

ユーザ間における属性の差異を用いた エリア推薦の検討

志賀圭駿, 上野秀剛

奈良工業高等専門学校 情報工学科

shiga@info.nara-k.ac.jp

概要: 膨大なコンテンツで構成されるご当地観光アプリからユーザの嗜好に合うコンテンツを発見することを支援するために推薦が良く用いられる。著者らは、新規のユーザに推薦を行う方法としてユーザの属性情報に着目し、その差によって訪問頻度が異なる場所を検出し推薦に用いる方法を検討している。本稿では、属性の違いによって訪れるエリアにどの程度の差異が発生するか検証する。提案手法では属性情報が異なる 2 群のユーザに対して、あるエリアを訪れたユーザの割合を求め、差を取ることでユーザ間の差異を算出する。提案手法を元にした推薦は、単にある入力情報のユーザが頻繁に訪れる場所をその入力情報のユーザに対して推薦することに比べて、セレンディピティのある推薦が期待できる。

Keyword: 情報推薦, ご当地観光アプリ, GPS, セレンディピティ

1. はじめに

個人が所有する携帯端末の普及と共に、特定地域の観光情報を提供するためのアプリケーション、ご当地観光アプリが全国で多数配信されている[1]。ご当地観光アプリをはじめとする観光情報サービスは様々なコンテンツから構成されているが、膨大なコンテンツの中からユーザが自身の嗜好に合うコンテンツを発見するのは難しい。このような問題を改善する技術として推薦がある。推薦は多量の情報からユーザにとって有用と考えられる情報を絞り込む用途に広く用いられ、膨大なコンテンツを扱う観光情報サービスを実現するうえでも重要である[2]。

現在の推薦に広く用いられる手法として協調フィルタリングがある。これはユーザの嗜好が反映されていると思われるデータ群（たとえば商品の購入履歴）からユーザの潜在的な嗜好を推測する手法である。ご当地観光アプリにおけるユーザの嗜好が反映されていると思われるデータ群としてはユーザの移動履歴がある。しかし、ご当地観光アプリは一度限りの使用を想定しているものが多く、ユーザの移動履歴を長期に渡って収集できない。よって、ご当地観光アプリには協調フィルタリングの適用が困難である。このような、新規ユーザに対する推薦が困難であることは **cold-start** 問題として知られている[3]。

著者らはこの問題に対して性別や年齢といっ

たユーザの属性情報と他ユーザの移動履歴を用いた推薦を検討している。一般に店舗・観光地は主たる客層を持つことがある。例として若い女性を対象とした店舗や、遠方地からきた（対象地域を初めて訪れた可能性が高い）観光客に人気の高い観光地などが挙げられる。客層は、属性が同じユーザが似た嗜好を持っていることを表しており、興味を持つ店舗・観光地も類似していると考えられる。そのため、属性情報でユーザを分類し、移動履歴を分析することでその属性の観光者の嗜好を反映した観光地を抽出できると考えられる。

一方で、属性が同一のユーザ群が良く訪れる観光地はユーザにとって既知である可能性が高い。ご当地観光アプリは推薦対象のエリア内において宣伝されるため、その場でインストールされる可能性が高い。そのため、同一属性のユーザがよく訪れる観光地はアプリの利用を開始する時点ですでに調査済みであると考えられる。

本研究ではユーザと属性が同一ではないが、近いユーザの情報を元に未知の観光地を推薦する手法を提案する。ユーザにとって未知であると考えられるアイテムの推薦は、一般にセレンディピティのある推薦であるとされる[4]。ユーザの属性に（たとえば若い男性）対して、属性が近い（若い女性や高齢の男性）ユーザ群の移動履歴に基づいた推薦することで、ユーザにとって未知で興味のある観光地を推薦できる可能性がある。本稿では、

以上の推薦手法がご当地観光アプリにおける推薦に有効であるかを検討する。その手段として、属性情報の違いによって訪れるエリアにどの程度の差異が発生するか、実際の観光アプリのデータを用いて検証する。

