

行動範囲と興味の同時推定モデルによる地域情報推薦

倉島 健^{1,a)} 岩田 具治^{2,b)} 星出 高秀^{1,c)} 高屋 典子^{1,d)} 藤村 考^{1,†1,e)}

受付日 2012年9月20日, 採録日 2013年1月6日

概要: ユーザの過去の行動履歴から次の行動を予測し推薦するためのジオトピックモデルを提案する. 提案モデルは, ユーザの行き先が, 1) 家, 勤務地, 学校などが存在するユーザの行動範囲からの近さと, 2) ユーザ自身の興味情報によって決まるという仮定に基づく確率的行動モデルである. ユーザの行動範囲が場所の選択に与える影響を明示的にモデリングすることで, 位置情報 (緯度, 経度) 以外の潜在的な場所特徴とそれに対するユーザの興味を効果的に学習することができる. 評価実験においては, ユーザの興味情報を推定することでユーザの行動を高精度に予測できることを, レストランとランドマークの訪問履歴を用いて示す. さらに, 人々のランドマーク選択を説明するためのユーザの興味対象が “アート”, “自然”, “建築物”, “知名度”, “眺めの良さ” などの潜在的なランドマーク特徴であることを明らかにした.

キーワード: 推薦, 行動モデリング, 情報抽出, トピックモデル

Joint Modeling of User's Activity Area and Interests for Local Recommendation

TAKESHI KURASHIMA^{1,a)} TOMOHARU IWATA^{2,b)} TAKAHIDE HOSHIDE^{1,c)}
NORIKO TAKAYA^{1,d)} KO FUJIMURA^{1,†1,e)}

Received: September 20, 2012, Accepted: January 6, 2013

Abstract: This paper proposes a method that analyzes the location log data of multiple users to recommend locations to be visited. The method uses our new topic model, called *Geo Topic Model*, that can jointly estimate both the user's interests and activity area hosting the user's home, office and other personal places. By explicitly modeling geographical features of locations and users, the user's interests in other features of locations, which we call latent topics, can be inferred effectively. Experiments are conducted using Flickr-based and Tabelog-based location logs to evaluate the recommendation performance of the proposed method in terms of the accuracy of predicting visit selections. We also show that our model can estimate latent features of locations such as art, nature, construction, popularity and great views from location logs of landmark visits, and describe each user's preference based on them.

Keywords: recommendation, behavior modeling, information extraction, probabilistic topic model

1. はじめに

近年, 写真共有サイト (Flickr [7], Google Picasa [8]), ソーシャルネットワークサービス (foursquare [1], Facebook [2]), 旅行サイト (TripAdvisor [3]), レストランレビューサイト (食べログ [4], yelp [5], ZAGAT.com [6]) などで, 多くの人々が自らの実世界での経験を発信するようになった. 情報推薦技術に代表される, 他人の経験を自らの行動決定に反映させるための技術は, 人間の賢い意志決定において欠かせないものとなっている.

本稿では, 行き先に関するユーザの意志決定プロセスを

¹ 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
NTT Service Evolution Laboratories, Yokosuka, Kanagawa
239-0847, Japan

² 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, Atsugi, Kanagawa
243-0198, Japan

^{†1} 現在, 大妻女子大学
Presently with Otsuma Women's University

a) kurashima.takeshi@lab.ntt.co.jp

b) iwata.tomoharu@lab.ntt.co.jp

c) hoshide.takahide@lab.ntt.co.jp

d) takaya.noriko@lab.ntt.co.jp

e) fujimura@otsuma.ac.jp

表現するジオトピックモデルに基づく場所推薦手法を提案する．ここで，ランドマークや建物，寺社，店舗など，人々の行動の目的地となる実世界の対象を“場所”と定義する．ジオトピックモデルは，ユーザの行き先は，1) 行動範囲からの近さと，2) 個人的な興味とに基づいて決定するという人間行動に関する仮定に基づく確率的行動モデルである．

我々は，2つの基準を設けて場所推薦手法を設計した．まず，ユーザにとって興味深い場所を推薦することは重要である．たとえば，ユーザがアートに興味を持っているのであれば，ミュージアムやアートギャラリーなどのアートに関連する場所が推薦されることが望ましい．近年，購買履歴や Web 閲覧履歴に対してトピックモデルを適用することにより，このような，興味に基づく推薦を実現しようとする研究がさかんである．トピックモデルは，各ユーザ（文書）はユーザ（文書）固有のトピック比率に基づいて潜在トピックを選択し，各潜在トピックは潜在トピック固有の選択確率によってアイテム（単語）を選択するとしたモデルである．実世界の行動履歴に対して適用することで，“アート”や“自然”などのユーザの興味対象が潜在トピックとして推定されることが期待できる．しかし，ユーザの行き先は，ユーザ自身の位置情報に大きく依存するという，実世界の行動履歴を扱う場合に特有な問題により，それが難しいことが分かった．ある場所を訪れるか決定するうえで，移動コストは考慮すべき重要な要素の1つであり，ユーザは物理的に近い場所ほど訪れやすい傾向がある．近い場所を選択する傾向が色濃く反映された行動履歴集合に対してトピックモデルを適用すると，位置が近い場所どうしをまとめるものとして潜在トピックが抽出され，同時に，各ユーザの行動範囲を表すものとして，ユーザ固有のトピック比率が学習されてしまう．つまり，ユーザの興味情報を既存のトピックモデルで推定することはできない．

次に，推薦リストと同時に，推薦のベースとなったユーザ自身の興味表現をユーザに提示することは重要である．最初に提示した推薦リストにユーザが満足しなかった場合，ユーザは推薦システムが自動推定した自身の興味表現を確認し修正する．ユーザフィードバックが得られれば，システムは推薦リストを再構築することができる．このようなユーザインタラクションを通じて，ユーザはより満足する推薦結果にたどり着くことができると考える．ユーザ*アイテム行列に基づく既存の協調フィルタリング技術においては，推薦理由の説明となる情報を付加的に提示することができないというデメリットがある．

本稿で提案するジオトピックモデルはこれらの基準を満たしている．行動履歴からユーザの興味情報を推定するためのキーとなるアイデアは，ユーザの行動範囲が場所の選択に与える影響を明示的にモデリングすることである．これにより，ユーザの興味対象を表現するものとして潜在トピックを学習することが可能となる．より直感的には，

ユーザ自身の行動範囲から近い場所の選択は，“アクセスのしやすさ”が理由で引き起こされたものとして説明し，逆に，ユーザ自身の行動範囲から離れた場所の選択は，“ユーザ固有の興味”によって引き起こされたものとして潜在トピックを用いて説明するモデルである．提案モデルに基づく推薦システムは，ユーザの行動範囲から近く，ユーザ自身の興味に合致した場所を推薦することができる．また，ユーザの現在地情報を行動範囲の代わりに用いて，ユーザ自身の興味に合致した周辺情報を推薦するといったことも可能である．さらに，各場所のメタデータ（テキストタグ）を用いて，潜在トピックに意味付け（タギング）を行うことで，“言語的解釈を与えた潜在トピックへの確率的重み”の形でユーザの興味を表現できる．ユーザは自身の興味情報を直感的に理解することができる．

評価実験において，ランドマークとレストランの訪問履歴を用い，行動予測の観点で提案法の有効性を定量的に示す．また，ユーザが過去に訪れたことのある地域（行動範囲近辺）に存在する場所を推薦する場合と，旅行先に存在する場所を推薦する場合とで手法の有効性を検証した．提案法が，場所に非依存な形式でユーザの興味情報を推定するため，ふだんの行動範囲内だけでなく，旅行先の行動も高精度に予測できることを示す．さらに，ユーザのランドマーク選択を説明するためのユーザの興味対象として，“アート”，“自然”，“建築物”，“知名度”，“眺めの良さ”などが抽出されたことを報告する．

本稿の構成を以下に示す．次章では関連研究を述べる．3章では問題定義，モデル，推薦手法について述べる．4章では提案法の有効性を実際に観測された行動履歴に基づいて検証した結果を，5章で結論を述べる．

