Vol.2012-CDS-4 No.18
2012/5/11

滞留場所間遷移途中の位置を利用した行動予測手法

西野 正彬[1,a)]　小林 透[1,b)]

概要：近年，多くの携帯電話端末に Global Positioning System(GPS) による測位機能が搭載されるなど，ユーザの位置を継続的に測位できる環境が整いつつある．位置データが継続的に取得されるようになると，蓄積されたデータからユーザの行動の規則性を抽出し，ユーザの未来の位置を予測することが可能となる．これまで様々な位置の予測手法が提案されてきているが，その多くは前処理としてユーザが滞留した場所を抽出するなどの離散化を行い，離散的な時系列データに変換した上で予測を行うものであった．しかし，離散化を行うことで処理が容易になり，長期間の予測精度も向上できる一方で，離散化の際に捨てられていたデータを予測に利用できないために，逐次的に位置データが取得される状況で，それに合わせて予測を更新することができないという問題があった．本稿では，既存手法では捨てられていた場所間を遷移する途中の位置を行動予測に用いることによって，予測の逐次更新性を高める手法を提案する．手法の有効性を実データを用いて検証した．また，予測時の計算時間を削減する手法も示した．

キーワード：位置情報，行動予測，ライフログ

A Location Predictor Using the Locations on Transitions between Stayed Places

Masaaki Nishino[1,a)]　Toru Kobayashi[1,b)]

Abstract:
In these days, many mobile phones are equipped with a GPS (Global Positioning System) location services, and it becomes much easier to continuously obtain the locations of a person who carries a mobile phone. With these obtained location data, we can predict the future locations of a person by extracting regularities from them. By predicting future locations, it is possible to deliver useful information related to the predicted locations to a mobile phone user more quickly. Many prediction methods first descretize location data to make descrete time-series sequence and then do predictions on these descrete data. These methods have a common drawback that they cannot revise predictions as the current location of a user is updated. In this paper, we propose a method to improve the frequency of updates of prediction by using the location data on transitions between the stayed places. We show the effectiveness of our method with experiments with real data. We also show a approximation method to reduce the calculation time needed to make predictions.

Keywords: location data, movement prediction, lifelogs

1. はじめに

近年，スマートフォンとよばれる様々な機能を搭載した携帯電話端末が広く普及するなど，人々が日々の生活で触れる情報機器はますます増加しつつある．そのような中で，各種端末等によって取得された機器の利用ログなど，ユーザの活動に関係する様々な行動ログ（ライフログ）を収集・蓄積することによって，より利便性の高いサービスの実現を目指す動きが活発になっている．ライフログを用いることによって，各個人の嗜好や状況を推定し，最適な情報を提供したり [1]，行動の振り返りを可能にできる [2][3] など，

1 日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation
a) nishino.masaaki@lab.ntt.co.jp
b) kobayashi.toru@lab.ntt.co.jp

ⓒ2012 Information Processing Society of Japan

その幅広い応用が期待されている．

様々な応用が期待されるライフログの代表的なものとして，ユーザの位置情報が挙げられる．近年，多くの携帯電話端末に GPS（Global Positioning System）が搭載されるなど，位置情報を取得し利用することは一般的になりつつある．携帯電話端末における位置情報の利用の代表的なものとして，端末で撮影した写真を整理するために位置情報を付与するジオタギングや，経路案内（ナビゲーション）などがある．

さらに，ユーザの位置の継続的な取得が一般的になると，蓄積された位置情報から行動の規則性を抽出することによって，ユーザが未来に訪れそうな場所の予測を行うことが可能となる．ユーザが訪れる場所を前もって予測することができれば，ユーザに対してより利便性の高い情報を提供することが可能になる．位置情報のログから未来の位置を予測する手法については，これまでも様々な手法が提案されてきている [4][5][6] [7][8][9] ．手法は大きく 2 種類に分けることができる．まず，位置情報のログに前処理を施すことによって，離散的な系列として表現し，それを対象として予測を行う手法がある [4][5][6] ．次に，そのような前処理を行わずに入力された位置データを対象として予測を行うための予測関数を推定し，それを用いて予測を行う手法がある [7][8]．前者の手法には，離散化によってデータの扱いが簡単になること，そしてユーザの蓄積された過去の行動ログを活用した予測と相性がよいという利点がある．一般に，人々の日々の位置の軌跡は，過去のそれと似ている．通勤や通学などはその最たる例であり，日々の活動で位置が大きく変わることはない．そして，こうした日々の行動は滞在した場所に着目することで取り出すことができる．例えば通勤であったら，自宅から職場に向かい，その後自宅に戻るという滞在した場所の列として表現することができる．以上より，離散化することによって繰り返し行われる活動を過去のデータより取り出して行動予測に利用できる．その一方で，前処理として離散化を行う過程で取得されたデータの一部を破棄しているため，ユーザの最新の位置情報が与えられたときに，それに応じて逐次的に予測内容を変化させることが難しい．後者の手法については，離散化を必要としないために，新たに入力される最新の位置に応じて逐次的に予測を変化させることは可能なものの，予想可能な軌跡が予測に用いる関数によって制限されるという問題がある．例えば最も素朴な予測として線形予測が挙げられるが，線形予測では直線的な移動しか予測することができない．しかし，人々の移動経路は道路や鉄道に大きく影響を受けるため一般には直線的ではないし，数時間後の位置の予測となるとなおのこと直線上を移動していると考えるのは現実的ではない．

