カープローブデータを用いた地域特性分析に関する研究

清武　寛†　幸島　匡宏†　松林　達史†　澤田　宏†

† 日本電信電話株式会社 NTT サービスエボリューション研究所
〒 239-0847 神奈川県横須賀市光の丘 1-1
E-mail: †{kiyotake.hiroshi,kohjima.masahiro,matsubayashi.tatsushi,sawada.hiroshi}@lab.ntt.co.jp

あらまし　近年のスマートフォンの普及などに伴い，人や車の移動に関する様々なデータが入手可能となった．本論文では，カーナビゲーションから得られたユーザの行動履歴を分析し，非負値行列分解を用いて，地域特性の抽出・分析を試みる．ユーザの滞在した地点を推定し，その地点での訪問人数と滞在時間を用いて類似している地点の抽出を行う．今回分析によって得られた結果について議論と報告を行う．

キーワード　カープローブデータ, 非負値行列分解, NMF, 地域特性分析

1. はじめに

スマートフォンの普及やセンサデバイスの発達に伴って人や車の移動に関する様々なデータが入手可能となり，データ分析による新たな知見やユーザの行動分析への期待が高まっている．特に近年では，観光振興による地域活性化などを目指して，新たなデータ分析手法とその応用事例が複数報告されている [4] [5] [2]．しかしながら，これらの位置情報データ分析では，特定の場所を訪問したユーザが次にどの場所を訪問しているか，ある場所を頻繁に訪問しているユーザはほかにどの場所を訪問しやすいか等の移動ルートの分析に着目した研究報告が多い．

また，近年ではユーザに対する行動推薦などを目的として，ユーザの移動履歴から地域特性を抽出しようとする試みも行われている [6]．地域特性抽出の研究では，ある場所における滞在人数やその訪問時刻を分析することで，例えば，ある地域のカフェは駅付近にあるため，出社・通学前のユーザが多く立ち寄り，混雑しやすいという特性が抽出可能である．しかしながら，ユーザの利用目的に即した行動推薦は，滞在人数や訪問時刻情報に加えて，滞在時間を考慮することによって，より確度の高い情報提供が可能になると考えられる．滞在時間は，時間帯やその滞在地のカテゴリ情報や利用目的によって大きく異なると考えられる．例えば，コンビニエンスストアでの滞在時間が５分以内など短いものが大半であろうし，遊園地などのテーマパークであれば滞在時間は３時間を超える長いものとなりうる．また，同じ飲食店であっても，休憩や語らいの場として利用される店舗では滞在時間は長く，単純に料理の購入の場所としてテイクアウト用の商品購入のために立ち寄られる店舗では滞在時間が短いと考えられる．従って，滞在時間に着目した場所の分析を行うことで，例えば，朝の時間がないユーザにはテイクアウト専用のカフェを推薦し，休日にゆっくり語らいたいユーザには滞在時間の長いカフェを推薦する等ということが可能となる．

そこで本稿では，店舗や施設，地図上を分割した地域などの“ 場所 ”に対するユーザの滞在時間に着目した分析による知見発見に取り組む．

2. 関連研究

位置情報を用いた研究の代表的なものの一つとして，ユーザの移動ルートを分析する研究が行われている．樋口ら [4] の研究では京都観光に訪れたユーザの行動履歴を分析することで，公共交通機関を利用する観光客は京都駅を起点とする観光ルート，車移動の観光客は金閣寺を起点とする観光ルートを組むという知見を得ている．熊谷ら [5] の研究では非負値複合テンソル因子分解技術によって訪日外国人観光客の行動履歴を分析し，日本人には馴染みが浅いが外国人観光客特有である回遊パターンが存在するという新たな知見を得ている．また，情報推薦に関する研究において，Zheng ら [2] は GPS 情報に加えて POI（Point of Interst）データ等を用いることでユーザに対しておすすめのスポットを推薦する研究を行っている．滞在時間情報を用いた研究として，西田ら [1] は，密集した地域での滞留点抽出という課題に対し，ユーザの滞在時間を考慮することで従来手法より高い精度での滞留点の抽出に成功している．このように，滞在時間情報を利用することは位置情報分析技術において有効であると考えられる．