2. 関連研究

2.1 情報推薦のための行動モデリング

協調フィルタリング技術は，類似ユーザの行動履歴に基づく（user-based collaborative filtering），もしくは類似アイテム集合に基づく（item-based collaborative filtering）行動推薦手法である．協調フィルタリングの一種である *latent factor models*，その中でも *matrix factorization* が高い行動予測精度を示すことが知られている [29]．*Matrix factorization* はユーザとアイテムをそれぞれ低次元ベクトル表現し，2つのベクトルの内積をとることでユーザがあるアイテムを好きか否かを決定する．しかし，*Matrix factorization* に代表される *latent factor model* の欠点は，低次元のベクトル表現の直感的な理解が人間には困難で，ユーザがアイテムを選択する理由が不明瞭なことである．

その一方，トピックモデルはアイテムの生成プロセス，本稿においては，場所の選択プロセスに関する，より直感的な理解と説明を可能とする．トピックモデルは，複数ユーザの行動履歴をもとに，アイテム（場所）に関する潜

在的な特徴（潜在トピック）を学習する。各潜在トピックは、該当する潜在的な特徴を持つアイテムを確率的に選択する箱のようなものである。それと同時に、ユーザがどのような特徴を好むかの興味情報を、潜在トピックに対する確率的な重みとして学習することができる。トピックモデルを用いて分析することで、“ユーザがある特徴に対して興味を持っており、その特徴を持つものとしてあるアイテム（場所）を選択した”といった、ユーザの行動決定プロセスの説明が可能となる。

最も一般的なトピックモデルは *Probabilistic Latent Semantic Analysis* (PLSA) [14], [15] と、*Latent Dirichlet Allocation* (LDA) [11] であり、情報検索、言語モデルからユーザの興味、行動モデリングまで、広範な分野で用いられている。たとえば、Das らは Google News のパーソナライゼーションのために、PLSA を含む複数の推薦モデルを線形結合する手法を提案した [12]。また、我々もユーザ興味を表現する潜在トピックの時系列的な分析を可能とする *Topic Tracking Model* を提案した [16]。Web 閲覧行動を PLSA を用いて分析した手法も存在する [17]。さらに、我々は、Flickr のジオタグ情報から生成した実世界の行動履歴をもとに PLSA を用いた場所推薦を提案した [18]。しかし、1 章で述べたように、位置的に近い場所どうして 1 つの潜在トピックが形成され、それに対する重みとして個人のプロフィールが学習されてしまう傾向にあり、ユーザの行動範囲に近い場所は推薦できても、ユーザの興味に基づく推薦は難しかった。本稿では、位置が近い場所どうしが選択されやすいという実空間の行動履歴を扱う場合に特有な問題に対処し、ユーザ興味に基づく推薦を実現する。

2.2 GPS 軌跡のマイニング

GPS 機能を搭載した携帯端末が一定間隔で取得した緯度、経度情報に基づき、ユーザの GPS 軌跡を分析する研究が存在する。その多くが、ユーザが次の時刻にどこに存在するかを予測する研究である。Ashbrook らは、マルコフモデルを用いて行動予測している [10]。Zheng らは、107 ユーザの GPS 軌跡をグラフマイニング手法を用いて分析し、主要な観光地や典型的なトラベルルートを抽出し提示する手法を示した [20], [21]。Leung らは、ユーザ行動-場所からなる三部グラフに基づく *CLM graph* を提案した [27]。“学校に行く”、“出社する”といったある特定の行動に起因し、時系列的に連続して観測される場所集合を行動情報と呼んでいる。これらの関連手法が、時空間的な連続性を扱うことで短期的な行動予測を得意とするのに対して、提案手法は個人的な興味を扱い、旅行先や週末の訪問先など、長期的な目的地予測を得意とする。

2.3 ローカル推薦システム

現在、個人に特化した旅行計画を推薦するシステムが数

多く存在する。Yahoo Travel は、空き時間と場所に対するレーティングを入力すると、行き先を推薦してくれるサービスである [9]。COMPASS はユーザプロフィール、スケジュール、ショッピングリスト、交通状況や天候を考慮して行き先を推薦する [24]。MobyRec はユーザの状況と興味情報とに基づいて推薦する [25]。これらの推薦システムは、最初に、検索クエリや興味情報をシステムに明示的に伝える必要があるが、我々は、ユーザの行動履歴のみから自動でユーザ興味を推定するシステムを実現する。

Bellotti らは、大規模なフィールドワークを東京都内で実施し、ユーザの行動（食べる、ショッピングなど）と興味に基づく推薦システムを提案した [23]。行動履歴からユーザの暗黙的な興味を推定する試みであるという点で我々の試みと類似しているが、彼らの推薦手法は、既存の推薦モデルをアドホックに組み合わせたものであり、新たな推薦モデルを提案したわけではない。

3. 提案モデル

問題定義、提案するジオトピックモデル、ジオトピックモデルに基づく推薦手法を本章で述べる。

3.1 ジオトピックモデル

N 人のユーザ集合を $U = \{u\}_{u=1}^N$ 、 I 個の推薦対象となる場所を $I = \{i\}_{i=1}^I$ と表現する。また、各場所は位置座標 r_i （緯度、経度情報）を持ち、各ユーザは行動履歴 $\mathbf{x}_u = \{x_{u1}, \dots, x_{uM_u}\}$ を持つ。ここで、 x_{um} はユーザ u が m 番目に訪れた場所であり、 $x_{um} \in I$ である。我々は、ユーザの行動履歴 $\mathbf{x}_u$ が与えられたもとで、ユーザが次に高確率で訪れる場所を推薦したい。本稿の中で用いる記号を表 1 にまとめた。

ユーザ行動をモデリングするために、我々は人間の行動に関する以下の仮定を設けた。

- (1) ユーザが訪れる場所はユーザ自身の行動範囲によって決まる。たとえば、勤務先が東京駅にある人は東京駅近辺の場所を訪れやすい。
- (2) ユーザが訪れる場所はユーザ個人の興味によって決まる。たとえば、アートに興味がある人は東京ドームよりも東京国立近代美術館を訪れやすい。

これらの仮定を場所生成プロセスに反映させる。各ユーザは、ユーザ自身の興味を表すユーザ依存の潜在トピック分布 $\theta_u = \{\theta_{uz}\}_{z=1}^Z$ を持つ。 $\sum_z \theta_{uz} = 1$, $\theta_{uz} \geq 0$ であり、 Z は潜在トピック数である。ユーザは θ_u に従ってある潜在トピックを選択した後、潜在トピック依存のパラメータ $\phi_z = \{\phi_{zi}\}_{i=1}^I$ と、過去に訪れた場所の位置座標集合 $\mathbf{R}_u = \{r_{x_{um}}\}_{m=1}^{M_u}$ とに基づいてある場所を選択する。ここで、 ϕ_{zi} は場所 i の潜在トピック z からの選ばれやすさを表すパラメータである。提案モデルのグラフィカルモデルを図 1 に示す。塗潰し円が観測変数を、中抜き円が潜在変

表 1 本稿で用いる記号
Table 1 Notation.