本稿では，こうした状況を踏まえて，未来の位置を利用したサービスを提供するため適した行動予測手法を提案する．具体的には，ユーザの位置情報が継続的に得られるという前提のもと，予測精度，予測の更新頻度，予測の効率性といった観点から必要となる要件を明らかにし，その要件を満たす手法を提案する．提案手法のポイントは，離散化による手法をベースとしながら，これまで離散化の際に捨てられていたデータを有効活用することによって，予測の更新性を向上させる点にある．これによって，離散化による手法の利点である，ユーザの行動に応じた予測を可能としつつ，かつ予測の逐次更新性を高めることができる．

本稿の主要な貢献は以下の 3 点である．

- 未来の位置を予測し，活用するサービスを想定したうえで，その際に必要となる要件をまとめた．
- 離散化に基づく予測手法において，離散化の際にこれまで捨てられていた遷移途中の位置を利用することによって，予測の逐次更新性を高める方法を考案した．
- 上記遷移途中の位置を活用する予測手法の処理性能を改善するための手法を提案した．

以下，2 節で，本稿で想定する予測がなされる条件とその条件のもとで必要となる機能要件について述べる．その後，3 節で関連研究について述べる．次に，4 節で本稿で提案する予測手法の基本的な考え方について述べた後に，5 節でその処理性能を改善するための手法について述べる．6 節で実際に収集した GPS データを用いて手法の有用性について検証したのち，7 節で結論を述べる．

2. 想定する条件と必要となる機能

手法についての具体的な説明に入る前に，行動予測手法が満たすべき条件について考察する．まず初めに本稿で想定する予測が利用される状況について説明し，次にそのような条件でどのような機能が必要となるかの検討を行う．

2.1 予測が行われる状況

まずは本稿で想定する予測が利用される状況について説明する．予測と一言で言っても，どのような状況で利用されると想定するかによって問題設定が変わってくる．以下で本稿で想定する予測の利用方法と，予測を行うための入力となる GPS データがどのように獲得されるかについて述べる．利用形態は現実的なものであるし，また GPS データの取得方法についても現在の状況に即したものとなっている．

ユーザの未来の位置の予測結果の利用方法は幅広いが，本稿では予測された位置に関連する情報の配信に利用されるものとする．配信の対象となる情報としては，例えば予測された先での交通情報であったり，そこで開催されるイベントといった，誰にでも関係するような情報から，その場所に関連するスケジュールなどの個人的なものも含まれるものとする．現在でもユーザの現在位置に関する情報を配信するサービスは多数存在するが，本稿で想定する利用

形態はその自然な拡張となっている．

ユーザの位置を取得する手法には様々なものがあるが，本稿では継続的に位置が測定され続けるものであることを前提とする．例えば携帯電話端末はユーザが常に持ち歩くものであり，近年では携帯電話端末を用いてユーザの位置を継続的に取得し続けることも一般的となってきている．例えば NTT ドコモのオート GPS サービス [10] では，携帯電話端末の位置を自動的に継続して取得することで，情報配信等の位置利用サービスに利用している．そのほか，カーナビゲーションシステムにおいてもユーザの位置は継続的に取得され，経路案内に利用されている．

以上のように，本稿では継続的に位置情報が取得されることを前提として，場所に関する情報配信を行うことを目的とした予測を行うこととする．次に今回定義した問題設定における行動予測において必要となる機能について議論する．

2.2 機能要件

上記の条件でどのような行動予測ができればよいかを議論する．まず，どの程度の未来を予測できればよいかを検討する．例えば数分後の近い未来であっても，行動を予測し未来の場所に関する情報を前もって配信することにメリットはある．しかし，例えば予測された場所の天候に応じて傘を準備する等，予測を行動に活用するためには，できるだけ遠い未来，数十分後から数時間後までのユーザの位置を正確に予測できることが必要となると考える．

次に，現実の予測において重要となるのは，誤った予測をできるだけ素早く更新することである．一般に行動予測は難しい問題であり，常に正しくユーザの未来の位置を予測できるものではない．この前提のもと，本稿で述べる手法では，予測の逐次更新性を重視する手法を考える．予測の逐次更新性とは，ユーザの位置情報が逐次的に入力としてあたえられたときに，それに応じて予測の精度を高めて行ける性質であると定義する．予測システムが逐次的に予測を更新できないとすると，予測に基づく情報配信サービスのユーザは古い位置情報に基づいた予測を受けることになる．一般に遠い未来ほど活動の予測は困難となるため，古い位置に基づいた予測は誤ったものになる可能性が高い．

以上より，本稿で提案する未来の位置の予測手法の要件を以下のように定める．

(1) 数十分〜数時間後の位置を正確に予測できること．
(2) 逐次更新性を備えていること．

3. 関連研究

カーナビゲーションや位置測位機能を備えた携帯電話端末の普及に伴い，この 10 年あまりでさまざまな予測手法が提案されてきている．ほとんどの手法は予測する対象の位置が継続的に取得可能であることを前提として，予測を行っている．以下，予測手法を (1) 取得された位置データの離散化を前提とする手法，(2) 離散化を行わなず，まず入力された位置データをもとにして予測関数を推定する手法の 2 種類に分類して先行研究について述べる．

表 1 予測方法の比較
Table 1 Comparation of rediction methods.

