

位置情報を利用した情報配信のための習慣度算出手法

Regular Behavior Measure for Location Based Services

林 亜紀[♡] 松林 達史[△] 澤田 宏^{*}

Aki HAYASHI Tatsushi MATSUBAYASHI
Hiroshi SAWADA

本研究では、行動が本人の習慣に沿っているかを示す「習慣度」を定義する。位置情報を利用した情報配信において、サービス利用者のニーズに合った配信内容・タイミングの決定が課題となっている。従来法では自宅からの距離や履歴頻度を用いた習慣/非習慣判別により、ニーズの把握を試みた。提案法では、曜日・時間帯を考慮した複数の多項分布の重み付き線形和を用いた習慣度定義により、自宅近くにおける非習慣も含めた詳細な行動理解を実現する。チェックインログを使用した検証により、本人履歴にない場所を推薦可能なタイミングの検出に加え、類似した習慣度で他ユーザが訪問した場所の抽出から、例えば出張者同士の履歴の考慮が可能になるなどの効果を示す。

We introduce a method that can measure the degree of regularity or irregularity of the behavior for enhancing the performance of location-based services (LBSs) such as *check-in*. It is still challenging for LBSs to determine the places to recommend that best suits the user's needs. Our aim is to identify the user's status (regular or irregular) of each check-in. Most previous studies approached this problem by acquiring usual locations (e.g., home or office) or assessing check-in frequency. We propose more effective measure by using a multinomial-distribution-based method that considers the periodic check-ins of the user on various time-scales. Our method can accurately identify irregular check-ins even in usual locations. We show that our method can find the best timing to recommend new places and it can also utilize other user's check-in history in similar situations.

1. はじめに

近年、モバイル端末の普及とソーシャルメディアの発展から、位置情報の取得・蓄積が容易になり、様々なサービスが提供されている。現在地情報の活用だけでなく、履歴頻度など蓄積ログの

[♡] 正会員 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
hayashi.aki@lab.ntt.co.jp

[△] 非会員 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
matsubayashi.tatsushi@lab.ntt.co.jp

^{*} 非会員 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
sawada.hiroshi@lab.ntt.co.jp

解析から、ユーザの行動を予測し、よりニーズにマッチした情報を配信する取り組みも行われており、i コンシェル、Google Now、Foursquare¹、Wityou²などのサービスでは、現在地や居住地、本人履歴を利用した情報配信が行われている。しかしながら、配信内容・タイミングの決定において以下に述べる未解決の課題がある。

情報配信におけるニーズには、

- 確実性：ユーザの行動傾向に沿った確度の高い情報
- 意外性：ユーザが従来では知り得なかつた情報

の 2 側面が存在すると考えられる。例えば、会社出勤日の昼食時など日常時には、よく行く店舗の割引クーポンなど確実性の高い情報が好まれるが、例え会社近くであっても休日の昼食時には、普段より高級で、訪れたことのない店のような意外性の高い情報が好まれる可能性がある。一方、出張時にはご当地料理を出す店舗が望まれる可能性や、普段からユーザが好んで食べているラーメンを現地で食べられる店舗が好ましい可能性もある。このように、飲食店情報配信において、どちらを重視して考慮すべきかは、時と場合に応じて異なり、好ましい考慮比率を定めることは困難である。既存サービスでは、履歴にある店が近隣にあればその店の情報を、出張時には、現在地近くでよく行かれる店舗の情報配信を行うことが多く、日常的に訪れる場所における未訪問店舗の推薦機会を逃す例や、遠方への出張時にご当地料理店よりも現地で一般的によく行かれる社員食堂が提示されてしまう例もある。情報配信では内容とタイミングが重要であり、ユーザの納得感に加え、店舗側の新規顧客開拓というニーズを考えた際に、いつどのユーザにどの店舗を提示すれば快く来店されるかという課題の解決が望まれる。我々は、時と場合に応じた 2 側面の適切な考慮によるユーザと情報提供者双方の満足度向上に向け、ユーザの行動理解技術の向上を目指す。

