Vol.2015-DBS-162 No.16
2015/11/26

ユーザ行動モデルに基づく スポット間確率ネットワークの構築

山岸 祐己[1,a)]　斉藤 和巳[1,b)]

概要：本論文では，観測されたデータを基に，ユーザの観光行動を確率モデル化することを試みる．我々はまず，各観光スポットの人気度を導入した Lévy flight 行動過程に基いた確率モデルと，観測されたユーザ行動データからモデルのパラメータを推定する効率的な学習アルゴリズムを提案する．そして，提案したモデルから得られた条件付き確率を用いて，二種類のスポットランキング手法を提案する．レビューサイトのデータセットから生成したユーザ行動データを用いた実験では，パラメータ推定に関する実験結果を述べ，提案したランキング手法とナイーブな人気度ランキングとの比較を行う．実験結果から，パラメータ推定結果は直感的に解釈可能であることを示し，提案ランキング手法は魅力的な地域に位置するスポットを自然と高く評価していることを示す．

キーワード：確率モデル，機械学習，レビューサイト

A Probability Network Construction of POIs Based on User Behavior Model

Yuki Yamagishi[1,a)]　Kazumi Saito[1,b)]

Abstract: We attempt stochastic modeling of travel behavior processes from the observed data. To this end, based on the Lévy flight behavior process combined with the popularity of each point of interest, we first propose a probability model and efficient method that estimates the model parameters from the observed user behavior data. Then, we propose two methods for POI ranking by using the probability obtained from our proposed model. In our experiments using user behavior data constructed from a review site dataset, we report our experimental results on parameter estimation, and examine the properties of POI ranking methods in comparison to a naive popularity ranking method. As our experimental results, we show that our parameter estimation results are intuitively interpretable, and as a favorable property, our ranking methods naturally give high ranks to POIs located in attractive regions.

Keywords: probability model, machine learning, review sites

1. はじめに

“TripAdvisor”[*1] に代表されるような観光に関するソーシャルメディアの出現によって，あらゆる観光スポットに対してレビューが投稿されるようになったため，それらのレビューを基に，大規模なユーザ行動データを取得できるようになった．このようなレビューは，ユーザの個々の意思決定に基いて常に生成され続けているが，人間行動の本質的な特性として，このような観光行動にはいくつかの統計的な規則性が自然と仮定できる．よって，巨視的分析の見地から，観光行動の統計的規則性を見出すことは可能なはずであり，それらの統計的性質に基いて計算モデルを構築すれば，観光行動における人間行動の正確な将来予測が

[1] 静岡県立大学
University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka, 422-8526 Japan
[a)] yamagissy@gmail.com
[b)] k-saito@u-shizuoka-ken.ac.jp
[*1] http://www.tripadvisor.com/

期待できる．特に，そのような予測能力は，社会動向や市場動向の先取りのためにも重要であると言える．従って我々は，観光行動における，基本的な人間行動メカニズムを解明することを目的として，計算モデルを用いた研究手法を提案する．

我々は，ユーザ行動のモデル化の先駆的研究として，Lévy flight 行動過程 [4] に焦点を当てている．よって，各観光スポットの人気度を導入した Lévy flight 行動過程に基いた確率モデルと，観測されたユーザ行動データからそのモデルのパラメータを推定する効率的な学習アルゴリズムを提案する．提案手法は，観測されたユーザの行動分布について，ユーザが次にどこの観光スポットへ移動するかという予測を最適化させることにより，このモデルのパラメータを推定する．より具体的には，観測されたユーザ行動データについての対数尤度関数を構築し，機械学習の枠組みで関数の最大化を行う [2]．提案学習アルゴリズムは，尤度関数の凸性を十分に利用した反復計算の仕組みを用いているため非常に効率的である [11]．更に，このユーザ行動の確率モデルの応用として，スポット間に条件付き確率を持つ有向リンクを張った観光スポットネットワークを構築する．また，ネットワーク上の重要ノードを見分ける研究 [15] と同じように，それらの条件付き確率に基づいた二種類のランキング手法を提案する．

提案モデルと提案ランキングの評価には，TripAdvisor のデータセットから生成したユーザ行動データを用いる．具体的には，まず，ユーザ行動データの基本統計量を示した後，観光スポットの人気度のスケールフリー性 [12] と行動データにおける移動距離のスケールフリー性 [4], [12] を調べる．そして，パラメータ推定に関する実験結果を述べ，提案したランキング手法とナイーブな人気度ランキングとの比較を行う．ここで我々は，提案モデルと提案ランキング手法は，orienteering problem [13] 等に対する他の手法を改善するためのコア技術になり得ると考えている．

本論文の構成は以下の通りである．まず，関連研究について述べた後，提案モデル，パラメータ推定法，ランキング手法について説明する．そして，TripAdvisor から取得したデータセットの調査結果を示し，パラメータ推定とランキング手法による実験結果を述べる．最後に，今回得られた主要な結果と今後の展開についてまとめる．

2. 関連研究

昨今の技術革新と高性能なモバイルデバイスの普及は，現代人のコミュニケーションスタイルを大きく変え，種々のソーシャルメディアは現代人の日常生活に実質的な影響を及ぼしている．よって，近年は推薦システム [10] の研究が注目されており，我々の研究内容はロケーションベースの推薦手法 [1] に該当すると言える．注目すべき研究としては，Zheng らが提案した GPS の軌跡から重要なスポットを発見する手法 [16] が挙げられる．しかし，殆どの既存法は，人間の行動原理を仮定することなく考案されているため，予測性能の改善には限界があるはずである．

人間行動の基本的な特性としては，人間の移動距離の分布は $p(\delta) \propto \delta^{\beta}$ のような冪乗則によって近似できることが報告されている [4]．ここで，δ と β はそれぞれ移動距離と指数パラメータである．これは，人間の移動パターンは，Lévy flight 行動過程によってモデル化が可能であることを意味している．本論文では，観測されたユーザ行動に対して，各観光スポットの人気度を導入した Lévy flight 行動過程に基いた確率モデルを提案する．