記号	説明
$\mathbf{U}$	ユーザ集合
u	ユーザ, $u \in \mathbf{U}$
$\mathbf{I}$	場所集合
i	場所, $i \in \mathbf{I}$
z	潜在トピック
M_u	ユーザ u の行動履歴に含まれる場所数
x_{um}	ユーザ u の m 番目の場所, $x_{um} \in \mathbf{I}$
$\mathbf{x}_u$	ユーザ u の行動履歴, $\mathbf{x}_u = \{x_{u1}, \dots, x_{uM_u}\}$
r_i	場所 i の位置座標 (緯度, 経度)
$\mathbf{R}_u$	ユーザ u の場所集合の位置座標集合, $\mathbf{R}_u = \{r_{x_{um}}\}_{m=1}^{M_u}$
θ_{uz}	ユーザ u の潜在トピック z の選択確率
$\boldsymbol{\theta}_u$	ユーザ u の潜在トピック選択確率集合, $\boldsymbol{\theta}_u = \{\theta_{uz}\}_{z=1}^Z$
ϕ_{zi}	場所 i が潜在トピック z から選択される確率
$\boldsymbol{\phi}_z$	潜在トピック z の場所選択確率集合, $\boldsymbol{\phi}_z = \{\phi_{zi}\}_{i=1}^I$
N	ユーザ数
I	場所数
Z	潜在トピック数
β	行動範囲を調整する bandwidth パラメータ

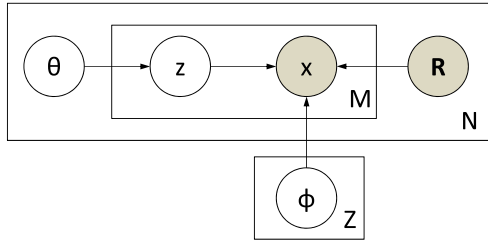


図 1 ジオトピックモデルのグラフィカル表現

Fig. 1 Graphic model representation of Geo Topic Model.

数を表している。

位置座標集合 $\mathbf{R}_u$ を持つユーザ u が場所 i を選択する確率は、

$$P(i|u, \mathbf{R}_u, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{z=1}^Z P(z|u, \boldsymbol{\Theta}) P(i|z, \mathbf{R}_u, \boldsymbol{\Phi}), \quad (1)$$

となる。なお、 $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_u\}_{u=1}^N$ 、 $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_z\}_{z=1}^Z$ 、 $P(z|u, \boldsymbol{\Theta}) = \theta_{uz}$ である。 $P(i|z, \mathbf{R}_u, \boldsymbol{\Phi})$ は、ユーザの位置座標集合 $\mathbf{R}_u$ を加味したうえで、場所 i が潜在トピック z から選択される確率である。我々は、過去に訪れた場所に地理的に近い他の場所もまた選ばれやすいと仮定している。 $P(i|z, \mathbf{R}_u, \boldsymbol{\Phi})$ は、

$$P(i|z, \mathbf{R}_u, \boldsymbol{\Phi}) = \frac{1}{C} \exp(\phi_{zi}) \sum_{r \in \mathbf{R}_u} \exp\left(-\frac{\beta}{2} \|r_i - r\|^2\right), \quad (2)$$

となる。なお、 $\|\cdot\|$ は位置座標空間におけるユークリッド距離、 $C = \sum_{i' \in \mathbf{I}} \exp(\phi_{zi'}) \sum_{r \in \mathbf{R}_u} \exp\left(-\frac{\beta}{2} \|r_{i'} - r\|^2\right)$ は正規化項である。 β は行動範囲の広さを決定する bandwidth パラメータであり、 β の値を小さく設定するほど、行動範囲をより広くとらえることができる。 $\exp(\phi_{zi})$ がユーザの興

味が場所の選択に及ぼす影響を決定する項である。第 1 項 ϕ_{zi} は、場所 i が潜在トピック z から選択される確率であり、PLSA などの既存のトピックモデルも同様のパラメータを含んでいる。もし、潜在トピック z がアートを表現するものとして推定されれば、場所 i として、ミュージアムや美術館などのアートに関する場所が選択されやすくなる。第 2 項 $\sum_{r \in \mathbf{R}_u} \exp\left(-\frac{\beta}{2} \|r_i - r\|^2\right)$ はユーザの行動範囲が場所の選択に及ぼす影響を表現しており、ユーザの行動履歴 $\mathbf{x}_u$ に含まれる場所の近くに存在する他の場所は選択されやすくなる。たとえば、ユーザがある駅の周辺に存在するレストランをよく利用する場合、その駅の周辺に存在する他のレストランも選択されやすくなる。場所 (とユーザ) に関する地理的な特徴を明示的にモデリングすることにより、地理的な特徴以外の潜在的特徴 (とそれに対するユーザの興味) を扱うものとして潜在トピックを推定する。より直感的に提案モデルを述べると、ユーザ自身の行動範囲から近い場所の選択はアクセスのしやすさが理由で引き起こされたものとして説明し、逆に、ユーザ自身の行動範囲から離れた場所の選択は、ユーザ固有の興味によって引き起こされたものとして潜在トピックを用いて説明する。

このように、提案モデルは既存のトピックモデル (PLSA) をベースにしている。協調フィルタリング手法ではなくトピックモデルベースで解析する 1 つの利点として、潜在トピックを用いた低次元表現により、本来であれば $N \times I$ の要素からなるユーザ * 場所行列を、より少ないパラメータ数で表現し保持できる点があげられる。提案手法は $N \times I$ の確率値をパラメータ数 $N(Z-1) + ZI$ ($Z \ll N$ かつ $Z \ll I$) で表現可能であり、これはパラメータ数 $N(Z-1) + Z(I-1)$ の PLSA と同じオーダーである。また、PLSA が多くの推薦システムに採用されていることをふまえると、計算量の面においても提案法は実用的であると考えられる。

3.2 パラメータ推定

最尤法を用いて提案モデルのパラメータを推定する。未知パラメータは潜在トピック選択確率集合 $\boldsymbol{\Theta}$ と、場所選択確率集合 $\boldsymbol{\Phi}$ であり、全未知パラメータ集合を $\boldsymbol{\Psi} = \{\boldsymbol{\Theta}, \boldsymbol{\Phi}\}$ で表現する。ユーザの行動履歴集合 $\mathbf{X} = \{\mathbf{x}_u\}_{u=1}^N$ が与えられたときのパラメータ集合 $\boldsymbol{\Psi}$ の対数尤度は

$$L(\boldsymbol{\Psi}|\mathbf{X}) = \sum_{u=1}^N \sum_{m=1}^{M_u} \log \sum_{z=1}^Z \theta_{uz} P(x_{um}|z, \mathbf{R}_u, \boldsymbol{\Phi}), \quad (3)$$

となる。事後確率を EM アルゴリズム [13] を用いて最大化することにより、未知パラメータ $\boldsymbol{\Psi}$ を推定する。事前確率を含む完全対数尤度は、

$$\begin{aligned} Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}) &= \sum_{u=1}^N \sum_{m=1}^{M_u} \sum_{z=1}^Z P(z|u, m; \hat{\boldsymbol{\Psi}}) \log \theta_{uz} P(x_{um}|z, \mathbf{R}_u, \boldsymbol{\Phi}), \end{aligned} \quad (4)$$

となる．ここで， $\hat{\Psi}$ は現在の推定値， $P(z|u, m; \hat{\Psi})$ は現在の推定値が与えられたときの，ユーザ u の m 番目の場所のトピック事後確率を表す．E ステップでは，ベイズ則に従い，トピック事後確率を計算する．

$$P(z|u, m; \hat{\Psi}) = \frac{P(z|u, \Theta)P(x_{um}|z, \mathbf{R}_u, \Phi)}{\sum_{z'=1}^Z P(z'|u, \Theta)P(x_{um}|z', \mathbf{R}_u, \Phi)}, \quad (5)$$

ここで， $P(x_{nm}|z, \mathbf{R}_u, \Phi)$ は式 (2) により計算する．M ステップでは， $\sum_{z=1}^Z \theta_{uz} = 1$ という制約のもと， $Q(\Psi|\hat{\Psi})$ を θ_{uz} に関して最大化することにより，トピック選択確率の推定値 $\hat{\theta}_{uz}$ を求める．

$$\hat{\theta}_{uz} = \frac{\sum_{m=1}^{M_u} P(z|u, m; \hat{\Psi})}{\sum_{z'=1}^Z \sum_{m=1}^{M_u} P(z'|u, m; \hat{\Psi})}, \quad (6)$$

