

エリア訪問の時刻と時間長を考慮した観光行動分析

西村 拓哉[†] 山本 修平[†] 戸田 浩之[†]

[†] 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: †{nishimura.takuya,yamamoto.shuhei,toda.hiroyuki}@lab.ntt.co.jp

あらまし 人々の一日単位の行動から頻出パターンを発見することは重要である。例えば観光客の行動データから頻出の行動パターンを発見することで、人気の観光ルートを知ることが出来る。本研究では一日単位と各エリアへの訪問単位の二階層の隠れ状態を仮定した確率的生成モデルによる頻出行動パターン発見手法を提案する。一日単位の隠れ状態はユーザの立てるスケジュールのパターンを表し、訪問単位の隠れ状態は訪れるエリアや訪問時刻、時間長の潜在的な分布を表す。この手法によって、訪問するエリアだけでなく、訪問時刻や時間長も併せて考慮した頻出行動パターンの発見が可能となった。福岡に訪れた訪日観光客の実際の行動データに対して提案手法を適用し、特定の場所への訪問が行われる時間帯やその前後の頻出行動内容についての分析などが可能であることを示した。

キーワード トピックモデル, 頻出パターン発見, 観光, GPS

1. はじめに

GPS を用いた移動軌跡の収集や、ソーシャルメディアを用いたチェックイン情報の記録など、人々の行動履歴情報を収集する様々な手段が普及してきている。行動履歴の記録形式として、滞在したエリアや Point of Interest (店や観光地などのスポットのこと) と、滞在開始と終了の時刻を一日単位で記録したものがあ。例を図 1 に示す。実際のサービスでは、Google Map のタイムライン機能^(注1) や Moves^(注2) などで用いられている形式である。本研究では、この形式で記述された行動履歴データから一日単位での頻出の行動パターンを発見することに取り組む。

行動履歴から頻出の行動パターンを発見する取り組みは、観光客の行動データから人気の観光地を発見することによる観光施策の決定などのために重要である。従来の頻出行動パターン発見は、訪問行動の記録からなる系列データから同一のパターンを発見することで行われてきた [1] [2]。これらの研究では同じ場所への訪問は同一の行動とみなされている。そのため、図 1 の行動データ 1 と行動データ 2 は同一のパターンであるとみなされる。しかしながら、実際には同一の場所であっても訪問する時刻や滞在時間によってその場所への訪問が持つ意味は異なる。例えば、繁華街に昼に 1 時間滞在した場合に行っていた行動内容としては昼食などが考えられるが、夜に 3 時間滞在した場合は宴会などが考えられる。また、一部に訪問の欠損や追加があってもその他の訪問内容が類似している行動データは類似したパターンであるとみなした方がよい。例えば図 1 の行動データ 2 は行動データ 3 のエリア C への訪問行動が欠損したものとみなせる。マルコフ過程ベースの手法ではこれらは異なるパターンとして認識されるが、エリア A と B への訪問時刻と滞在時間長は同一であり、一日単位の行動パターンとしては

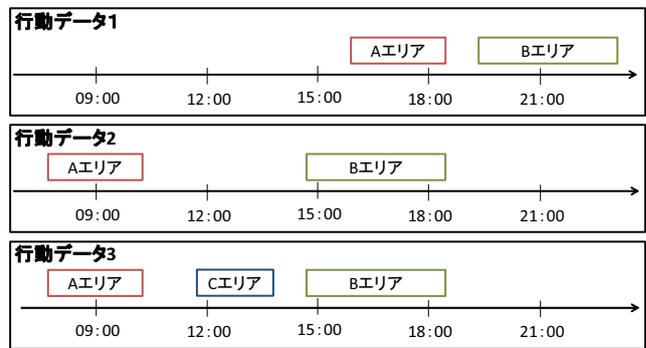


図 1 本研究で扱う、一日単位の行動データの模式図。長方形の左端と右端が滞在の開始時刻と終了時刻を示している。従来の頻出パターン発見を用いた分析では、図に示す行動データ 1 とデータ 2 を同じ行動パターンだとしてきたが、実際には同じエリアでも訪問時刻や滞在時間長によって行動の意味は異なる。また行動データ 2 と 3 は従来では異なる行動パターンだとされてきたが、エリア A と B への訪問時刻と滞在時間長は同一であり、それらを考慮すると一日単位の行動パターンとしても類似している。