Lévy flight 行動過程において必要となる β のようなモデルパラメータを推定するために，我々は対数尤度関数に基づく統計的機械学習手法 [2] を採用し，パラメータについて関数を最大化するために非線形最適化における反復アルゴリズム [7] を利用する．ここで，我々のモデルは，目的関数の凸性によって大域最適解を持つことが保証されている [11] ことに注意されたい．

大規模な複雑ネットワークの構造や機能に関する研究は，社会学，生物学，物理学，コンピュータ科学等の様々な分野で注目されている [9]．特に，これらのネットワークにおけるスケールフリー性は幅広く研究されており [5], [12]，次数相関 [14] 等のより複雑な特徴が提案されてきた．本論文では，ネットワークが持つとされるこれらの特徴に着目し，データセットと実験結果の分析を行う．

与えられたネットワークにおいて，様々な側面から重要ノードを発見することは基礎的問題とされており，ソーシャルネットワーク分析の分野では，次数中心性，近接中心性，媒介中心性といった，中心性の指標が幅広く研究されている [15]．一方，Web 情報検索の分野では，PageRank [3] と HITS [6] によるノードランキングが広く認識されている．これらの手法の中でも，入次数と PageRank は今回の確率ネットワークに適用することができるので，我々はこの二手法に基づいたランキングを提案する．実験では，ナイーブな人気度ランキングと比較して，提案ランキング手法がどのような特性を持っているのかを検証する．

3. 提案手法

まず，ユーザ行動に対する確率モデルを提案する．ユーザ集合を $\mathcal{U} = \{u, v, w, \cdots\}$，スポット集合を $\mathcal{S} = \{q, r, s, \cdots\}$ とし，それぞれの要素数を $M = |\mathcal{U}|$, $N = |\mathcal{S}|$ とする．ここで，二つのスポット r，s 間の距離を $d(r, s)$ として表す．そして，指数パラメータ θ_1 による Lévy flight 行動過程に従えば，ユーザ u がスポット r を訪れた後にスポット s を訪れる条件付き確率 $p_1(s \mid r; \theta_1)$ は，$d(r, s)^{-\theta_1}$ に比例することが仮定できる．即ち，その関係は次式となる．

$$p_1(s \mid r; \theta_1) = \frac{d(r, s)^{-\theta_1}}{\sum_{q \in \mathcal{S}} d(r, q)^{-\theta_1}}. \quad (1)$$

後のデータセットの分析で明らかにするが，今回のデータのスポット人気度はスケールフリー性を持っている [12] ため，指数パラメータ θ_2 を用いたスポット $s \in \mathcal{S}$ の人気度を $f(s)$ とすれば，ユーザ u がスポット s に訪れる確率 $p_2(s;\theta_2)$ は $f(s)^{\theta_2}$ に比例することが仮定できる．即ち，その関係は次式となる．

$$p_2(s;\theta_2) = \frac{f(s)^{\theta_2}}{\sum_{q\in\mathcal{S}} f(q)^{\theta_2}}. \tag{2}$$

よって，これらの確率 $p_1(s \mid r;\theta_1)$，$p_2(s;\theta_2)$ を組み合わせれば，ユーザの基本行動モデルとして，以下の条件付き確率を得ることができる．

$$\begin{aligned} p(s \mid r;\boldsymbol{\theta}) &= \frac{p_1(s \mid r;\theta_1)p_2(s;\theta_2)}{\sum_{q\in\mathcal{S}} p_1(q \mid r;\theta_1)p_2(q;\theta_2)} \\ &= \frac{d(r,s)^{-\theta_1} f(s)^{\theta_2}}{\sum_{q\in\mathcal{S}} d(r,q)^{-\theta_1} f(q)^{\theta_2}}. \end{aligned} \tag{3}$$

$\boldsymbol{\theta}$ は $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ であり，$\boldsymbol{a}^T$ はベクトル $\boldsymbol{a}$ の転置を意味する．ここで，我々のモデルは，他の要因に基づく訪問確率 $p(s \mid \theta)$ を導入することによって，容易に拡張が可能であることを強調しておきたい．

次に，パラメータベクトル $\boldsymbol{\theta}$ を推定する学習アルゴリズムについて述べる．ユーザ $u \in \mathcal{U}$ がスポット $s \in \mathcal{S}$ に時刻 t で訪れたことを (u,s,t) で表せば，観測されたユーザ行動データは $\mathcal{D} = \{\cdots, (u,s,t), \cdots\}$ のように書ける．観測データ $\mathcal{D}$ から，ユーザ u が m 番目に訪問したスポットが分かるため，それを $s(u,m) \in \mathcal{S}$ として表す．ここからは，$M(u)$ をユーザ u が訪れたスポット数，$N(s)$ をスポット s に訪れたユーザ数とする．$\mathcal{D}$ についての $\boldsymbol{\theta}$ を推定するために，標準的な機械学習アプローチ [2] に基づいて，最大化する目的関数として以下の対数尤度関数を考える．

$$L(\boldsymbol{\theta};\mathcal{D}) = \sum_{u\in\mathcal{U}} \sum_{m=1}^{M(u)-1} \log p(s(u,m+1) \mid s(u,m);\boldsymbol{\theta}). \tag{4}$$

そして，$\boldsymbol{x}(r,s) = (-\log d(r,s), \log f(s))^T$ のように定義されたベクトルを新たに導入すれば，式 (3) より，式 (4) は以下のように変形できる．

$$\begin{aligned} L(\boldsymbol{\theta};\mathcal{D}) = \sum_{u\in\mathcal{U}} \sum_{m=1}^{M(u)-1} \Bigg(& \boldsymbol{\theta}^T \boldsymbol{x}(s(u,m), s(u,m+1)) \\ & - \log \sum_{q\in\mathcal{S}} \exp(\boldsymbol{\theta}^T \boldsymbol{x}(s(u,m), q)) \Bigg). \end{aligned} \tag{5}$$

