2008－UBI－20 (10)
2008／11／13

滞在地遷移情報からの行動パターン抽出方式の検討

西野正彬[†1] 瀬古俊一[†1] 青木政勝[†1]
山田智広[†1] 武藤伸洋[†1] 阿部匡伸[†1]

本報告ではGPSデータロガを用いて取得したGPSデータから，その利用者の行動パターンを抽出するための方式について述べる．行動パターンの抽出は，まずGPSデータを滞在した場所を並べた離散的な系列に変換したのちに，その系列に対して頻出なパターンを選択する頻出系列マイニング処理を適用し，さらに特徴パターン選択処理を行うことで達成される．本報告では特徴パターン選択処理として3種類の方式を提案する．計測したGPSデータに3種類の特徴パターン選択方式を適用し，得られた特徴パターンの比較を行って，各方式の特徴を明らかにした．

A Study on Extracting Movement Patterns from Transition Data

Masaaki Nishino ,[†1] Shunichi Seko ,[†1] Masakatsu Aoki ,[†1]
Tomohiro Yamada ,[†1] Shinyo Muto [†1] and Masanobu Abe[†1]

In this report, we propose a method for extracting movement patterns of a person from location data obtained by carrying a GPS data logger. First we translate GPS data to a discrete sequence by aligning the places the user had stayed in time series order, then we use sequential frequent pattern mining technique and obtain frequent movement patterns. Moreover, we propose three methods to select feature movement patterns from those mined ones. We evaluate these methods by applying representative movement patterns to location data from GPS and show features of each method.

1. はじめに

近年，Web上のオンラインショッピングサービスにおいて，商品のリコメンデーション技術に代表されるような個人適応化技術が発展し，サービスの利用者は余分な負担を強いられることなく，自身に適応したサービスを受けることができるようになりつつある．

いっぽう，こうした個人適応化技術の有用性はWeb上の世界に限ったものではなく，実世界の活動においても等しく有用なものである．例えばGPSを利用して位置情報を取得し．その位置に応じた情報を提供するiエリア[*1]といったサービスは既に実現されている．しかし，より的確に利用者に適応したサービスを実現するためには，たとえば位置情報だけでなく，利用者の移動手段も推定するなど，センサデータからより複雑な情報を推定する必要がある[7]．

本報告では，利用者の行動を推定するためにGPSデータを利用する．職場で仕事をする，デパートで買い物をするなど，利用者がとる行動はその位置に影響を受けるが，GPSを用いれば，屋外であれば利用者の位置を誤差10m程度[5]で測定することが可能である．また，長時間GPSデータを取得することが可能なGPSデータロガも安価で入手することができるため，GPSデータを利用したサービスの実現性は高い．

本報告ではGPSデータロガによって記録したGPSデータをもとにして，利用者の行動に特有な特徴行動パターンを抽出する方式について述べる．行動パターンという言葉には様々な意味が与えられることがあるが，本研究では利用者がどのように滞在場所を変化させたかという遷移行動を行動パターンとして捉える．利用者の行動パターンを知ることで，利用者の現在の行動から，次にどのような行動をとりそうかを予測し，その予測に沿ったサービスを提供したり，ある行動が過去の行動パターンとどのように異なっていたかを抽出することで，非日常的な行動の検知に繋げるなどの応用が考えられる．

1.1 関連文献

近年，記録媒体の低価格化などの影響で，個人に関する様々な情報をデジタル化し，蓄積することで個人

†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
*1 http://www.nttdocomo.co.jp/service/location/iarea/

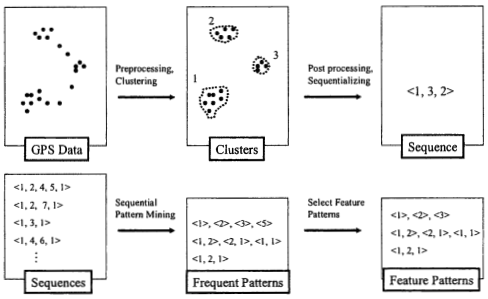

図 1 特徴パターン抽出処理の概要

の生活に役立てるというライフログに関するの研究が行われている．デジタル化される情報はドキュメントやイメージなど様々であるが，GPS に代表される位置情報は重要な情報のひとつである．GPS データによって獲得できる位置情報に限っても，複数人の同行判定や[6]，行動手段の判定を行うことが可能であったりと[8]，その実用的価値は高い．

いっぽう，GPS データには計測方式の特性上ノイズが含まれるため，ノイズがあっても比較的安定に獲得できる利用者の滞在場所を抽出して利用する方式が提案されている．Brett らは GPS データから利用者の滞在場所を抽出し，それをもとにして利用者にとって重要な場所を抽出し，さらに複数利用者間の関係の抽出を行っている[1]．また，Ashbrook らは滞在地を抽出し，そこから利用者の遷移をマルコフモデルとして表現することで，行動予測が行えるとしている[3]．本研究では，抽出したパターンを用いて，とくに利用者のある日の行動が日常とどのように異なるかを評価することを目的とする．