	長期間の予測	逐次更新性
離散化法	○	×
関数法	×	○
提案手法	○	○

3.1 離散化を行う手法

これまで位置データの履歴を利用して活動予測をする際には，処理に先立って位置データを離散化したのちに，離散データを対象として処理を行う手法がよくとられている．すなわち，位置情報データに対して，クラスタリング等の処理を適用することにより，ユーザが滞留した場所および車で走行した道路を離散系列とした上で処理を行う手法がとられている．

Ashbrook らは，GPS データのログに対してクラスタリングアルゴリズムを適用して，滞留した場所を抽出したのちに，マルコフモデルによってユーザの遷移先を予測する手法を提案している [4]．Monreale らは，Trajectory Pattern[11] とよばれる，時間情報を含んだ系列パターンを利用した行動予測手法を提案している [5]．Trajectory Pattern の作成においても，GPS データから Region of Interest とよばれる，特徴的な場所を抽出する離散化処理を行っている．Liao らは，Dynamic Bayesian Networks によって，未来の位置と移動手段とを同時に予測する手法を提案している [6]．Liao らの手法では，GPS データは近傍の道路に対応づけられて，処理が行われる．これも一種の離散化と考えられる．

なお，離散化した位置データと他の種類のライフログを組み合わせて行動を予測する手法もこれまでに提案されてきている．著者らはスケジューラデータと GPS データとを組み合わせて，Dynamic Bayesian Networks を用いて行動を予測する手法を提案している．[12]．また，携帯電話端末によって取得できる基地局 ID のログと Blutooth ID のログを併用して，隠れマルコフモデルで活動を予測する手法もいくつか提案されている [13], [14]．

位置情報の履歴を離散状態の系列に変換することで，系列データから規則性を抽出するデータマイニング・機械学習のアルゴリズムを適用してユーザの行動の長期的な規則性を抽出することが可能となる．結果として，数時間後の遠い未来（長期間）の行動予測が可能となる．その一方で，離散化の際に，ある離散状態から別の離散状態に遷移する間の情報が捨てられてしまうため，予測を逐次的に更新す

ⓒ2012 Information Processing Society of Japan

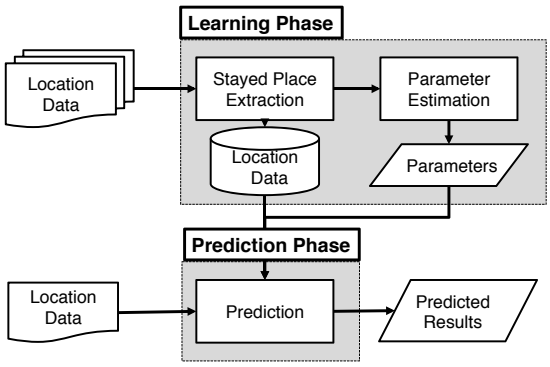

図 1　予測フロー
Fig. 1　The flow chart of the prediction process.

ることが困難になるという問題がある．

3.2　予測関数を用いる方法

離散化を経由せず，入力された位置データをすべて利用して予測を行う手法も提案されている．Tao らは，入力された過去の観測データをもとに，線形に経路を予測して未来の位置を予測する手法 [7]，および非線形な関数を同定して予測を行う手法を提案している [8]．これら，関数による予測を行う手法では，離散化の際に発生していたデータの欠落が発生しないため，常に最新の入力をもとにして予測を行うことが可能となる．すなわち，予測の逐次更新性が高い．その一方で，関数による予測方法では直近の過去の位置情報のみを入力として予測を行うため，数分後の予測はできても数時間語の予測は難しいという問題がある．

先行研究と提案手法との関係を表 1 に示す．提案手法では離散化に基づいた予測を行うことによって，数時間後の予測を可能としている．また，遷移途中の位置を予測に用いることによって，予測の逐次更新性を高めることにも成功している．これらの点から，携帯電話端末などによって人の位置が逐次的に取得可能な状況においては，提案手法が有効であると言える．

4.　遷移途中の位置を用いた予測手法

以下，本稿で提案する予測手法について説明する．予測手法のフローを図 1 に示す．予測手法は大きく分けて，入力された位置データからパラメータを推定する学習フェーズ (Learning Phase) と，実際に逐次的に位置データが入力される状況で予測を行う予測フェーズ (Prediction Phase) とに 2 分される．それぞれについて順に説明する．

4.1　学習フェーズ

学習フェーズは，位置データの前処理としての離散化と，離散化されたデータを入力とした遷移確率の推定の 2 つの処理に分けられる．ここで，前処理では生の GPS データから滞留した場所をクラスタリングによって抽出し，ユーザの行動を滞留した場所の系列（離散系列）に変換する．

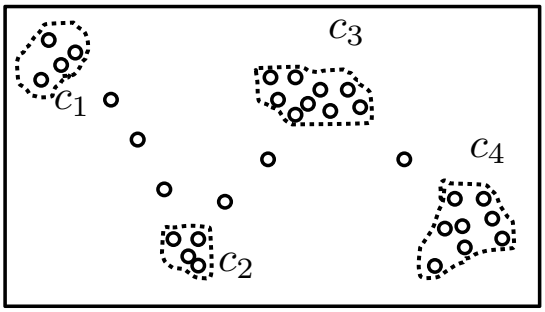

図 2　クラスタリングによる滞留場所抽出の概念図
Fig. 2　Illustration of extracting stay place from GPS data.

