

観光レビューデータから構築した 確率ネットワークによる地域分析 A Regional Analysis in Stochastic Net- work Constructed by Tourism Review Data

山岸 祐己[♡] 齊藤 和巳[◇]

Yuki YAMAGISHI Kazumi SAITO

本論文では、観測されたレビューデータを基に、ユーザー観光行動の確率モデル化を行う。我々は、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 に基づいた確率モデルと、観測されたユーザー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。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ー行動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う。実験では、パラメータ推定結果は直感的に解釈が可能であること、更に、提案ランキング手法は地域毎の特性を見出す指標として有用であることを示す。

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 our ranking method is useful as an indicator to find characteristics of each region.

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、近年、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。インターネット上のレビューというものは、ユーザーの個々の意思決定に基いて常に生成され続けているため、大規模なレビューデータに対しては、人間行動の本質的な特性、すなわち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、人間の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確率モデルを構築すれば、人間の観光行動における精緻な予測が期待できる。特に、そのような行動予測を利用した地域分析は、社会動向や市場動向の調査のためにも重要であると言える。従って我々は、人間の行動特性を利用した観光レビューデータの分析手法を、カテゴリー評価の観点から提案する。

我々は、ユーザー行動のモデル化の先行研究として、Lévy flight [1, 2, 3, 4] に焦点を当てている。実際、今回提案する確率モデルは Lévy flight の特性を大いに利用しており、観光スポットの人気度もこの特性に関連付けて扱っている。よって、各観光スポットの人気度を導入した Lévy flight に基いた確率モデルと、観測されたユーザー行動データから、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、観測されたユーザーの観光行動について、ユーザーが次にどこ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ものである。より具体的には、観測されたユーザー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 [5]。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、効率的に大域最適解を求めることが可能である [6]。更に、このユーザー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、観光スポット間の確率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 [7] と同じように、それらの条件付き確率に基づいた 2 種類のランキング手法を提案する。各ランキング結果は、Mann-Whitney の統計量 [8] に基づき、八地方区分及び都道府県区分によって評価する。

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ー行動データを用いる。具体的には、まず、ユーザー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 [9] と行動データにおける移動距離のスケールフリー性 [1, 9] 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 [10] 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多大な影響を及ぼしている。よって、近年は情報推薦システム [11] の研究が注目されており、本論文の研究内容はロケーションベースの推薦手法 [12] に該当すると言える。特に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 [13] が挙げられる。しかし、ほとんどの既存法は、人間の行動特性を仮定することなく考案されているため、予測性能の改善は困難なことが予期される。

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 [1, 9]。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 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、観測されたユーザー行動に対して、各観光スポットの人気度を導入した Lévy flight に基いた確率モデルを提案する。Lévy flight において必要

[♡] 学生会員 静岡県立大学大学院経営情報イノベーション研究科
yamagissy@gmail.com

[◇] 正会員 静岡県立大学経営情報学部
k-saito@u-shizuoka-ken.ac.jp

¹http://www.tripadvisor.com/

となる β のようなモデルパラメータを推定するために、ここでは対数尤度関数に基づく統計的機械学習手法 [5] を採用し、この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 [14] を使用する。本論文におけるモデルは、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 [6] ことに注意されたい。

提案モデルの推定パラメータが求まれば、精緻な予測性能を備えた観光スポット間の大規模ネットワークを構築することが可能となる。かねてより、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 [15]。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 [9, 16]、次数相関 [17] 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。

更に今回は、構築した確率ネットワークを用いて、日本の観光における地域毎の重要度を評価することを試みる。一般に、様々な側面から重要ノード群を発見することは、ネットワーク分析において基礎的な問題とされている。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おり [7]、一方、Web 情報検索の分野では、PageRank [18] と HITS [19] 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、次数に関する指標と PageRank は今回の確率ネットワークに適用することができるため、本論文ではこの 2 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によって地域毎の評価がどのように変化するかを検証する。

3. 提案手法

まず、ユーザー行動に対する確率モデルを提案する。ユーザー集合を $\mathcal{U} = \{u, v, w, \dots\}$ 、スポット集合を $\mathcal{S} = \{q, r, s, \dots\}$ とし、それぞれの要素数を $N = |\mathcal{U}|$, $M = |\mathcal{S}|$ とする。ここで、2 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 に従えば、ユーザー 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 $p_1(s | r; \theta_1)$ は、 $d(r, s)^{-\theta_1}$ に比例することが仮定できる。すなわち、その関係は次式となる。

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

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 [9] 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、ユーザー 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。すなわち、その関係は次式となる。

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

よって、これらの確率 $p_1(s | r; \theta_1)$, $p_2(s; \theta_2)$ を組み合わせれば、ユーザー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。