2. GPS データからの行動パターン抽出

2.1 パターン抽出プロセス概要

本報告では，滞在地以外も含めた行動パターン抽出も考慮し，以下のアプローチを提案する．つまり，行動パターン抽出を

(1) GPS データからの離散系列の作成

(2) 離散系列の集合からの特徴滞在パターンの抽出

という 2 段階の処理とする．図 1 は本手法の概要を表しており，図上段は GPS データからの離散系列作成処理，下段は蓄積された離散系列からの特徴パターン抽出処理を表している．

上段では，まず入力として与えられた GPS データの緯度・経度をもとに前処理とクラスタリング処理を実行し，GPS データからクラスタを抽出する．その後，クラスタリング結果から利用者がある場所にどのように滞在したかを示す離散系列を作成する．この系列を滞在系列とよぶ．GPS.データをクラスタリングし離散的な系列で表すことで，GPS データに含まれるノイズなどの影響を低減することができる．詳細は 2.2 節で述べる．

図下段では，上段の処理で得られた滞在系列を蓄積し，そこから系列マイニングおよび特徴滞在パターン選択処理によって特徴滞在パターンを得る過程を表している．詳細は 2.3 節および 3 節で述べる．

2.2 滞在系列の作成

入力として与えられる GPS データは，利用者が GPS データロガを携行し，自身の位置（緯度・経度）を一定時間間隔で取得したものとする．

GPS を用いた測位には，原理的に屋内で実行不能であり，屋外であっても測位結果には常に誤りが含まれ，ビルの影などではその測定精度はさらに悪化することもある．いっぽう、各測定点での正確な位置情報が取得できなかった場合においても、家あるいはオフィスといった滞在場所を識別することにより、利用者の行動推定を行うことができる。

これらをふまえ，まず GPS データから利用者が滞在した位置を抽出し，それを時系列順に並べた系列によって利用者の行動を表現するアプローチをとる．GPS データロガは数秒から数分の一定間隔で利用者の位置を測定するが，利用者がある場所に滞在した場合にはその場所で多数の観測が行われることになる．そのため，GPS データを参照して多数の観測が行われている場所を利用者の滞在地として抽出することができる．

2.2.1 GPS データのクラスタリング

利用者の滞在地の検出には[1] と同様にクラスタリングアルゴリズム DBSCAN[4] を利用する．DBSCAN は，入力データ中のデータ密度が高い集団をそれぞれひとつのクラスタとして取り出し，いずれのクラスタにも属さない点をノイズとして除去するアルゴリズムである．K-means 法のような代表的なクラスタリングアルゴリズムと比べ，DBSCAN アルゴリズムにはノイズを除去することができる，あらかじめクラスタ数を定めないでよいといった特長がある．1 日分集積した GPS データに DBSCAN を適用することで，その日に利用者が滞在した場所を抽出することができる．

DBSCAN アルゴリズムの詳細は[4] を参照されたい．ここでは DBSCAN の振舞いとその性能に影響を及ぼすパラメータについて簡単に説明する．DBSCAN アルゴリズムはふたつのパラメータ ϵ と Num を受けとり，入力として与えられたデータ間の距離に基づいて

クラスタリングを行う．入力データを x, y などと表し，2 点の間の距離を $\mathrm{dist}(x, y)$ と表す．

DBSCAN アルゴリズムは以下の手順で実行される．まず，入力データ中から一点を無作為に選択し（x とする），x から距離 ϵ 以内にあるすべての点の集合 $\mathrm{Ne}_\epsilon(x) = \{y|\mathrm{dist}(x, y) \leq \epsilon\}$ を計算する．もし $|\mathrm{Ne}_\epsilon(x)| \geq \mathrm{Num}$ ならば，x と $\mathrm{Ne}_\epsilon(x)$ に含まれる点をひとつのクラスタとし，そうでないならば x をノイズとし，別の点を選択して計算を続行する．クラスタと判定されたならば，$y \in \mathrm{Ne}_\epsilon(x)$ である y に対しても同様に $|\mathrm{Ne}_\epsilon(y)| \geq \mathrm{Num}$ かどうかを判定し，条件を満たすならば $\mathrm{Ne}_\epsilon(y)$ の点もクラスタに追加する．この処理をクラスタに追加する点がなくなるまで行い，最終的にひとつのクラスタが形成される．すべての入力データがいずれかのクラスタに含まれるかノイズであるかが判定された時点で処理を終了し，最終的にクラスタ群を得ることができる．