離散系列への変換までは既存手法でも同等の処理を行なっているが，本稿で提案する手法では，既存手法では離散化において捨てられていた滞留場所間の遷移途中の状態についても予測に利用するための前処理を行う．

4.1.1　位置データの前処理 (Stayed Place Extraction)

学習フェーズで，入力としてあたえられる位置データの集合を $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{D}|}$ とする．ここで，各 $\mathbf{x}_i$ は 2 次元の実数ベクトル $\mathbf{x}_i = (x_i, y_i)$ とし，2 次元空間での座標（緯度・経度）を表す．また，$\{t_i\}_{i=1}^{|\mathcal{D}|}$ を $\mathbf{x}_i$ が観測された時刻を表し，$1 \leq i, j \leq |\mathcal{D}|$ において，$i < j$ なら $t_i < t_j$ であるとする．なお，位置 $\mathbf{x}_i$ は数十秒から数分程度の時間間隔で定期的に取得されているものとする．

$\mathcal{D}$ に対する前処理として，[15], [9] を参考に位置データ集合 $\mathcal{D}$ にクラスタリングアルゴリズム DBSCAN[16] を適用して，対象が滞在した場所を抽出する．DBSCAN アルゴリズムは，データの集合を密度によってクラスタリングすることが可能なクラスタリングアルゴリズムである．DBSCAN には任意の形状のクラスタを抽出することができる，データに含まれるノイズの影響を受けにくい，抽出されるクラスタの個数を事前に定めなくてよいなどの特徴がある．DBSCAN には 2 種類のパラメータ ϵ，NumPoints があり，それぞれ抽出されるクラスタのサイズ，抽出されるクラスタの密度のしきい値である．DBSCAN アルゴリズムを位置データに適用することによって，対象が滞留した場所を表すクラスタの集合 $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$ を得る．ここで各 c_i は対象が滞留した 1 つの滞留場所を示す．各 c_i は 2 次元空間上でそれぞれのクラスタに対応する固有の領域をもつとし，これを $A(c_i)$ とする．$1 \leq i, j \leq |\mathcal{C}|$ である i, j について，$i \neq j$ ならば $A(c_i) \cap A(c_j) = \emptyset$ とする．図 2 は，DBSCAN によって滞留場所を抽出する例を示している．図中の点は学習フェーズで与えられた位置データである．図中の c_1〜c_4 がクラスタとして抽出される滞留場所である．

クラスタ集合 $\mathcal{C}$ が求まると，各 $\mathbf{x}_i \in \mathcal{D}$ について，それがあるクラスタ c_k について $\mathbf{x}_i \in A(c_k)$ であるかどうかが判定できるようになる．$x_i \in A(c_k)$ のとき，時刻 t_i で c_k に滞

ⓒ2012 Information Processing Society of Japan

留したとよぶ．また，$\alpha \leq i \leq \beta$ の各 i について $x_i \in A(c_k)$ のときに，時刻 t_α から t_β までの間連続して c_k に滞留したとよぶことにする．$\mathcal{D}$ の全データについて，連続して滞留した場所を時刻順に並べた系列を $\mathcal{S} = \{s_i\}_{i=1}^{|\mathcal{S}|}$ とする．各 s_i はいずれかの滞在場所であり，すなわち $s_i \in \mathcal{C}$ である[*1]

$\mathbf{x}_i \in \mathcal{D}$ がすべての $c_k \in \mathcal{C}$ $1 \leq k \leq |\mathcal{C}|$ について $\mathbf{x}_i \notin A(c_i)$ であるとき，$\mathbf{x}_i$ を遷移途中であるという．遷移途中である観測点を集めた $\mathcal{D}$ の部分集合を $\mathcal{T} \subseteq \mathcal{D}$ とする．

4.1.2 学習 (Parameter Estimation)

学習フェーズでは，パラメータとして各滞留場所間の遷移確率 $P(c_i|c_j)$ を求める．$P(c_i|c_j) = \sigma_{ij}$ とすると，最尤推定によって σ_{ij} を $\mathcal{S}$ から以下のように求めることができる．

$$\sigma_{ij} = \frac{\sum_{k=1}^{|\mathcal{S}|-1} I((s_k = c_j) \wedge (s_{k+1} = c_i))}{\sum_{k=1}^{|\mathcal{S}|-1} I(s_k = c_j)} \tag{1}$$

ここで，$I(\cdot)$ は引数が真のときに 1，そうでないときに 0 を返す関数である．

次に，各 $\mathbf{x} \in \mathcal{T}$ について，その直前に滞留していた滞留場所を before($\mathbf{x}$)，直後に滞留していた滞留場所を after($\mathbf{x}$) とし，$\mathbf{x}$ とともに記憶する．

4.2 予測フェーズ

予測は，新しい点 $\mathbf{x}$ が入力として与えられた際に，対象が次に滞留すると考えられる場所 c' を返す手続きである．この手順をアルゴリズム 1 に示す．