一方，近年では位置情報データから地域特性を抽出しようとする試みも行われている．李ら [6] は，ユーザが情報発信した場所における滞在人数や人口の流入・流出情報を位置ベース SNS を用いて取得し，その地域が「食べる」,「買う」,「遊ぶ」,「暮らす」,「働く」という５つのジャンルのうちどの割合が高いかを分析することで地域の特徴づけを行っている．このように，近年，データ分析による新たな知見発見やユーザの行動分析への期待が高まっている．

3. データ前処理

今回分析には, 各種ナビサービス（NAVITIME）を展開するナビタイムジャパン社の携帯カーナビアプリから得られたカープローブデータを用いる．ユーザの移動履歴から訪問した場所，滞在時間を得ることができれば，ユーザの訪問した場所での行動パターンの把握ができると期待される．そこで本章では，与えられたデータセットの中から，ユーザがどこにどれくらい滞

分析場所	横浜駅付近
分析期間	2015 年 4 月 18 日から 2015 年 5 月 17 日
ユーザ ID 数	67,947
経路 ID 数	130,915

表 1　研究に用いたデータセット

在したかというデータを得るための手法について記述する．

まず，カープローブデータの内容を以下に記述する．

・ユーザ ID: ユーザを識別するために割り振られる ID.

・経路 ID: 同一ユーザの一行程に割り振られる ID.

・位置情報: 1 秒単位のユーザの緯度経度情報.

本研究で用いたデータセットの概要を表 1 に示した．

経路 ID に関しては, 10 分以上同じ場所に滞在した場合，もしくはアプリケーションの ON/OFF を行った場合にその都度付与される．そのため今回我々は, ユーザの滞在時間に着目した分析を行うために経路 ID の切り替わりが発生する場所に注目した．

経路 ID 切り替わりの条件を考えると, 切り替わりの地点でユーザが 10 分以上の滞在をしていることがわかる．そのため, 以下の 3 つの条件に従って訪問地点の抽出を行った．

・条件 1: 経路 ID が切り替わっている

・条件 2: 経路 ID の切り替わり地点の前後で 10 分以上経過している

・条件 3: 経路 ID の切り替わり地点の前後の距離が 1km 未満

条件 2 は電波障害や遮蔽物により経路 ID は変わっているが, 移動を続けているような地点を除くために設けている．また, 条件 3 はアプリの ON/OFF により経路 ID が変わってしまっているが, アプリを OFF の状態で移動したため, ログが残っておらず休憩していないような地点除くために設けている．各地点での滞在時間は，経路 ID の切り替わり前後の時間差とした．

上記 3 つの条件を同時に満たす地点は 14184 地点あり, それらの地点に対して Mean-Shift を行うことで 1492 地点にまとめた．これら 1492 地点には複数のユーザの情報が含まれており，その情報を非負値の行列と同一視することで, NMF(non-Negative Matrix Factorization:非負値行列因子分解) [3] を用いてクラスタリングを行なった．

4.　NMF によるクラスタリング

NMF の入力データとして, 行が Mean-shift により抽出した地点, 列が平日・休日合わせた 48 時間を 3 時間ごとに区切った時間帯に対応する行列を 2 つ作成する．1 つ目は要素の値がその地点・場所における訪問人数を表す行列 $\boldsymbol{X}_1 \in \mathbb{R}^{1492\times 16}$, 2 つ目は要素の値がその地点・場所における平均滞在時間を表す行列 $\boldsymbol{X}_2 \in \mathbb{R}^{1492\times 16}$ である．これ以降 2 つの行列に適用する処理は同じものであるため, どちらも $\boldsymbol{X}$ と書く．つまり，$\boldsymbol{X}$ は $\boldsymbol{X}_1$ もしくは $\boldsymbol{X}_2$ のどちらかを表すものとする．また, 行列 $\boldsymbol{X}$ の行数, 列数をそれぞれ I, J, 行列の要素を x_{ij} と書く．図 1 に示すように NMF の適用により, $\boldsymbol{X}$ の因子分解の結果である I 行 R 列の因子行列 $\boldsymbol{A} \geqq 0$ と J 行 R 列の因子行列 $\boldsymbol{B} \geqq 0$ を得る．なお, R は分解の際の因子数を表し, NMF の適用前に事前