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

θ は $\theta = (\theta_1, \theta_2)^T$ であり、 $\mathbf{a}^T$ はベクトル $\mathbf{a}$ の転置を意味する。ここで、本論文のモデルは、他の要因に基づく訪問確率 $p(s | \theta)$ 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 θ を推定する学習アルゴリズムについて述べる。ユーザー $u \in \mathcal{U}$ 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、観測されたユーザー行動データは $\mathcal{D} = \{\dots, (u, s, t), \dots\}$ のように書ける。観測データ $\mathcal{D}$ から、ユーザー 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、 $M(u)$ をユーザー u が訪れたスポット数、 $N(s)$ をスポット s に訪れたユーザー数とする。 $\mathcal{D}$ についての θ を推定するために、標準的な機械学習アプローチ [5] 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。

$$L(\theta; \mathcal{D}) = \sum_{u \in \mathcal{U}} \sum_{1 \leq m < M(u)} \log p(s(u, m+1) | s(u, m); \theta). \quad (4)$$

そして、 $\mathbf{x}(r, s) = (-\log d(r, s), \log f(s))^T$ 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。

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

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。

$$\begin{aligned} \frac{\partial L(\theta; \mathcal{D})}{\partial \theta} &= \sum_{u \in \mathcal{U}} \sum_{1 \leq m < M(u)} \left(\mathbf{x}(s(u, m), s(u, m+1)) \right. \\ &\quad \left. - \sum_{q \in \mathcal{S}} p(q | s(u, m); \theta) \mathbf{x}(s(u, m), q) \right), \end{aligned} \quad (6)$$

$$\begin{aligned} \frac{\partial^2 L(\theta; \mathcal{D})}{\partial \theta \partial \theta^T} &= \sum_{u \in \mathcal{U}} \sum_{1 \leq m < M(u)} \left(- \left(\sum_{q \in \mathcal{S}} p(q | s(u, m); \theta) \mathbf{x}(s(u, m), q) \mathbf{x}(s(u, m), q)^T \right. \right. \\ &\quad \left. - \left(\sum_{q \in \mathcal{S}} p(q | s(u, m); \theta) \mathbf{x}(s(u, m), q) \right) \right. \\ &\quad \left. \left(\sum_{q \in \mathcal{S}} p(q | s(u, m); \theta) \mathbf{x}(s(u, m), q) \right)^T \right). \end{aligned} \quad (7)$$

ここで、このヘス行列は想定される入力 ($d(r, s) \geq 1$, $f(s) \geq 1$) において負定置となることから、目的関数が上に凸な単峰関数であることが分かるため、本手法のモデルが大域最適解を持つことが保証される [6]。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 [14]。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。

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

終了条件として $\epsilon = 10^{-8}$ を設定すれば、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 $\theta_v \leftarrow \mathbf{0}$ と初期化する。
2. 式 (8) で修正ベクトル δ を計算し、もし $\|\delta\| < \epsilon$ となれば反復を終了する。
3. パラメータベクトルを $\theta \leftarrow \theta + \delta$ 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 $\hat{\theta} = \arg \max_{\theta} L(\theta; \mathcal{D})$ 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 $p(s | r; \hat{\theta})$ 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 $p(s | r; \hat{\theta})$ 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（今回の場合は観光スポット）を探索することである。よって、ここでは入次数と PageRank を参考にした 2 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。

$$ind(s) = \sum_{q \in \mathcal{S}} p(s | q; \hat{\theta}). \quad (9)$$

すなわち、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 $\sum_{q \in \mathcal{S}} p(q | s; \hat{\theta}) = 1$ 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 [18] に従えば、スポット $s \in \mathcal{S}$ の訪問確率 $prk(s)$ は以下のように考えられる。

$$prk(s) \leftarrow (1 - \alpha) \sum_{q \in \mathcal{S}} p(s | q; \hat{\theta}) prk(q) + \frac{\alpha}{M}. \quad (10)$$

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 [18]。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 $prk(s)$ を得ることができる。

4. データセット

現実データとして、“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、レビューの順序に基いてユーザー行動データ $\mathcal{D}$ 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、 $N = 52,355$ ユーザー、 $M = 19,827$ スポットを有しており、レビュー時刻は 2007/02/07 から 2015/10/21 迄である。よって、ユーザー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 年からの 1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、ユーザー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、ユーザー行動データ $\mathcal{D}$ 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ー行動のスケールフリー性 [16] を検証する。与えられた整数 i に対し、ユーザーの度数 $ud(i)$ とスポットの度数 $sd(i)$ を以下のように定義する。

$$\begin{aligned} ud(i) &= |\{u \in \mathcal{U} : M(u) = i\}| \\ sd(i) &= |\{s \in \mathcal{S} : N(s) = i\}|. \end{aligned} \quad (11)$$