Algorithm 1 予測アルゴリズム

Input: 現在の位置 $\mathbf{x}$，直前の滞留場所 c_j
Output: 遷移確率が最大となる次の滞留場所 $c' \in \mathcal{C}$

```
if x ∈ A(c_k) となる c_k ∈ C が存在 then
    c' = arg max_{c_i ∈ C} P(c_i|c_k)
else
    c' = arg max_{c_i ∈ C} P(c_i|c_j, x)
end if
return c'
```

予測時は，$\mathbf{x}$ がいずれかの滞留場所に含まれるか，そうでないかで処理を分けている．もし $\mathbf{x}$ がある滞留場所 c_i について $\mathbf{x} \in A(c_i)$ であるなら，(1) で求めた遷移確率に基づいて確率最大となる $c' \in \mathcal{C}$ を次の滞留場所の候補として出力する．そうでない場合，すなわち $\mathbf{x}$ が遷移途中の点であるならば，$\mathcal{T}$ および，$\mathbf{x}$ の直前に滞留していた場所 c_j を利用して以下のように遷移確率 $P(c_i|c_j, \mathbf{x})$ を定義する．

[*1] 屋内では GPS による測位は不可能である場合が多いため，GPS データをそのままクラスタリングしても屋内での滞留は取得できない場合がある．本稿では [15] で示されている方法を用いて，屋内に滞在時の GPS データを補間した後に DBSCAN を適用する．

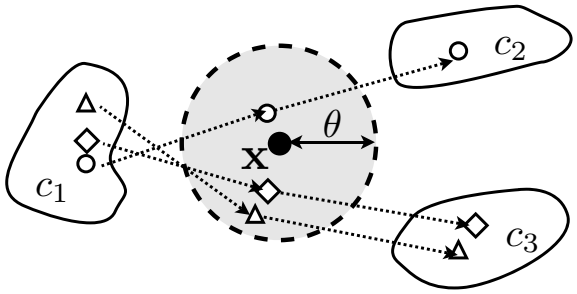

図 3 遷移確率計算の例
Fig. 3 An example of calculating transition probabilities.

$$P(c_i|c_j, \mathbf{x}) = \frac{\sum_{\mathbf{x}' \in T_{ij}} k_\theta(\mathbf{x}, \mathbf{x}')}{\sum_{\mathbf{x}' \in T_j} k_\theta(\mathbf{x}, \mathbf{x}')} \tag{2}$$

ここで T_{ij}, T_j はそれぞれ $T_{ij} = \{\mathbf{x} | \mathbf{x} \in \mathcal{T}, \mathrm{before}(\mathbf{x}) = c_j, \mathrm{after}(\mathbf{x}) = c_i\}$, $T_j = \{\mathbf{x} | \mathbf{x} \in \mathcal{T}, \mathrm{before}(\mathbf{x}) = c_j\}$ と定義する．また，$k_\theta(\mathbf{x}, \mathbf{x}')$ はカーネル関数であり，$||\mathbf{x} - \mathbf{x}'|| \leq \theta$ ならば $k_\theta(\mathbf{x}, \mathbf{x}') = 1$，そうでなければ 0 とする．

式 (2) では，ある滞留場所 c_j から別の滞留場所 c_i に向かう際に，同じ経路を用いることが多いという仮定のもと，訓練データ中での $\mathbf{x}$ の近傍の点を利用している．例えば電車での通勤であれば，その経路は常に同じであるし，車での移動であっても同じような道を利用することが多い．

提案手法を用いた予測の例を図 3 に示す．図中の c_1, c_2, c_3 は滞留場所であり，白い点 (△，◇，○) はそれぞれ過去の滞留場所 c_1 から c_2 または c_3 への遷移の軌跡を表している．対象が滞留場所 c_1 に滞留したのちに，点 $\mathbf{x}$ で最新の位置が得られたとする．このとき，$\mathbf{x}$ から距離 θ 以内にある近傍の点を用いて，遷移確率 $P(c_2|c_1, \mathbf{x})$，$P(c_3|c_1, \mathbf{x})$ がそれぞれ 1/3, 2/3 として計算される．例えば対象が c_3 へ向かっている途中であるとすれば，c_3 に近づくにつれ $P(c_3|c_1, \mathbf{x})$ はさらに大きくなる．

5. 効率の改善

前節で示した予測方法では，(2) の計算において，毎回 $\mathcal{T}$ に含まれるすべての点にアクセスする必要がある．この計算は，空間データへの高速なアクセスを可能とする R* 木 [17] のようなデータ構造を用いることで効率化できるが，それでも訓練データが蓄積されるにつれアクセスするデータの個数が増えることには変りない．逐次的な予測の更新を考えたときには，一度の予測にかかる計算量が小さいことが望まれるので，ここでは近似的に解を与える方法について述べる．

式 (2) では，任意の位置 $\mathbf{x}$ について遷移確率を与えていた．これを，前もって定めておいた有限個の点 $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$ について計算した遷移確率によって近似することを考える．各点における遷移確率は (2) で求める．今回，$\mathbf{z}_i$ は図 4 中の白い点のように格子状に配置するものとす

ⓒ2012 Information Processing Society of Japan

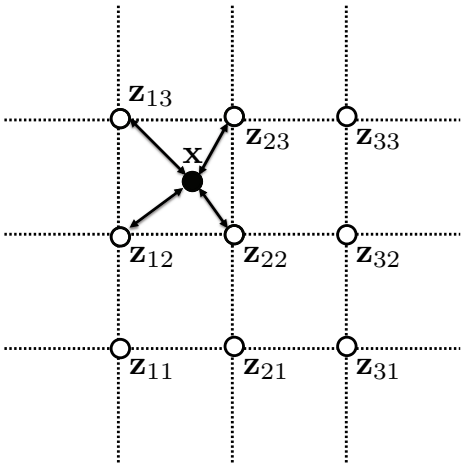

図 4　近似解の計算の概念図
Fig. 4　An Illustration of approximated inference.