類似度が高いと考えられる。

上記を踏まえ、本研究では以下の要件を満たす頻出行動パターン発見の実現に取り組む。

- 要件 (1): 複数の一日単位の行動データから、頻出行動パターンを発見可能である。
- 要件 (2): 訪問した場所だけではなく、訪問時刻と訪問時間長も併せて連続値として考慮できる。
- 要件 (3): 一部の訪問に追加や欠損があってもその他の訪問内容が類似している行動データは類似しているとみなす。

近年行われている行列分解やテンソル分解を用いた行動パターン分析 [3] は時刻情報などを考慮した分析が可能だが離散値に変換して分析する必要があり、その過程で元々のデータに含まれる連続値としての情報が損なわれる問題があるため要件 (2) を満たさない。訪問地予測などの研究で用いられているマルコフ過程をベースとした手法 [4] [5] では訪問時刻や訪問間隔

(注1) : <https://www.google.com/maps/timeline>

(注2) : <https://moves-app.com/>

を考慮しながらユーザの行動をモデリングしている。しかしこれらは訪問順序によって次の訪問場所が決定されるという仮定を置いており、要件 (3) を満たさない。

我々は上記を満たす頻出行動パターン発見を可能にする、トピックモデルを拡張した新しい確率的生成モデルを提案する。ユーザの一日の行動内容は、まず一日単位でのスケジュールが決定され、次にそれに従うように各訪問行動が行われると仮定する。例えばまず一日の観光方針を決定し、次にその方針通りに行動することである観光地に 12 時から 14 時まで訪問し、次に違う観光地に移動して 15 時から 17 時まで訪問するといった訪問行動を実際に行われると考える。この過程に基づき、一日単位の隠れ変数から各訪問行動についての隠れ変数が生成され、そこから最終的な訪問場所や時刻・時間長が生成されるモデルを考える。このモデルのパラメータを最適化すると、一日単位の隠れ変数が頻出行動パターンの種類を表し、訪問場所の生成などに関わるパラメータがその行動パターンの内容を示すため、要件を満たした頻出行動パターン発見を実現できる。

本研究の貢献は以下の二点である。

- 複数の日次行動データから、訪問した場所に加えて時刻と時間長を考慮しながら頻出行動パターンを検出できる確率的生成モデルを提案した。
- 福岡に訪れた訪日観光客の実際の行動データに対して提案手法を適用し、特定の場所への訪問が行われる時間帯やその前後の頻出行動内容についての分析などが可能であることを示した。

2. 関連研究

系列データを分析する目的で、与えられた系列データ中に頻出するパターンを抽出する研究には Prefixspan [6] などが存在する。人々の行動パターン分析においてもそれらの知見を用いた分析を行う研究が存在する。Giannotti らは人の移動軌跡情報から、典型的なパターンを発見することに取り組んだ [1]。Yang らは移動軌跡からまずユーザが一か所に留まっていた部分を抽出してからそれらをクラスタリングすることで重要な場所を明らかにした後、場所間の遷移についての頻出パターンの発見を行っている [2]。これらの手法は場所が同じであればそれらの持つ意味は同じであるとみなしている。そのため、訪問の時刻や時間長によって同じ場所でも異なる意味を持つことが考慮されておらず、1 章で示した要件 (2) を満たさない。

時間に関する情報などを考慮してユーザの訪問行動をモデリングする研究は、マルコフ過程や点過程をベースにしたもの、行列分解、テンソル分解をベースにしたもの、トピックモデルをベースにしたものに大別できる。

マルコフ過程や点過程は観測データの順序の生成確率をモデリングする手法である。これを用いて次に訪問するエリアや Point of Interest (店や観光地などのスポットのこと) を予測する研究には、訪問順序に加えて訪問時刻を考慮するもの [7] [4] や訪問間隔を考慮するもの [5] がある。日次行動データからの頻出行動パターン分析においても訪問順序を考慮することは重要である。しかし、これらのモデルは一日単位での頻出パター

