ソフトクラスタリングを用いた推薦システムのための類似ユーザ抽出手法の提案

清水 拓馬† 齊藤 史哲†

†千葉工業大学先進工学部 〒275-0016 千葉県習志野市津田沼２丁目１７−１

E-mail: †s19c3063ug@s.chibakoudai.jp

あらまし 宿泊レビューなどの複数の項目を持つ Web 上の評価では，回答者の嗜好や評価の傾向などの個人差が含まれやすい．本研究はこの個人差を活用し，類似したユーザにとって有効なパーソナライゼーション手法の確立を目指すものである．こだわりが強い項目とそうでない項目の評価傾向を定量化し，その値を特徴量としたクラスタリングを行うことでユーザを類別化する枠組みを提案した. 楽天トラベルの 5 スター評価データに適用し，解析することで，提案方法の挙動を確認した.

キーワード ユーザ類別化，5 スター評価，データ可視化，ソフトクラスタリング，顧客満足度

1. はじめに

Web サービスの普及により，ホテルなどの宿泊施設を予約できるサイトが数多く登場している．ホテル予約サイトでは実際に利用したことのあるユーザによるレビューが閲覧可能であり，これらのレビューや評価値は，まだ宿泊したことのないユーザがホテルの利用を検討する際の参考にされている．これらのレビューには様々な評価項目が 5 スター形式で付与されており，それぞれの項目に対してユーザが宿泊施設を評価している．ここで得られた評価値は宿泊施設にとっては強みや弱みを表す指標になり，宿泊客にとっては自身のニーズに合致した宿泊先を選ぶ基準になる．

その一方で，これらは評価者の主観評価であるがゆえに評価者が重視する項目や点数の付け方の傾向には個人差がある．例えば，立地条件に対してこだわりが強い顧客は立地に対しての評価は厳しくなりがちであるが，こだわりのない他項目に対しては評価が甘くなる可能性が高い．また，少しでも満足していれば 5 点をつけるような甘い評価するユーザもいれば，少しの満足であれば 3〜4 点をつけるような厳しめなユーザもいる．このように 5 スター評価を参考にしてホテルを調べる際は個人差があることを前提として活用する必要がある．近年では，ユーザの特性を最大限に活用した推薦であるパーソナライゼーションが注目を集めているが，宿泊施設の推薦などを検討する際はこのようなユーザの評価特性を最大限に活用したアプローチも有効であると考えられる．

先行研究では，ユーザの甘辛度を定義して美容院の評価に適用した研究がある．これは，評価傾向の個人差を Web 上の評価データに拡張したという意味で先駆的な試みであると言える．本研究ではこの視点を参考に，ユーザの評価の仕方を定量化し類似したユーザの類別化を試みる．類似した評価傾向を持つ嗜好が近しい顧客を類別化することで，類似したユーザが高く評価する宿泊施設を推薦できるシステムの構築を将来的な目標として，そのためにユーザを類別化する枠組みの構築を目的とする．なおこの先行研究の対象は美容院であり，美容院業界では美容師などとの信頼関係や立地などの拘束条件が強く，本研究が目指す評価傾向のみでの推薦には向かないと判断し，本研究では宿泊業界のデータを利用している．

2. 解析対象について

本研究では，楽天グループ株式会社より提供された楽天トラベルのレビューデータセットにおけるユーザが利用した宿泊施設に対する評価データを解析対象とする. このデータには, あらかじめ設定された「立地」「部屋」「食事」「風呂」「サービス」「設備」「総合」の７つの評価項目について，ユーザが五段階で評価したデータなどが格納されている．このデータにはユーザ ID が紐づけられていることから, 他の評価履歴と比較することができる．このため，ユーザごとの宿泊施設に対する評価の傾向を調べることが可能であり，類似ユーザを抽出するためのクラスタリングに有用な特徴量を構築することができる．

3. 解析手法

先述のとおり，本研究における解析対象は評価方法の個人差を考慮したユーザデータである．個人差はあいまいさを含むデータであるため，類別化した際にその差異を明確に区分けできるとは限らない．そこで本研究ではソフトクラスタリングによって個人差のあいまいな違いを可視化しつつ，ユーザの類別化を行える方法が望ましい．そこで，本研究ではクラスタの所属度を数値で示す代表的なソフトクラスタリングである Fuzzy C-means を用いる．また，その帰属度の微妙な違

いを可視化するためにデータを t-SNE で低次元マッピングしたものに対して帰属度で彩色することでユーザセグメントを可視化する.

4. ユーザ評価傾向の定量化

類似ユーザを求める過程において，重視する項目の傾向を示す評価視点を数値化する必要がある．ここでは，本稿で提案する計算手順を以下に示す.

Step1:全ユーザ評価データを対象とし，宿泊施設ごとに各評価項目の平均を算出する.

Step2:Step1 で算出した各項目の平均点と各ユーザが実際に評価した値の差を算出する.

Step3:Step2 で抽出した各項目の値をユーザーごとに集計して平均を計算することで,各ユーザの評価視点に該当する項目を検出する.