前述の例のように、ユーザが本人の習慣に沿った行動を行っている状態では確実性を重視する場合が多く、一方で非習慣的な行動をしている状態では意外性を考慮する必要がある。必ずしも二者択一的な考慮が適切とは限らないが、適切な考慮比重とその内容を決定するための行動理解の一手段として、ユーザが習慣に沿った行動をしているのか、非習慣的な行動をしているのか (*) を推定する技術が必要であると考えられる。そこで、提案手法では、各ユーザが持つ「習慣」を抽出した上で、(*) を推定する習慣度という指標を新しく定義・算出する。なお、「習慣」は例え「Aさんは平日の朝は品川駅付近の会社にいることが多い」などの行動傾向であり、ユーザ情報・時間情報・位置情報から構成する。

提案する習慣度指標により、例えば出張中などの非習慣状態を検出し、本人履歴に加え、他人履歴を考慮できるタイミングを見極めることが期待される。また、配信内容について、出張者同士など状況が類似した他人履歴の抽出が必要であり、どのような他人履歴の考慮が効果的かを見極めることが求められる。そこで、提案する習慣度の要件として以下の 2 つを設定する。

- (1) 直前の行動の習慣度が直後の行動でも持続しやすい
(2) 同一場所は複数ユーザから類似した習慣度で訪問されやすい

要件 (1) により、直前の行動の習慣度が低ければ、直後の行動においても非習慣的状態が続くと予測して、他人履歴を考慮した意外性の高い情報を重視することが可能になる。また、要件 (2) により、出張者同士の履歴の考慮などが可能になる。

以下、2 章では、位置情報を利用した行動予測・情報配信に関する既存研究の概要と本研究の位置づけを述べる。3 章では、提案する習慣抽出モデルと習慣度算出手法について詳細に述べる。4 章では、提案手法を定性的・定量的に検証し、5 章で提案法を活用した情報配信について例を挙げて検討する。最後に、6 章にまとめと今後の課題を示す。

¹<https://ja.foursquare.com/>

²<http://wityou.jp/>

表 1: チェックインデータの例

User ID	年月日	曜日	時間帯	Location ID	緯度	経度
58186	2010/01/08	月	21	12587	39.XX	-105.XX
58186	2010/05/20	火	21	12587	39.XX	-105.XX
58187	2010/06/12	日	07	31544	41.XX	-95.XX

2. 関連研究

習慣度の算出

行動が習慣的であるか、非習慣的であるかを検出する研究として、自宅や勤務場所など訪問頻度の高い場所から遠いほど非習慣であるとするもの [6, 3, 11]、移動距離が通常時の 2 倍以上であれば非習慣とするもの [4]、移動時間の長短で判別するもの [10] がある。単位時間内の訪問場所の多様性から旅行者という非習慣を検出する手法 [8] も提案されている。しかしながら、これらのルールベースの手法では、習慣抽出時に時間情報を考慮しないため、例えば「会社近くにおける家族との休日の買物」「いつもより遅い/早い帰宅」などの、日常的によく行く場所への弱い非習慣を含めた詳細な習慣度は定義されていない。

他人履歴を考慮した行動予測・情報配信

他人履歴を活用した行動予測・情報配信を行う研究として、ユーザ別場所頻度分布 [14, 2] や、時間を考慮したユーザ別場所頻度分布 [12] を元にしたユーザのクラスタリングによる協調フィルタリングが提案されている。習慣時と非習慣時のログが混在した状態で類似ユーザを抽出するため、普段類似した場所に行くユーザの行動が非習慣時にも類似するとは限らず、適切な推薦が実現できない場合がある。Malmi ら [9] は訪問時間分布を元にした場所間のクラスタリングにより、他人ログを活用した行動予測によるコールドスタート問題の解決を図っているが、習慣・非習慣の区別なく訪問傾向を抽出している。

一方、ログの習慣/非習慣性を考慮した情報配信も行われているが、現在地にログが多いユーザの習慣を強く考慮する手法 [13] では、社員食堂や自宅が推薦されてしまう可能性がある。倉島ら [7, 6]、Leung ら [8] は非習慣行動時の習慣を反映した情報配信を提案しているが、前述のように習慣/非習慣の抽出がルールベースとなっていることに加え、習慣行動時の履歴は考慮しないため、ユーザが普段好んでいるラーメン店を出張先でも提示する、などの嗜好の反映は想定していない。