ンの分類などを行うには適しておらず、また直前の状態によって次の状態が決定されるという仮定を置くマルコフ過程や点過程は一部の訪問に追加や欠損があることを想定できないため、1 章で示した要件 (3) も満たさない。

行列分解、テンソル分解は関係データの分析手法として近年注目されており [8] [9]、ユーザとエリア、POI の関係性を、時刻やユーザ属性などの情報を考慮しながらモデリングすることに取り組んでいる。熊谷らは非負複合テンソル因子分解を用いて、外国人観光客の観光行動について、訪問エリアと訪問時刻、また観光客の属性などの観点から分析している [3]。しかし、テンソル分解を用いた分析は連続値を扱えないため、連続値である時間情報を扱うためには離散化する必要があり、情報量が失われてしまう。これは 1 章で示した要件 (2) を満たすことが出来ない。

LDA [10] に代表されるトピックモデルは元々は文書の分析を目的とした手法であるが、近年では訪問行動のパターン分析をはじめ様々な分野に応用されている。トピックモデルを訪問行動の分析に用いる場合はユーザを文書、訪問行動を単語とみなして適用する。Kurashima らはユーザの居住地を考慮して訪問する場所を予測することに取り組んでいる [11]。トピックモデルは連続値も連続値に関する分布として自然に扱いつつ、確率的な生成過程によって一部訪問行動の追加や欠損も扱うことが出来る。しかし、通常のトピックモデルでは生成される変数が一種類だけであり、訪問行動におけるエリア・時刻・時間長のような三つの変数が同時に生成されるようなデータを扱うためには拡張が必要である。また、トピックモデルは頻出パターンの発見を目的としたモデルでもないため、頻出パターン発見を可能とするような拡張も必要である。

以上のように、既存研究では 1 章で示した要件を満たす頻出行動パターン発見が実現できない。本研究ではトピックモデルをベースに頻出行動パターン発見を可能にする形に拡張した、確率的生成モデルによる分析を提案する。

3. 提案手法

最初に本研究で扱うデータ形式についての定義を述べた後、提案モデルを示し、最後にモデルのパラメータ最適化の手法について述べる。本研究での記号の定義を表 1 に示す。

3.1 用語定義

本研究で扱う用語の定義について述べる。

[定義 1] (日次行動データ) 日次行動データ $s = \{a_1, a_2, \dots, a_{A_s}\}$ は訪問行動データ a の集合である。

[定義 2] (訪問行動データ) 訪問行動データ a は訪問場所 p 、中心時刻 t 、時間長 d の集合である。

3.2 提案モデル

1 章で示した要件を満たす頻出行動パターン発見のために、トピックモデル [10] を発展させた日次行動データ分析のための確率的生成モデルを提案する。提案モデルのグラフィカルモデルを図 3 に示す。各訪問行動データは一つの隠れ変数から生成されると仮定する。この隠れ変数を行動トピックと呼ぶ。訪問行動データは訪問エリア、訪問時刻、訪問時間長の三つの属性