なお，DBSCAN の前処理として欠落した GPS データの補完を行う．たとえばある時刻に利用者が屋内にいるため位置計測ができなかったとしても，前後の時刻の観測点を用いてその時刻データを補完することによって，クラスタリング時に建物に滞在していたことを判定することが可能になる．補完処理として計算が容易な線形補間を用いる．

2.2.2 クラスタからの滞在系列の作成

クラスタリングによってある時刻の GPS データがどのクラスタに属しているかが分かるため，データ中に出現するクラスタを時系列順に並べることで，滞在した場所を並べた系列を得ることができる．ここで，後述する特徴滞在パターン抽出を行うためには，利用者のある日の GPS データを入力として作成した系列中のある滞在地が，他の日の GPS データから得られた系列中のどの滞在地と等しいかを判定する必要がある．そこで，クラスタリングによって得られた滞在地をテーブルに記録しておき，新しい GPS データに対してクラスタリングを行うごとにそのテーブルを参照して共通である滞在地の判定を行う．

入力としてクラスタ C が与えられたときの滞在地の判定手順は以下のようになる．ここで，テーブルに記録されている滞在地を $l_1, l_2, \ldots, l_k$ であるとし，各滞在地の座標とともに，処理の時点で滞在地 l_i に属するとされた点の数も保管しておく．

Step 1 新しく入力されたあるクラスタ C に含まれる点の重心を求める．

Step 2 滞在地 $l_1, l_2, \ldots, l_k$ のうち，C の重心との距離がしきい値 δ 以下のものがあるかを調べる．

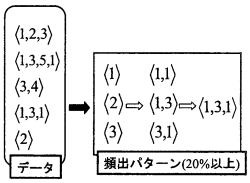

図 2 頻出系列マイニングの実行例

そのような l_i が複数存在すれば，C の重心との距離が最小のものをひとつ選択する．あれば **Step 3** へ，なければ **Step 4** へすすむ．

Step 3 クラスタ C の滞在地が **Step 2** で選択された滞在地とし，滞在地の位置を更新する．

Step 4 クラスタの重心を新たに l_{k+1} として登録し，C の滞在地を l_{k+1} とする．

なお，**Step 3** の滞在地の更新では，滞在地の新しい位置を，滞在地に属する点にクラスタ C の点を加えて重心を計算したものとする．

各クラスタに対応する滞在地が定まったら，あとは GPS データに出現している滞在地を時系列順に並べれば滞在系列を作成することができる．つまり，クラスタリング処理を適用した GPS データを時系列順に眺めていき，あるクラスタ C に含まれる点が出現したときから，次に異なるクラスタに含まれる点が出現する時点までを，そのクラスタ C が対応する滞在地に滞在していたものとして滞在系列を作成する．このとき，あるクラスタ C に含まれる点が出現してから，次のクラスタに含まれる点が出現するまでの時間を，C における滞在時間とよぶ．

滞在時間は，系列作成の後処理として滞在時間が短い滞在を系列から削除するために用いられる．すなわち，滞在時間に対して数分から十数分の間でしきい値を設定し，滞在時間がしきい値以下であったならばその滞在を系列から削除する．

ある GPS データが入力として与えられたときに，上記の手順によって最終的に作成される系列を $s_i = \langle u_{i1}, u_{i2}, \ldots, u_{iM_i} \rangle$ とする．ここで u_{ij} は滞在地集合に含まれるひとつの滞在地に対応し，つまり $u_{ij} \in L$ であり，GPS データにおけるひとつの滞在を表す．また M_i は系列 s_i に含まれる滞在の個数であり，s_i の系列長ともよぶ．複数の GPS データが入力として与えられたときに，それらの GPS データに対応して作成される系列をそれぞれ $s_1, s_2, \ldots$ と表し，すべての系列の集合を $S = \{s_1, s_2, \ldots, s_K\}$ とする．K は系列の個数である．

2.3 頻出系列マイニングの適用

前節の手法により，GPS データを入力として系列

を作成することができた．次にこの系列から利用者の特徴的な行動パターンを抽出することを試みる．その第一歩として，まず頻出系列マイニング[2]を適用する．頻出系列マイニングは，系列の集合が与えられたときに，そこからあるパラメータ θ 以上の頻度で出現しているすべてのパターンを高速に抽出するための手法であり，もとは顧客の購買データから頻出な購買パターンをすべて抽出するための手法として提案された．利用者の行動は，買い物であれば複数の店に滞在したり，通勤であれば自宅，勤務先，駅にしか滞在しないといったように，行動によって滞在地の数は一定ではないため，パターンの長さに依存せずに頻出なパターンをすべて選択できる頻出系列マイニングは有用である．

