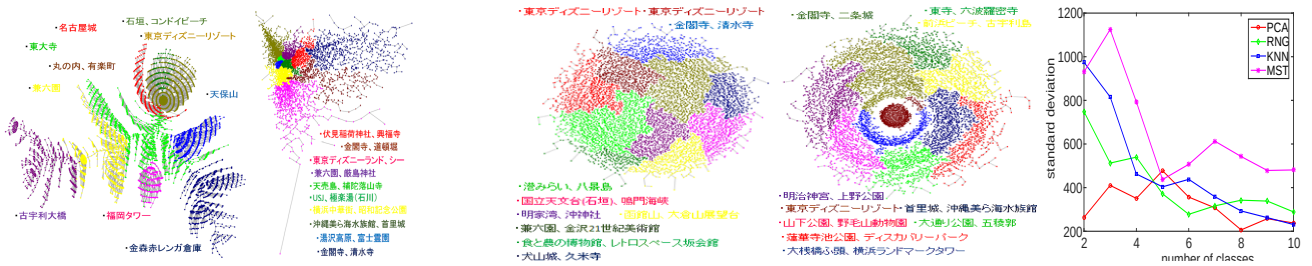
$$Z_{k,\,l} = \frac{a_{k,\,l} - |V_k| p_l}{\sqrt{|V_k| p_l (1 - p_l)}} \tag{2}$$

$|V_k|$ は部分集合 V_k に属するオブジェクト数を示す．本論文では，Z-スコアが非負値の属性のみを対象とする．既存研究 [Kobayashi14] では単一の属性情報のみを扱ったが，本論文では多様な視点での多種属性情報を用いる.

Evaluating characteristics Annotated Visualization Results by using different Embedding Methods

†Keisuke Oohata　†Kazumi SAITO

†University of Shizuoka



(a) MST(10) (b) PCA(10) (c) RNG(10) (d) KNN(10) (e) 各手法のバラつき度

図 1: 埋め込み手法の違いによるアノテーション付き可視化の特性評価

step5 では，二次元平面上に最小全域木をプロットし，リンク集合 E_K より全域木を部分集合毎に彩色し，各部分集合に対するアノテーションをプロット図に記述する．

3 実験設定

本実験では, レビューサイト tripAdvisor(www.tripadvisor.jp) から収集した観光スポットのデータを用いて，アノテーション付き可視化法によるユーザ行動分析を試みた．本実験では，ユーザは訪問スポット数が 10 以上のユーザに限定し，これらユーザが訪問したスポットを用いたとこ，ユーザ数は 6,693 で，スポット数は 11,621 となった．アノテーション付き可視化法の適用では，ユーザをオブジェクトとし，各ユーザが訪問したスポットを属性とした．

4 評価実験

図 1 に提案法による可視化結果，及び各部分集合に対するアノテーションを示す．オブジェクトノード，アノテーションは所属する部分集合によって色分けされている．

図 1（a),(b),(c) 及び (d) には, それぞれ MST,PCA,RNG 及び KNN で可視化し,10 個に分割した結果を示す. これら結果より, 東京ディズニーリゾートは全ての可視化法でのアノテーションとして出現した. また, 都道府県ごとに見ると, 東京, 千葉, 北海道, 沖縄が全ての可視化結果のアノテーションとして出現したということでどの結果を取っても多くのユーザが訪れている主要な観光地として抽出することができた. 他にも清水寺, 金閣寺, 東大寺のような有名な寺院が多くある京都・奈良や日本三名園の 1 つである兼六園がある石川県など人気な観光地がアノテーションとして抽出される結果となった. RNG の可視化結果は, プロットした任意の点間の重なりが少なく, 全体的に一様に広がっていることより, 最もブラウズに適していると考える．RNG 法と KNN 法は似たような可視化結果となった. 図 1 (e) には, 各手法のバラつき度をグラフにまとめた結果を示す. この結果より 2 個に分割したときは PCA 法が 1 番バラつきが少ないことが分かるが,10 個に分割したときには MST 法を除く PCA,RNG,KNN の 3 つの手法においてはバラつきの差はほとんどないことが分かる. 観光スポットをアノテーションとして付与することにより，どの位置にどのような観光スポットを訪問したユーザが配置されているか，視覚的に把握しやすくなり，提案法による可視化結果を用いれば，ユーザ行動分析に貢献しうると考えられる．以上の結果より RNG を作成する手法が有望であると考える．

5 おわりに

本研究では，観光レビューサイトである tripAdvisor のデータを基に MST,PCA,RNG,KNN の 4 手法を用いたアノテーション付き可視化での比較研究を行い, どのような特徴が見られるかを検証した. そして, プロットした任意の点間の重なりが最も少なく、ブラウズに適している RNG 法が今回の研究の結果, 最も有望であると考える．

謝辞 本研究は, 総務省 SCOPE(No.142306004), 及び, 科学研究費補助金基盤研究 (C)(No.15K00429) の助成を受けた.

参考文献

[Lee07] J.A. Lee and M. Verleysen, “Nonlinear Dimensionality Reduction,” Springer, (2007).

[Sammon69] J.W. Sammon, “A nonlinear mapping algorithm for data structure analysis.,” IEEE transactions on Computers, CC-18(5):401-409, (1969).

[Tenenbaum00] J.B. Tenenbaum, V. de Silva, and J.C. Langford. “A global geometric-framework for nonlinear dimensionality reduction.,” Science, 290(5500):2319-2323, (2000).

[Torgerson58] W. Torgerson, “Theory and methods of scaling,” Proc. of Wiley New York, (1958).

[Kobayashi14] 小林 えり，斉藤 和巳，池田 哲夫，大久保 誠也，“可視化結果へのツリー分割による アノテーション付与法，”ネットワークが創発する知能研究会（JWEIN2014），(2014).

[Oohata 16] 大畑 圭佑，斉藤 和巳，”アノテーション付き可視化によるユーザ行動分析，”第 15 回情報科学技術フォーラム（FIT2016），Sep.2016.