3C4-1

行動クラスタリングと隠れマルコフモデルを用いたユーザの位置予測モデル

User Location Prediction Based on Behavior Pattern Clustering and Hidden Markov Model

和田　計也 *1　増田　早紀 *2　梅林　泰孝 *2
Wada Kazuya　Masuda Saki　Umebayashi Yasutaka

*1株式会社サイバーエージェント　技術本部
Technical Department, CyberAgent Inc.

*2株式会社サイバーエージェント　アドテクスタジオ
AdTech Studio, CyberAgent Inc.

The growth of smartphones has made it easy to get enormous amount of user location data. If we can predict the next locations that a user is going from current and past location data, we can improve user experiences in several applications/services. In this paper, we propose a method to predict the next locations of a user based on behavior pattern clustering and Hidden Markov Model. Our experiment results show that location prediction by Hidden Markov Model achives higher precision than Naive Bayes method and combining user clustering with Hidden Markov Model possibly improves the performance of the prediction algorithm.

1. 背景と目的

スマートフォンの普及に伴い, 位置情報の取得が簡単になって来ている. もしユーザの同意が得られたら位置情報を取得することができ, アプリケーション作成者や各種のサービスプロバイダーがこの位置情報を分析してサービスのユーザ体験 (UX) を改善することが出来る. 分析の中にはユーザの活動地域の予測や近い将来の位置予測などが含まれており, 本研究ではユーザの将来の位置予測に焦点を当てる.

例えば, 従来の広告ターゲティングにおいて, 図 1a で示すように, 現在位置周辺の広告をユーザに対して表示することが多かった. 現在位置によるターゲティングでは消費行動がそのエリアに対して行われないケースだとコンバージョンまで結びつかないことが多く, 本当にユーザの欲している情報は次に行くエリアの情報であることが多々ある. 例えば, 昼頃にお店の予約をする際に昼の時間帯で予約を取るケースは稀で, 今夜のお店を予約するケースのほうが多いことは消費行動として至極当然である. そこで, ユーザが次に行く場所を予測できたら, 図 1b のように, ユーザに適切な広告配信などが出来る. ユーザの位置情報予測を通じてユーザの本当に必要としている深層心理に対して広告の訴求に繋げることが本研究の背景である.

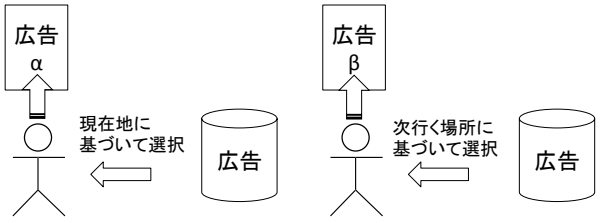
本研究では位置情報データの履歴から, 次にユーザが訪れる位置予測手法を提案する. 次に行動パターンを抽出して似た行動パターンを持つユーザ同士をクラスタリングする手法を提案する. また, クラスタリングの情報を用いることで, HMM を用いたユーザの行動予測精度を向上する手法を提案する. 実際のシステムでは，ユーザの位置しか取れずそれ以上のタグやコンテキストを取得するのは難しいことが多いので，本研究が応用できる範囲が広い.

連絡先: 和田　計也，株式会社サイバーエージェント　技術本部，Email: wada_kazuya@cyberagent.co.jp
連絡先: 増田　早紀，株式会社サイバーエージェント　アドテクスタジオ，Email: masuda_saki@cyberagent.co.jp
連絡先: 梅林　泰孝，株式会社サイバーエージェント　アドテクスタジオ，Email: umebayashi_yasutaka@cyberagent.co.jp

以下では, 第 2. 章に関連研究を延べ, 第 3. 章で提案手法を延べる. 第 4. 章で実験による評価を延べ, 最後に第 5. 章で結論を述べる.

2. 関連研究

通信会社をはじめとして GPS 位置情報を扱う研究が従来より進められてきた.



(a) 位置予測なしの位置情報に基づくターゲティング　(b) 位置予測ありの位置情報に基づくターゲティング

図 1: 位置情報を用いたターゲティング広告

山田らは測位誤差のある GPS データから正確な移動推移を推定する手法を提案した [山田 11].Wesley らは隠れマルコフモデルを用いて未来の位置予測をする手法を提案した [Mathew 12].