る．長方形の領域に $N \times M$ 個の点を並べ，それぞれ $\mathbf{z}_{11}, \ldots, \mathbf{z}_{1M}, \ldots, \mathbf{z}_{N1}, \ldots, \mathbf{z}_{NM}$ と表記する．

5.1　学習フェーズ

図 1 のフローの学習フェーズ (Learning Phase) では，クラスタリングによる前処理の後，パラメータ σ_{ij} の推定とともに，$\mathbf{z}_{11}, \ldots, \mathbf{z}_{1M}, \ldots, \mathbf{z}_{N1}, \ldots, \mathbf{z}_{NM}$ 点での確率分布 $P(c'|c, \mathbf{z}_{ij})$ を推定する．なお，点 $\mathbf{z}_{ij}$ で記憶されている遷移確率分布 $P(c'|c, \mathbf{z}_{ij})$ を，以下では $P_{ij}(c'|c)$ とも表す．なお，4 節の手法では学習フェーズで訓練データを記憶していたが，本節で示す手法では訓練データは記憶せず，遷移確率を定めるパラメータのみを記憶する．

5.2　予測フェーズ

予測フェーズ (Prediction Phase) では学習された分布を用いて，最新の位置 $\mathbf{x}$ での確率分布を近似する．図 4 のように配置すると，任意の入力 $\mathbf{x}$ に対し，近傍の点が 4 つ定まる．これらの 4 点をそれぞれ $\mathbf{z}_{ab}, \mathbf{z}_{a,b+1}, \mathbf{z}_{a+1,b}, \mathbf{z}_{a+1,b+1}$ とすると，$P(c_i|c_j, \mathbf{x})$ は以下のように近似される．

$$P(c_i|c_j, \mathbf{x}) = \frac{1}{W} \sum_{k \in \{0,1\}} \sum_{l \in \{0,1\}} w_{kl} P_{a+k,b+l}(c_i|c_j) \quad (3)$$

ここで，w_{kl} は以下の式によって定義される，$\mathbf{x}$ と $\mathbf{z}_{a+k,b+l}$ との距離を反映した重み係数である．

$$w_{kl} = \exp\left(\frac{-||\mathbf{x} - \mathbf{z}_{a+k,b+l}||^2}{\gamma^2}\right) \quad (4)$$

ここで γ は距離による重み付けを調整するパラメータである．また，W は正規化項であり，

$$W = \sum_{k \in \{0,1\}} \sum_{l \in \{0,1\}} w_{kl} \quad (5)$$

として定義する．

各 $\mathbf{z}$ において遷移確率をあらかじめ計算しておくことによって，予測時に必要となる計算量を削減することが可能となる．各 $\mathbf{z}$ の分布の学習は事前に行う必要があるが，これも効率よく行うことができる．式 (2) での定義のように，確率分布 $P_{ij}(c'|c)$ の推定は，近傍の点についてカウントを行うだけである．このため，新たに訓練データが与えられたときの確率分布 $P_{ij}(c'|c)$ の更新は，(2) の分母，分子の値をそれぞれ整数として記憶しておくことで容易に行うことができる．

表 2　検証用データセットの概要
Table 2　Information about data set.

	被験者 A	被験者 B
日数 (日)	48	54
滞留場所の種類	118	79
GPS データの個数	143,586	230,355
遷移時間の中央値 (秒)	1,405	310

6.　検証

6.1　検証設定

提案手法の評価のため，実データを用いた検証を行う．検証用のデータは，被験者 2 名が実際に GPS データロガーを持ち歩くことで収集した GPS データを用いた．GPS データ収集には，GlobalSat 社の DG-100 GPS データロガ，および NTT docomo 社から発売されている HT-03A 携帯電話端末を用いた．図 5 にこれらの端末を示す．GPS データは平日を中心に，1 秒〜数十秒に一度の頻度で測定したものを，前処理によって 10 秒周期のデータにサンプリングしたものを用いた．なお，DBSCAN アルゴリズムによるデータの離散化の際には，事前検証によって DBSCAN のパラメータをそれぞれ $\epsilon = 100$m，NumPoints $= 18$ とした．これはおよそ半径 100m 以内の領域に 3 分以上滞在したときに，それをクラスタとして取り出すことを意味する．また，パラメータ γ は $\gamma = \theta$ と設定した．検証で利用するデータセットのデータ総量，抽出されたクラスタの個数，データ収集日数を表 2 に示す．