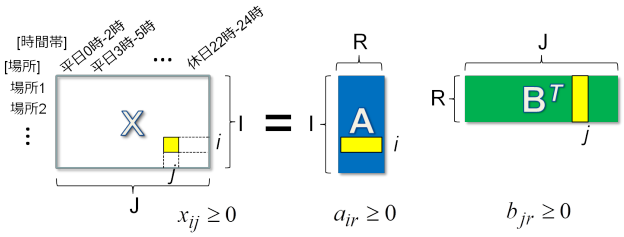

図 1　NMF の定式化

図 2　訪問人数をもとにしたクラスタリング結果

に設定する値である．因子数 R は, データから抽出するパターンの数に相当する．

行列 $\boldsymbol{A}, \boldsymbol{B}$ の推定は次の最適化問題を解くことで得られる．

$$\arg\min_{\boldsymbol{A},\boldsymbol{B}} \sum_{(i,j)\in\Omega} \left(x_{ij} - \sum_{r=1}^{R} a_{ir} b_{jr}\right)^2 \quad s.t.\ \boldsymbol{A}, \boldsymbol{B} \geqq 0.$$

なお, Ω は行列 $\boldsymbol{X}$ 中における値が定義された要素全体を表す．また, 行列 $\boldsymbol{A}, \boldsymbol{B}$ の全ての要素が 0 以上であることを $\boldsymbol{A}, \boldsymbol{B} \geqq 0$ と書いた．この問題を解くアルゴリズムは複数存在するが，ここでは実装の容易さから利用されることの多い乗法更新則に基づくアルゴリズムを紹介する．このアルゴリズムは, 行列 $\boldsymbol{A}, \boldsymbol{B}$ をランダムな非負値で初期化したのち, 次の更新式に従い交互に $\boldsymbol{A}, \boldsymbol{B}$ の更新を行うで分解結果を得るものである．

$$a_{ir} \leftarrow a_{ir} \frac{\sum_{j=1}^{J} x_{ij} b_{jr}}{\sum_{j=1}^{J} \hat{x}_{ij} b_{jr}}, \quad b_{jr} \leftarrow b_{jr} \frac{\sum_{i=1}^{I} x_{ij} a_{ir}}{\sum_{i=1}^{I} \hat{x}_{ij} a_{ir}}.$$

5.　適用結果と考察

まず，訪問人数を値に持つ行列 $\boldsymbol{X}_1$ に対してクラスタリングを行った結果を図 2 に，平均滞在時間数を値に持つ行列 $\boldsymbol{X}_2$ に対してクラスタリングを行った結果を図 3 を示す．丸印はその地点に訪問したユーザを表し，星形の印はその場所の中心を表す．また，同じクラスタに属する地点は同じ色で表示している．各クラスタに属する地点のうち，そのクラスタ属性を最もよく表す地点を 2 点から 3 点表示している．図 2，図 3 にて抽出された地点の特徴，各クラスタリング手法におけるクラスタ番号を表 2 に記す．

訪問人数を考慮したクラスタリング結果の図 2 を見ると，場所 A, 場所 B，場所 C が属するクラスタ 1-1，場所 F・場所 G が属する飲食店やショップが含まれるクラスタ 1-2，場所 D・