即ち、 $ud(i)$ は i 種類のスポットを訪問したユーザー数を意味しており、 $sd(i)$ は i 種類のユーザーが訪問したスポット数を意味している。図 3, 4 はユーザー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、

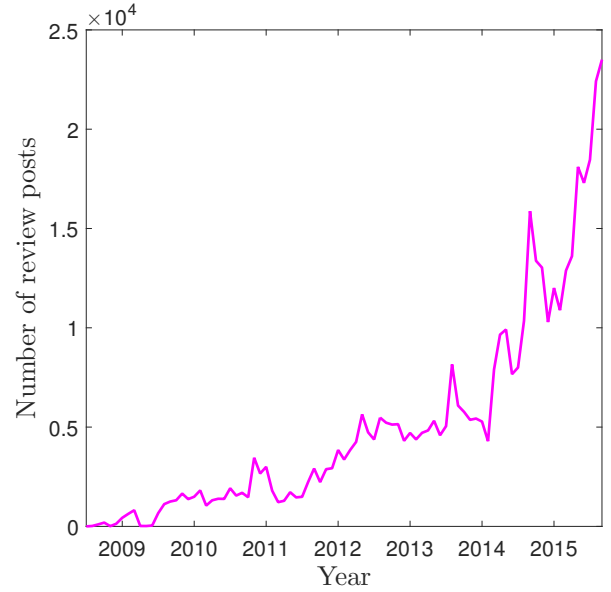


図 1: レビュー投稿数の推移

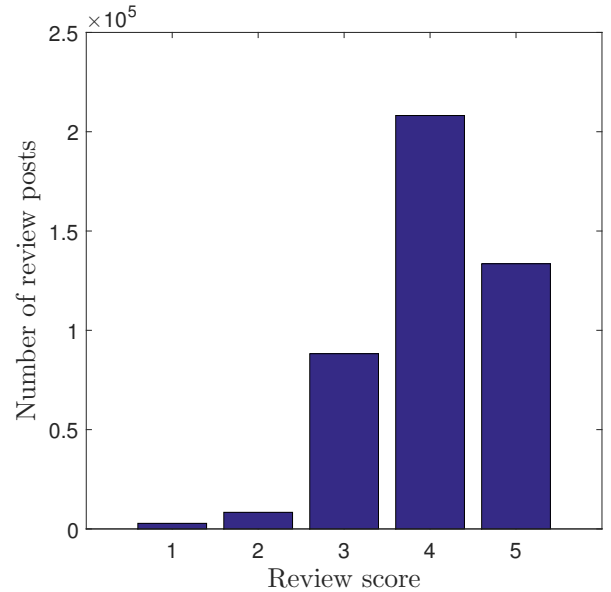


図 2: レビュー評点の度数分布

ユーザーが投稿したレビューの殆どは高評価であり、1点や2点のレビューは僅かであるため、スポットに投稿されたレビュー数を、単純にそのスポットの人気度と考えても問題はないように思われる。よってここからは、スポット s に訪れたユーザー数 $N(s)$ を、スポットの人気度 $f(s) = N(s)$ として定義する。

最後に、ユーザー行動データ $\mathcal{D}$ における移動距離のスケールフリー性 [1, 9] を検証する。与えられた距離 Δ に対し、移動距離の度数 $dd(\Delta)$ を以下のように定義する。

$$\begin{aligned} dd(\Delta) &= |\{u \in \mathcal{U} \mid \exists 1 \leq m < M(u) : (u, m) : \\ &\quad \Delta \leq d(s(u, m), s(u, m+1)) < \Delta + \epsilon\}|. \end{aligned} \quad (12)$$