プログラムはすべて C++で実装し，実験は Xeon CPU(2.67GHz)，48GB メモリを搭載した Linux マシン上で実行した．

6.2　予測正答率の検証

逐次更新性の検証のため，提案手法によって遷移途中の状態を利用して予測正答率を高められるかどうかを評価した．なお，検証においてはデータセットのうち 1 日分をテストデータ，残りを訓練データとする交叉検定法を用いて正答率を評価した．比較対象として遷移途中の状態を考慮しないマルコフモデル (MM) を用いた．提案手法として，パラメータ θ の値を $\theta = 100\text{m}, 1000\text{m}$ と変化させた手法 (kernel) と，前節で述べた近似手法 (mesh) を用いた．

ⓒ2012 Information Processing Society of Japan

図 5　GPS データ取得に利用した端末 (左: GlobalSat DG-100, 右: HT-03A)

Fig. 5　Devices used for GPS data collection. Left: GlobalSat DG-100, Right: HT-03A.

各データセットでの予測結果を図 6，7 に示す．各図での横軸は，滞留場所間遷移途中における相対位置を表している．すなわち，ある滞留場所 A から別の滞留場所 B に向かっているとしたときに，A を 0, B を 1 としたときの位置を表す．例えば A と B のちょうど中間ならば，相対位置は 0.5 とする．各図での縦軸は，予測の Top-3 正答率を示している．すなわち，実際に被験者が次に滞留した場所が，遷移確率が最も高く出力された 3 つの滞留場所のいずれかに該当したなら予測が成功したとする．図より，いずれの被験者についても遷移途中の位置を用いることによって既存手法よりも正答率高く次の滞留場所を予測できていることが分かる．また，相対位置が 1 に近づくにつれ，予測正答率が向上する傾向も見られる．以上より，提案手法によって予測の逐次更新性を高められたことが示される．

5 節で導入した効率化手法を用いた場合，どちらのデータセットにおいても正答率が元の手法を下回る結果となった．特に，メッシュの間隔を 1000m に設定したときは大きく正答率が低下した．一方，メッシュの間隔を 100m に設定した場合には，正答率の低下は最大でも 10%以内にとどまった．今回の結果より，メッシュ間隔が 1000m だと近似として粗いことが分かる．メッシュ間隔が狭いと近似を行わないときに近い精度を得られるが，一方でメッシュ間隔を狭めるとあらかじめ記憶しておく必要がある確率分布の数も増大するため，実際に利用する際にはその点も考慮したうえで適切な値を定める必要がある．

近似を用いない手法同士を比較した場合，全般的に $\theta = 100$m の方が 1000m よりも正答率が高い傾向が見られた．これは，θ が大きい場合，予測を行う点の近傍の，別の目的地に向かったときの位置データを参照してしまうからと考えられる．

6.3　処理性能の検証

処理性能の検証として，提案手法とメッシュを用いた近似法とを用いて予測を行い，予測に要した時間を測定した．なお，提案手法では訓練データを R*-tree に格納している．検証結果を図 8 に示す．提案手法では近傍の GPS データの個数が増えるにつれ，予測に必要な時間が線形に増加しているのに対し，メッシュを利用した手法では定数時間にとどまっている様子が分かる．この結果より，メッシュを用いた近似方法を用いることによって，データ量の増大に伴う計算量の増加に対処できていることが示された．

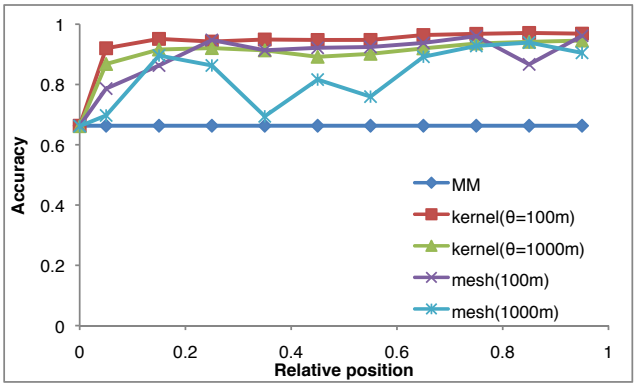

図 6　予測正答率 (被験者 A)

Fig. 6　Prediction accuracy of subject A.

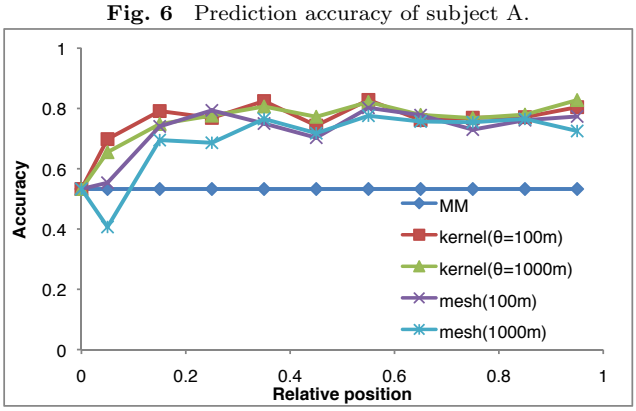

図 7　予測正答率の比較 (被験者 B)

Fig. 7　Prediction accuracy of subject B.