表 1 本研究での主な記号と定義

記号	定義
S	日次行動データ数
S_j	スケジュールトピック j が割り当てられた日次行動データ数
A_s	日次行動データ s に含まれる訪問行動データ数
A_j	スケジュールトピック j が割り当てられた日次行動データに含まれる訪問行動データ数の合計
A_k	行動トピック k が割り当てられた訪問行動データ数の合計
A_{jk}	スケジュールトピック j が割り当てられた日次行動データに含まれる、行動トピック k が割り当てられた訪問行動データ数の合計
J	スケジュールトピック数
K	行動トピック数
Q	訪問場所の候補数
θ_j	スケジュールトピック j の行動トピック生成の多項分布
θ_{jk}	スケジュールトピック j の行動トピック生成の多項分布の k 次元目の値
α	θ の事前分布のハイパーパラメータ
ϕ_{pk}	行動トピック k における訪問場所に関する多項分布
ϕ_{pkq}	行動トピック k における訪問場所に関する多項分布の場所 q に相当する次元の値
ϕ_{tmk}, ϕ_{tvk}	行動トピック k における訪問中心時刻に関するガウス分布の平均と分散
ϕ_{dmk}, ϕ_{dvk}	行動トピック k における訪問時間長に関するガウス分布の平均と分散
β_p	ϕ_p の事前分布のハイパーパラメータ
$\beta_{tv_a}, \beta_{tv_b}$	ϕ_{tv} の事前分布のハイパーパラメータ
$\beta_{dv_a}, \beta_{dv_b}$	ϕ_{dv} の事前分布のハイパーパラメータ
y_s	日次行動データ s のスケジュールトピック
z_{sa}	日次行動データ s 中の訪問行動データ a の行動トピック
p_{sa}	日次行動データ s 中の訪問行動データ a の訪問場所
t_{sa}	日次行動データ s 中の訪問行動データ a の訪問時刻
d_{sa}	日次行動データ s 中の訪問行動データ a の訪問時間長

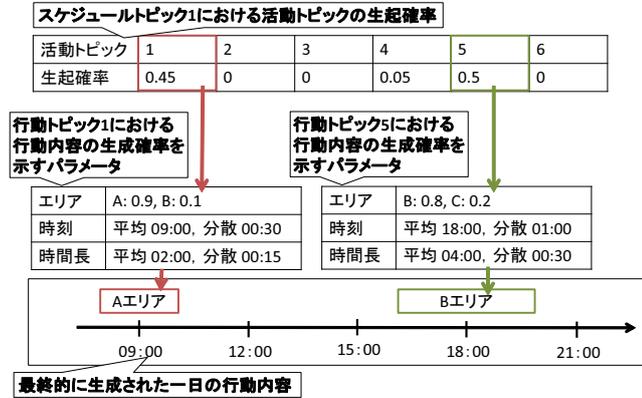


図 2 本研究で提案する確率的生成モデルが仮定する、日次行動データの生成手順の模式図。各日次行動データは一つのスケジュールトピックから生成される。スケジュールトピックの持つ多項分布から行動トピックが生成される。行動トピックは訪問する場所に関する多項分布と、中心時刻と時間長に関するガウス分布の平均と分散を保持している。それらのパラメータに基づいて訪問行動が生成され、一日の行動内容が決定される。

値を持つため、これら三つについての観測値が一つの行動トピックから生成されることになる。さらに、各日次行動データが一つの隠れ変数から生成されるとする。この隠れ変数をスケジュールトピックと呼ぶ。一つの日次行動データは一つもしくは

Algorithm 1 各行動の生成プロセス

```

for スケジュールトピック  $j = 1$  to  $J$  do
    スケジュールトピック  $j$  における行動トピック分布を生成
     $\theta_j \sim \text{Dirichlet}(\alpha)$ 
end for
for 行動トピック  $k = 1$  to  $K$  do
    訪問場所に関する多項分布を生成  $\phi_{pk} \sim \text{Dirichlet}(\beta_p)$ 
    訪問中心時刻に関するガウス分布の平均を生成  $\phi_{tmk} \sim U$ 
    訪問中心時刻に関するガウス分布の分散を生成  $\phi_{tvk} \sim \text{Gamma}(\beta_{tv_a}, \beta_{tv_b})$ 
    訪問時間長に関するガウス分布の平均を生成  $\phi_{dmk} \sim U$ 
    訪問時間長に関するガウス分布の分散を生成  $\phi_{dvk} \sim \text{Gamma}(\beta_{dv_a}, \beta_{dv_b})$ 
end for
for 日次行動データ  $s = 1$  to  $S$  do
    日次行動データ  $s$  におけるスケジュールトピックを生成  $y_s \sim U$ 
    for 訪問行動データ  $a = 1$  to  $A_s$  do
        行動トピックを生成  $z_{sa} \sim \text{Categorical}(\theta_p)$ 
        訪問場所を生成  $p_{sa} \sim \text{Categorical}(\phi_{pz_{sa}})$ 
        訪問中心時刻を生成  $t_{sa} \sim \text{Normal}(\phi_{tmz_{sa}}, \phi_{tvz_{sa}})$ 
        訪問時間長を生成  $d_{sa} \sim \text{Normal}(\phi_{dmz_{sa}}, \phi_{dvz_{sa}})$ 
    end for
end for