ϕ_z の推定値は閉形式で書くことができない．そのため，準ニュートン法 [19] などの最適化法を用いて $Q(\Psi|\hat{\Psi})$ を最大化することにより推定する．準ニュートン法で必要となる $Q(\Psi|\hat{\Psi})$ の ϕ_z に関する勾配ベクトルは，

$$\frac{\partial Q}{\partial \phi_z} = \sum_{u=1}^N \sum_{m=1}^{M_u} P(z|u, m; \hat{\Psi}) - \sum_{u=1}^N \sum_{m=1}^{M_u} P(z|u, m; \hat{\Psi})P(x_{um}|z, \mathbf{R}_u, \Phi), \quad (7)$$

となる．収束するまで E ステップと M ステップを交互に繰り返すことにより， Ψ の最適解を推定することができる．なお，潜在トピック数と行動範囲の bandwidth パラメータ β は交差検定を用いて推定する．

3.3 モバイルユーザのための推薦

3.1 節で，ユーザの場所選択プロセスを説明するジオトピックモデルについて述べた．提案モデルは，ユーザが過去に訪れたことのある地域近辺で，次に訪れる可能性が高い場所を推定し，推薦することができる．ここでは，その応用として，モバイル環境におかれたユーザを想定し，現在地に基づく場所推薦手法を示す．携帯電話などのモバイル端末の多くは GPS 機能を搭載しており，現在地を検索要求の一部として用いることが比較的容易になってきている．特に，旅行やビジネス出張など，精通していない地域を訪問した際に，現在地周辺の情報を推薦してほしい状況が起こる．このようなユーザに対して，1) 現在地からすぐにアクセスでき，2) 個人の興味に合致した場所を推薦したい．前者は，ジオトピックモデルの行動範囲情報の代わりに現在地情報を用いることで，後者は，ジオトピックモデルで推定した Φ と Θ を利用することで実現できる．

ユーザの現在地の緯度，経度情報を c とする．ユーザの過去の行動履歴 $\mathbf{x}_u$ と現在地 c が与えられたときに，ユーザが訪れそうな場所を推定し推薦することがタスクである． c にいるユーザ u が場所 i に訪れる確率値を

$$P(i|u, c, \Theta, \Phi) = \sum_{z=1}^Z P(z|u, \Theta)P(i|z, c, \Phi), \quad (8)$$

で求める．潜在トピック z と現在地 c が与えられたときの場所 i の選択確率は，

$$P(i|z, c, \Phi) = \frac{\exp(\phi_{zi}) \exp\left(-\frac{\beta}{2}\|r_i - c\|^2\right)}{\sum_{i' \in I} \exp(\phi_{zi'}) \exp\left(-\frac{\beta}{2}\|r_{i'} - c\|^2\right)}, \quad (9)$$

である．ある場所の選択確率は，現在地からのユークリッド距離（緯度，経度座標空間における距離）に依存して決まる．

3.1 節で述べたジオトピックモデルが扱う潜在トピックは，アートや建築物といった，位置とは独立した興味の対象を表現するものとして推定できる．したがって，1 度，ユーザ依存の潜在トピック出現確率 θ_u （と場所出現確率 Φ ）を行動履歴から学習すれば，初めて訪れた地域に関する場所を，位置とは独立した形式で保持したユーザ興味をもとに推薦することが可能である．たとえば，居住地である東京で，アートに関する場所をよく訪問するユーザに対して，旅行先である京都の京都国立近代美術館や京都市美術館を推薦することができる．また，確率値 $P(i|u, c, \Theta, \Phi)$ は，式 (8) と式 (9) を用いてパラメータ推定なしに求めることができるため計算コストが低い．現在地に応じてリアルタイムに推薦リストを生成できるため，モバイル推薦システムにおいて重要な計算コストの問題をクリアしている．

4. 評価実験

本章では，ランドマークとレストランの訪問履歴を用いて提案法の有効性を定量的に示す．また，ユーザが過去に訪れたことのある地域に存在する場所を推薦する場合（4.1 節）と，旅行で初めて訪れる地域に存在する場所を推薦する場合（4.2 節）とで提案法の有効性を示す．最後に，場所に付与されたメタデータを用いて提案法が推定する潜在トピックに対して意味付けを行い，既存のトピックモデルと異なる観点で場所の特徴（とそれに対するユーザの興味情報）を推定可能なことを示す（4.3 節）．

4.1 実験 1：行動範囲での推薦

4.1.1 データセット

食べログに投稿されたレビュー情報をもとに 4 種類の行動履歴データを作成した．東京都，大阪府，東日本（東京都，神奈川県，埼玉県），西日本（大阪府，京都府，兵庫県）に存在する店舗の訪問履歴（レビュー履歴）である．同時に，各地域に存在する店舗の位置座標も収集した．また，写真共有サイト Flickr に投稿された写真に付与されたジオタグ（緯度，経度）情報を Flickr API を利用して収集した．ユーザ単位にジオタグ情報を集約することで，個人

表 2 実験 1 に用いた行動履歴

Table 2 Location log data used in the first experiment.

地域	ユーザ数	場所数	行動履歴の平均長
東京	6,257	910	14.872
大阪	2,304	763	15.995
東日本	7,326	1,067	14.871
西日本	3,531	956	15.245
ニューヨーク	3,363	967	9.778
サンフランシスコ	4,433	1,030	12.231
ロサンジェルス	2,848	943	11.686

の行動履歴を再現することができる。収集した Flickr データから、ニューヨーク、サンフランシスコ、そしてロサンジェルスに関する 3 種類の行動履歴データを作成した。最初に、Flickr API を利用して各都市の中心部から 50 km 以内で撮影されたジオタグ情報を取得した。次に、得られたジオタグ集合から多くの人々が撮影したランドマークを抽出する。ランドマークとは、観光地、店舗、建物、橋、アウトレットなどである。ランドマーク抽出問題に関して、すでに、Crandall らが Mean-shift 法の有効性を示している [26]。そのため、我々も Mean-shift 法でランドマーク抽出を行った。Mean-shift 法は、ランドマーク（の緯度、経度情報）集合と、観測された各ジオタグ情報がどのランドマークに属するかの推定値を出力する。この出力と、ユーザ単位に集約したジオタグ情報をもとに、各ユーザのランドマーク訪問履歴を生成した。実験に用いる計 7 種類の行動履歴について表 2 にまとめた。なお、観測データの少ないユーザ（訪問場所数が 5 未満のユーザ）は削除している。また、推薦対象となる場所数が各データセットで均一となるように、訪問ユーザ数が 100, 40, 100, 50, 10 未満の場所を東京、大阪、東日本、西日本、Flickr データセットからそれぞれ削除した。

4.1.2 モデル比較

推薦研究の最も一般的な評価指標は予測精度であるとサーベイ論文 [22] に述べられているため、行動予測精度で提案モデルの妥当性を示す。我々は提案モデルを以下の 4 つの確率モデルと比較した。

多項分布モデル： 人気度に基づく推薦。人気度は場所に関する多項分布 $P(i)$ で計算する。

カーネルモデル： ユーザの行動範囲に基づく推薦。過去に訪れた場所の近辺に存在する他の場所を推薦する。

$$P_K(i|u) = \frac{\sum_{r \in R_u} \exp\left(-\frac{\beta}{2}\|r_i - r\|^2\right)}{\sum_{i' \in I} \sum_{r \in R_u} \exp\left(-\frac{\beta}{2}\|r_{i'} - r\|^2\right)}.$$
(10)

トピックモデル： PLSA が推定したユーザの興味情報に基づく推薦。

$$P_T(i|u) = \sum_{z \in Z} P(z|u)P(i|z).$$
(11)

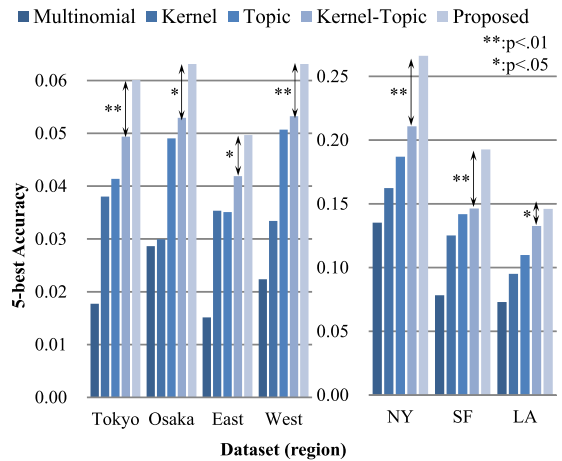


図 2 ユーザの行動範囲近辺での行動予測精度（5-best prediction accuracy）の比較。X 軸にデータセットをプロット

Fig. 2 Comparisons of 5-best prediction accuracy of daily-life behavior. X-axis plots the datasets.