その他の既存研究では，位置予測問題を位置レコメンド問題として扱っている研究が多い [Ge 11, Zhuang 11]．Ying らは Geographic-Temporal-Semantic（位置，時間，コンテキスト）の頻出パターンマイニングアルゴリズムを利用して次の位置を予測する手法を提案した [Ying 13]．また，本研究でも Ying らの研究と同じように Stay Point を定義し，Stay Point を対象にユーザの次の訪問位置を推測する.

3. 提案手法

3.1 位置情報データの正規化

SDK で取得できるデータは GPS のように決まった時間間隔で緯度経度データを取得できるものではなく, ユーザが一定

距離移動したタイミングで緯度経度データを取得できる仕組みとなっている. 取得したデータは緯度, 経度, 日時が格納されているためユーザ毎に次点のデータとの差分を滞在時間とした. ユーザの位置情報の予測を行うに際し, データの正規化を行い各データがユーザの停留した緯度経度を表すようにした. 実際の広告配信のサービス化を考えた場合, ユーザの次に訪れる位置周辺の店舗広告でいいので, 本研究では地理情報を 1km のグリッドに変換した. 例えば図 2 に示す (2),(3) は同一グリッドに含まれているために同じ位置であると定義した. したがってユーザの行動は四, 五, 三, 九の順となり, 五の滞在時間は (4) と (2) 時刻の差分となる.

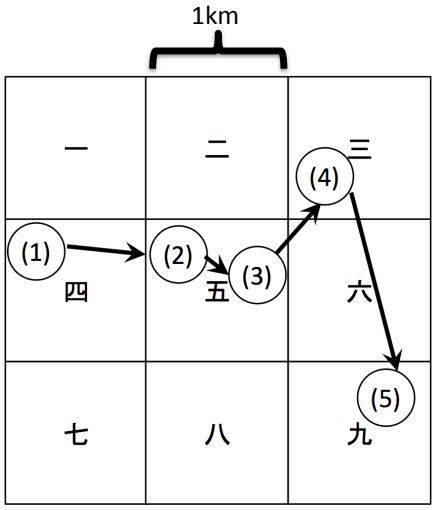


図 2: 位置データの正規化

3.2 Stay Point の定義

ユーザの位置情報は行動と滞在からなると考えるのが自然である. 滞在はユーザにとってその場所が目的地である一方, 移動は目的地へ行くための手段である. 本研究ではユーザの目的地を予測したいという観点から Ying らの提案 [Ying 13] を参考に, 滞在している場所を Stay Point とし以下のように定義した.

i) 全ユーザ全期間の位置情報を 1km のグリッドに変換した少なくとも 2 回以上の訪問ログが出ていること. ユーザが目的を持って滞在する位置であるべきと考えると, 利用データ中の 9000 人ほどのユーザが 1ヶ月で 1 度しか訪れていない位置はノイズのようなものだと考え除去した.

ii) 滞在時間が 20 分以上であること. 徒歩で移動していたとしても電車やバスを待っているとしても 1km グリッドであるなら 20 分未満で別なグリッドに移動するだろうと考え除去した.

3.3 個人の位置データを用いた HMM による位置予測

ユーザごとに Stay Point に絞った位置情報データが得られたら, 図 3 のように, 緯度経度の組み合わせペアを離散値とした離散隠れマルコフモデルによりモデリングを行う.

離散隠れマルコフモデルでは, 隠れ状態（図では楕円で表現されている）と観測値（長方形）がある. この場合、観測値はユーザの位置を表す. モデルのパラメータは各状態間の遷移確率と状態から観測値を出力する確率がある. これらのパラメータを学習できたら, ユーザが最も行きやすい位置を下記の式で計算出来る:

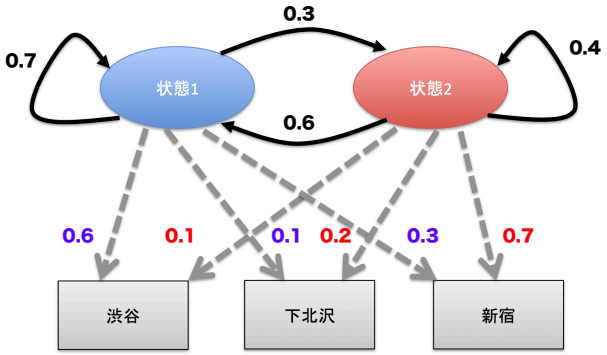


図 3: 離散隠れマルコフモデルの例

$$\text{next\_location} = \text{argmax} P(next_{s_n} | current_{s_n}, s_1 \cdots s_n) \quad (1)$$