```

は複数の訪問行動データを持つため、一つもしくは複数の訪問行動トピックが一つの日次行動データから生成されることになる。

提案モデルでの各日次行動データおよび訪問行動データの生成プロセスを述べる。まず各日次行動データ s について、一日全体のパターンを示す変数であるスケジュールトピック y_s が生成される。次に各訪問行動データ a について行動トピック z_{sa} が生成され、そこから訪問場所、中心時刻、時間長が生成される。各日次行動データおよび訪問行動データの生成プロセスを Algorithm 1 に、生成プロセスの模式図を図 2 に示す。

提案モデルのパラメータ最適化の結果得られるスケジュールトピックが、各日次行動データが属するパターンの種類とみなすことができる。各スケジュールトピックにおける行動トピックの分布と行動トピックの持つ訪問エリア、中心時刻、時間長に関する分布がそれぞれのパターンの内容であるとみなせる。行動トピックの生成は確率的に扱われるため、一部の訪問行動データが欠損している日次行動データ同士でも類似しているとみなすことができる。

次に、生成プロセスの詳細について述べる。

まず日次行動データ s 毎にスケジュールトピック y_s が生成される。生成確率は一様分布に従う、すなわちどのスケジュールトピックになる確率も等しいものとする。

次に日次行動データ s 中の訪問行動データ sa 毎に行動トピック z_{sa} が生成される。生成確率はスケジュールトピック毎の多項分布に従う。

$$z_{sa} \sim \text{Categorical}(\theta_p) \quad (1)$$

訪問行動の訪問場所 p_{sa} は多項分布から、中心時刻 t_{sa} と時

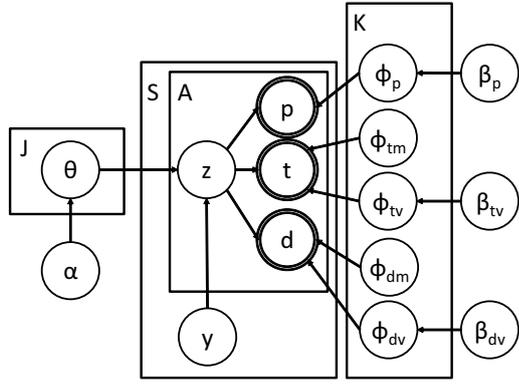


図3 提案手法のグラフィカルモデル。

間長 d_{sa} はガウス分布からそれぞれ生成される。

$$p_{sa} \sim \text{Categorical}(\phi_{pz_{sa}}) \quad (2)$$

$$t_{sa} \sim \text{Normal}(\phi_{tmz_{sa}}, \phi_{tvz_{sa}}) \quad (3)$$

$$d_{sa} \sim \text{Normal}(\phi_{dmz_{sa}}, \phi_{dvz_{sa}}) \quad (4)$$

パラメータ最適化において過学習を避けるため、式 (2) 中の多項分布のパラメータ $\phi_{pz_{sa}}$ と、式 (3) と式 (4) 中のガウス分布の分散のパラメータ $\phi_{tvz_{sa}}$, $\phi_{dvz_{sa}}$ についてはそれぞれ下記の事前分布に従うとする。

$$\phi_{pk} \sim \text{Dirichlet}(\beta_p) \quad (5)$$

$$\phi_{tvk} \sim \text{Gamma}(\beta_{tv_a}, \beta_{tv_b}) \quad (6)$$

$$\phi_{dvk} \sim \text{Gamma}(\beta_{dv_a}, \beta_{dv_b}) \quad (7)$$

3.3 パラメータ推定

ギブスサンプリング [12] を用いて未知のパラメータ ($y_s, z_{sa}, \theta_j, \phi_{pk}, \phi_{tmk}, \phi_{tvk}, \phi_{dmk}, \phi_{dvk}$) を推定する。推定は下式で示される尤度を最大化することを目的として行われる。

$$\prod_{s \in S} \prod_{a \in A_s} P(p_{sa}, t_{sa}, d_{sa} | z_{sa}, \phi) P(z_{sa} | y_s, \theta) \quad (8)$$