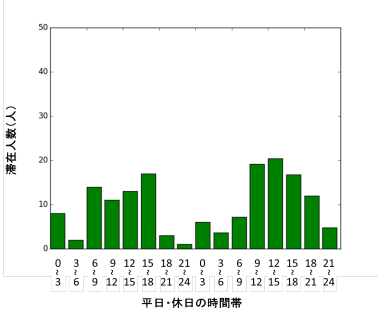

図 4　場所 A における訪問人数

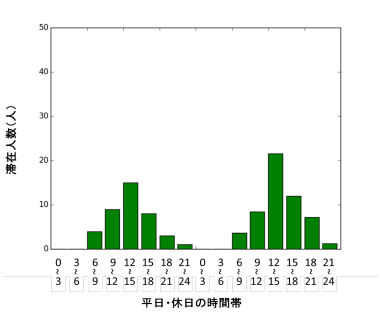

図 5　場所 B における訪問人数

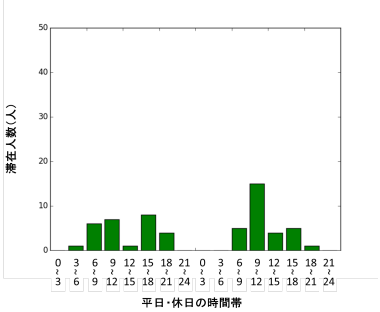

図 6　場所 C における訪問人数

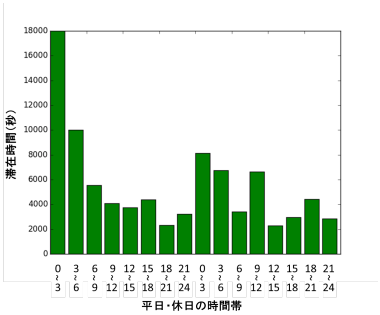

図 7　場所 A におけるユーザの平均滞在時間

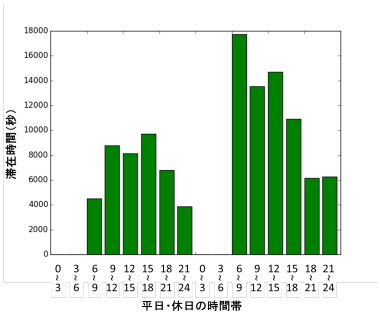

図 8　場所 B におけるユーザの平均滞在時間

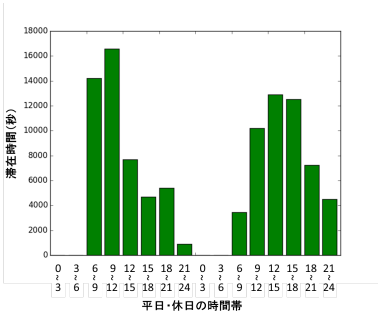

図 9　場所 C におけるユーザの平均滞在時間

図 3　平均滞在時間をもとにしたクラスタリング結果

場所 E が属する大型ショップを含むクラスタ 1-3 等が抽出されている．また，滞在時間を考慮したクラスタリング結果の図 3 を見ると，場所 A, 場所 H が属するクラスタ 2-2，場所 B，場所 C, 場所 I が属する飲食店やショップが含まれるクラスタ 2-1, 場所 D・場所 E が属する大型ショップを含むクラスタ 2-3 等が抽出されている．この結果からクラスタ 1-2 とクラスタ 2-2 やクラスタ 1-3 とクラスタ 2-3 のように似ている，もしくは同じ地点を含むクラスタが抽出されている．しかし，訪問人数を考慮したクラスタリングでは場所 A，場所 B，場所 C が同じクラスタ 1-1 に属することに対し，滞在時間を考慮したクラスタリングでは場所 B, 場所 C がクラスタ 2-2 に属しているが，場所 A はクラスタ 2-1 に属しているといったように異なるクラスタに属している．今回は特に，クラスタリングの結果が異なったこれらの地点 A,B,C について考察を行う．