本人履歴と他人履歴の両方を考慮した行動予測・情報配信

上述した行動予測・情報配信手法では、本人履歴と他人履歴のどちらをどのようなタイミングで優先させるかは考慮せず、一律に他人履歴の一部を考慮しており、1 章で述べた確実性の考慮が不十分な場合が生じると考えられる。習慣時には習慣に沿った場所を提示し、非習慣行動時には友人の訪問場所を推薦する手法 [3] が提案されているが、習慣として抽出する場所が自宅・会社など少数であること、必ずしも非習慣時に友人の影響を受けるとは限らないことが懸念される。

習慣の抽出

位置情報に基づく習慣を抽出する研究として、一次マルコフモデルにより場所遷移の習慣を抽出するもの [1]、曜日や時間帯に依存しないユーザの定常的な興味を抽出するもの [7, 6] が挙げられるが、曜日や時間帯に依存する習慣の抽出は未実現であった。一方、Cho ら [3] は時間を考慮して各ユーザが頻繁に訪問する場所を抽出したが、自宅・勤務地など限られた場所のみを抽出していた。Gao ら [5] は様々な場所の曜日・時間帯別の訪問頻度変遷をガウス分布を用いて表現した習慣を抽出し、マルコフモデルと掛け合わせることにより、行動予測精度を向上させたが、「月曜 8 時台」など曜日と時間帯の両方を考慮した習慣は抽出していなかった。

表 2: 入力情報の表現

入力情報	表現	入力情報	表現
User ID	u	曜日 [月曜-日曜]	d
Location ID	l	時間帯 [1-24]	t

提案手法との比較

提案手法では、様々な場所に対して、曜日・時間帯依存性を考慮した複数の尺度を持つ多項分布の重み付き線形和により、習慣をモデル化する。従来は検出できなかった、よく行く場所への非習慣状態を含めて習慣/非習慣状態を検出できる。提案手法では連続値を取る指標として習慣度を定義する。この習慣度算出により、他人の履歴を考慮すべきタイミングの決定を支援することが可能になるだけでなく、適切に個人履歴と他人履歴の両方を考慮することができると期待される。習慣度の値から、どの程度本人の履歴から次の行動が予測できない状態であるかを把握できることによる。また、他人履歴の考慮内容について、各行動の習慣度を参照することにより、非習慣行動同士のログの考慮、習慣時の嗜好の考慮などが可能になり、満足度の向上が期待される。加えて、昼食時には飲食店情報を配信する、などの配信タイミングの向上も可能になると考えられる。

3. 提案手法

3.1 提案手法概要と使用データ

本稿では、位置情報を用いた情報配信において、効果的な配信内容・タイミング決定を支援する習慣度算出手法を提案する。手法の概要は以下の通りである。

Step1: 確率モデルによる習慣の抽出

個人毎の周期的習慣(曜日・時間帯を考慮した各訪問場所への頻度分布)の抽出

Step2: 習慣度の算出

Step1 で抽出した個人周期的習慣をもとに、各行動の習慣度(習慣への当てはまり度合い)を算出

使用データの概要、実験設定を述べた後、3.2 節で Step1、3.3 節で Step2 の詳細を説明する。

本稿では、Stanford 大学が公開しているチェックインログ³への適用により提案手法を評価する。Gowalla, Brightkite の 2 つのサイトのログが公開されており、Cho ら [3] の研究でも用いられている。問題を簡潔にするため、チェックインデータを使用して検証を行うが、GPS など連続的な位置情報ログへの適用も可能である。ただし、チェックインデータの場合は、能動的な記録作業が必要となるため、GPS データに比べて普段行く場所における習慣的なチェックインが減り、出張や旅行など普段あまり行かない場所でのチェックインが増える可能性がある。そのため、本稿の実験ではログ数が一定以上のユーザを抽出して実験を行うが、将来的に無意識的な記録が可能な GPS データを用いた実験も行う予定である。データの例を表 1 に示す。User ID、チェックイン日時、Location ID、緯度経度を属性として持つ。同一の緯度経度の組を持つ場所に対し、同一の Location ID が割り振られており、提案手法による習慣抽出では Location ID を場所として用いている。入力データから得られる情報を、表 2 のように表現する。全ての情報を Gowalla, Brightkite のチェックインログから得ることができる。公開データに含まれるログのうち、本稿の評価実験