2. 関連研究

ユーザの入力情報を用いる推薦手法として、小野らは被験者の入力情報と状況、映画への評価からベイジアンネットモデルを作成した。また、それに基づいてユーザが自らの情報を入力するとおすすり映画とその理由を提示する映画推薦システムを提案した[5]。精度評価の結果、既存の推薦手法に比べて良い予測精度を示した。

この研究では、ユーザの入力情報に基づき精度の高い推薦を行っている。しかし、これらの推薦手法は推薦を行うアイテムについて事前に大量のデータを取得せねばならず、サービス提供側は多大な労力を要する。また、モデル内のデータを更新せずに使用した場合、新しい映画の推薦は難しくなると考えられる。これは観光地推薦においても同様であり、上記の推薦手法では新しい観光地や流行りの観光地を推薦することは難しいと考えられる。よって長期の利用を想定する際には再調査が必要であり、継続的に使用する際のランニングコストが大きい。本研究ではユーザの位置情報と入力情報、観光地の緯度経度のみから推薦を行うことを想定している。アプリの初回起動時にユーザが情報を入力すれば推薦に必要なデータが取得できるため、サービス提供側の労力は小さい。また、日時の新しい位置情報のみを用いることで、時流に沿った推薦が可能であると考えられる。

3. 属性間の差異

属性の異なるユーザ群の訪れるエリアにどの程度の差異があるかを検証する。分析に必要なデータはユーザが入力した1種類以上の属性情報とご当地観光アプリの使用中に携帯端末から発信された位置情報とし、位置情報はどのような属性のユーザが発信したか分かるものとする。属性が異なる2群に対してある地点を訪れたユーザの割合を求め、その差を取ることでユーザ間の差異を算出する。 n 人のユーザの集合 $U = \{u_1, u_2, \dots, u_x, \dots, u_n\}$ において、各ユーザは m 個の属性情報 $P = \{p_1, p_2, \dots, p_y, \dots, p_m\}$ を持つ。処理手順を以下に示す。

1. ユーザの位置情報の緯度経度を小数点以下4桁で四捨五入する。

2. 1ユーザが同じ緯度経度を示す位置情報を2つ以上残している場合、1つを残して他の位置情報を削除する。本手法では滞在時間を考慮しないため、同じ箇所に長時間滞在したユーザの影響がでないようにするために行う。
3. 差異の分析対象とする属性 p_y を1つ以上選択する。選択した p_y の値 v が同一な u_x の集合であるユーザ群 U_g は以下のように定義される。ここで、 g は選択した入力情報 p_y の値 v の組合わせを表す。

$$U_g = \{u_x | p_y = v\} \quad (1)$$

4. 2群のユーザそれぞれに、緯度経度を軸として長方形のメッシュに区切った $Map(U_g)$ を作成する。各メッシュに存在する位置情報の数を $GPS(U_g, lon, lat)$ とする。
5. 全メッシュに対して、対応する p_y を持つユーザ数で除算を行い、各メッシュにおいて U_g が訪れた割合 $mesh(U_g, lon, lat)$ を算出する。

$$mesh(U_g, lon, lat) = \frac{Gps(U_g, lon, lat)}{|U_g|} \quad (2)$$

6. 2群のユーザから以下に定義するユーザ間の差異 $Diff(U_g, U_{g'})$ を算出する。

$$Diff(U_g, U_{g'}) = Map(U_g) - Map(U_{g'}) \quad (3)$$

差異が大きいメッシュはこの値が大きい。値が正であれば U_g が $U_{g'}$ に比べて頻繁に訪れるメッシュであり、値が負であればその逆である。

また、属性間に全体としてどの程度の差異があるのかを示す指標を定義する。 $Diff(U_g, U_{g'})$ の絶対値が大きい正負の a 番目から b 番目までの絶対値を加算し、加算結果における正負の組数である $b - a + 1$ で除算を行う。この操作により $Diff(U_g, U_{g'})$ の平均をとった値を指標 $Diff_{a-b}(U_g, U_{g'})$ とする。この指標は属性間の差異を1つの値で表現し、属性間に全体としてどの程度の差異が発生したのかを簡易に検証できる。