頻出系列マイニングの概要を紹介する．詳細については[2]を参照されたい．はじめにいくつかの用語を定義する．まず，行動パターンは $\langle v_1, v_2, \ldots, v_m \rangle$ といった形式で定義される．ここで $v_i \in L$ $(i = 1, 2, \ldots, m)$ であり，つまり行動パターンとは滞在地の系列として表現される．

また，あるパターン $\langle v_1, v_2, \ldots, v_m \rangle$ とある滞在系列 $l_a = \langle u_{a1}, u_{a2}, \ldots, u_{aM_a} \rangle$ に対して，整数の列 $i_1 < i_2 < \cdots < i_m$ が存在し，$u_{ai_1} = v_1, u_{ai_2} = v_2, \cdots, u_{ai_m} = v_m$ を満たすならば，パターンが系列に出現しているとよぶ．あるパターンと滞在系列の集合 $S = \{s_1, s_2, \ldots, s_k\}$ があったとき，そのパターンの出現頻度を定義することができる．すなわち，あるパターンの出現頻度は，滞在系列のうち，そのパターンが出現しているものの比率である．

図 2 は頻出系列マイニングの実行例であり，図左側の系列を入力として与え，頻度しきい値を 0.2 としたときに図右側のパターンが得られる．Agrawal らの頻出系列マイニングの手法を用いると，系列長の短いパターンから順にパターンを列挙していき，計算量を減らしつつすべての頻出パターンを列挙することが可能である．

3. 特徴行動パターンの抽出

頻出系列マイニングを適用することで，滞在系列の集合から出現頻度がしきい値以上である滞在パターンをすべて抽出することができた．ある滞在パターンの出現頻度が高いことは，利用者がその滞在パターンにそった行動を頻繁に行っていることを意味するため，利用者の主要な行動を表していると考えることができる．しかし，特徴滞在パターンを得るためにはデータの性質，すなわち利用者の行動の特性に適合した適切な出現頻度パラメータを決定する必要があり，処理の前にパラメータを定める必要のある頻出系列マイニングのみでは利用者の行動を反映するような特徴パターンを得ることができない．以下，特徴パターン抽出の問題点とその解決法について説明する．

3.1 頻出系列マイニング利用の問題点

まず，理想的な特徴滞在パターンの集合とはどのようなものであるかを説明する．今回，特徴的なパターンを抽出する目的として，(i) 利用者の行動を理解するための，行動の要約を作成すること，(ii) 利用者の行動をパターンと比較し，どのように日常の行動と異なるかを評価する，という 2 種類を想定している．特徴滞在パターンの集合がこれらの条件を満たすためには，集合が利用者の多様な行動を反映するパターンを含み（**多様性**），かつ集合に含まれる各パターンが日常的な行動を表した，安定して出現するパターンである必要がある（**安定性**）．一般に多様性と安定性にはトレードオフの関係があり，特徴パターン集合に出現頻度が低いパターンも含め多量のパターンを含めれば多様性は上がるが，安定性は下がる．逆に出現頻度が高いパターンを集めれば安定性は上がるが，そのようなパターンは一般に少量のため，多様性は下がる．

これらの性質は系列マイニングにおける出現頻度しきい値 θ に依存する．θ が大きくなると特徴滞在パターンの数は減少するため多様性は下がるが，パターンの出現頻度は上がるため安定性は上がる．θ が小さくなるとると出現する滞在パターン数は増加し，多様性は上がるが集合の安定性は下がる．

以上より，適切な特徴滞在パターンの集合を得るためには，系列マイニングにおいて適切な頻度しきい値を設定する必要があることが分かる．しかし系列マイニングにおいてはデータの性質を反映してしきい値を決定することが不可能であるから，本報告ではあらかじめ低い出現頻度しきい値で系列マイニングを行ったのちに，得られた滞在パターンおよびデータの性質から改めて出現頻度のしきい値を決定して，そのしきい値以上の出現頻度をもつ滞在パターンを特徴滞在パターンを選択する手法をとる．

3.2 特徴パターン選択のためのしきい値決定手法

本節では特徴滞在パターンを抽出するための出現頻度しきい値決定手法について説明する．手法は 3 種類準備し，それぞれ

(1) 滞在数を利用した手法
(2) 滞在パターン数を利用した手法
(3) 被覆率を利用した手法

とよぶ．滞在数を利用た手法では，滞在数という尺度

を導入して出現頻度しきい値の変化に対する頻出パターン集合の安定性を評価する．滞在パターン数を利用した手法では，出現頻度しきい値と滞在パターン数の分布にひとつの形状を仮定し，その形状に最の沿うしきい値を選択する．被覆率を利用した手法は被覆率という新たな尺度を導入して，パターン集合の多様性を評価する．いずれの手法も滞在系列の集合もしくは系列マイニングによって抽出された滞在パターンの集合に基づいて出現頻度しきい値を決定する．