スポット間の距離は、スポットの緯度と経度を用いて、GRS80 [20] に基づく測地系によって算出しており、距離間隔 ϵ は 1 km とし

²<http://www.tripadvisor.com/>

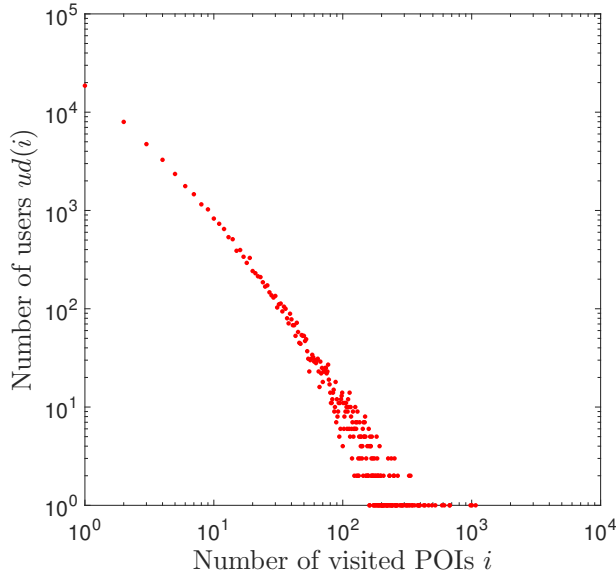


図 3: ユーザーの度数分布

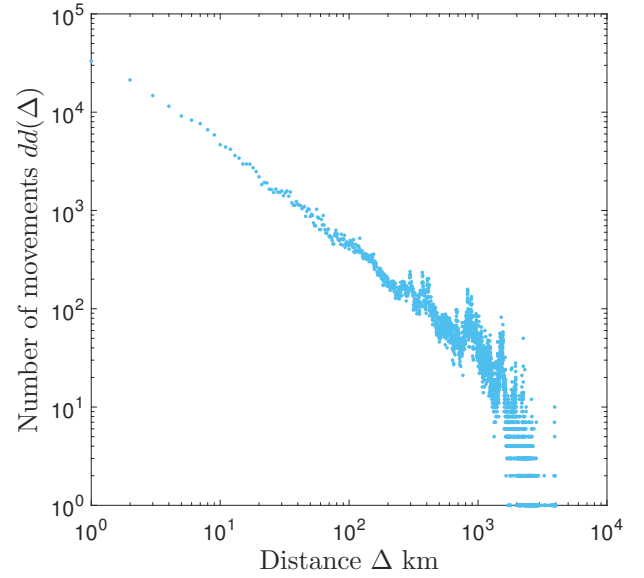


図 5: 移動距離の度数分布

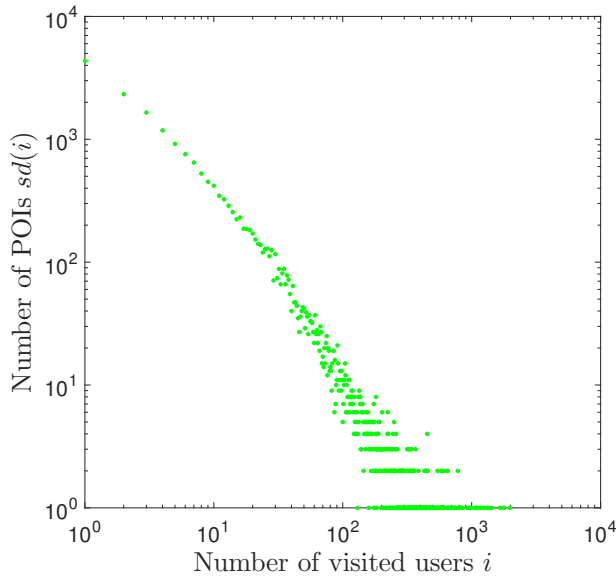


図 4: スポットの度数分布

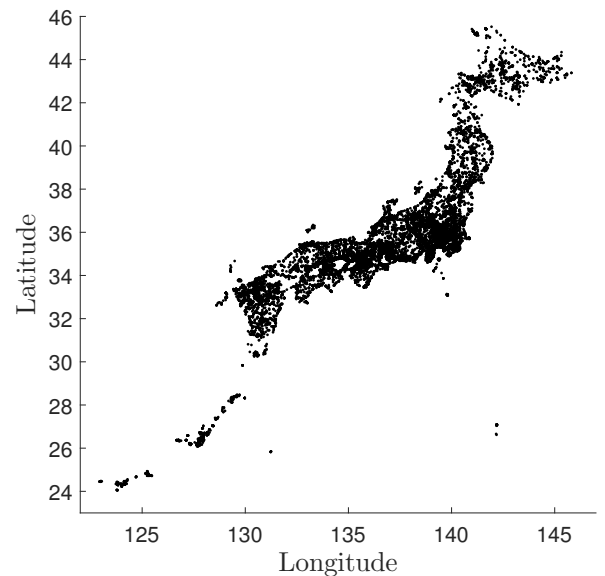


図 6: 実際のスポットの位置

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，冪乗則に近似しているため，*Lévy flight* を仮定している提案手法の妥当性は高いと言える．今回，提案モデルにおける条件付き確率 $p_1(s|r;\theta_1)$ 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 100 m 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 $M = 14,319$ となった．実際のスポットの位置は図 6 のようになっている．

5. 実験結果

5.1 パラメータ推定結果

まず，訪問スポット数 $M(u)$ 毎にユーザーを選別し，パラメータ θ_1, θ_2 を推定する．訪問スポット数 $M(u)$ の閾値 τ を導入したときの新たなユーザー行動データ $\mathcal{D}$ は以下となる．