以下に推定手順を示す。

(1) 各日次行動データにスケジュールトピックをランダムに割り当てる。各訪問行動データに行動トピックをランダムに割り当てる。

(2) 各スケジュールトピックの θ と、各行動トピックの $\phi_p, \phi_{tm}, \phi_{tv}, \phi_{dm}, \phi_{dv}$ を更新する。 $\theta, \phi_p, \phi_{tv}, \phi_{dv}$ については MAP 推定を行う。 ϕ_{tm}, ϕ_{dm} については最尤推定を行う。

$$\theta_{jk} = \frac{A_{jk} + \alpha}{A_j + \alpha \times K} \quad (9)$$

$$\phi_{pkq} = \frac{\sum_{s \in S} \sum_{a \in A, z_{sa}=k, p_{sa}=q} 1 + \beta_p}{A_k + \beta_p \times P} \quad (10)$$

$$\phi_{tmk} = \frac{\sum_{s \in S} \sum_{a \in A, z_{sa}=k} t_{sa}}{A_k} \quad (11)$$

表2 実験で用いたデータ

項目	概要
取得開始日	2014.10.14
取得終了日	2015.01.28
分析対象ユーザ数	1500
分析対象領域 北西端	[33.685632, 130.292963]
分析対象領域 南東端	[33.380168, 130.632165]
日次行動データ数	2589
訪問行動データ数	5239
総エリア数	101

$$\phi_{tvk} = \frac{\beta_{tv_a} + \frac{1}{2}A_k}{\beta_{tv_b} + \frac{1}{2}A_k \sum_{s \in S} \sum_{a \in A, z_{sa}=k} (t_{sa} - \phi_{tmk})^2} \quad (12)$$

$$\phi_{dmk} = \frac{\sum_{s \in S} \sum_{a \in A, z_{sa}=k} d_{sa}}{A_k} \quad (13)$$

$$\phi_{tdk} = \frac{\beta_{dv_a} + \frac{1}{2}A_k}{\beta_{dv_b} + \frac{1}{2}A_k \sum_{s \in S} \sum_{a \in A, z_{sa}=k} (d_{sa} - \phi_{dmk})^2} \quad (14)$$

(3) 各日次行動データのスケジュールトピックをサンプリングする。

$$y_s \sim \text{Multinomial}(\prod_{a \in A_s} P(z_{sa} | \theta)) \quad (15)$$

各訪問行動データの行動トピックをサンプリングする。

$$z_{sa} \sim \text{Multinomial}(P(p_{sa} | \phi_p) P(t_{sa} | \phi_{ta}, \phi_{tv}) P(d_{sa} | \phi_{da}, \phi_{dv})) \quad (16)$$

(4) ステップ (2) と (3) を一定回数繰り返す。繰り返した中で最も尤度が高かった時のパラメータを推定結果とする。

4. 実験

提案手法の有効性を検討するために、実際の観光客の行動データを基に典型的な観光行動パターンを抽出することを試みる。用いたデータおよび前処理の結果得られたデータの詳細について表2に示す。

4.1 データ収集

分析に用いたのは、福岡を訪れた外国人観光客に向けて配布したスマートフォンアプリケーションから得られた位置情報である。アプリケーションをインストールしたユーザのうち居住国設定が日本以外であったユーザの中から、1500人をランダムに選出し、彼らの約3111日分の位置情報データを用いた。位置情報はGPSかWi-Fiで取得されており、基本的には5分に1度の間隔で測位されている。しかし数分から数時間単位での欠損も多い。

4.2 前処理

従来手法 [2] にならない、以下の手順で移動軌跡情報から日次行動データを取得する。

まず各移動軌跡から、滞留点と呼ばれる一箇所に留まった部分を抽出する。滞留点抽出には西田らが提案したD-Starアルゴリズム [13] を用いる。D-Starアルゴリズムでは時間閾値と