よって，式 (4) で定義された目的関数の勾配ベクトルとヘス行列が以下のように計算できる．

$$\begin{aligned} \frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial \boldsymbol{\theta}} = & \sum_{u\in\mathcal{U}} \sum_{m=1}^{M(u)-1} \Bigg(\boldsymbol{x}(s(u,m), s(u,m+1)) \\ & - \sum_{q\in\mathcal{S}} p(q \mid s(u,m);\boldsymbol{\theta}) \boldsymbol{x}(s(u,m), q)) \Bigg), \end{aligned} \tag{6}$$

$$\begin{aligned} \frac{\partial^2 L(\boldsymbol{\theta};\mathcal{D})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = & \sum_{u\in\mathcal{U}} \sum_{m=1}^{M(u)-1} \\ & - \Bigg(\sum_{q\in\mathcal{S}} p(q \mid s(u,m);\boldsymbol{\theta}) \boldsymbol{x}(s(u,m), q) \boldsymbol{x}(s(u,m), q)^T \\ & - \Bigg(\sum_{q\in\mathcal{S}} p(q \mid s(u,m);\boldsymbol{\theta}) \boldsymbol{x}(s(u,m), q)) \Bigg) \\ & \Bigg(\sum_{q\in\mathcal{S}} p(q \mid s(u,m);\boldsymbol{\theta}) \boldsymbol{x}(s(u,m), q)) \Bigg)^T \Bigg). \end{aligned} \tag{7}$$

ここで，このヘス行列は穏やかな条件下で負定値となることから，目的関数が上に凸な単峰関数であることが分かるため，我々のモデルが大域最適解を持つことが保証される [11]．従って，任意の初期パラメータ値から始まるような反復計算を用いることが可能である [7]．実験では，次式の修正ベクトルによるニュートン法を用いる．

$$\boldsymbol{\delta} = -\frac{\partial L(\boldsymbol{\theta};\mathcal{D})}{\partial \boldsymbol{\theta}} \left(\frac{\partial^2 L(\boldsymbol{\theta};\mathcal{D})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1}. \tag{8}$$

定数 $\epsilon = 10^{-8}$ を用いることにより，学習アルゴリズムは以下のように要約できる．

(1) パラメータベクトルを $\boldsymbol{\theta}_v \leftarrow \mathbf{0}$ と初期化する．
(2) 式 (8) で修正ベクトル $\boldsymbol{\delta}$ を計算し，もし $|\boldsymbol{\delta}| < \epsilon$ となれば反復を終了する．
(3) パラメータベクトルを $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \boldsymbol{\delta}$ と更新し，手順 (2) に戻る．

最後に，提案行動モデルに基づいた応用について述べる．学習アルゴリズムによる推定パラメータ値を $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathcal{D})$ とすると，スポット r からスポット s に訪れる条件付き確率 $p(s \mid r;\hat{\boldsymbol{\theta}})$ を得ることができる．従って，各リンク $(r,s) \in \mathcal{S} \times \mathcal{S}$ に条件付き確率 $p(s \mid r;\hat{\boldsymbol{\theta}})$ を割り当てたスポット間ネットワーク $G = (\mathcal{S}, \mathcal{S} \times \mathcal{S})$ を考えることができる．先に述べたように，複雑ネットワークにおける基礎的問題の一つは，与えられたネットワークに対して有用な指標を使い，重要ノード（スポット）を探索することである．この目的を実現するべく，ここでは入次数と PageRank を参考にした二種類の指標を考える．入次数に基づく手法（入次数法）では，各スポット $s \in \mathcal{S}$ の指標を以下のように定義する．

$$id(s) = \sum_{q\in\mathcal{S}} p(s \mid q;\hat{\boldsymbol{\theta}}). \tag{9}$$

即ち，この手法では，他のスポットからの訪問確率の合計値が大きいスポットほど上位となる．ここで，他のスポットへの訪問確率の合計は $\sum_{q\in\mathcal{S}} p(q \mid s;\hat{\boldsymbol{\theta}}) = 1$ になることに注意されたい．一方，PageRank に基づく手法（PageRank

法）では，ランダムウォーク過程下での訪問確率が高いスポットほど上位となる．PageRank アルゴリズム [3] に従えば，スポット $s \in \mathcal{S}$ の訪問確率 $pr(s)$ は以下のように考えられる．

$$pr(s) \leftarrow (1-\alpha)\sum_{q\in\mathcal{S}} p(s \mid q;\hat{\boldsymbol{\theta}})pr(q) + \frac{\alpha}{M}. \qquad (10)$$

ここで，α は一様ジャンプ確率であり，一般的な値として $\alpha = 0.15$ と設定した [3]．上記のランダムウォークシミュレーションを行うことによって，ランキングの指標となる定常状態値 $pr(s)$ を得ることができる．

4. データセット

我々は，“TripAdvisor”[*2] から日本の観光スポット，及びそれらに投稿された日本語のレビューを取得し，レビュー ID の順序に基いてユーザ行動データ $\mathcal{D}$ を構築した．まず，このデータセットの基本統計量について述べる． このデータセットは，441,087 レビュー，$M = 52,355$ ユーザ，$N = 19,827$ スポットを有しており，レビュー時刻は 2008/07/12 から 2015/10/01 迄である．よって，ユーザが投稿したレビュー数の平均は 8.4，スポットに投稿されたレビュー数の平均は 22.2 である．また，レビュー評点は整数の 1 から 5 であり，全評点の平均点は 4.0 である．図 1 は 2008 年 7 月からの一ヶ月毎のレビュー投稿数の推移を示している．数が大きく変動しているが，投稿数は着実に増加していることが図から見て取れる．図 2 はレビュー評点の度数分布を示している．図から，ユーザは訪問したスポットに対して殆どの場合高得点をつけていることがわかる．ユーザのレビュー順序が，実際のスポット訪問順序に一致していると仮定すれば，全てのレビューデータを用いてユーザ行動データ $\mathcal{D}$ を構築することができる．