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

図 7 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， θ_1 にはさして変化が見られないが， θ_2 は τ が大きくなるにつれて明らか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ーにとっては，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 [17] を考える。

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

ここで、 $N(s)$ はスポット s に訪れたユーザー数であり、スポットの人気度であることに注意されたい。図 8 は次数相関の検証結果を示しており、横軸はユーザーが訪れたスポット数を、縦軸はユーザーが訪れたスポットの人気度の平均をそれぞれ表す。図より、次数相関の観点から見ても、図 7 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ーは、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。

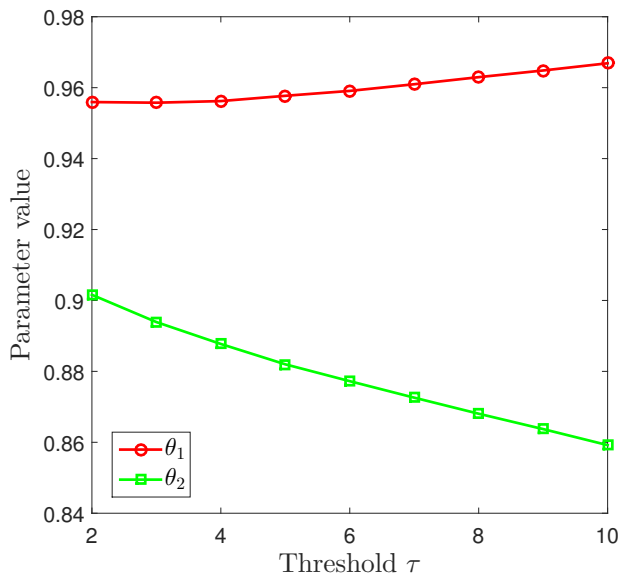


図 7: 閾値 τ 毎のパラメータ推定結果

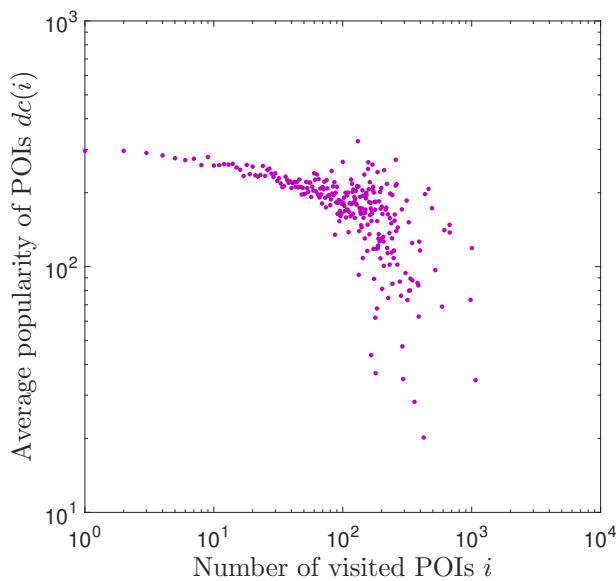


図 8: 次数相関の散布図

5.2 ランキング結果

次に、PageRank 法によるランキング (prk) の特性を、人気度によるランキング (以降 pop) と入次数法によるランキング (ind) 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 $rs(k; x)$ を用いて比較評価を行う。

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

ここで、 $x \in \{pop, ind\}$ である。図 9 はランキング類似度 $rs(k; x)$ 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 10, 11, 12 は、上位 1000 位のスポットを順位で色付けしてプロットし、ランキング結果を可視化したものである。図 10 は人気度、図 11 は入次数法、図 12 は PageRank 法による可視化結果をそれぞれ表す。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 2 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集中的に上位を与えていると考えられる。