距離閾値を設定することで、一定の速度以下で移動した部分の移動軌跡をノイズや欠損を考慮ながら抽出することができる。本研究では、測位間隔が5分に一度と疎であることと、観光においては散策活動など低速での移動も行われることを考慮し、距離閾値を250mに、時間閾値を15分に設定した。この設定は、15分以上にわたって平均移動速度が時速1kmを下回っている部分が滞留点として抽出されることを意味する。

次に、抽出された各滞留点の重心を密度ベースクラスタリングであるDBSCAN [14]を用いてクラスタリングする。DBSCANは距離閾値と個数閾値を指定し、点 x から距離閾値内に個数閾値以上のクラスタ c に属する点がある場合、点 x はクラスタ c に属するとするクラスタリング手法である。距離閾値を示すパラメータである ϵ は150mに、個数閾値を示すパラメータであるMinPtsは20とした。これによって得られた各クラスタを一エリアとみなし、訪問行動データにおける訪問場所とみなした。各エリアの範囲を凹包 [15]を用いて可視化したものの一部を図4に示す。数字はエリアIDを示し、エリア0は博多駅周辺、エリア1は中州周辺、エリア7は天神駅周辺などとなっている。

4.3 分析結果

前処理によって得られた日次行動データに対して提案手法を適用する。スケジュールトピック数は20、行動トピック数は30とした。

提案手法の適用によって得られた結果の例を図5に示す。スケジュールトピック1は代表的な行動パターンの一つを示しており、午前中に博多に長期的に滞在した後、夕方に天神、夜に中洲に行くパターンが存在することが分かる。午前中の滞在が560分間と非常に長いですが、これは前日から継続してホテルに宿泊して同じ場所に留まっているユーザーが含まれているためであると考えられる。博多、天神、中州は福岡において代表的な観光エリアであることが知られており、それら3つを回る行動パターンについて、それぞれの訪問時刻と時間長を明らかにする形で抽出することが出来た。

スケジュールトピック2は大宰府への訪問を含むパターンであり、大宰府に訪問したユーザーの多くは夜に博多に行くことが分かる。行動トピック4とトピック5は時間的に被っているが、これは行動トピックの生成は確率的であり、同じスケジュールトピックから生成された日次行動データでも大宰府を訪問する場合とそうでない場合の両方が存在していることを示している。行動トピック4スケジュールトピック1では20時を中心とした時間帯に訪問するエリアとしては中州が最も多かったが、大宰府に訪問した場合は博多に行くことが多く、また時間長に関してはそれほど変わらないことが分かった。

上記で示したように、提案手法は訪問エリアやその遷移関係に加えて訪問時刻や時間長も考慮した頻出行動パターン発見を実現していることが確かめられた。

5. おわりに

本研究では訪問の場所に加えて中心時刻と時間長を考慮した日次行動データからの頻出行動パターン発見をめざし、確率的

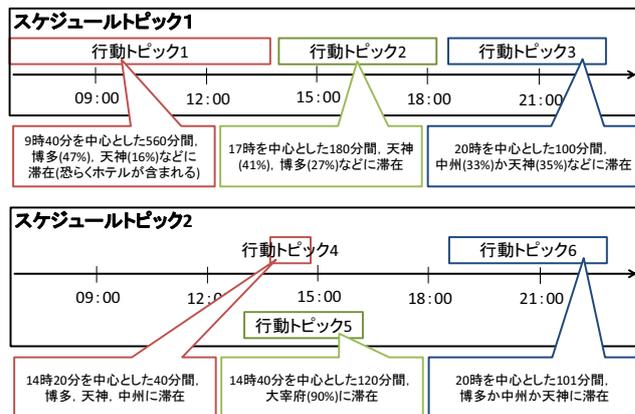


図5 提案手法によって得られた頻出行動パターンの例。長方形の左端と右端が訪問の開始時刻と終了時刻を示している。

生成モデルを用いた分析手法を提案した。提案モデルは各日次行動データが一つのスケジュールトピックを持ち、スケジュールトピックが行動トピックを生成し、行動トピックが訪問行動データを生成するものと仮定する。行動トピックは訪問の場所、中心時刻、時間長についてのパラメータを持っており、これによって時刻や時間長を考慮できる。また、行動トピックの生成が確率的であることから、一部の訪問行動データに欠損があっても類似したパターンとして抽出することを実現している。