3.2.1 滞在数を利用した手法

まず，系列マイニングを適用することによって得られた滞在パターンの集合から滞在数を計算し，それを利用して滞在パターンの出現頻度しきい値を決定する手法について述べる．

はじめにいくつかの定義を行う．まず，滞在系列の集合 S に対してある出現頻度しきい値を設定して系列マイニングを行った結果，得られた滞在パターンの集合を $P = \{p_1, p_2, \ldots, p_O\}$ とする．ここで p_i $(1 \leq i \leq O)$ はそれぞれひとつの滞在パターンを表しており，$p_i = \langle v_{i1}, v_{i2}, \ldots, v_{iN_i} \rangle$ である．$v_{ij} \in S$ であり，ひとつの v_{ij} はパターン中に含まれるひとつの滞在を表す．つぎに，P の部分集合として，P に含まれる滞在パターンのうち，その出現頻度が ϕ 以上であったものの集合を Q_ϕ とする．定義より $Q_\phi \subseteq P$ である．これらを利用し，滞在数を出現頻度しきい値 ϕ の関数として

$$\mathrm{stay}(\phi) = \sum_{i=1}^{K} \sum_{j=1}^{M_i} c_\phi(u_{ij}) \qquad (1)$$

と定義する．ここで c_ϕ は滞在 u_{ij} を引数とする関数であり，

$$c_\phi(u) = \begin{cases} 1 & \text{if } \exists i, j (u = v_{ij}, p_i \in Q_\phi) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

と定義する．c_ϕ は引数として与えられた滞在が，集合 Q_ϕ に含まれるいずれかのパターンの構成要素となっているときに 1 を返し，それ以外のときには 0 を返す．つまり，ある滞在 u について，それが Q_ϕ に出現しているかどうかを調べている．式 (1) より，滞在数は S に含まれるすべての滞在系列のすべての滞在について c_ϕ を適用し，それを足しあわせたものであるから，つまり滞在数とは S に出現する滞在のうち，Q_ϕ に含まれるパターンに出現しているものの総数である．

滞在数を用いた出現頻度しきい値決定手法について説明する．滞在数は出現頻度しきい値に対して定義される関数であるから，まず頻度しきい値を ϕ とし，ϕ を $[0, 1]$ の範囲で一定間隔で変化させ，それぞれの ϕ に対応する滞在数を記録する．つぎに，記録された滞在数のうち，出現頻度しきい値の変化に対して滞在数が不変であるような出現頻度しきい値の区間をすべて探し出す．その後，抽出された区間のうち区間幅が最大となるものをひとつ選択する．このとき出現頻度の最大値である 1 を含む区間は除く．また，区間幅が最大となるような区間が複数存在するならば，それらのうち最も値が小さいものを選択する．最後に，得られた区間の最小の出現頻度を取り出し，それを特徴パターン選択のための出現頻度しきい値とする．

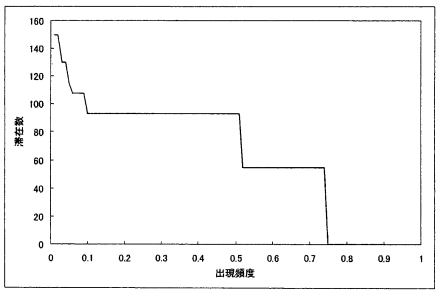

図 3 出現頻度しきい値と滞在数のプロット

GPS 実データに対して実際に系列作成および系列マイニングの処理を行い，それをもとに滞在数を計算し，プロットしたものを図 3 に示す．この図から特徴滞在パターン選択のための出現頻度しきい値を求めると，滞在数が不変な最長の区間は $[0.1, 0.5]$ であるから，特徴滞在パターンの集合は出現頻度 0.1 以上であるような滞在パターンの集合であるとして定まる．

3.2.2 パターン数を利用した手法

次に，系列マイニングによって抽出された滞在パターンの総数を用いて出現頻度しきい値を求める手法について述べる．まずはじめに出現数のときと同様に出現頻度しきい値 ϕ を $[0, 1]$ の区間で一定間隔で変化させ，それぞれの値に対する滞在パターン総数を計算する．なお，出現頻度しきい値 ϕ に対する滞在パターンの総数は $|Q_\phi|$ と表される．

つぎに，出現頻度と出現頻度とパターン総数の関係を近似する，出現頻度しきい値をパラメータとする式を導入し，この式との二乗誤差が最小となる値を特徴滞在パターン選択のためのしきい値とする．この式を $f_\phi(x)$ として

$$f_\phi(x) = \begin{cases} \frac{|Q_\phi| - |Q_0|}{\phi - 0}(x - \phi) + |Q_0| & \text{if } x < \phi \\ \frac{|Q_1| - |Q_\phi|}{1 - \phi}(x - \phi) + |Q_\phi| & \text{ohterwise} \end{cases} \qquad (3)$$