続いて，観光スポットの人気度とユーザ行動のスケールフリー性 [5] を検証する．与えられた整数 i に対し，ユーザの度数 $ud(i)$ とスポットの度数 $sd(i)$ を以下のように定義する．

$$ud(i) = |\{u \in \mathcal{U} \ : \ M(u) = i\}|,$$
$$sd(i) = |\{s \in \mathcal{S} \ : \ N(s) = i\}|. \qquad (11)$$

即ち，$ud(i)$ は i 種類のスポットを訪問したユーザ数を意味しており，$sd(i)$ は i 種類のユーザが訪問したスポット数を意味している．図 4, 3 はユーザとスポットの度数分布を示している．両図から，どちらの度数分布も適度に冪乗則に近似していることがわかるため，スポットの人気度はスケールフリー性を持つという提案手法の仮定は妥当であると言える．図 2 から分かるように，ユーザが投稿したレビューの殆どは高評価であるため，スポットに投稿されたレビュー数を，スポットの人気度そのものとみなしても，

[*2] http://www.tripadvisor.com/

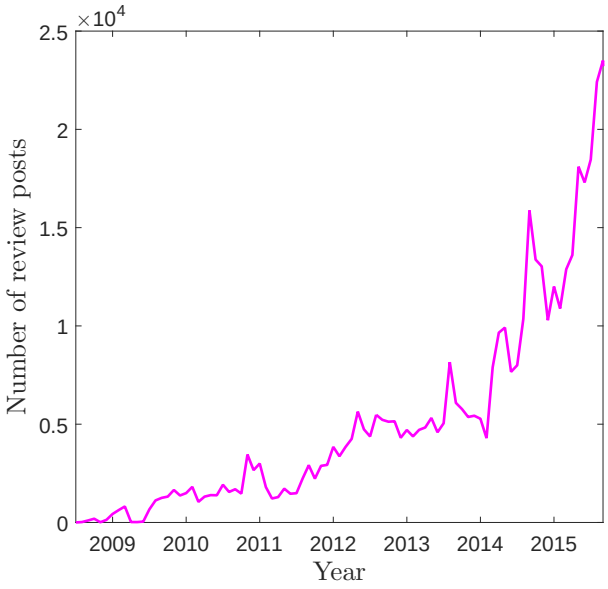

図 1 レビュー投稿数の推移

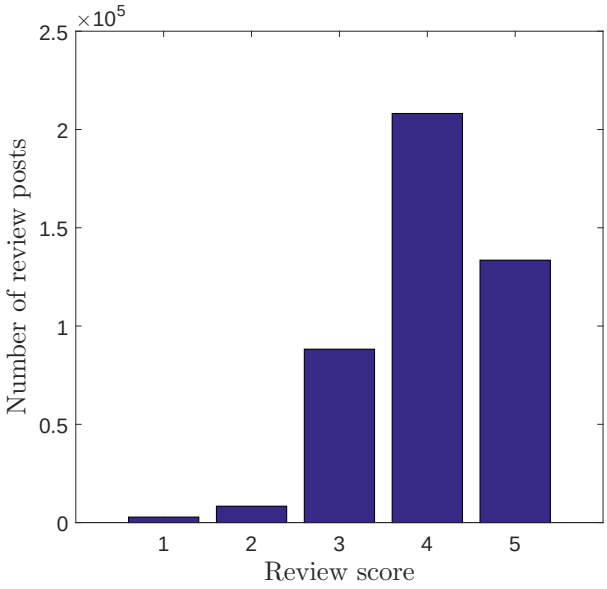

図 2 レビュー評点の度数分布

ほぼ相違がないことは自然と考えられる．よってここからは，スポット s に訪れたユーザ数 $N(s)$ をスポットの人気度 $f(s)$ として定義する．

最後に，ユーザ行動データ $\mathcal{D}$ における移動距離のスケールフリー性 [4], [12] を検証する．与えられた距離 $\mathcal{D}$ に対し，移動距離の度数 $dd(\delta)$ を以下のように定義する．

$$dd(\delta) = |\{\cup_{u\in\mathcal{U}} \cup_{1\le m<M(u)} (u,m) \ : \ \delta \le d(s(u,m), s(u,m+1)) < \delta + \epsilon\}|. \qquad (12)$$

ここで，スポット間距離は，スポットの緯度と経度を用いて GRS80 [8] に基づく測地系によって算出しており，距離間隔パラメータ ϵ は 1km とした．図 5 は移動距離の度数分布を示している．移動距離の度数分布も冪乗則に近似し