カーネルトピックモデル： ユーザの行動範囲と（PLSA が推定した）興味情報とに基づく推薦。カーネルモデルとトピックモデルの線形結合、

$$P_{KT}(i|u) = \alpha P_K(i|u) + (1 - \alpha) P_T(i|u), \quad (12)$$

で計算する。 α は 2 つのモデルの影響をコントロールするパラメータである。

各ユーザが最後に訪れた場所を予測する。最初に、全ユーザの行動履歴から最後に訪れた場所を除いたデータを学習データとして用いてモデルを学習する。その後、各ユーザが最後に訪れた場所の予測テストを、全ユーザに対して行う。したがって、テスト回数は表 2 に示したユーザ数と等しくなる。評価指標は 5-best prediction accuracy を用いた。各ユーザに対して 5 つの場所を出力し、正解データが含まれた場合に 1、含まれない場合に 0 とする指標である。最終的に、1 を出力したユーザの全テストユーザに対する割合で各モデルを比較する。その比較結果を図 2 に示す。X 軸はデータセット、Y 軸は 5-best prediction accuracy をプロットしている。すべてのデータセットにおいて提案法が最も高精度であり、符号検定の結果、その差が有意であることが認められた（両側検定：p 値 < 0.05）。提案法とカーネルトピックモデルは、ユーザの行動範囲と興味情報とに基づく推薦を行うモデルである。それぞれ、潜在トピックとして扱う情報が異なることが、予測精度における優位性につながったといえる。トピックモデルとジオトピックモデルがそれぞれ推定する潜在トピックについての定性的な考察は実験 3 の中で行う。

なお、各手法のベストパラメータを 5 分割交差検定で決定した。提案法は潜在トピック数 Z と bandwidth パラメータ β 、カーネルモデルは bandwidth パラメータ β 、トピックモデルは潜在トピック数 Z 、カーネルトピックモデルは潜在トピック数 Z 、bandwidth パラメータ β と線形結

表 3 実験 1 で用いた各手法のパラメータ. β は $10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8$ の中から, Z は 3 から 50 の中から選択

Table 3 Best parameter values as identified in the first experiment. β was chosen from $10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7$ and 10^8 . Z was set in the range of 3 to 50.

モデル	地域	β	Z	α
カーネルトピックモデル	東京	10^5	22	0.2
	大阪	10^6	13	0.1
	東日本	10^5	15	0.3
	西日本	10^6	19	0.1
	ニューヨーク	10^2	5	0.9
	サンフランシスコ	10^3	20	0.5
	ロサンジェルス	10^2	16	0.7
提案モデル	東京	10^4	8	-
	大阪	10^4	7	-
	東日本	10^3	9	-
	西日本	10^3	10	-
	ニューヨーク	10^3	9	-
	サンフランシスコ	10^2	7	-
	ロサンジェルス	10^3	6	-

合パラメータ α を持つ. 提案法と他の比較手法とに共通するパラメータとして Z と β があるが, 各手法でその最適な値は異なるを考える. たとえば, 提案法の Z は人々の興味対象の数を表現するために用意された変数である. 既存のトピックモデルの Z も同様の目的のために用意された変数だが, 実際は, 位置が近い場所どうしのまとまりの数を表現するものとして用いられる. つまり, トピックモデルは, カーネルモデルと同様, ユーザの行動範囲に基づいて場所を推薦する. その組合せであるカーネルトピックモデルは, 双方のモデルがユーザの行動範囲を補完的に説明するように, 最適な Z と β と α の組合せを求めている. このように, 各手法においてパラメータの用いられ方が異なり, 最適な値も異なるため, 各手法にとってのベストパラメータを求めた後に精度比較している. 各手法に関して得られたベストパラメータを表 3 にまとめた. β に大きな値を設定するほど, 推定するユーザの行動範囲が狭くなる. 東京都 (大阪府) の行動履歴に関して推定した β が, より広い範囲である東日本 (西日本) よりも大きな値をとるのはそのためである. また, 領域が広いほど, より多くの種類の場所が存在するようになるため, 東京都 (大阪府) よりも東日本 (西日本) のほうが, 潜在トピック数が大きな値となる.

4.2 実験 2: 旅行先での推薦

提案法は, 位置に非依存な形式でユーザ興味を推定可能なため, 特に, 初めて訪れた地域に関する情報を高精度に推薦できるという特徴がある. 本節では, 3.3 節で述べたモバイルユーザのための推薦手法を用いて旅行先で周辺情報を推薦するという状況における提案手法の有効性を示す. 本実験に用いたのは食べログのレビューデータから生成し

表 4 実験 2 で用いた行動履歴

Table 4 Location log data used in the second experiment.

居住地-旅行先	場所数 (居住地/旅行先)	ユーザ数	行動履歴の平均長
東日本-西日本	1,424 (1,103/321)	1,558	24.335
西日本-東日本	1,424 (321/1,103)	820	19.162

た行動履歴である. 最初に, 食べログユーザ自身が記述したユーザプロフィールに基づき東日本に居住しているユーザを収集し, 各ユーザの東日本での行動履歴と, 旅行先である西日本での行動履歴を収集する. ただし, 居住地, 旅行先の行動履歴がともに存在するユーザに限定した, つまり, 各ユーザの行動履歴には, 必ず東日本, 西日本, 両方の店舗が含まれている. また, その逆のデータセット (居住地が西日本, 旅行先が東日本) も生成した. 本実験では, 居住地域の行動履歴に基づき, 旅行先での店舗選択を予測する. なお, 多くの人々にとっての旅のスタート地点であるという理由から, 旅行先が西日本の場合は現在地を大阪駅に, 東日本の場合は東京駅を各ユーザの現在地と見なした. 実験 2 で用いる行動履歴の詳細を表 4 に示す. なお, 観測データが少ない, 訪問数が 5 未満のユーザ, 訪問ユーザ数が 100 未満の場所は各データセットから削除している.

我々は, ユーザの現在地情報をモデルに組み込むことができるカーネルモデル, カーネルトピックモデル, そして提案法を比較した. カーネルモデルは現在地から近い場所から順に推薦する. カーネルトピックモデルは, 現在地からの近さに加え, PLSA で推定したユーザ興味に基づいて推薦する. 実験 1 と同様, 5-best prediction accuracy を評価指標として用いた. ただし, 今回の実験は Leave-one-out 交差検定を用いて行った. つまり, あるユーザの行動履歴だけを抜き出してテストデータとし, 残りを学習データとして用いる. これを全ユーザが 1 回ずつテストされるように繰り返した. つまり, 表 4 に示したユーザ数と同じ回数のテストを行う. 5-best prediction accuracy を全テストユーザに関して平均した結果を図 3 に示す. すべてのデータセットにおいて提案法が最も高精度であり, 符号検定の結果, その差が有意であることが認められた (両側検定: p 値 < 0.01). この結果は, 潜在トピック (とそれに対するユーザ興味) が位置に非依存な情報として推定できるため, 初めて訪れた地域, 旅行先に関する情報を提案手法が高精度に推薦可能なことを示すものである.