と定義する．$f_\phi(x)$ は 3 つの点 $(0, |Q_0|)$，$(\phi, |Q_\phi)|)$，

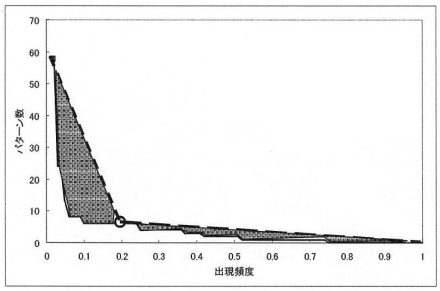

図 4 パターン総数と近似関数のプロット

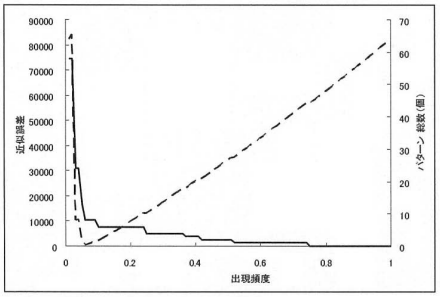

図 5 パターン総数と近似関数の近似誤差のプロット

$(1, |Q_1|)$ をこの順番に結ぶ 2 本の直線によって表現される．

図 4 は出現頻度を変化させてパターン総数をプロットしたもの（実線）と，$\phi = 0.2$ に設定して近似関数 $f_\phi(x)$ をプロットしたもの（破線）である．総パターン数の変化に着目すると，出現頻度しきい値が小さいときに総パターン数はしきい値の変化に対して急な変化を見せ，しきい値が大きなときはそれほど変化しないことが分かる．この，パターン総数の変化が大きなところでは，パターンの安定性が低いため，特徴的なパターンとするには好ましくない．いっぽう，出現頻度しきい値を高くすると出現するパターン数の変動がなだらかになり，出現しているパターンが安定であるが，パターン数が少ないため多様性は小さいと考えられる．そこで，パターン数の変動が急な領域と緩やかな領域のちょうど境界を出現頻度しきい値として出力することを考える．関数 $f_\phi(x)$ で 2 本の直線によって近似を行うのは，それぞれの領域を 1 つの直線で近似し，その近似性能を二乗誤差によって評価することで 2 つの領域の境界を探るためである．

図 5 は滞在パターン総数（実線）と近似関数 $f_\phi(x)$ のパラメータ ϕ を変化させて二乗誤差を計算したもの（破線）をプロットしたものである．この図から，近似誤差がパターン総数の変化の速度が緩やかになった直後あたりで最小値をとることが分かる．

3.2.3 被覆率を利用した手法

最後に，滞在パターンの集合を評価する新たな尺度を導入して，それに基づいて特徴滞在パターンを選択するための出現頻度しきい値を設定する手法について述べる．導入する尺度は

$$\mathrm{cover}(\phi) = \frac{\mathrm{stay}(\phi)}{\sum_{i=1}^{K} M_i} \tag{4}$$

として，これを**被覆率**と定義する．被覆率は 3.2.1 節で定義した滞在数を S に含まれる滞在系列の滞在の総数で割ったものであり，S に含まれる滞在のうち，パターンの集合 Q_ϕ に含まれるものの割合を表している．近似的にではあるが，被覆率は滞在パターンの集合が滞在系列集合の多様性をどれだけ反映することができているかを表すために用いることが可能である．なお，$\phi_1 < \phi_2$ に対して $\mathrm{cover}(\phi_1) \geq \mathrm{cover}(\phi_2)$ であることから，被覆率は広義単調減少である．

被覆率は，それ自身が特徴滞在パターン集合の多様性の尺度として用いられる．つまり，被覆率にあらかじめあるしきい値 η を設けて，出現頻度しきい値 ϕ を変化させながら被覆率を計算し，被覆率がしきい値を下回らない最大の ϕ を出現頻度しきい値として出力する．被覆率に対してしきい値を定め，特徴滞在パターンを得ることで，パターンの集合が利用者の行動の多様性を反映することを保証できる．

4. 提案手法の評価

4.1 評価手法

4.1.1 評価項目

提案手法の評価として，まず実際に被験者実験によって得られた GPS データに対して，3 節で述べた 3 種類の特長滞在パターン抽出手法と，基準としての頻度しきい値を用いた手法とを適用し，特徴滞在パターンを抽出する．得られた特長滞在パターン集合の性質を 2 つの項目に基づいて定量化し評価する．つぎに，実験結果をふまえて各手法の特徴を 3 つの評価項目に基づいて比較・検討する．