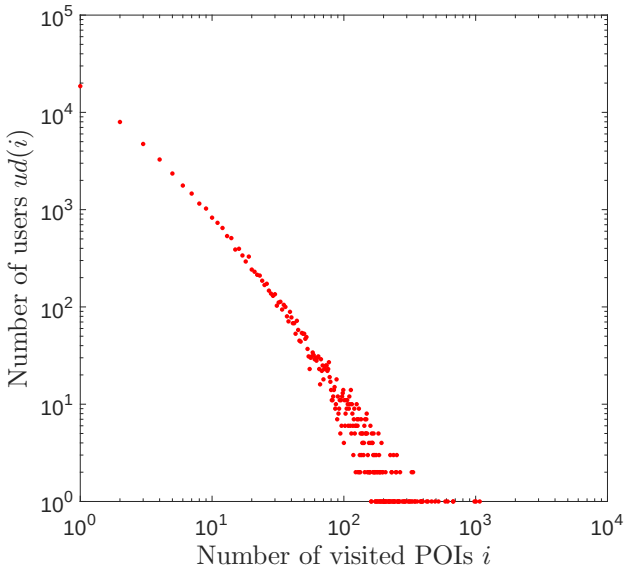

図 3 ユーザの度数分布

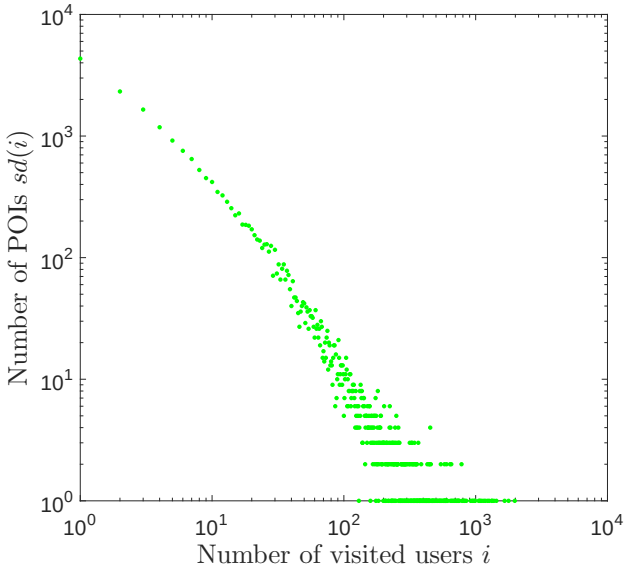

図 4 スポットの度数分布

ていることがわかるため，Lévy flight 行動過程を仮定している提案手法の妥当性が説明できる．今回，提案モデルにおける条件付き確率 $p_1(s \mid r; \theta_1)$ が，殆ど距離が離れていないスポット間に対して極めて有効に働くため，レビュー数が多い順に各スポットから半径 100m を探索していき，探索範囲内に存在する近接スポットのレビューを，探索元スポットのレビューと統合する処理を行った．この処理により，対象スポット数は最終的に $N = 14,319$ となった．

5. 実験結果

まず，訪問スポット数 $M(u)$ によってユーザを選別し，パラメータ θ_1, θ_2 を推定する．閾値 τ を導入したときの新たなユーザ行動データ $\mathcal{D}_\tau$ は以下となる．

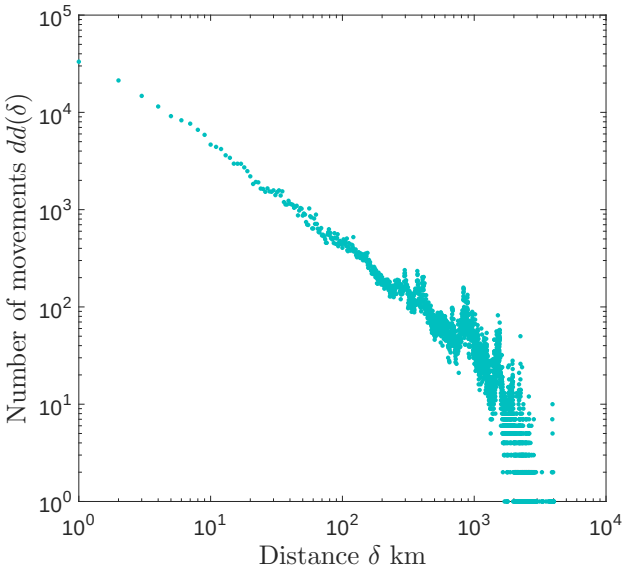

図 5 移動距離の度数分布

$$\mathcal{D}_\tau = \{(u, v, t) \in \mathcal{D} \ : \ M(u) \geq \tau\}, \tag{13}$$

図 6 はパラメータの推定結果を示したものであり，横軸は閾値 τ を，縦軸はパラメータ値をそれぞれ表す．図より，τ が大きくなるにつれて，θ_1 はわずかに増加しており，θ_2 は大いに減少していることが見て取れる．この結果は，観光スポットを多く訪問するユーザにとっては，スポットの人気度はそれほど重要な要因ではないということを示唆している．この結果の妥当性を裏付けるため，以下のように定義される次数相関 [14] を考える．

$$dc(i) = \frac{1}{ud(i)} \sum_{\{u \in \mathcal{U} \ : \ M(u)=i\}} \frac{1}{i} \sum_{m=1}^{i} N(s(u, m)), \tag{14}$$

ここで，$N(s)$ はスポット s に訪れたユーザ数であることを再度述べておく．図 7 は次数相関の検証結果を示しており，横軸はユーザが訪れたスポット数を，縦軸はユーザが訪れたスポットの人気度の平均をそれぞれ表す．なお，この検証については，近接スポットを統合していない元のデータの状態で行った．図より，次数相関の観点から見ても，図 6 と同様のことが示唆される．即ち，比較的少数の観光スポットしか訪問していないユーザは，概して比較的人気度が高い観光スポットに訪れるということがわかる．

次に，PageRank 法によるランキングの特性を，人気度によるランキングと入次数法によるランキングとの比較から考察する．いま，PageRank 法によるランキングの上位 k スポットを $R(k)$ とし，人気度と入次数法も同様に，それぞれ $R_{pop}(k), R_{ind}(k)$ とする．そして，以下のランキング類似度 $rs(k; x)$ を用いて比較評価する．