Step1 では，全ユーザーの評価データを対象に宿泊施設ごとに評価データを集計し，各評価項目の平均値を算出している．同じ宿泊施設に対してのユーザー評価を集計することで，その宿泊施設がどの程度評価されているのかを算出することができる.

Step2 では，Step1 で算出した各項目の平均点と各ユーザーが実際に評価した値の差を算出している．例えば，立地の評価平均が 3 の宿泊施設に対し，ユーザーA が星 5 つの評価を行ったとすると，ユーザーA はホテルの評価平均よりも高い評価を行っているため，ユーザーA はホテルの立地に対し，他の人よりも優しめの評価を行っているということがわかる.Step2 では，評価の値の差を算出することで，そのユーザーが特定のホテルの評価平均に対してどのような評価をしているかを定量的に表現している.

Step3 では，Step2 で抽出した評価の差をユーザーごとに集計してその平均を計算することで，各ユーザーが持つ評価傾向を検出している．あるユーザーが評価を行ったすべての宿泊施設の評価データについて集計を行い，項目ごとに平均を算出することで，そのユーザーがどの評価項目に対してこだわりを持っているか，また，全体的に評価が厳しいのか，それとも優しいのかなどといったことを定量的に表すことができている.以上の手順を行い，ユーザーそれぞれの評価傾向を定量化することができた.

5. データ解析

5.1. 解析の設定

Fuzzy c-means によるクラスタへの所属度の曖昧さを決定するハイパーパラメータ m は 1.5，クラスタ数は変数の数に合わせて 7 としている．このクラスタリングの結果に t-SNE を適用し，所属度で彩色することで結果の可視化を行った.

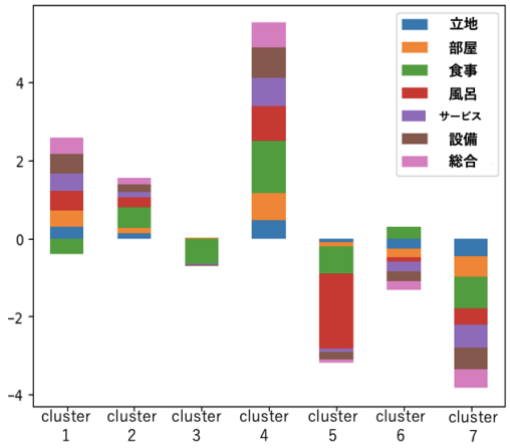

図 1 各クラスタの評価項目平均の積み上げ棒グラフ

5.2. データ解析

解析結果の一部を図 2~図 5 に示している.ここでは，散布図上に配置された各ドットが t-SNE によって 2 次元空間上に配置された各ユーザ情報を表している．これは 7 つのクラスタの内一つのクラスタに対する所属度を青色の濃淡で表しており，色が濃くなるほどそのクラスタに所属していることを表している.すなわち，クラスタの中心に近づくほど所属度が高いことが確認できる.

また，図 1 は各クラスタ内における各評価項目の平均値を積み上げ棒グラフで表したものである．グラフが縦軸の 0 より上に位置していれば，その項目はクラスタの中で優しい評価をされているということを表しており，逆に 0 より下に位置する場合は，その項目はクラスタの中では厳しい評価をされているということを表している.

5.3. まとめと考察

ここでは，図 1 の結果より，各クラスタの内容について考察していく．図 1 の積み上げ棒グラフと解析結果を見ると，第 3 クラスタは食事の項目に対して厳しい評価を行う傾向にあるユーザーが集まっているということがわかる．また，第 4 クラスタには，ユーザーの中で評価が優しい傾向にあるユーザーが集まっていると考えられる．第 5 クラスタについては，風呂の項目に対して厳しい評価を行うユーザーが集まっているように見えるが，宿泊施設によっては風呂は評価を行う対象にはならない場合もあるので，今後さらに解析を続け，この部分を明確にしていきたい．第 7 クラスタには，第 4 クラスタとは逆に，全体的に厳しい評価を行うユーザーが集まっていると考えられる.

ユーザが注視する項目を定量化したことで，同一クラスタ内には宿泊において価値観や嗜好が近いユーザが含まれている．同一クラスタ内のユーザが高評価す

る対象を推薦することで推薦の質の向上が期待できる．

図 2〜5 のように，各クラスタの境界線を曖昧にすることで，クラスタへの所属度が高いユーザーと低いユーザーを明確に分別することが可能になり，ユーザーそれぞれの評価傾向を考慮した推薦が可能になる．色が濃い，つまり，クラスタへの所属度が比較的高いユーザーについては，そのクラスタで重視されている項目をもとに推薦を行うことになるが，色が薄い，クラスタへの所属度が低いユーザーについては，他にどのクラスタへの所属度を有しているかを調査し，そのユーザーの持つ評価傾向を明らかにした上で，その情報をもとにした推薦を行うことが，推薦のパーソナライゼーションに繋がると考えられる．