まず，定量化の手法について述べる．3.1 節で特徴的な滞在パターンは多様性と安定性をもつとしたが，ここでそれぞれの性質を定量化するための尺度として，それぞれ**系列被覆率**と**パターン利用率**を導入する．系列被覆率はパターンの集合 Q と滞在系列 s に対して定義され，Q に含まれるパターンが系列 s にどれだけ一致しているかを表す．Q の要素で系列 s に出現しているパターンのうち，系列長が最長のもの p をひとつ選択し，その系列被覆率を $|p|/|s|$ と定義する．もし s

に出現している $p \in Q$ が存在しないならば系列被覆率は 0 とする．

パターン利用率も同様にパターンの集合 Q と滞在系列 s に対して定義され，パターン集合 Q のすべての要素が平均的に滞在系列に出現しているかどうかを表す．系列 s に出現しているすべてのパターンからなる，Q の部分集合を Q_s とすると，パターン利用率は $|Q_s|/|Q|$ として定義される．$Q = \emptyset$ の場合は 0 とする．

定量化に際し LOOCV(Leave one out cross validation) 法を用いる．すなわち，ある被験者の GPS データから得られた滞在系列の集合 S から系列をひとつ取り除いたものに対して各種手法を適用し特徴滞在パターン集合を抽出する．その後，取り除かれた系列に対して，各種スコア算出を行う．この操作を取り除く滞在系列を変化させながら $|S|$ 回くり返し．最後にスコアの平均値を求め，それを出力とする．

次に比較検討法について説明する．手法の評価項目は，3.1 節の内容をもとに，得られる特徴滞在パターンの安定性，多様性，そして手法そのものの**適応性**を挙げる．適応性とは，利用者や状況の違いによって，性質の異なる GPS データが与えられたとしても，上記の安定性・多様性をもった特徴滞在パターンを抽出できるかどうか示す尺度である．適応性は多様な GPS データから特徴滞在パターンを得るために必要な性質である．これらの項目について，手法の性質および前述の実験結果をもとに比較・検討を行う．

4.1.2 測定条件

実験環境について説明する．GPS データの取得には GlobalSat 社の DG-100 GPS Data Logger を用いる．GPS データロガは 10 秒間隔で利用者の位置（緯度・経度）および測定時刻を記録する．被験者はデータロガを起床時から就寝時まで持ち歩き，得られた GPS データを 1 日分の行動を GPS データとして記録する．記録された各データは 2.2 節で示した滞在系列作成手法によって，それに対応する滞在系列に変換される．なお，今回被験者は 3 人とし（それぞれ A, B, C とする），A, B, C それぞれ 75，14，13 日分の GPS データを収集した．

DBSCAN アルゴリズムのパラメータとして $\epsilon = 0.0003$, Num $= 30$ を用いる．パラメータ ϵ は同一のクラスタに属すると認める範囲の大きさを表している．Num はクラスタに最低限含まれる点の数を表しており，つまりクラスタとして検出される最低限の滞在時間を表している．今回の設定では，データ取得間隔が 10 秒で Num $= 30$ であるから，およそ 300 秒

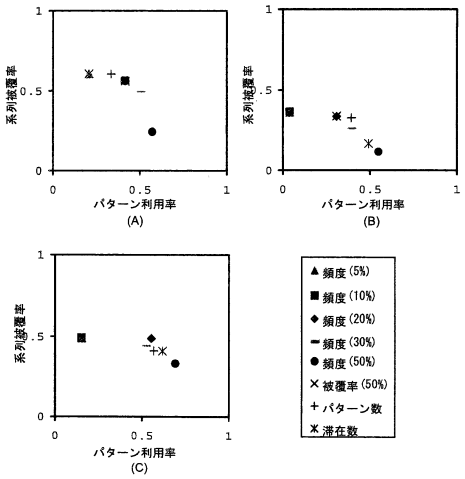

図 6 特徴滞在パターン集合の比較

(5 分) 以上滞在した場所を滞在地としてクラスタとして扱う．クラスタの重心座標より滞在地テーブル中の滞在地を選択する処理では，$\delta = 0.0006$ と設定する．また，被覆率を用いた特徴滞在パターン決定手法におけるしきい値は 50%に設定する．比較対象として，出現頻度しきい値を 5%，10%，20%，30%，50%とした頻出系列マイニングを用いる．

4.2 評価結果

図 6 は，各出現頻度しきい値決定手法によって得られた特徴滞在パターンの集合について，4.1 節で用いた評価軸によって定量化したものである．各グラフの横軸はパターン利用率に，縦軸は系列被覆率にそれぞれ対応する．グラフ下の添字は GPS データが対応する被験者番号を表している．

出現頻度しきい値を変化させたときに注目すると，系列被覆率とパターン利用率がトレードオフの関係にあることと，基本的に頻度しきい値を高くするほどパターン利用率が上昇し，系列被覆率が低下する関係にあることが分かる．また，データによって値の出現する範囲および出現頻度しきい値を変化させたときの値の変化幅が異なることが分かる．