$$rs(k; x) = \frac{|R(k) \cap R_x(k)|}{k}, \tag{15}$$

ここで，$x \in \{pop, ind\}$ である．図 8 はランキング類似

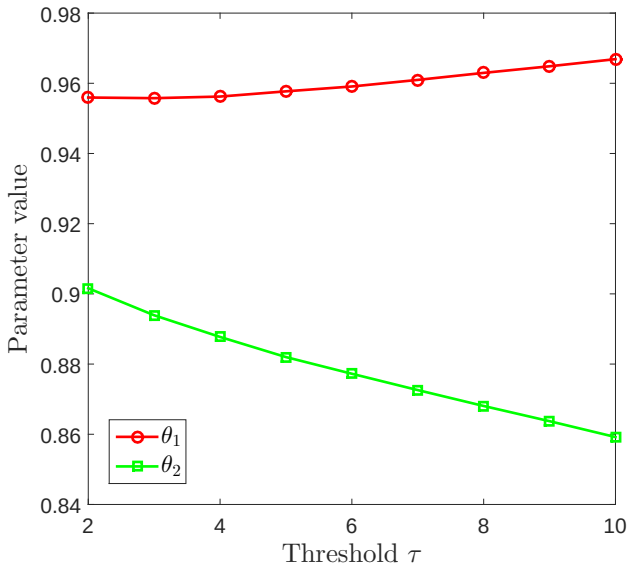

図 6 閾値 τ 毎のパラメータ推定結果

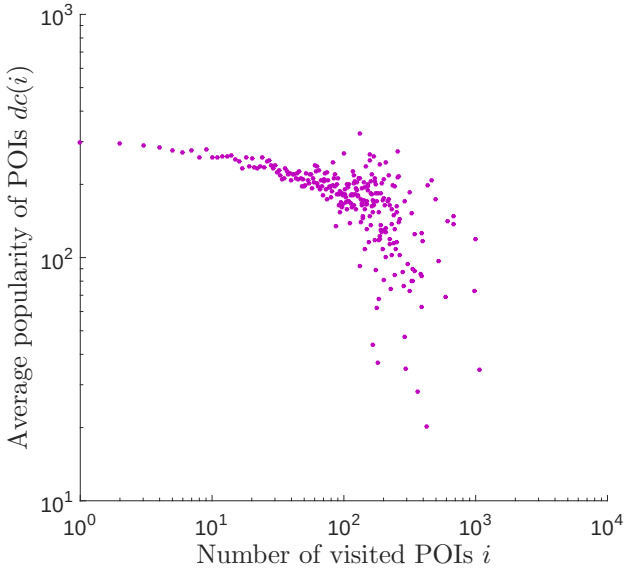

図 7 次数相関の散布図

度 $rs(k;x)$ による評価結果を $k = 1,000$ まで示したものである．図から，人気度とのランキング類似度は平均して 0.70 程度，入次数法とのランキング類似度は平均して 0.80 程度であることがわかる．即ち，これらの手法はそれぞれ，実質的に異なるランキングを生成していると言える．図 9, 10, 11 は，上位 1000 位のスポットを順位で色付けしてプロットし，ランキング結果の違いを可視化したものである．図 9 は人気度，図 10 は入次数法，図 11 は PageRank 法による可視化結果をそれぞれ示す．人気度の上位は比較的日本全体に散らばっているが，PageRank 法の上位は一部の地域に限定して分布しており，入次数法の上位はそれらの中間に位置することが見て取れる．これらの実験結果から，PageRank 法によるランキングは，他の二手法によるランキングの上位を多く含むいくつかの地域内のスポットに対して上位を与えていると考える事ができる． 更に綿密な調査をするため，上位 10 スポットに

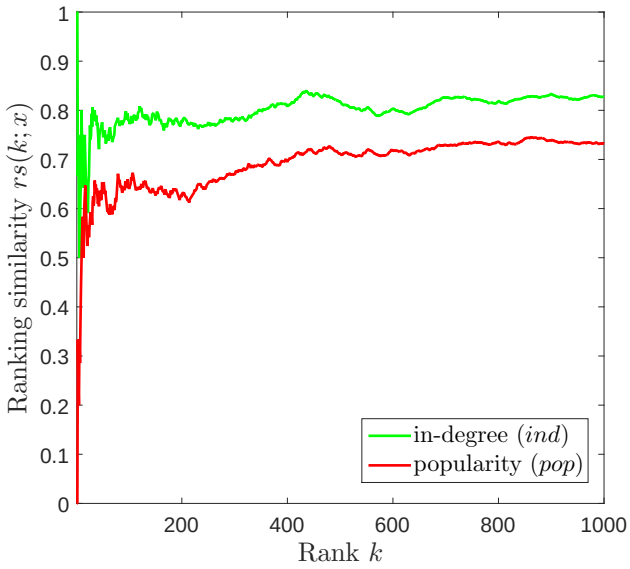

図 8 ランキング類似度の評価

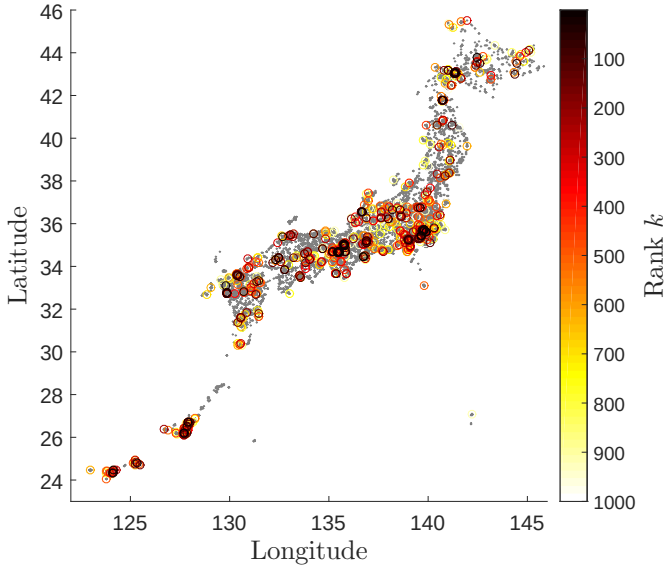

図 9 人気度ランキングの可視化結果

絞ってランキング結果の比較を行う．表 1, 2, 3 はそれぞれ人気度，入次数法，PageRank 法の上位 10 スポットを示したものである．人気度の結果は，沖縄と広島のスポットが出現していること，また，東京のスポットが 1 つしかないことが特徴的である．入次数の結果は，南関東，大阪，京都といった，代表的な都市部のスポットでバランス良く構成されていることが特徴的である．PageRank 法の結果は，東京のスポットが 4 つもあること，北海道のスポット