二つの手法において異なるクラスタに分類された場所 A, 場所 B, 場所 C について考察を行うため，それぞれの場所の訪問人数と平均滞在時間を図 4 から図 9 に示す．図 4 から図 6 も見て取れるように，場所ごとに訪問人数について大きな違いは見られない．しかし, 平均滞在時間のグラフを見ると場所 A に関しては深夜 0 時から深夜 3 時ごろの滞在時間が場所 B, 場所 C での同時間帯の滞在時間に比べ，非常に長いことがわかる．これは, 場所 A はパーキングエリアであることから，運転に疲れたユーザが長時間過ごすためだと推測できる．一方，図 7 から図 9 を見ると場所 B，C に関し，特に休日の滞在時間のヒス

抽出地点	特徴	訪問人数クラスタ	滞在時間クラスタ
場所 A	大型のパーキングエリア．また，横浜ベイブリッジ等が見れる景色のよい場所．	1-1	2-2
場所 B	カップヌードルミュージアムや赤レンガ倉庫．観光やショッピングを楽める商業施設．	1-1	2-1
場所 C	横浜中華街．観光地でもあり食事処が多くある．	1-1	2-1
場所 D	ホームセンターやファッションセンターが存在．	1-3	2-3
場所 E	ファッションセンターが多く並ぶ場所．	1-3	2-3
場所 F	みなとみらい周辺のショッピングセンター．	1-2	-
場所 G	みなとみらい周辺のショッピングセンター．	1-2	-
場所 H	横浜市中央卸売市場．早朝からセリなどが行われる．	-	2-2
場所 I	赤レンガ倉庫付近のショッピングセンター	-	2-1

表 2　クラスタリングにて抽出された地点

トグラムの形が似ていることがわかる．これは，休日にショッピングや昼食に訪れ，周囲の観光地等を見ることで長時間滞在するユーザが多いためだと推測できる．また，午前中からきたユーザは夕方の時間帯に来たユーザに比べて滞在時間が長いという傾向も見ることができた．

これらのクラスタリング結果から, ユーザの訪問人数という情報では見つけることができなかったクラスタを滞在時間という情報を用いることで発見出来ることがわかる．ユーザの利用シーンに応じたクラスタリングが行われていることが確かめられた．

6. ま と め

本論文では，カープローブデータを用いて利用シーンの類似した地域を抽出するために,『滞在時間』に着目した地域特性分析を行った．実験では，ある地点における時間帯ごとのユーザの滞在時間を非負値のベクトルと同一視することで NMF を適用し，時間帯ごとに特徴的な行動を行うクラスタを抽出した．結果として，ある地点におけるユーザの滞在人数の情報だけでは把握することができなかったクラスタを抽出することができた．今後の課題としては、ユーザの滞在人数・滞在時間のどちらも考慮するような手法の検討を行う。また，今回は「平日と休日」のユーザの滞在時間用いて場所を特徴づけたが，休日を別途考慮することで休日特有の行動、例えば，観光目的で訪れられることが多い場所などのクラスタを発見する手法の検討も行う．

文　　献

[1] Kyosuke Nishida, Hiroyuki Toda, Takeshi Kurashima, and Yoshihiko Suhara. Probabilistic identification of visited point-of-interest for personalized automatic check-in. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pp. 631–642. ACM, 2014.

[2] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World wide web*, pp. 1029–1038. ACM, 2010.

[3] 澤田宏. 非負値行列因子分解 NMF の基礎とデータ/信号解析への応用. 電子情報通信学会学会誌, Vol. 95, No. 9, 2012.

[4] 樋口彰, 服部宏充. プローブカーデータに基づいた京都市観光者の観光行動分析. 人工知能学会全国大会論文集, Vol. 28, pp. 1–4, 2014.

[5] 熊谷雄介, 今井良太, 松林達史, 佐藤吉秀, 堀岡力. 非負値複合テンソル因子分解を用いた訪日外国人観光客の回遊行動分析. 信学技報, Vol. 115, No. 112, pp. 15–19, 2015.

[6] 李龍, 若宮翔子, 角谷和俊ほか. Tweet 分析による群衆行動を用いた地域特徴抽出. 情報処理学会論文誌: データベース (TOD54), Vol. 5, No. 2, pp. 36–52, 2012.