6P-06

多種情報を利用したアノテーション付き可視化法

後藤　裕†　　小林　えり‡　　斉藤　和巳†　　大久保誠也†　　池田哲夫†
†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

近年，大規模なデータを対象として，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，本研究ではアノテーションを用いた可視化法に着目する．アノテーションを用いた可視化法とは，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である．既存の研究 [Kobayashi14] では，属性情報としてオブジェクトの該当ジャンル等の単一の項目にのみ採用していたが，オブジェクトには様々な属性情報があり，複数の情報を取り入れることで，新たな知見発見につながると考えられる．そこで，本論文では，多種属性情報を用いて部分集合の特徴的属性をアノテーションとして付与する手法を提案する．

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　オブジェクト間の類似度を計算し，最小全域木を生成.
step2　その最小全域木と属性情報から特徴的部分集合を抽出.
step3　Z-スコアでその部分集合に対しアノテーションを付与.
step4　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，特徴ベクトル間のコサイン類似度を用いて，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.

step2 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．

step3 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．

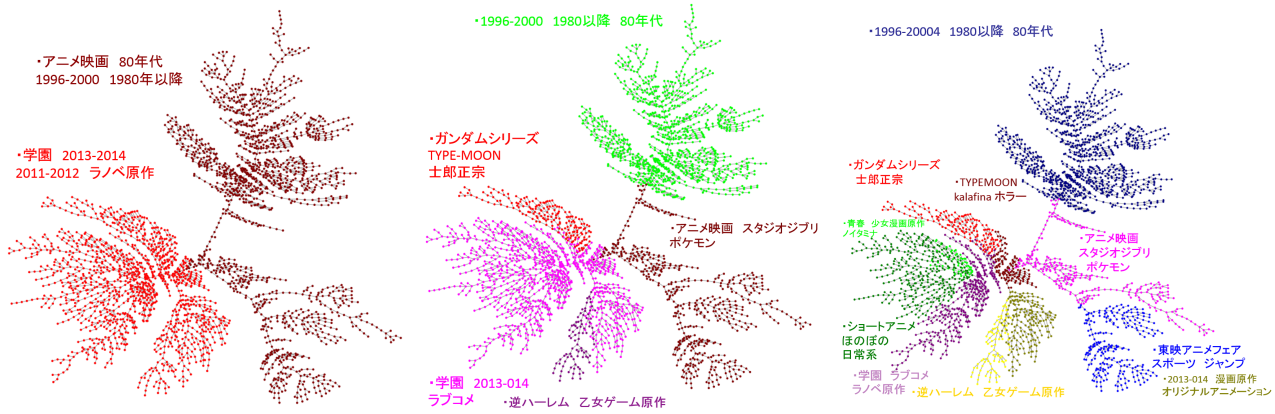
$$Z_{k,\ l} = \frac{a_{k,\ l} - |V_k| p_l}{\sqrt{|V_k| p_l (1 - p_l)}} \tag{2}$$

$|V_k|$ 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．

step4 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

Visualization with annotation using multi-types of information
†Yu GOTO　‡Eri KOBAYASHI　†Kazumi SAITO　†Seiya OKUBO
†Tetsuo IKEDA
†University of Shizuoka

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.



(a) 2 色　(b) 5 色　(c) 10 色

図 1: アニコレを用いたアノテーション付き可視化法

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

実験データとして，アニメのレビューサイトである，アニコレを用いた. 2014 年 4 月 4 日にクロールして収集したものを対象とし，オブジェクト数は 2,808，一度でもレビューしたユーザ数 M は 32,281 である．最小全域木生成時に用いた特徴ベクトル $\mathbf{z}$ は，M 次元で表記され，アニメ（オブジェクト）v_n にユーザ u_m がレビューした場合，$z_{n,m} = 1$ とし，さこなければ 0 となる ($z_{n,m} \in \{0, 1\}$)．よって，近傍に位置するオブジェクト同士は似たユーザがレビューをしているアニメ同士となる．属性ベクトルとして，アニメに付与されたタグ（ジャンルや原作，監督等の情報），放映時期，種類（テレビ放送，映画等）を統合したものを用いた．今回は，オブジェクト集合を 2 個，5 個に分割，10 個に分割したときのアノテーションが妥当なものであるかを定性的に評価する．

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）より，2 個に分割した場合，2000 年を境に分割され，放送時期が強く影響していると分かる．図 1（b）より，5 個に分割した場合，2000 年以降の古いアニメ作品群だった部分集合は「テレビアニメ」と「映画」作品に分かれ，アニメの種類による分割が確認できた. 一方，最新のアニメ群だった部分集合は「SF 作品」,「学園・ラブコメ」,「女性向けアニメ」と，ジャンルによる分割が確認できた．図 1（c）より，10 個に分割した場合，2000 年以前のアニメは分割されず，最新のアニメ群がジャンルや種類によってより詳細に分割される結果となった．

アノテーションを付与することにより，どの位置にどのようなアニメ作品が配置されているか，視覚的に把握しやすくなり，提案法は可視化結果の読解性向上に貢献しうる手法だと言える．また，本実験では，ユーザの嗜好（ユーザのレビューの有無）を用いて類似度を計算しアニメ作品をプロットしているため，ユーザは放映時期やジャンルにこだわってアニメ視聴していることが分かった．

5　おわりに

本研究では，可視化結果の読解性向上を目的に，複数の属性を用いたアノテーションを付与する可視化法を提案した．アニメレビューデータを用いて，複数の属性を考慮したオブジェクト集合の分割を確認し，提案法の有効性を検証した．今後は，他のデータでの検証を行う．

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).

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.