そこで, 本手法では, ユーザの過去の位置推移データから上記の隠れマルコフモデルのパラメータを学習して, 位置予測を行う. 図 3 の例だと隠れ状態を 2 つとしているが実際はユーザ毎に最適な隠れ状態数は未知数であるため, ユーザ毎に隠れ状態数を 2~20 の場合で HMM でモデルを作り, モデルの適合度を表す指標でよく用いられている AIC(Akaike's Information Criterion) を算出して最小となる隠れ状態数の HMM をそのユーザの最終的なモデルとして採用した.

また, 位置予測手法として一番単純に考えられるのは Naive Bayes で, 下記のような式であり学習データから頻度の比例で算出可能である.

$$\text{next\_location} = \text{argmax} P(next | current) \quad (2)$$

この Naive Bayes をベースライン手法とし提案手法との比較に利用した.

3.4 ユーザ毎の代表的な行動パターンの抽出

ユーザ毎の代表的な行動パターンの抽出を行う方法は以下の手順で行った.

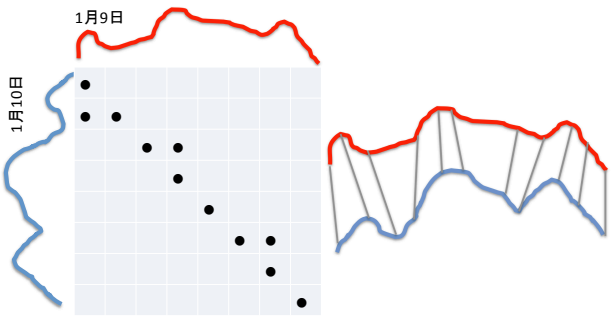


図 4: Dynamic Time Warping 法の概念

i) 1 ユーザ毎に, 一日の切り替わりを深夜 2 時として日ごとの位置情報に分割した後に, 各日の最初の位置データが欠落しているため前日の最終位置データで補間した.

ii) 1 ユーザ毎に, 分割した日毎の行動パターン間の距離を Dynamic Time Warping 法を用いて全ての組み合わせで算出した. 距離の算出は式 (3) にある通りで, ある一日の系列を s_a, 別な一日の系列を s_b とするとその二系列間の距離 d(s_a, s_b) は DTW でマッチした箇所 $s_{a\_matched}$, $s_{b\_matched}$ のユークリッド距離 dist を log 変換後の合計である.

$$d(s_a, s_b) = \sum \log(dist(s_{a\_matched}, s_{b\_matched})+1)+\delta* \text{ n} \quad (3)$$

ある程度以上距離が離れていた二点に関しては「比較的離れている」という評価でよく, それより比較的近くの離れてる距離をより厳密に評価したかったので log 変換を行っている. また,DTW で二系列間のアライメントをした際の挿入数を n とし, 挿入ペナルティとしての定数 δ を乗算した値を足している.

iii) 算出した距離を用いて,1 ユーザ内での日別行動パターンを Ward 法による階層クラスタリングを行った.

iv) 階層クラスタリングの結果を元に, 最も一致度の高い日ペアからはじめて 1 日ずつ日ごとの行動パターンを追加しながら整列していき, 最終的に 1 ユーザ内での複数日行動パターンの多重整列を得た.

v) 多重整列を表 1 のように，含有率 20%以上かつ 2 日以上で存在する位置情報という条件でフィルタリングを行い最終的には多数決で該当ユーザの代表的な行動パターンを抽出した. なお, 表 1 中の数値は緯度, 経度を表している.

表 1: アライメント後に代表的な行動パターンを抽出する例

含有率	40%	40%	20%	80%	20%	80%
1 月 9 日	36.21 138.62			35.98 139.12		35.77 136.34
1 月 10 日	36.20 138.62	36.58 138.44			35.92 139.00	
1 月 11 日			36.08 140.12	35.98 139.12		35.77 136.34
1 月 12 日		36.58 138.46		35.97 139.10		35.77 136.34
1 月 13 日				35.95 139.11		35.77 136.34
代表パターン	36.21 138.62	36.58 138.46	-	35.98 119.12	-	35.77 136.34

3.5 ユーザ間のクラスタリング

似た行動パターンを持つユーザ同士をグルーピングする目的でユーザ間でのクラスタリングを行った．クラスタリングに用いたデータは上記で抽出したユーザ毎の代表的な行動パターンであり,3.4 節の時と同様にして DTW (Dynamic Time Warping) 法でユーザ同士の距離を算出後に Ward 法による階層クラスタリングを行った.

