F-011

データ対象と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
Evaluating characteristics Annotated Visualization Results by using different Embedding Methods and Data sets

大畑　圭佑 †　　斉藤 和巳 †
Keisuke Oohata　　Kazumi Saito

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]．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, Oohata 17] では,MST,RNG, またはKNNを作成しこれらネットワークをバネモデルで可視化する手法とPCAで可視化する合計 4 手法をトリップアドバイザーのデータを用いて比較検証してきた．本研究では, データ対象として Y!movie を追加し,2 つのデータと 4 手法の可視化法を用いて比較検証を行う．

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．詳細は，[Kobayashi14] を参照．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．詳細は, [Kobayashi14] を参照．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

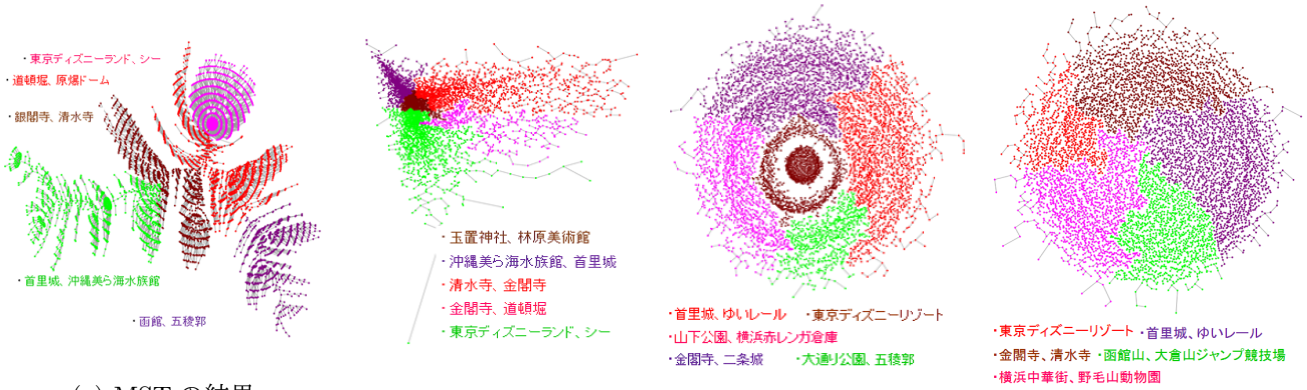
本実験では，レビューサイト tripAdvisor (www.tripadvisor.jp) と Y!movie (www.yahoo.com/movies) から収集した観光と映画のデータを用いて，アノテーション付き可視化法によるユーザ行動分析を試みた．本実験では，tripAdvisor のユーザは訪問スポット数が 10 以上のユーザに限定し，ユーザ数は 6,693 で，スポット数は 11,621 となった．対して,Y!movie のユーザも 10 以上の映画をレビューしたユーザに限定し, ユーザ数は 8,103 で, 映画数は 25,251 となった．アノテーション付き可視化法の適用では，ユーザをオブジェクトとし，tripAdvisor では各ユーザが訪問したスポットを属性とし, Y!movie では各ユーザがレビューした映画を属性とした．

4. 評価実験

図 1 と 2 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．

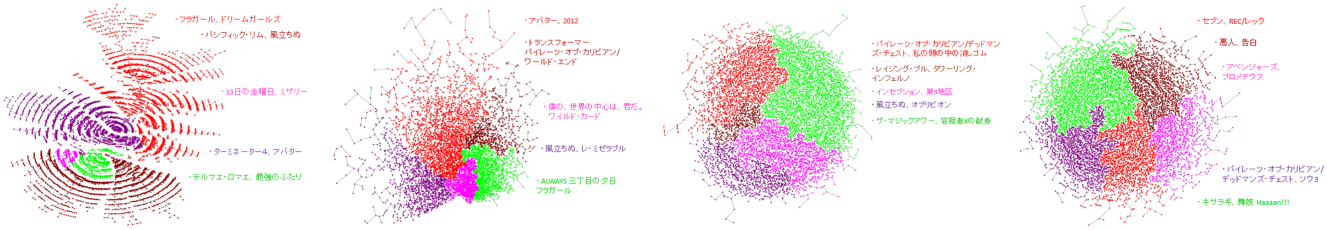
図 1 (a),(b),(c) 及び (d) には,trpAdvisor のデータを用いて, それぞれ MST,PCA,KNN 及び RNG で可視化し,5 個に分割した結果を示す．これら結果より, 観光地のアノテーションとして東京ディズニーリゾートは全ての可視化結果で抽出され、また金閣寺・首里城のような人気の観光地も多くの可視化結果のアノテーションと

†静岡県立大学 経営情報学部



(a) MST の結果　(b) PCA の結果　(c) KNN の結果　(d) RNG の結果

図 1: tripAdvisor データでの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結果 (5 個分割時)



(a) MST の結果　(b) PCA の結果　(c) KNN の結果　(d) RNG の結果

図 2: Y!movie データでの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結果 (5 個分割時)

して抽出された. 観光のデータを用いて 5 個に分割した際には, どの可視化結果を見ても共通のアノテーションが多く抽出される結果となった. 観光におけるユーザ行動はアノテーションを見てみると可視化手法の違いにかかわらず同じような特徴が見られた. 可視化結果を比較してみると,RNG 法はまとまりがあり, 任意の点が分散した図となり,KNN 法と少し似たような可視化結果が抽出された. 図 2 (a),(b),(c) 及び (d) には,Y!movie のデータを用いて, それぞれ MST,PCA,KNN 及び RNG で可視化し,5 個に分割した結果を示す. これら結果より, 映画のアノテーションとしては観光のデータとは異なり, 全ての可視化結果に共通したアノテーションは抽出されず, 個人の嗜好が観光データより現れ, 多くの映画作品がアノテーションとして抽出された. 可視化結果を比較してみると, 観光データの時よりも KNN 法と RNG 法の可視化図がより近似するようになった. データを変えてみても RNG 法の可視化結果は変わることなく抽出され, 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.

5. おわりに

既存の研究 [Kobayashi14, Oohata 16, Oohata 17] では,MST,RNG, または KNN を作成しこれらネットワークをバネモデルで可視化する手法と PCA で可視化する合計 4 手法をトリップアドバイザーのデータを用いて比較検証してきた. 本研究では, データ対象として Y!movie を追加し,2 つのデータと 4 手法の可視化法を用いて比較検証を行った. そして,trpAdvisor を用いた時の RNG 法の可視化結果において, 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りがあることでブラウズに適している. よって RNG 法が一番有望な可視化手法だと考察した既存研究の結果が本研究のようにデータを変えても同様の結果が得られ, 有意義な研究となった.

謝辞 本研究は, 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).

[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.

[Oohata 17] 大畑 圭佑, 斉藤 和巳, ”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, ”第 79 回情報処理学会全国大会（IPSJ2017）, Sep.2017.