数時間, 数十分で今すぐに行ける場所を探したいときなど, 即時的な推薦をしてほしい状況においては, ユーザの興味よりも現在地からの近さが支配的に行動決定に影響してくる考える. 今回の実験においては, 東京駅 (大阪駅) にいるユーザに対する周辺情報として, 広範囲な東日本 (西日本) に存在する店舗を推薦する状況を想定して行った. つまり, 半日, 1 日といったある程度の移動時間を許

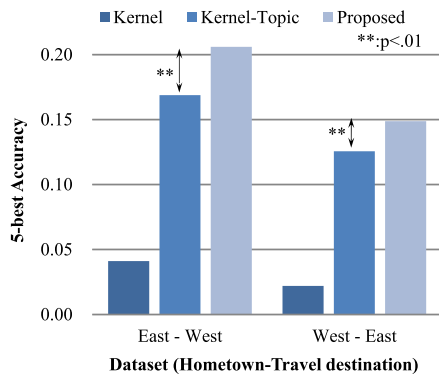


図 3 旅行先での行動予測精度 (5-best prediction accuracy) の比較. X 軸にデータセットを, Y 軸に全テストを通じた 5-best prediction accuracy の平均値をプロット

Fig. 3 Comparisons of 5-best prediction accuracy of travel behavior. The X-axis plots datasets and the Y-axis plots the 5-best prediction accuracy averaged across all runs.

表 5 実験 2 で用いた各手法のパラメータ. β は 10, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 , 10^8 の中から, Z は 5 から 50 まで 5 刻みの中から, α は 0.1 から 0.9 まで 0.1 刻みの中から選択

Table 5 Best parameter values in the second experiment. β was chosen from 10, 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , 10^7 and 10^8 . Z was set from 5 to 50 with interval of 5. α was chosen from 0.1 to 0.9 with interval of 0.1.

モデル	居住地-目的地	β	Z	α
カーネルトピックモデル	東日本-西日本	10^2	5	0.9
カーネルトピックモデル	西日本-東日本	10^2	10	0.7
提案モデル	東日本-西日本	10^3	10	-
提案モデル	西日本-東日本	10^3	10	-

容し, 個人の興味が行動決定に与える影響が高い状況の中で, 提案手法が推定した興味が目的地予測に寄与することを示したといえる. 提案手法の β は推薦してほしい店舗の地理的な範囲に関するパラメータであり, ユーザの空き時間, 考慮する範囲に応じて最適に設定する必要がある. なお, 今回の実験では, 各手法のベストパラメータを 5 分割交差検定で決定した. 各手法ごとに求めたベストパラメータを表 5 にまとめた. 提案法については, Z と β のすべての組合せ, カーネルトピックモデルは, Z と β と α のすべての組合せの中から最適なパラメータを選択した.

4.3 実験 3: トピック抽出

本節では, 既存のトピックモデルである PLSA との比較により, 提案モデルで推定可能な潜在トピックの特徴を定性的に示す. まず, 提案法, PLSA がそれぞれ推定した潜在トピックにおいて選択確率が高かった場所を地図上にマッピングした結果を図 4 に示す. データセットは Flickr のジオタグ情報に基づくニューヨークとサンフランシスコのランドマーク訪問履歴である (実験 1 で用いたものと同じ). また, アイコンの色が潜在トピックの種類を表し, 各潜在トピックから選択確率が高かった上位 30 件のランドマ

クをマッピングしている. 提案法と比較し, 同色のアイコンが近い位置に配置されていることから, 既存のトピックモデル (PLSA) は位置が近い場所をまとめるものとして潜在トピックが推定されていることが分かる. 一方で, 提案法は, 同色のアイコンが比較的離れた位置に配置されている. 位置とは異なる何らかの場所特徴を扱うものとして潜在トピックが推定できていることが分かる. 次に, 地図上にマッピングされた 30 件のランドマークについて位置座標空間における平均地点を求め, 各ランドマークと平均地点との分散を求めた結果を図 5 に示す. ニューヨーク, サンフランシスコの各エリアについて, PLSA が提案法よりも分散の平均値が小さいことから, PLSA が地理的な性質が近い場所をまとめるものとして潜在トピックを推定していることが定量的にも分かる.

それでは, 提案法の潜在トピックはどのような場所特徴を扱っているのだろうか. この疑問に答えるために, Flickr にアップロードされた写真に付与されたテキストタグ情報を用いて潜在トピックに意味付けを行った. テキストタグは, 地名, 被写体, 行動内容など, 様々な観点でユーザによって付与されたものであるが, この中から各潜在トピックを代表するタグ集合を自動で取得する.

代表タグ抽出方法は以下のとおりである. タグを g , タグ集合を G とする. 各潜在トピックに特徴的に出現するタグを取得するために, 潜在トピック z が与えられたときのタグ g の $lift$ 値を,

$$Lift_z(g) = \frac{P(g|z)}{\sum_{z' \in Z} P(g|z')P(z')}, \quad (13)$$

で計算する. $Lift$ は相関ルール抽出や頻出パターン抽出において用いられる指標であり, $P(g|z)$ と, 全トピックにおけるタグ g の出現確率との比である [28]. 全トピックに共通して出現する “nyc” や “sf” などのタグに対して低い値を与え, 潜在トピック z に相対的に多く出現するタグ g に対して高い値を与える. 潜在トピック z が与えられた場合のタグ g の出現確率 $P(g|z)$ は,

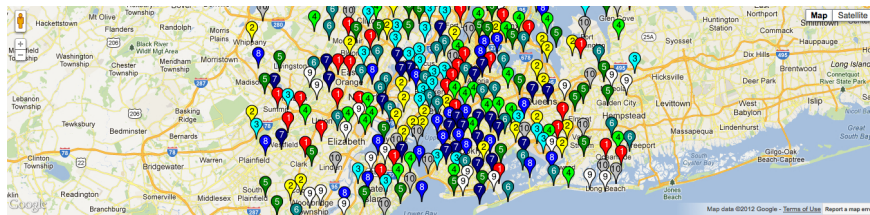
$$P(g|z) = \frac{1}{C} \sum_{i \in I} \exp(\phi_{zi}) P(g|i), \quad (14)$$

で求めることができる. $C = \sum_{i' \in I} \exp(\phi_{zi'})$ は正規化項である. $P(g|i)$ はランドマーク i が与えられた場合のタグ g の出現確率であり,

$$P(g|i) = \frac{H(g,i)}{H(i)}, \quad (15)$$

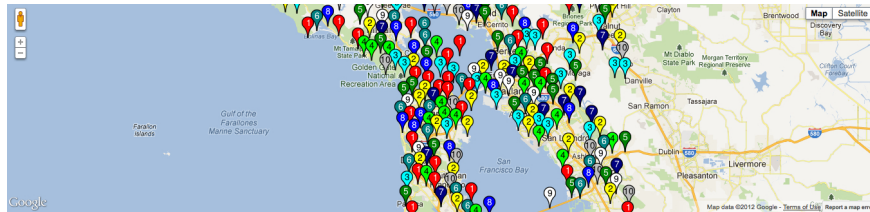
で求める. $H(i)$ はランドマーク i を訪問したユーザ数であり, $H(g,i)$ はランドマーク i にタグ g を付与したユーザ数である.