3.6 クラスタリング結果を組み合わせたモデルによる位置予測

ユーザ間でのクラスタリング後に階層クラスタのリンクの高さが 2 倍以上になっている箇所でリンクを切断してクラスタの分離を行った. 例えば図 5 の場合では高さ 2.0 のリンクと高さ 1.0 のリンクの箇所が高さ 2 倍以上の箇所であるため, そこで切断してクラスタ 1 を形成している. この操作によりユーザは特定の 1 つのクラスタに所属するもしくはいかなるクラスタにも所属しないという状態となる．このクラスタ毎に連続

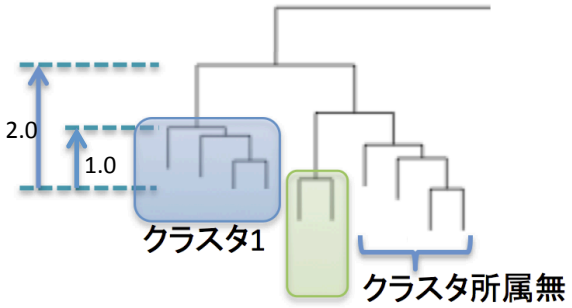


図 5: クラスタの作り方

型の HMM モデリングを行いクラスタ毎の予測も行った．そして 1 ユーザの位置情報の中で今までそのユーザが訪れたことの無い位置情報だった場合に, ここで算出したクラスタでの予測で補間する方法をとった.

4. 実験による評価

4.1 実験データ

実験データには 2014 年 12 月 10 日〜2015 年 1 月 29 日の期間に SDK を利用して収集された位置情報データから, 一回以上関東地方近郊で位置情報が取得できたユーザを対象にした. 関東地方にいる人の行動はもっとも複雑で予測しにくいと考えられる. そこで, 関東地方のユーザの次の位置が予測できたら, 提案手法は他のエリアにも適用できるので, まずは関東地方に絞って実験を行った. この段階で 9034 人のユーザに絞られておりこのユーザ群で Stay Point の検出は行った. 位置予測に関してはこの 9034 人から更に 256 人をランダム抽出したユーザ群を用いた. SDK から送られてくるログデータは世界測地系の緯度経度を表す座標データとデータの取得日時, ユーザ固有の識別子から成っており, 各ユーザ平均で 76 箇所の位置データを期間内に飛ばしていた.

ユーザ毎に Stay Point と判定された地点の情報のみを利用し, 位置データの正規化を行った. その後, ユーザ毎に最終日の位置情報を評価用データセットとしそれ以外をトレーニング用データセットとした.

4.2 ユーザ毎の位置データからの予測結果

第 3.3 節に述べたように, 隠れマルコフモデルを用いた位置予測とベースライン手法（Naive Bayes）との性能比較を行った．その比較結果を表 2 に示す. Naive Bayes を用いたベースライン手法は正解率が 21.2%であったのに対して, 隠れマルコ

表 2: ベースライン手法と HMM との予測性能の比較

手法	正解数	不正解数	正解率
Naive Bayes(ベースライン)	201	745	21.2%
隠れマルコフモデル	320	626	33.8%

フモデルでは 33.8%の正解率となり，大きく性能が改善されたことが分かる．

4.3 階層型クラスタリング結果

ユーザ間でクラスタリングを行い 1 つのクラスタを切り出した結果の一例を図 6 に示す．

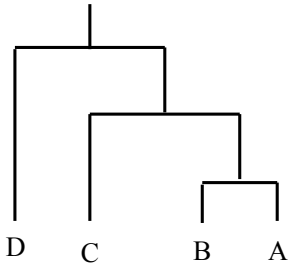


図 6: 4 ユーザのヒストグラム例

このクラスタには 4 人のユーザ A,B,C,D が含まれており,A と B の距離が一番近く, 同じ階層にある. 一方 A と D の距離は一番遠く別の階層にある. このクラスタのユーザの行動パターンを実際に地図の上にプロットした結果を図 7 に示す. 図 7 では異なるユーザは色・形状の違いにより表現されている. この図を見ると A と B は千葉と東京を往復しているので同じ行動パターンをしている. 一方で A と D はあまり行動パターンが似ていないことがわかる. 従ってクラスタリングアルゴリズムはこの例では良く機能しているといえる.