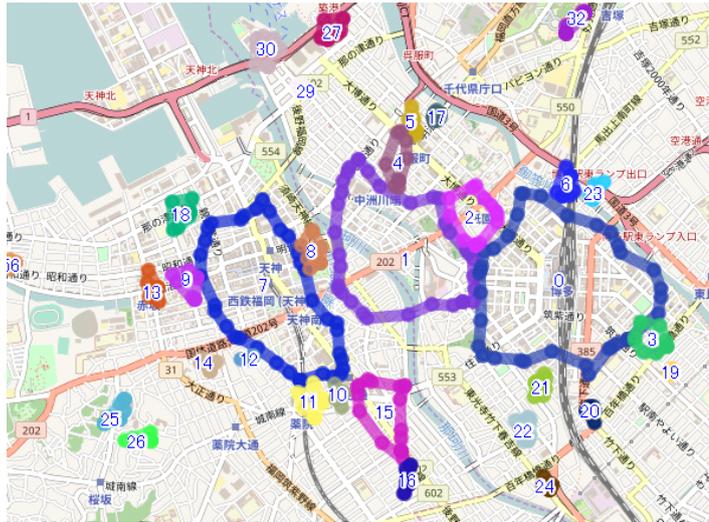
実際に訪日観光客から収集された日次行動データに対して提案手法を適用することによって、観光施策の決定などに有益な情報が抽出できることが確認された。

今後は抽出された行動パターンとユーザーの属性情報などの外部情報を関連付けた分析を行い、より有益な知見を発見することに取り組む。また、得られた知見を基に情報推薦を行うなどして得られた知見の有用性についても確かめていく。

文 献

- [1] Fosca Giannotti, Mirco Nanni, and Dino Pedreschi. Efficient mining of temporally annotated sequences. In *SDM*, pp. 348–359. SIAM, 2006.
- [2] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. Mining individual life pattern based on location history. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pp. 1–10. IEEE, 2009.
- [3] 熊谷雄介, 今井良太, 松林達史. 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.
- [4] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090. ACM, 2011.
- [5] Kim Hideaki, Iwata Tomoharu, Fujiwara Yasuhiro, and Ueda Naonori. Read the silence: Well-timed recommendation via admixture marked point process. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [6] Jiawei Han, Jian Pei, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and MC Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *proceedings of the 17th international conference on data engineering*, pp. 215–224, 2001.

博多・天神などの市街地



福岡空港



大宰府



©OpenStreetMap Contributors

図 4 福岡周辺の観光客の移動軌跡データから得られた。数字はエリアの ID を示す。各エリアを囲う枠の色はランダムである。

- [7] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 363–372. ACM, 2013.
- [8] 幸島匡宏, 松林達史, 澤田宏. 複合データ分析技術と NTF (1) 複合データ分析技術とその発展. 電子情報通信学会誌, Vol. 99, No. 6, pp. 543–550, 2016.
- [9] 松林達史, 幸島匡宏, 澤田宏. 複合データ分析技術と NTF (2) テンソルデータの因子分解技術と実応用例. 電子情報通信学会誌, Vol. 99, No. 7, pp. 691–698, 2016.
- [10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.
- [11] Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya, and Ko Fujimura. Geo topic model: joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 375–384. ACM, 2013.
- [12] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, Vol. 85, No. 410, pp. 398–409, 1990.
- [13] Kyosuke Nishida, Hiroyuki Toda, and Yoshimasa Koike. Extracting arbitrary-shaped stay regions from geospatial trajectories with outliers and missing points. In *proceedings of the 8th ACM SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS 2015)*, pp. 1–6. ACM, 2015.
- [14] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, Vol. 96, pp. 226–231, 1996.
- [15] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, Vol. 29, No. 4, pp. 551–559, 1983.