提案法, PLSA の各潜在トピックをテキストでラベリングした結果を, 図 4 中の各地図の下に示す. 各モデルの $P(z) = \sum_{u \in U} P(z|u)P(u)$ の降順で潜在トピック 1 から



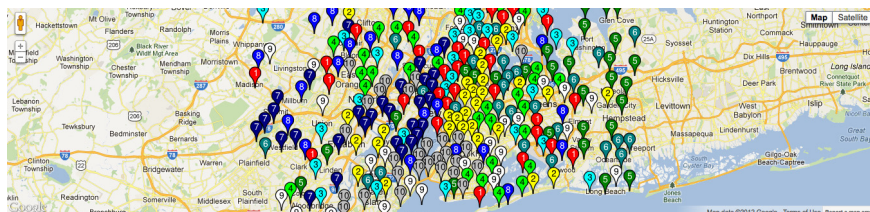
topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
centralpark	ferry	construction	art	yankeestadium	cityscape	garden	livemusic	beach	spring
guggenheim	bird	architecture	bar	coneyisland	liberty	flowers	graffiti	zoo	bike
met	sunset	chelsea	streetart	baseball	batterypark	prospectpark	music	vacation	bicycle
fifthavenue	lake	show	soho	summer	panorama	green	party	animals	color
rockefeller	boat	urban	dumbo	mets	skyline	nature	waterfront	garden	library

Geo Topic Model - New York City



topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
goldengate	woods	ship	store	landscape	graffiti	concert	sausalito	bike	coast
palaceoffinearts	goldengatepark	alcatraz	hotel	marin	art	live	urban	bicycle	sailing
downtown	trees	sunrise	house	sunrise	color	fireworks	cafe	bus	summer
presidio	muirwoods	sea	dinner	sea	mural	show	bar	ship	bakerbeach
chinatown	park	bridge	marketstreet	coast	haightashbury	livemusic	restaurant	birds	bike

Geo Topic Model - San Francisco



topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
met	parkslope	greenwichvillage	newark	coneyisland	yankees	jerseycity	libertyisland	harlem	jerseycity
guggenheim	prospectpark	soho	ice	neon	mets	airport	ellisland	upperwestside	bridge
columbuscircle	astoria	eastvillage	jersey	beach	bronx	statenisland	statueofliberty	library	statenisland
centralpark	williamsburg	jfk	nature	island	queens	ferry	statenisland	broadway	astoria
fifthavenue	eastvillage	lowereastside	tree	parade	castriver	newark	batterypark	centralpark	trees

PLSA - New York City



topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
prison	palaceoffinearts	alameda	giants	goldengatepark	airport	muirwoods	oceanbeach	marinheadlands	mission
alcatraz	presidio	oakland	embarcadero	castro	sfo	sausalito	pacificocean	coast	mural
island	marina	eastbay	soma	museum	berkeley	marincounty	museum	marin	hill
pier39	fishermanswharf	berkeley	baseball	oceanbeach	baybridge	marin	soma	marincounty	house
presidio	pier39	train	baybridge	garden	baseball	marinheadlands	alameda	pacificocean	bart

PLSA - San Francisco

図 4 提案モデルと PLSA が推定した潜在トピックの代表タグ抽出結果. β の値は実験 1 と同じ, 潜在トピック数は 10 とした. 潜在トピックごとにタグを $Lift_z(g)$ でソート

Fig. 4 The top five representative tags for topics estimated by our model and PLSA. Bandwidth parameter β of proposed model is set to the best parameter as estimated in the first experiment. The number of topics is 10. For each topic, z , tags are ordered according to $Lift_z(g)$.

表 6 提案モデルが推定した潜在トピックの代表タグ抽出結果. 食べログデータの東京に関する行動履歴を利用. β の値は実験 1 と同じ, 潜在トピック数は 10 とした

Table 6 The top five representative tags for topics estimated by our model. We use the location logs of Tokyo. The number of topics is 10, and bandwidth parameter β of proposed model is set to the best parameter as estimated in the first experiment.

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
ラーメン	そば	カフェ	イタリアン	フレンチ	焼き肉	ダイニングバー	ハンバーガー	インドカレー	魚介料理
つけ麺	うなぎ	洋菓子	フレンチ	うなぎ	もつ鍋	焼酎	サンドイッチ	インド料理	スペイン料理
餃子	うどん	チョコレート	ワインバー	天ぷら	焼き鳥	日本酒	ステーキ	カレー	ベトナム料理
担々麺	もつ鍋	パン	パスタ	寿司	鳥料理	居酒屋	ピザ	タイ料理	洋食
中華麺	天ぷら	ベーグル	ピザ	そば	串焼き	ワインバー	洋食	タイカレー	タイ料理

表 7 PLSA が推定した潜在トピックの代表タグ抽出結果. 食べログデータの東京に関する行動履歴を利用. 潜在トピック数は 10 とした

Table 7 The top five representative tags for topics estimated by PLSA. We use the location logs of Tokyo. The number of topics is 10.

topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10
カフェ	ラーメン	フレンチ	ラーメン	焼き肉	カフェ	ラーメン	ラーメン	そば	フレンチ
洋菓子	冷麺	イタリアン	洋食	イタリアン	洋菓子	イタリアン	冷麺	うどん	洋食
パン	インドカレー	寿司	冷麺	フレンチ	イタリアン	フレンチ	うどん	カレー	カフェ
チョコレート	インド料理	ワインバー	カレー	中華料理	ハンバーガー	冷麺	餃子	洋食	鳥料理
ベーグル	カレー	中華料理	餃子	カフェ	フレンチ	ピザ	インドカレー	カフェ	ハンバーガー

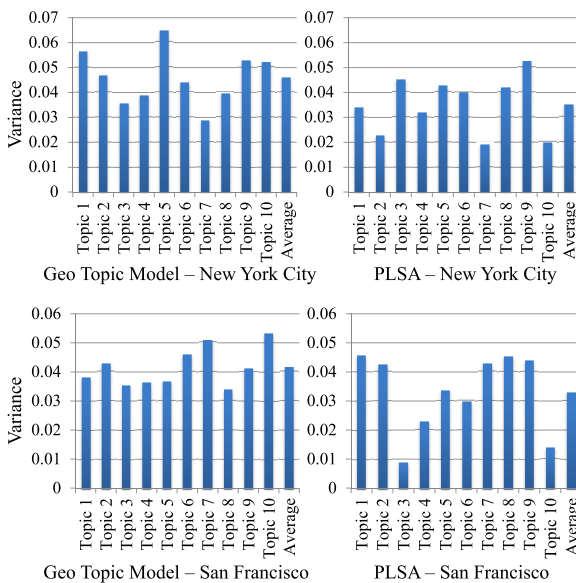


図 5 提案モデルと PLSA が推定した潜在トピックに含まれる場所集合の地理的な分散

Fig. 5 Variances of locations contained in each topic estimated by our model and PLSA.

10 とした. 2つの都市において, 提案法が抽出した最も影響力の大きい潜在トピックは有名な観光地をまとめるものであった (トピック 1). このトピックに含まれる場所集合は, 観光客にとって興味深いものである. トピック 2 から 10 は, 自然, アート, エンターテインメント, スポーツ, 眺めの良さなどの特徴を扱い, これらの特徴を持つ場所をまとめる役割を果たしていることが分かる. 一方で, PLSA

は地名を多く含んでおり, 場所の位置的特徴しか扱えていないことが分かる.

このように, 提案手法は場所の潜在的な特徴を潜在トピックとして抽出し, 潜在トピックに対して言語的な意味解釈を与えることができる (図 4). さらに, 各潜在トピックは, 自動で推定した θ_u に基づき個人ごとに並べ替え, 推薦する場所情報とともに, ユーザに提示することもできる. 提示情報はユーザインタラクションを発生させ, 推薦の満足度を全体として向上させることが期待できる. たとえば, 最初に推薦された場所情報にユーザが満足しなかった場合, ユーザはシステムが自動で推定した自身のユーザプロフィールを確認し, 明示的に自らの興味に合致した特定の潜在トピックを選択する. 選択された潜在トピック z 依存の場所選択確率 ϕ_z に基づき, システムが推薦リストを再提示するため, ユーザは最終的に満足のいく推薦結果をインタラクティブに引き出すことができる.