パターン数を利用した手法と滞在数を利用した手法，被覆率を利用した手法によって得られた特徴滞在パターン集合間を比較すると，いずれの被験者のデータに対しても滞在数を用いた手法がパターン利用率が最大，系列被覆率が最小であり，被覆率を利用した手法でパターン利用率が最小，系列被覆率が最大，パターン数を利用した手法はその中間であるという結果が得られた．

表 1　しきい値決定手法の比較

	頻度	滞在数	パターン数	被覆率
適応性	×	○	○	×
安定性	-	○	△	×
多様性	-	×	△	○

4.3 考　　察

実験結果をもとに各手法を適応性・安定性，多様性の軸で評価する．まず，適応性については，出現頻度をあらかじめ定める手法では，出現するパターンの多様性，安定性をが制御できないため，適応性はない．また，被覆率に基づく手法は事前にパラメータを定める必要があり，そのパラメータによって結果が大きく変化するので，こちらも適応性は乏しい．

安定性については，滞在数を利用した手法でつねに他の手法よりもパターン使用率が高いという結果が得られた．パターン数を用いた手法も，滞在数を利用した手法に次いでパターン使用率が高いという結果が得られた．これら 2 種類の手法は安定性の高いパターンを取り出すことができると考えられる．いっぽう被覆率を利用した手法では，他の手法よりも安定性が低くなった．これは被覆率による出現頻度しきい値決定手法において安定性を考慮していないことに起因すると考えられる．

多様性については，滞在数を用いた手法では系列被覆率が他の手法よりも低いという結果が得られた．いっぽうパターン数を利用した手法では常に滞在数を用いた手法よりは高いという結果が得られた．被覆率を利用した手法では，しきい値の決めかたによって系列被覆率に差はあるが，被覆率は多様性を直接決定する尺度となっているため，多様性を保証したパターンの集合を得ることが可能である．

以上の結果を表 1 に示す．ここでは出現頻度を用いた手法の安定性および多様性はしきい値に依存するため評価していない．各手法は目的に応じて使い分けることができる．たとえば，利用者がどのような行動をとっているかを広く知りたい場合には，しきい値を高い値に定めた被覆率による手法を用いればよいし，行動がいつもとどのように異なるかという差分情報を求めたければ，行動の多様性と安定性を備えたパターン数による手法を用いればよい．差分情報をより安定に求めたいならば，滞在数による手法を用いることもできる．

5. おわりに

本報告では利用者が GPS データロガを持ち歩くことによって得られた GPS データから利用者の特徴的な滞在パターンを抽出するための手法について述べた．特徴滞在パターン抽出のため，まず GPS データを利用者が滞在した場所を時系列順に並べた系列として表現し，それを蓄積したものに対して頻出系列マイニング処理を実行することで，頻出な滞在パターンの集合を獲得することができる．得られた滞在パターン集合に対して，3 種類の特徴パターン決定手法を適用して，得られたパターンの違いについて比較検討を行った．

今後の展開として，得られた特徴的なパターンを利用した実践的なアプリケーションの開発を検討中である．利用者の特徴的な行動パターンを知ることができれば，ある日の行動がどういったパターンに一致するのかを調べることができ，利用者の状況に応じたレコメンドなどのサービスにつながると考える．

また，本報告の手法では利用者が 1 人であることを想定していたが，実践的な応用を考えた場合にグループでの行動パターンを考慮することは重要である．よって，今後はグループを対象として GPS データから行動パターンを抽出する手法についても検討したい．

参 考 文 献

1) Adams, B., Phung, D. and Venkatesh, S.: Extraction of social context and application to personal multimedia exploration, *MULTIMEDIA '06*, pp.987–996 (2006).
2) Agrawal, R. and Srikant, R.: Mining Sequential Patterns, *ICDE '95*, pp.3–14 (1995).
3) 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).
4) Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *KDD '96*, pp.226–231 (1996).
5) 安田明生：GPS 技術の展望，電子情報通信学会論文誌 B，Vol.84, No.12, pp.2082–2091 (2001).
6) 瀬古俊一，西野正彬，青木政勝，山田智広，武藤伸洋，阿部匡伸：誤差情報を考慮した同行判定手法，情報処理学会 第 20 回ユビキタスコンピューティングシステム研究会 (2008).
7) 井上順子：購買行動における同伴者の影響 – 母娘ショッピングの観点から–，産研アカデミック・フォーラム，pp.29–40 (2005).
8) 青木政勝，瀬古俊一，西野正彬，山田智広，武藤伸洋，阿部匡伸：GPS 未計測区間における移動手段判定手法の検討，情報処理学会第 20 回ユビキタスコンピューティングシステム研究会 (2008).