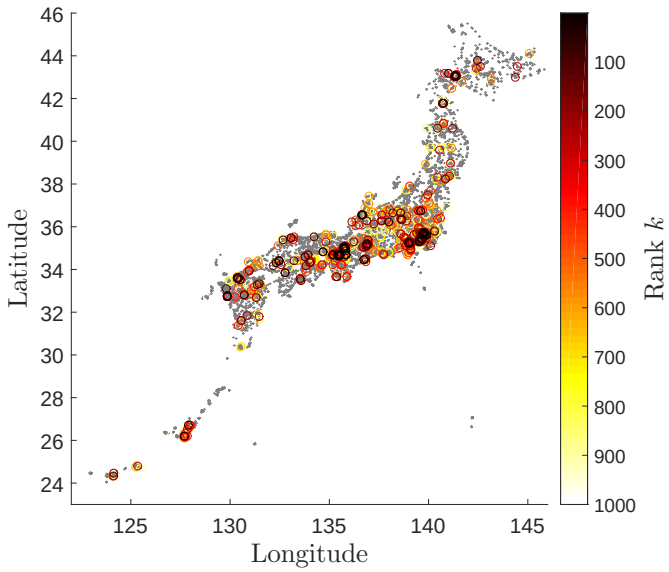

図 10 入次数法ランキングの可視化結果

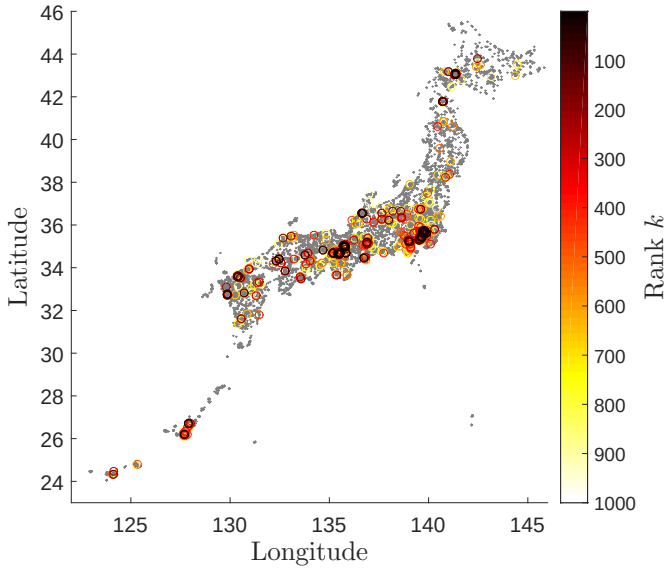

図 11 PageRank 法ランキングの可視化結果

表 1 人気度ランキングの上位 10 スポット

Rank	$N(s)$	POI Name	Prefecture
1	2277	Okinawa Churaumi Aquarium	Okinawa
2	1777	Shuri Castle	Okinawa
3	1682	Tokyo Skytree	Tokyo
4	1658	Universal Studios Japan	Osaka
5	1550	Itsukushima Shrine	Hiroshima
6	1450	Kiyomizu-dera	Kyoto
7	1436	Tokyo Disneyland	Chiba
8	1388	Dotonbori	Osaka
9	1328	Fushimi Inari-taisha	Kyoto
10	1298	Osaka Castle	Osaka

が出現していることが特徴的である．これらの結果は，先程の可視化結果からの考察を裏付けるために重要であると言える．

表 2 入次数法ランキングの上位 10 スポット

Rank	$id(s)$	POI Name	Prefecture
1	74.4131	Tokyo Skytree	Tokyo
2	66.7609	Dotonbori	Osaka
3	65.5231	Universal Studios Japan	Osaka
4	63.4598	Kiyomizu-dera	Kyoto
5	58.9946	Osaka Castle	Osaka
6	55.3797	Fushimi Inari-taisha	Kyoto
7	53.7213	Tokyo Tower	Tokyo
8	53.4037	Yokohama Chinatown	Kanagawa
9	52.3849	Tokyo Disneyland	Chiba
10	47.2586	Senso-ji	Tokyo

表 3 PageRank 法ランキングの上位 10 スポット

Rank	$pr(s)$	POI Name	Prefecture
1	0.00665	Tokyo Skytree	Tokyo
2	0.00613	Dotonbori	Osaka
3	0.00568	Kiyomizu-dera	Kyoto
4	0.00518	Yokohama Chinatown	Kanagawa
5	0.00509	Tokyo Tower	Tokyo
6	0.00487	Senso-ji	Tokyo
7	0.00471	Tokyo Station	Tokyo
8	0.00434	Odori Park	Hokkaido
9	0.00417	Tokyo Disneyland	Chiba
10	0.00409	Universal Studios Japan	Osaka

最後に，各リンク $(r,s) \in \mathcal{S} \times \mathcal{S}$ のうち，ある程度確率が低いリンク $p(s \mid r; \hat{\boldsymbol{\theta}}) < \mu$ を除去し，リンク除去の影響を平均リンク数とランキング類似度の側面から調査する．リンク除去率 μ は，$\mu \in \{0.0001 \times 2^{\eta} \ : \ \eta = 0, \cdots 6\}$ のように設定した．図 12 はリンク除去に伴うノードの平均リンク数の推移を示している．図より，ネットワークの各リンクに割り当てられた確率は非常に低いことがわかる．ここで，リンク除去率 μ を適用したネットワークにおける PageRank 法の上位 k スポットを $R_{\mu}(k)$ とし，式 15 のランキング類似度 $rs(k;\mu)$ を同様に計算する．図 13 は $k = 1000$ までの評価結果を示したものである．リンク除去率が $\mu < 0.0016$ のとき，ランキング類似度は平均して 0.90 を超えているため，これらのネットワークは図 8 に示された結果と比較して頑健であると言える．