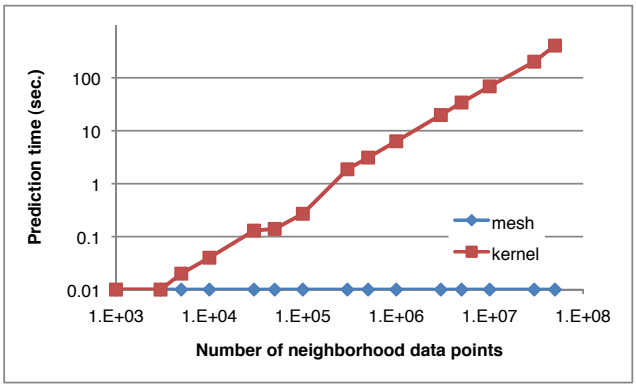

図 8　処理時間の比較

Fig. 8　The result of comparing execution time.

7.　おわりに

本稿では，ライフログの一種である位置情報を蓄積し，それをもとに未来の位置を予測する手法を提案した．提案手法は，近年一般的になりつつある，携帯端末で位置情報が逐次的に取得できる状況を想定し，その状況で必要とな

ⓒ2012 Information Processing Society of Japan

る要件を満たすことを目標としている．すなわち，予測の逐次更新性を重視した手法となっている．本稿ではこれまで離散化ベースで行われていた予測手法に，離散化の際に捨てられていた，滞留場所間を移動する途中の位置データを活用する手法を組み合わせることで，逐次更新性を高めることに成功した．提案手法には訓練データが増加するにつれ，予測にかかる計算時間がそれに比例して大きくなるという問題があった．そこで，遷移途中のデータを利用する際に発生する計算時間の問題を解くための近似法も提案した．近似によって，予測を行うのに必要な計算時間が削減できることを実験によって確認した．

参考文献

[1] 田中克明, 堀浩一, 山本真人：個人行動履歴に基づく情報推薦システムの開発, 人工知能学会論文誌, Vol. 23, No. 6, pp. 392 – 401 (2008).

[2] Gemmell, J., Bell, G. and Lueder, R.: MyLifeBits: a Personal Database for Everything, *Communications of the ACM*, Vol. 49, No. 1, pp. 88 – 95 (2006).

[3] Kalnikaitė, V., Sellen, A., Whittaker, S. and Kirk, D.: Now let me see where I was: Understanding how lifelogs mediate memory, in *In Proc. CHI'10*, pp. 2045 – 2054 (2010).

[4] Ashbrook, D. and Starner, T.: Using GPS to learn significant locations and predict movement across multiple users, *Personal and Ubiquitous Computing*, Vol. 7, No. 5, pp. 275 – 286 (2003).

[5] Monreale, A., Pinelli, F., Trassarti, R. and Giannotti, F.: WhereNext: a location predictor on trajectory pattern mining, in *In Proc. KDD'09*, pp. 637 – 645 (2009).

[6] Liao, L., Patterson, D. J., Fox, D. and Kautz, H.: Learning and inferring transportation routines, *Artificial Intelligence*, Vol. 171, No. 5, pp. 311 – 331 (2007).

[7] Tao, Y., Papadias, D. and Sun, J.: The TPR∗-Tree: An Optimized Spatio-Temporal Access Method for Predictive Queries, in *In Proc. VLDB'03*, pp. 790 – 801 (2003).

[8] Tao, Y., Faloutsos, C., Papadias, D. and Liu, B.: Prediction and Indexing of Moving Objects with Unknown Motion Patterns, in *In Proc. SIGMOD'04*, pp. 611 – 622 (2004).

[9] Jeung, H., Liu, Q., Shen, H. T. and Zhou, X.: A hybrid prediction model for moving objects, in *In Proc. ICDE' 08*, pp. 70 – 79 (2008).

[10] NTT ドコモ：オート GPS, `http://www.nttdocomo.co.jp/service/information/auto_gps/`.

[11] Giannotti, F., Nanni, N., Pedreschi, D. and Pinelli, F.: Trajectory Pattern Mining, in *In Proc. KDD'07*, pp. 330 – 339 (2007).

[12] Nishino, M., Nakamura, Y., Yagi, T., Muto, S. and Abe, M.: A Location Predictor based on Dependencies Between Multiple Lifelog Data, in *In Proc. LBSN'10*, pp. 11 – 18 (2010).

[13] Eagle, N. and Pentland, A.: Reality mining: sensing complex social systems, *Personal and Ubiquitous Computing*, Vol. 10, pp. 255 – 268 (2006).

[14] Farrahi, K. and Gatica-Perez, D.: Learning and Predicting Multimodal Daily Life Patterns from Cell Phones, in *In Proc. ICMI-MLMI'09*, pp. 277 – 280 (2009).

[15] Adams, B., Phung, D. and Venkatesh, S.: Extraction of Social Context and Application to Personal Multimedia Exploration, in *In Proc. MM'06*, pp. 987 – 996 (2006).

[16] Ester, M., Kriegel, H. P., Sander, J. and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise, in *In Proc. KDD'96*, pp. 226 – 231 (1996).

[17] Beckmann, N., Kriegel, H.-P., Scheneider, R. and Seeger, B.: The R*-tree: An Efficient and Robust Access Method for Points and Rectangles, in *In Proc. SIGMOD'90*, pp. 322 – 331 (1990).

ⓒ2012 Information Processing Society of Japan