Flickr でつけられているタグには地名が含まれやすいため, 特に, PLSA の潜在トピックのラベリング結果に出現しやすい傾向にあった. そこで, 食べログデータの料理カテゴリに関するタグを用いた実験により, 地理的性質とは独立した場所の特徴, ここでは料理カテゴリに近い店舗をまとめるものとして潜在トピックが推定できることを示す. 具体的には, 食べログデータから生成した東京都に関する店舗訪問履歴から, 各潜在トピックを代表するタグ集合を取得した. 提案法, PLSA の各潜在トピックのラベリング結果を表 6 と表 7 に示す. タグのスコアリング方法

はランドマーク訪問履歴のものと同様である。表 7 に示すように、PLSA に関する各潜在トピックのラベリング結果は類似している。たとえば、トピック 2, 4, 7, 8 は上位にタグ“ラーメン”を含んでいる。その一方、表 6 に示すように、提案法は味覚、食材、雰囲気などを表現するものとしてトピックが推定できている。たとえば、トピック 3 は甘い味付けが好みの人々の、トピック 1 や 9 は塩辛い味付けが好みの人々の興味対象である。トピック 4, 5, 7 は雰囲気が良い店舗を含むトピックである。トピック 2 や 6 は日本に固有の料理を提供する店舗が含まれる傾向にある一方で、トピック 8 や 10 は他の国の料理を提供する店舗が含まれる。

5. 結論

ユーザの行動範囲と興味情報を同時推定するジオトピックモデルを提案した。地理的に記述できるユーザの行動特徴を明示的にモデリングすることで、地理的な特徴以外の潜在的な場所特徴（とそれに対するユーザの興味）を扱うものとして潜在トピックを推定できる。提案モデルは、ユーザの行動範囲から近く、ユーザの興味に合致する場所を推薦することもできれば、ユーザの現在地情報を行動範囲の代わりに用いることで、ユーザ自身の興味に合致した周辺情報を推薦することも可能である。

提案するジオトピックモデルに基づく推薦手法は、ユーザの興味に合致した場所を推薦できる点、推定した興味情報をユーザにカスタマイズしてもらうために提示できる点で実用的な推薦システムが満たすべき要件を満たしている。評価実験においては、ユーザの興味を効果的に推定することで、他の確率モデルと比較して提案モデルが予測精度の観点で優れていることを示した。また、人々のランドマーク選択を説明するための潜在的なランドマーク特徴を明らかにした。

本稿では、ランドマーク推薦とレストラン推薦という主要な推薦サービスを念頭に提案モデルの妥当性を評価した。今後は、他の推薦対象、データセットへの適用可能性も検討したい。また、今回のような過去の行動履歴を用いたオフラインの評価も重要であるが、実フィールドでの検証実験を通して提案法の真の有効性を検証していきたい。

参考文献

- [1] foursquare, available from <http://foursquare.com/>.
- [2] Facebook, available from <http://www.facebook.com/>.
- [3] TripAdvisor, available from <http://www.tripadvisor.com/>.
- [4] Tabelog, available from <http://tabelog.com/>.
- [5] yelp, available from <http://www.yelp.com/>.
- [6] ZAGAT.com, available from <http://www.zagat.com/>.
- [7] Flickr, available from <http://www.Flickr.com/>.
- [8] Google Picasa, available from <http://picasa.google.com/>.

- [9] Yahoo! TRAVEL, available from <http://www.travel.yahoo.com/>.
- [10] 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).
- [11] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
- [12] Das, A.S., Datar, M., Garg, A. and Rajaram, S.: Google news personalization: Scalable online collaborative filtering, *WWW '07*, pp.271–280 (2007).
- [13] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical*, Vol.39, No.1, pp.1–38 (1977).
- [14] Hofmann, T.: Probabilistic latent semantic analysis, *UAI '99*, pp.289–296 (1999).
- [15] Hofmann, T.: Collaborative filtering via Gaussian probabilistic latent semantic analysis, *SIGIR '03*, pp.259–266 (2003).
- [16] Iwata, T., Watanabe, S., Yamada, T. and Ueda, N.: Topic tracking model for analyzing consumer purchase behavior, *IJCAI '09*, pp.1427–1432 (2009).
- [17] Jin, X., Zhou, Y. and Mobasher, B.: Web usage mining based on probabilistic latent semantic analysis, *KDD '04*, pp.197–205 (2004).
- [18] Kurashima, T., Iwata, T., Irie, G. and Fujimura, K.: Travel route recommendation using geotags in photo sharing sites, *CIKM '10*, pp.579–588 (2010).
- [19] Liu, D.C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, Vol.45, No.3, pp.503–528 (1989).
- [20] Zheng, Y., Zhang, L., Xie, X. and Ma, W.Y.: Mining interesting locations and travel sequences from GPS trajectories, *WWW '09*, pp.791–800 (2009).
- [21] Zheng, V.W., Zheng, Y., Xie, X. and Yang, Q.: Collaborative location and activity recommendations with GPS history data, *WWW '10*, pp.1029–1038 (2010).
- [22] Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T.: Evaluating collaborative filtering recommender systems, *ACM Trans. Information systems*, Vol.22, No.1, pp.5–53 (2004).
- [23] Bellotti, V., Begole, B., Chi, E.H., Ducheneaut, N., Fang, J., Isaacs, E., King, T., Newman, M.W., Partridge, K., Price, B., Rasmussen, P., Roberts, M., Schiano, D.J. and Walendowski, A.: Activity-based serendipitous recommendations with the magitti mobile leisure guide, *CHI '08*, pp.1157–1166 (2008).
- [24] van Setten, M., Pokraev, S. and Koolwaaij, J.: Context-aware recommendations in the mobile tourist application COMPASS, *AH '04*, pp.235–244 (2004).
- [25] Ricci, F. and Nguyen, Q.N.: Acquiring and revising preferences in a critique-based mobile recommender system, *IEEE Intelligent Systems*, Vol.22, No.3, pp.22–29 (2007).
- [26] Crandall, D., Backstrom, L., Huttenlocher, D. and Kleinberg, J.: Mapping the world's photos, *WWW '09*, pp.761–770 (2009).
- [27] Leung, K.W.T., Lee, D.L. and Lee, W.C.: CLR: A collaborative location recommendation framework based on co-clustering, *SIGIR '11*, pp.305–314 (2011).
- [28] Geng, L. and Hamilton, H.J.: Interestingness measures for data mining: A survey, *ACM Computing Surveys*, Vol.38, No.3 (2006).

- [29] Salakhutdinov, R. and Mnih, A.: Probabilistic matrix factorization, *NIPS '08*, pp.1257-1264 (2008).



倉島 健 (正会員)

2004 年同志社大学工学部卒業。2006 年京都大学大学院情報学研究科修士課程修了。同年日本電信電話(株)入社。2011 年京都大学大学院情報学研究科博士後期課程入学。現在, NTT サービスエボリューション研究所所

属。Web マイニング, 行動モデリング, 情報推薦の研究開発に従事。電子情報通信学会, 日本データベース学会各会員。



藤村 考 (正会員)

1989 年北海道大学大学院工学研究科情報工学専攻博士課程修了。同年日本電信電話(株)入社。電気通信大学大学院情報システム学研究科客員教授, NTT サイバーソリューション研究所主幹研究員を経て, 2012 年より大妻女

子大学社会情報学部教授。ソーシャルメディアからの知識抽出, 情報可視化, インタラクティブメディア等の研究開発に従事。工学博士。電子情報通信学会, 日本データベース学会各会員。

(担当編集委員 前田 亮)



岩田 具治 (正会員)

2001 年慶應義塾大学環境情報学部卒業。2003 年東京大学大学院総合文化研究科広域科学専攻修士課程修了。同年日本電信電話(株)入社。2008 年京都大学大学院情報学研究科システム科学専攻博士課程修了。博士(情報学)。

現在, NTT コミュニケーション科学基礎研究所研究主任。機械学習, データマイニング, 情報可視化の研究に従事。2004 年船井ベストペーパー賞, 2009 年情報処理学会論文誌論文賞等受賞。電子情報通信学会会員。



星出 高秀 (正会員)

1993 年九州大学大学院総合理工学研究科修士課程修了。同年日本電信電話(株)入社。現在, NTT サービスエボリューション研究所主任研究員。e ラーニングシステム, マイニングの研究開発に従事。



高屋 典子 (正会員)

1987 年学習院大学卒業。1995 年日本電信電話(株)入社。1995~2003 年ネットワークサービスのマーケティング戦略担当, 2004~2009 年 NTT レゾナントにてマーケティング部門主幹, 2010 年より NTT サービスエボリューション研究所主任研究員。EC における購買行動モデリン

グの研究開発に従事。