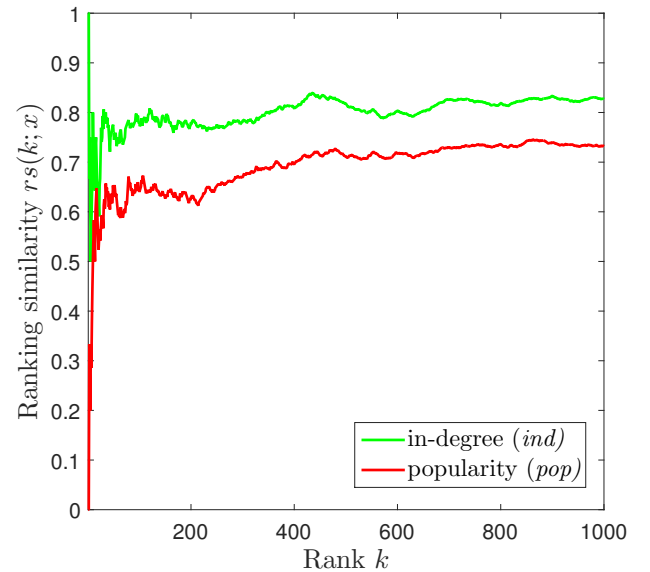


図 9: ランキング類似度の評価

更に詳細な考察をするため、ランキング結果を用いた地域毎の評価を行う。与えられたスポット集合 S の地域集合を J とする。ここで、地域集合の要素数は $J = |\mathcal{J}|$ とし (スポット集合は $M = |S|$)、各要素は整数と同一視されるとする。つまり、 $S = \{1, \dots, s, \dots, S\}$ および $\mathcal{J} = \{1, \dots, j, \dots, J\}$ である。また、スポット s が属す地域を $j = f(s)$ で表し、各地域に属するスポット数を $M_j = |S_j| = |\{s : j = f(s)\}|$ とする。各スポット s に対し、そのランキングは $1 \leq r_s \leq M$ で与えられているとする。ただし、同順位が起こるケースでは、 r_s は平均順位で補正されるとする。ここでの目的は、地域集合とランキング付きのスポット集合が与えられたとき、ランキングの高い、または逆に低いスポットが有意に多く含まれる地域を定量的に評価する指標の構築である。以下には Mann-Whitney の統計量 [8] に基づく自然な拡張法を示す。