4. 実験

本研究ではご当地観光アプリ「ならたん」における収集データを分析する。ならたんは2017年1~2月にかけて奈良県が実証実験を実施しており、主に奈良市の観光情報が掲載されている。実験時点でダウンロード数は4053件であった。実証実験ではユーザが所有する携帯端末へアプリを

インストールした後、利用開始時に自身の属性情報を入力し、観光を行う。アプリは一度起動されるとバックグラウンドで動作を続け、ユーザの位置情報を5分毎に発信する。ユーザが入力する属性情報は性別、年齢、居住都道府県である。本研究ではユーザとして観光者を想定しているため、奈良県に住むユーザは分析の対象外とする。

実験の対象となるデータは、位置情報が1点以上存在し、属性情報が入力されていた378名のデータで、位置情報の総数は87169件だった。

5. 結果と考察

5.1. 差分の可視化

結果を可視化するため、 $Diff(U_g, U_g')$ の絶対値が高い正負それぞれの上位30件のmeshを地図上にプロットした。なお、本稿では主に奈良市街を観光する観光客についてユーザ間の差異を検証するため、奈良市街から離れた、近隣駅などに残されていた移動履歴は削除した。男性のユーザ群 U_M ($n=181$)と女性のユーザ群 U_F ($n=197$)の比較結果を図1に示す。図の丸印は男性が女性より、三角印は女性が男性より訪れている地点を表す。 U_M は U_F に比べて駅周辺の大通りや転書門付近に多く滞留している。一方で、 U_F は U_M に比べて東大寺や奈良公園などの有名な観光地やショッピング街に多く滞留している。

比較のために図2に U_M 、図5.3に U_F のGPS(U_g, lon, lat)が大きい上位30のmeshを示す。図2、図3ともに、差分の算出はしていないため、それぞれのユーザ群が良く訪れた地点が示されて

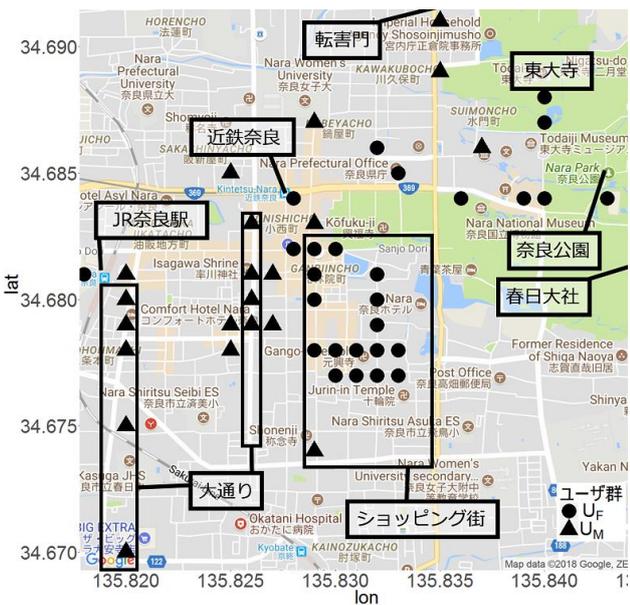


図1 U_M (▲)と U_F (●)の差異

いる。図からは U_M と U_F の間に大きな差異は見られない。これは、属性に依らず多くのユーザが訪れるエリアは似通っているためと考えられ、セレンディピティのある推薦には適さない。一方で図1では U_M と U_F の差異が現れている。これは、属性に依らず多くのユーザが訪れるエリアが提案手法によって除外されるためである。

この結果は提案手法を用いることで各属性よく訪れる場所をその入力情報のユーザに推薦する方法に比べて、セレンディピティのある推薦が可能になる可能性を示している。

5.2. 属性間の差異の比較

3章で述べた $Diff_{a-b}(U_g, U_g')$ を用いて上位30件の属性間の差異を算出した。表1に比較する属性数が1つである場合の $Diff_{1-30}(U_g, U_g')$ を示す。表中の $U_L \cdot U_O$ は居住地が近畿圏内のユーザと近畿圏外のユーザ、 $U_L \cdot U_H$ は年齢が全ユーザの平均より