6. おわりに

本論文では，観測されたデータを基に，ユーザ観光行動の確率モデル化を試みた．モデル化実現のために，各観光スポットの人気度を導入した Lévy flight 行動過程に基づいた確率モデルと，観測されたユーザ行動データからモデルのパラメータを推定する効率的な学習アルゴリズムを提案した．また，提案したモデルから得られた条件付き確率を用いて，二種類のスポットランキング手法を提案した．レビューサイトのデータセットから生成したユーザ行動

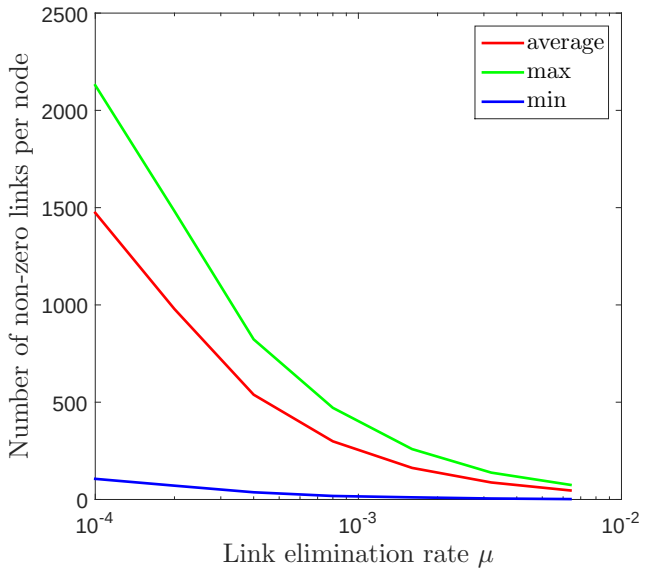

図 12 ノードの平均リンク数の推移

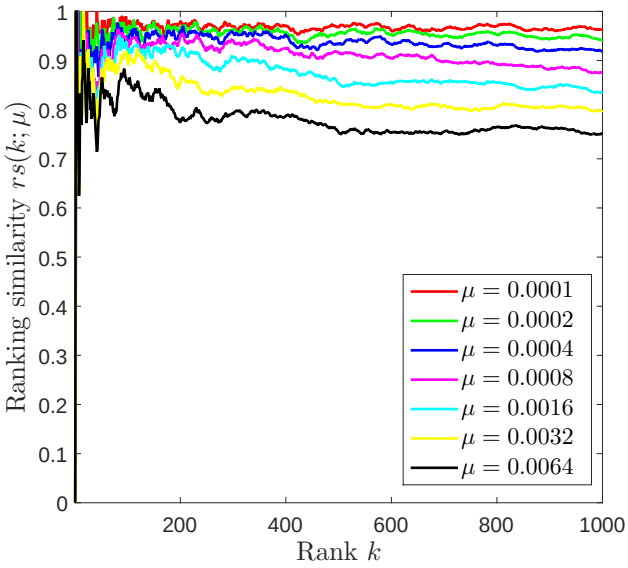

図 13 リンク除去によるランキングへの影響

データを用いた実験では，パラメータ推定に関する詳細な検証と，提案したランキング手法とナイーブな人気度ランキングとの比較を行った．実験結果としては，パラメータ推定結果は直感的に解釈が可能であることが示され，提案ランキング手法は魅力的な地域に位置するスポットを自然と高く評価していることが示された．今後は，多種多様なデータセットに対して更なる実験と検証を行い，提案手法を評価したいと考えている．

謝辞

本研究は，総務省 SCOPE (No.142306004)，及び科学研究費補助基金基盤研究 (C) (No.15K00311) の支援を受けて行ったものである．

参考文献

[1] Bao, J., Zheng, Y., Wilkie, D. and Mokbel, M.: Recommendations in location-based social networks: a survey, *GeoInformatica*, Vol. 19, No. 3, pp. 525–565 (2015).
[2] Bishop, C. M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer (2010).
[3] Brin, S. and Page, L.: The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems*, Vol. 30, pp. 107–117 (1998).
[4] Brockmann, D., Hufnagel, L. and Geisel, T.: The scaling laws of human travel, *Nature*, Vol. 439, pp. 462–465 (2006).
[5] Easley, D. and Kleinberg, J.: *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, Cambridge University Press, New York, NY, USA (2010).
[6] Kleinberg, J.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).
[7] Luenberger, D. G.: *Linear and Nonlinear Programming: Second Edition*, Kluwer Academic Publishers (2003).
[8] Moritz, H.: Geodetic Reference System 1980, *Journal of Geodesy*, Vol. 74, No. 1, pp. 128–133 (2000).
[9] Newman, M.: The structure and function of complex networks, *SIAM Review*, Vol. 45, pp. 167–256 (2003).
[10] Ricci, F., Rokach, L., Shapira, B. and Kantor, P.: *Recommender Systems Handbook*, Springer-Verlag New York, Inc, New York, NY, USA (2011).
[11] Seber, G. A. F. and Wild, C. J.: *Nonlinear Regression*, John Wiley & Sons (1989).
[12] Song, C., Koren, T., Wang, P. and Barabási, A.-L.: Modelling the scaling properties of human mobility, *Nature Physics*, Vol. 6, pp. 818–823 (2010).
[13] Vansteenwegen, P., Souffriau, W. and Oudheusden, D.: The orienteering problem: a survey, *European Journal of Operational Research*, Vol. 209, pp. 1–10 (2011).
[14] Vázquez, A.: Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations, *Physical Review*, Vol. 67, No. 5, p. 056104 (2003).
[15] Wasserman, S. and Faust, K.: *Social network analysis*, Cambridge University Press, Cambridge, UK (1994).
[16] Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y.: Mining Interesting Locations and Travel Sequences from GPS Trajectories, *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, New York, NY, USA, ACM, pp. 791–800 (2009).