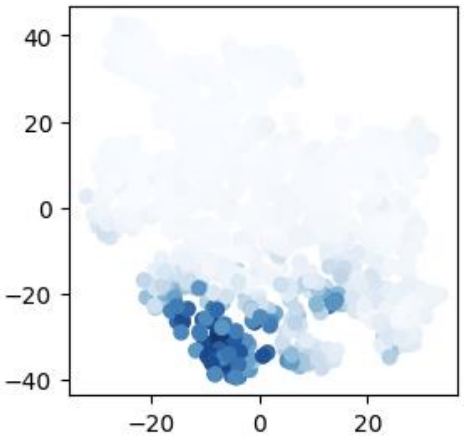

図 2 第 3 クラスタ
（食事の項目に厳しいユーザーが集まっている）

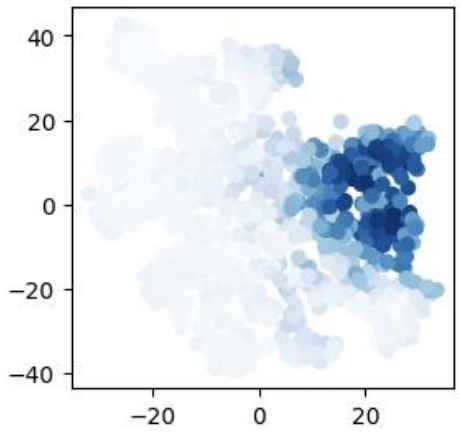

図 3 第 4 クラスタ
（評価が甘いユーザーが集まっている）

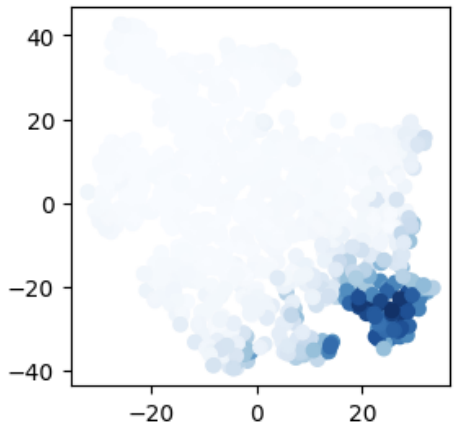

図 4 第 5 クラスタ
（風呂の項目に厳しいユーザーが集まっている？）

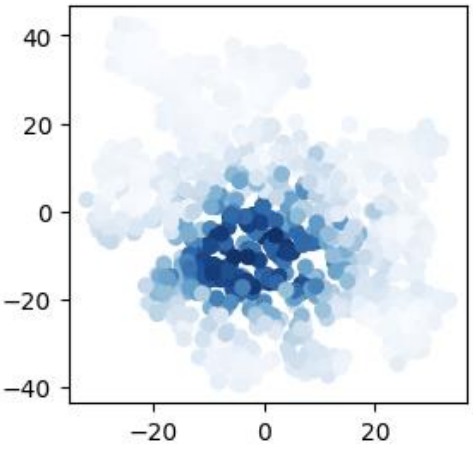

図 5 第 7 クラスタ
（評価が厳しいユーザーが集まっている）

6. おわりに

本研究では評価方法の個人差に着目し，ユーザの嗜好に応じて評価の傾向が異なるという性質を利用したユーザの類別化手法を提案した．ユーザによって評価の甘さや重要視するポイントは個人差があることから，この傾向を定量化する方法を提案し，それを特徴量としたクラスタリングを行った．ソフトクラスタリングによってクラスタリングし，所属度で彩色した t-SNE によって個人差のあいまいさを考慮しつつデータの可視化を実現した．ユーザ個人の嗜好に合致した推薦が求められる昨今において提案法は有効なアプローチであると考えている．

将来的にはこのユーザの類別化結果を推薦タスクに拡張することを目指している．また，クラスタ数や perplexity といったハイパーパラメータの設定はアドホックな決定にゆだねられたため，これらのより適切な設定に関する詳細な議論も今後の課題としたい．

謝辞：本研究では，国立情報学研究所 IDR データセット提供サービスにより楽天グループ株式会社様から提供された「楽天データセット」を利用しました．厚く御礼申し上げます. また本研究は科学研究費（基盤 C）19K04887 より支援いただきました．

参　考　文　献

[1] 日高加奈，豊田哲也，大原剛三，“レビュー投稿者の甘辛度を考慮した美容院評判情報と利用者情報の可視化”，情報処理学会第 79 回全国大会，pp.521-522, 2017

[2] 宮本定明，“ファジィクラスタリングの有用性について”，知能と情報（日本知能情報ファジィ学会誌）Vol.21，No.6，pp.1008-1017，2009

[3] 楽天グループ株式会社 (2014): 楽天データセット,楽天トラベルデータ. 国立情報学研究所情報学研究データリポジトリ. (データセット). https://doi.org/10.32130/idr.2.0

[4] Laurens van der Maaten, Geoffrey Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, Vol.9, pp. 2579-2605, 2008.