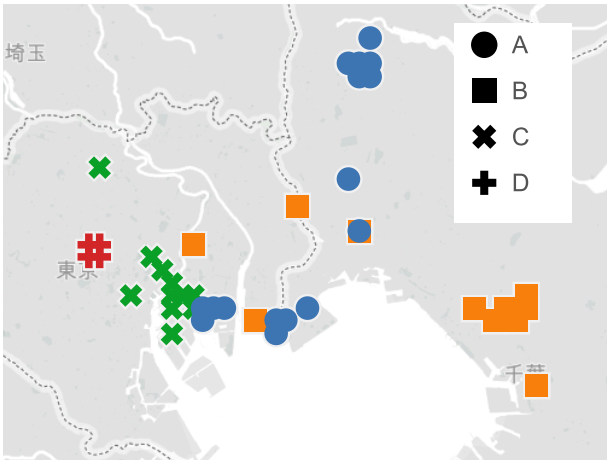


図 7: 同一クラスタに含まれるユーザ例

4.4 クラスタリングと HMM による位置予測結果

第 3.3 節で, ユーザの未訪問場所では, 遷移確率が 0 のため, 位置予測が出来ない. そこでユーザクラスタリングにより付与されたクラスタを一人のユーザとして扱い, 第 3.3 節で記述した手法を適用して, そのユーザクラスタの次の位置を予測する. そして, 次の位置が予測出来ないユーザについて, ユーザクラスタの次の予測位置を当てはまる.

このクラスタの位置を利用する時と利用しない時の予測結果を表 3 に示す. クラスタリング結果を組み合わせた予測モデルは正解率が 34.0%であり, クラスタリングなしのときと比べ,0.2%と若干の精度向上が見られた. しかしクラスタリング結果を組み合わせない手法からの明らかな有意性は認められなかった (フィッシャーの正確確率検定による片側検定で p 値 = 0.481).

表 3: ユーザクラスタリングありとなしの場合の性能比較

手法	正解数	不正解数	正解率	p 値
クラスタリングなし	320	626	33.8%	-
クラスタリングあり	322	624	34.0%	0.481

5. 結論

本稿ではユーザの過去の位置情報データを使い, 隠れマルコフモデルによる位置予測手法を提案した. 提案手法はベースライン手法と比べて, 大きく予測精度を改善出来た.

また, ユーザの未訪問場所も予測できるように, 行動パターンを抽出しクラスタリングを行った. クラスタリングを用いることで,0.2%の精度向上に繋がった. 今後は更なるクラスタリング手法の検討や, クラスタリング結果を適切に組み入れたモデルにすることで精度向上を目指していきたい.

参考文献

[Ge 11] Ge, Y., Liu, Q., Xiong, H., Tuzhilin, A., and Chen, J.: Cost-aware travel tour recommendation, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, San Diego, CA, USA, August 21-24*, pp. 983–991 (2011)

[Mathew 12] Mathew, W., Raposo, R., and Martins, B.: Predicting future locations with hidden Markov models, in *Proceedings of the 14th ACM International Conference on Ubiquitous Computing, Ubicomp 2012, Pittsburgh, PA, USA, September 5-8*, pp. 911–918 (2012)

[Ying 13] Ying, J. J., Lee, W., and Tseng, V. S.: Mining geographic-temporal-semantic patterns in trajectories for location prediction, *ACM Transactions on Intelligent Systems and Technology (ACM TIST)*, Vol. 5, No. 1, pp. 2:1–2:33 (2013)

[Zhuang 11] Zhuang, J., Mei, T., Hoi, S. C. H., Xu, Y., and Li, S.: When recommendation meets mobile: contextual and personalized recommendation on the go, in *Proceedings of the 13th ACM International Conference on Ubiquitous Computing, UbiComp 2011, Beijing, China, September 17-21*, pp. 153–162 (2011)

[山田 11] 山田 直治, 礒田 佳徳, 南 正輝, 森川 博之：屋外行動支援のための GPS 搭載携帯電話を用いた移動経路の逐次的精錬手法, 情報処理学会論文誌, Vol. 52, No. 6, pp. 1951–1967 (2011)