

## アノテーション付き可視化によるユーザ行動分析 User Behavior Analysis by Visualization with Annotation

大畑 圭佑<sup>†</sup>  
Keisuke Oohata

齊藤 和巳<sup>†</sup>  
Kazumi Saito

### 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する研究が活発に展開されている。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。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 [小林 14, 後藤 16] 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。先行研究 [小林 14] 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た。一方、先行研究 [後藤 16] 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した。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。本研究では、観光スポット訪問履歴データへの適用を通して、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 [小林 14] 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、最小全域木を生成する。
- 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する。
- 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する。
- 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する。

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  $V = \{1, \dots, N\}$  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  $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$  と、属性ベクトル群  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$  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  $E_K = \{e_1, \dots, e_K\} \subset E$  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  $\{V_1, \dots, V_{K+1}\}$  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。

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

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の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。

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

$|V_k|$  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 [小林 14] 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。

<sup>†</sup>静岡県立大学 経営情報学部

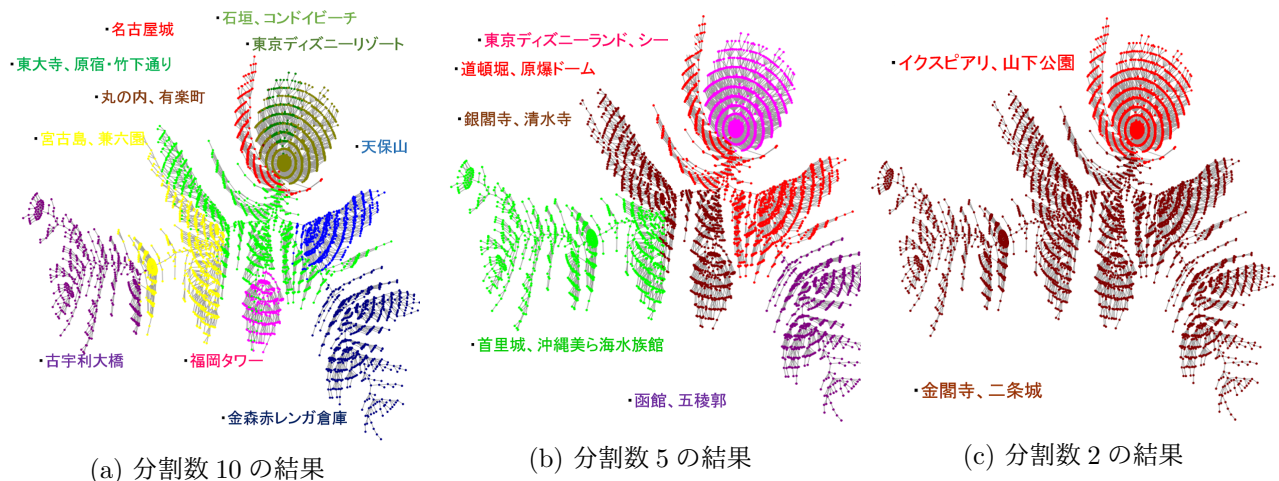


図 1: アノテーション付き可視化法による結果

### 3. 実験による評価

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ー行動分析を試みた。本実験では、ユーザーは訪問スポット数が 10 以上のユーザーに限定し、これらユーザー訪問したスポットを用いたところ、ユーザー数は 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、ユーザーをオブジェクトとし、各ユーザーが訪問したスポットを属性とした。

図 1 に、アノテーション付き可視化法による分析結果を示す。図 1(a) では、ユーザー行動を 10 のグループに区切って分類した結果を示す。赤色で区分されるのは名古屋城周辺、茶色は東京の丸の内・有楽町といった地域、ピンク色は福岡周辺の福岡タワー、紫色は沖縄方面の古宇利大橋、黄緑色は原宿・竹下通り、緑色は石垣のコンドイビーチ、黄色は宮古島の兼六園、茶色は東京ディズニーリゾート、青色は大阪の天保山、濃い青色は北海道の金森赤レンガ倉庫、といったようにユーザーを観光地別に 10 分割したグループに分類されたことを示している。図 1(b) では、ユーザー行動を 5 のグループに区切って分類した結果を示す。赤色で区分されるのは大阪の道頓堀や広島原爆ドーム、茶色は京都の銀閣寺・清水寺、ピンク色は東京ディズニーリゾート、紫色は函館、五稜郭、黄緑色は沖縄の首里城や美ら海水族館といったようにユーザーを観光地別に 5 分割し、グループごとに分類されたことを示している。図 1(c) では、ユーザー行動を 2 のグループに区切って分類した結果を示す。赤色で区分されるのはイクスピアリや山下公園、茶色は金閣寺や二条城のように観光地別に 2 分割したグループに分類されたことを示している。

図 1 より、妥当な観光スポットがアノテーションとして付与されていることを確認できる。また、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ー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、ユーザー行動分析に貢献しうると考えられる。さらに、図 1 に示したように、カット数  $K$  を変化させることで、多様な分類粒度でユーザー行動分析の可能性も期待できる。

### 4. おわりに

本研究では、ユーザー行動分析を目的に、観光スポット訪問履歴データにアノテーション付き可視化法を適用し評価した。レビューサイト tripAdvisor から収集した観光スポットのデータを用いた可視化結果において、妥当な観光スポットがアノテーションとして付与されていることを検証し、アノテーション付き可視化法の有効性ととも、ユーザー行動分析に役立つツールとなり得ることを確認した。今後は、ユーザー属性なども考慮し多様なデータでのユーザー行動分析の可能性を探索する。

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

- [小林 14] 小林 えり, 齊藤 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.
- [後藤 16] 後藤 裕, 小林 えり, 齊藤 和巳, 大久保 誠也, 池田 哲夫, "多種情報を利用したアノテーション付き可視化法," 情報処理学会第 78 回全国大会 (IPSJ2016), 2016.