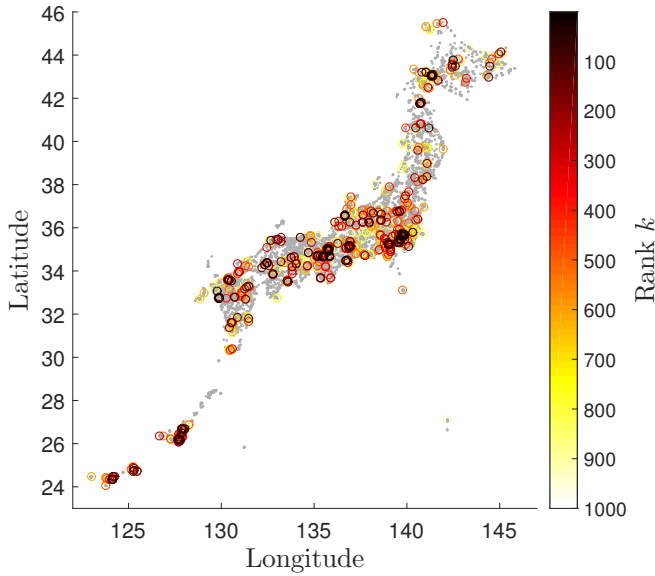


図 10: 人気度ランキングの可視化結果

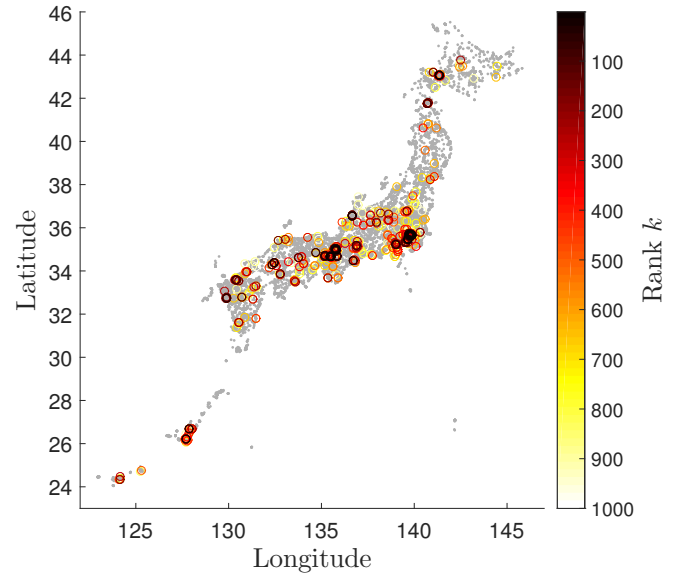


図 12: PageRank 法ランキングの可視化結果

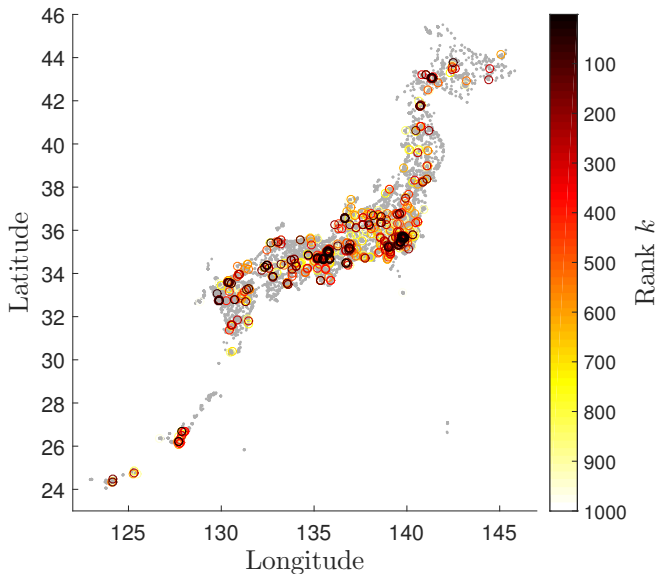


図 11: 入次数法ランキングの可視化結果

Mann-Whitney の 2 群順位統計量を多群に拡張して適用する方法について述べる。いま、地域 j に着目すれば、この地域に属すスポット集合 S_j と、それ以外のスポット集合 $S \setminus S_j$ の 2 群に分割することができる。ここで、 $\cdot \setminus \cdot$ は集合差を意味する。よって、Mann-Whitney の 2 群順位統計量に従い、次式により、地域 j に対し z -score z_j を求めることができる。

$$z_j = \frac{u_j - \mu_j}{\sigma_j}, \quad (16)$$

$$u_j = M_j(M - M_j) + \frac{M_j(M_j + 1)}{2} - \sum_{s \in S_j} r_s, \quad (17)$$

$$\mu_j = \frac{M_j(M - M_j)}{2}, \quad (18)$$

$$\sigma_j^2 = \frac{M_j(M - M_j)(M + 1)}{12}. \quad (19)$$

ただし、同順位が起こるケースでは、標準偏差 σ_j は標準的な方法で補正されるとする。よって、式 16 で求まる z -score z_j により、各地域 j がランキングの高い、または逆に低いスポットを有意に多く含むか定量的に評価することができる。この多群順位統計量は、基本的には 2 クラス分類器の SVM (Support Vector Machine) [21] を多クラス分類器に拡張するときに利用される one-against-all と類似した考え方となる。

まず、地域集合 J を八地方区分としたときの z -score z_j を図 13 に示す。北海道地方と九州地方は、人気度のスコアが他の地方よりも明らかに高いが、入次数のスコアは 0 に近く、PageRank のスコアに至っては負値となっている。よって、この 2 地方は、観光地としての人気はあるものの、あくまで一時的に滞在する地域であるということが推察される。それとは逆に、PageRank のスコアが正値となっている関東地方と近畿地方は、観光地というより、拠点地としての役割が強いことが分かる。一方、東北地方、中国地方、四国地方は、人気度のスコアが軒並み低く、入次数のスコアと PageRank のスコアは更に低くなっている。この 3 地方は、観光地として人気が低いだけでなく、拠点地としての役割も弱いことがうかがえる。また、中部地方は、どのスコアにおいても平均的であり、強いといえば入次数のスコアが比較的高いため、拠点と拠点を繋ぐ中継地点としての役割が強いように思える。

続いて、地域集合 J を都道府県区分としたときの、各ランキングにおける順位統計量の z -score 上位 10 都道府県を表 1, 2, 3 に示す。表 1 より、人気度ランキングにおいては、沖縄県が頭一つ抜けてスコアが高く、次いで北海道、東京都が高くなっているため、ここでの上位の地域は、観光地としての人気の高さがうかがえる。表 2 より、入次数法ランキングにおいては、東京都が 1 位に上がっており、京都府や愛知県も上位に入っているため、人の流れが多いと思われる地域の z -score が増加していることが見て取れる。表 3 より、PageRank 法ランキングにおいては、東京都が突出してスコアが高く、次いで神奈川県、京都府が高くなっており、大阪府も新たに出現していることから、ここでの上位の地域は、日本全国の観光スポットネットワークにおける優秀な拠点地であると言える。