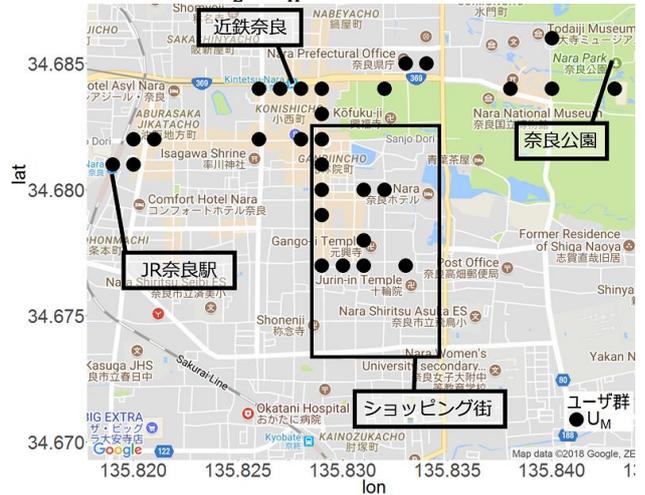


図2 U_M のGPS(U_M, lon, lat)が大きいmesh

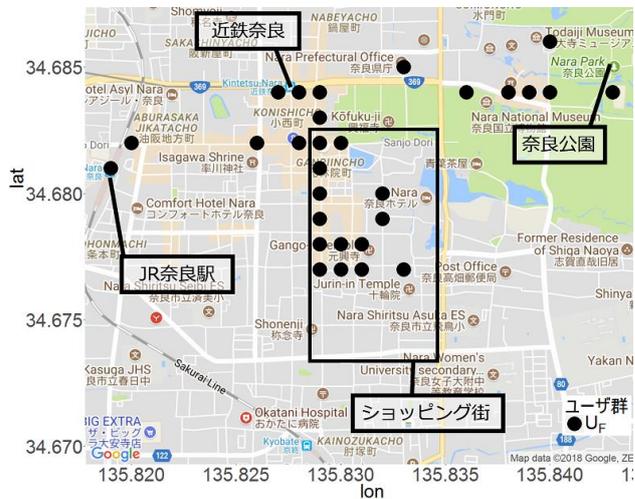


図3 U_F のGPS(U_F, lon, lat)が大きいmesh

表 1 属性数が 1 種類の $Diff_{1-30}(U_g, U_{g'})$

| 比較した属性 | 属性間の差異 |
|-------------------------|--------|
| $Diff_{1-30}(U_M, U_F)$ | 8.207 |
| $Diff_{1-30}(U_I, U_O)$ | 7.563 |
| $Diff_{1-30}(U_L, U_H)$ | 5.761 |

り低いユーザと高いユーザである。表 1 から、性別による属性間の差異が最も大きく、次いで居住地、年齢の順に差異が小さいことが分かる。多重比較 (Ryan 法) の結果、 $Diff_{1-30}(U_L, U_H)$ と $Diff_{1-30}(U_M, U_F)$ 、および、 $Diff_{1-30}(U_L, U_H)$ と $Diff_{1-30}(U_I, U_O)$ の間には $p < 0.01$ の有意差が見られた。属性によって差異が異なることは、推薦する際の属性の選択が重要であることを示している。

表 1 の結果は比較する属性数が 1 の場合を指名しているが、比較する属性数が $Diff(U_{ML}, U_{FH})$ のように 2 の場合やそれ以上の場合に $Diff_{1-30}(U_g, U_{g'})$ が大きくなる可能性がある。比較する属性数の違いによって $Diff_{1-30}(U_g, U_{g'})$ の値に差があるか確認するため、属性数ごとの $Diff_{1-30}(U_g, U_{g'})$ について Welch の t 検定を実施した。なお、検定には $Diff_{1-30}(U_g, U_{g'})$ の算出に用いた $Diff_1(U_g, U_{g'})$ から $Diff_{30}(U_g, U_{g'})$ の全ての値を使用した。これは $Diff_{1-30}(U_g, U_{g'})$ を用いて検定を行った場合、分散が極端に小さくなってしまい、正しい結果が得られないためである。検定の結果、比較する属性数が多ければ $Diff_{1-30}(U_g, U_{g'})$ が大きい傾向にあり、比較する属性数が 1 種類と 2 種類である場合には有意差 ($p < 0.01$) が見られた。