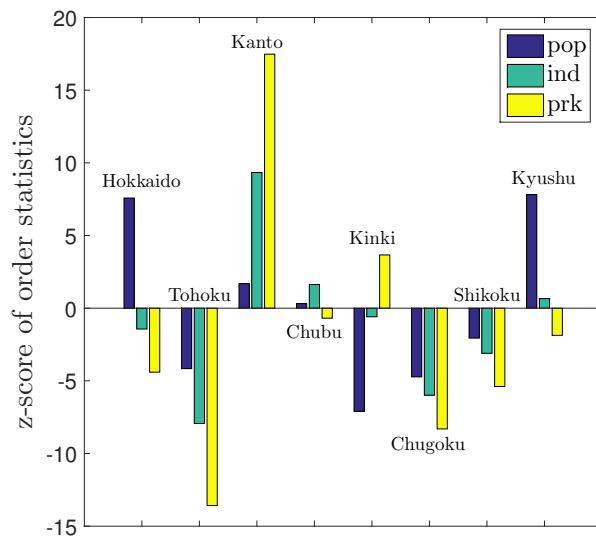


図 13: 八地方区分としたときの各ランキングの順位統計量 z-score

表 1: 人気度ランキングにおける順位統計量の z-score 上位 10 都道府県

Rank	z_j	Prefecture
1	14.477	Okinawa
2	7.5794	Hokkaido
3	5.3568	Tokyo
4	2.6809	Kanagawa
5	2.6258	Shizuoka
6	2.2201	Kagoshima
7	2.1930	Nagano
8	1.9683	Chiba
9	1.6848	Ishikawa
10	1.5340	Oita

6. おわりに

本論文では、観測されたデータを基に、ユーザー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 に基づいた確率モデルと、観測されたユーザー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、2 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ー行動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、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地域毎の特性を見出す指標として有用であることが示された。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。

表 2: 入次数法ランキングにおける順位統計量の z-score 上位 10 都道府県

Rank	z_j	Prefecture
1	12.605	Tokyo
2	6.5979	Okinawa
3	6.2236	Kanagawa
4	4.4648	Kyoto
5	3.0965	Nagano
6	2.9030	Shizuoka
7	2.2823	Aichi
8	2.1821	Yamanashi
9	1.7685	Chiba
10	1.0773	Ishikawa

表 3: PageRank 法ランキングにおける順位統計量の z-score 上位 10 都道府県

Rank	z_j	Prefecture
1	22.075	Tokyo
2	10.186	Kanagawa
3	9.4279	Kyoto
4	8.7675	Okinawa
5	3.1010	Chiba
6	2.8854	Shizuoka
7	2.6034	Osaka
8	2.0494	Nagano
9	1.9495	Yamanashi
10	1.8663	Ishikawa

【謝辞】

本研究は、JSPS 特別研究員奨励費 15K00311 の支援を受けて行ったものである。

【文献】

- [1] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, Vol. 439, pp. 462–465, 2006.
- [2] B. Jiang, J. Yin, and S. Zhao. Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, Vol. 80, p. 021136, 2009.
- [3] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, Vol. 453, pp. 779–782, 2008.
- [4] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Trans. Netw.*, Vol. 19, No. 3, pp. 630–643, 2011.
- [5] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2010.
- [6] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, 1989.

- [7] S. Wasserman and K. Faust. *Social network analysis*. Cambridge University Press, Cambridge, UK, 1994.
- [8] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, Vol. 18, No. 1, pp. 50–60, 03 1947.
- [9] C. Song, T. Koren, P. Wang, and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, Vol. 6, pp. 818–823, 2010.
- [10] P. Vansteenwegen, W. Souffriau, and D. V. Oudheusden. The orienteering problem: a survey. *European Journal of Operational Research*, Vol. 209, pp. 1–10, 2011.
- [11] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer-Verlag New York, Inc, New York, NY, USA, 2011.
- [12] J. Bao, Y. Zheng, D. Wilkie, and M. Mokbel. Recommendations in location-based social networks: a survey. *GeoInformatica*, Vol. 19, No. 3, pp. 525–565, 2015.
- [13] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, pp. 791–800, New York, NY, USA, 2009. ACM.
- [14] D. G. Luenberger. *Linear and Nonlinear Programming: Second Edition*. Kluwer Academic Publishers, 2003.
- [15] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, Vol. 45, pp. 167–256, 2003.
- [16] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.
- [17] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review*, Vol. 67, No. 5, p. 056104, 2003.
- [18] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol. 30, pp. 107–117, 1998.
- [19] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.
- [20] H. Moritz. Geodetic reference system 1980. *Journal of Geodesy*, Vol. 74, No. 1, pp. 128–133, 2000.
- [21] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

山岸 祐己 Yuki YAMAGISHI

静岡県立大学大学院経営情報イノベーション研究科博士課程在学中。データマイニングの研究に従事。人工知能学会，情報処理学会，日本データベース学会各学生会員。

斉藤 和巳 Kazumi SAITO

静岡県立大学経営情報学部教授。1985年慶応義塾大学理工学部数理科学科卒業，同年日本電信電話(株)入社。1998年東京大学博士(工学)。複雑ネットワーク，機械学習の研究に従事。情報処理学会，電子情報通信学会，人工知能学会，日本神経回路学会，日本応用数理学会，日本行動計量学会，日本データベース学会各会員。