次に、属性数が 2 種類である場合の $Diff_{1-30}(U_g, U_{g'})$ について、異なる属性が多ければ $Diff_{1-30}(U_g, U_{g'})$ が大きくなるのかを検証するため、異なる属性が 1 種類である場合と 2 種類である場合について Welch の t 検定を実施した。検定の結果、異なる属性が 2 種類である場合、1 種類である場合に比べて $Diff_{1-30}(U_g, U_{g'})$ が大きい傾向にあり、比較する属性数が 1 種類と 2 種類である場合には有意差 ($p < 0.05$) が見られた。

この結果から、ユーザの入力する属性を増やせば細分化した嗜好が顕在化し、異なる属性がよく訪れるエリアを推薦すれば、よりセレンディピティのある推薦を行える可能性がある。本稿で行った分析方法を用いる推薦では、属性の一部が異なるユーザ群がよく訪れるエリアを推薦する。そのため、アプリ利用開始時点で既に調査済みである既知のエリアの推薦が少ないと考えられ、セレンディピティの観点で優位である。また、異なる属性が多いユーザがよく訪れるエリアの推薦に比べて、訪れるエリアの差異が小さいために嗜好の差

異が小さいと考えられ、興味の観点においても優位であると考えられる。セレンディピティのある推薦と、推薦に対する興味とはトレードオフの関係にあると思われるため、推薦に用いる入力情報の種類や異なる属性数について、最適な値を今後明らかにする必要がある。

6. おわりに

本稿ではご当地観光アプリ「ならたん」で収集されたデータを用いて、属性の異なるユーザ間で発生する訪問エリアの差異を検証した。本稿では属性をもとにユーザをグループ分けし、その差異を算出することで、各属性特有の訪問エリアを抽出する分析を試みた。実験結果から、提案手法により算出したユーザ間の差異は、ご当地観光アプリにおいてセレンディピティと興味を両立させた推薦に有効である可能性を示した。本研究の結果に基づく推薦では推薦対象ユーザが自身の情報を入力するだけで推薦が可能となるため、新規ユーザに対する cold-start 問題が発生しないという優位性がある。

今後の課題として、本研究で算出した差異の実際の推薦における活用法や、精度評価の検討が挙げられる。また、本研究ではユーザ間の差異のみを算出したが、ユーザ間の類似度を算出し、両者の関係性を検証することで、新たな観点からより有効な推薦手法を提案できる可能性もある。

謝辞

本分析にあたり、奈良県から「ならたん-きょうから奈良へ」アプリにかかるデータの提供を受けました。

参考文献

- [1] 倉田陽平, 青木美岬, 相尚寿, “日本国内のご当地観光アプリの概要把握”, 観光情報学会第 12 回全国大会論文集(2015).
- [2] 松原仁, 山本雅人, 川村秀憲, 鈴木恵二, “観光情報学—情報処理技術の観光への適用”, システム/制御/情報, Vol.57, No.8, pp.316-322(2013).
- [3] 神畷敏弘, “推薦システムのアルゴリズム(2)”, 人工知能学会誌, Vol.23, No.1, pp.89-103(2008).
- [4] 神畷敏弘, “推薦システムのアルゴリズム(1)”, 人工知能学会誌, Vol.22, No.6, pp.826-837(2007).
- [5] 小野智弘, 黒川茂莉, 本村陽一, 麻生英樹, “ユーザ嗜好の個人差と状況差を考慮した映画推薦システムの実現と評価”, 情報処理学会論文誌, Vol.49, No.1, pp.130-140(2008).