³<http://snap.stanford.edu/data/#locnet>

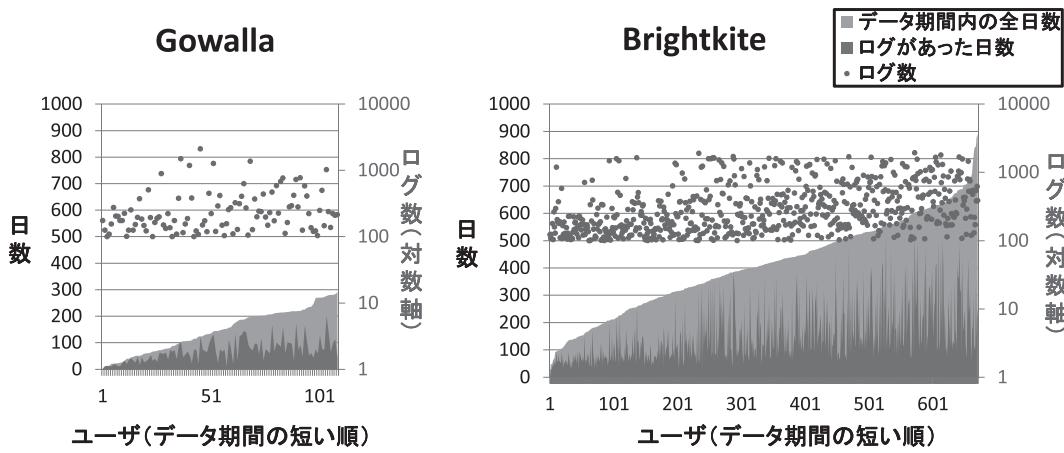
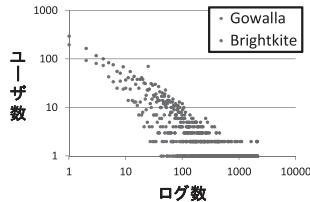


図 2: 各ユーザのデータ期間とログ記録日数

表 3: データの規模

項目	Gowalla	Brightkite
データ期間	2009/06-2010/10	2008/04-2010/10
ログ数	50,785 ログ	332,716 ログ
場所数	14,837 箇所	115,276 箇所
ユーザ数	1,164 人	3,194 人
平均ログ数/ユーザ	約 50.0 ログ	約 104 ログ

ログ数とユーザ数の関係



ログ数と場所数の関係

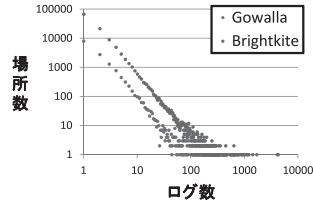


図 1: ユーザ・場所毎の利用ログ数

では日本国内のデータだけを抜粋する。国内データの規模を表 3 に示し、図 1 にログ数とユーザ数の関係（左）とログ数と場所数の関係（右）を示す。前述の通り、本稿では、100 ログ以上のユーザを抽出して評価実験を行う。100 ログ以上のユーザは Gowalla では 109 人、Brightkite では 671 人であった。同一ユーザの同一時間帯における同場所への連続的な訪問は、初回のみを残した。100 ログ以上のユーザのデータ期間内の全日数とログが記録されている日数を図 2 に示す。横軸はデータ期間の短い順にユーザを並べた。縦軸は日数について左側の軸、参考のために載ったログ数は右側の対数軸に対応する。Gowalla、Brightkite 両方で、多くのユーザについて、ログ数に関わらず一定以上の期間・日数のログが確保されていることが確認できた。

3.2 Step1: 確率モデルによる習慣の抽出

本手法では情報配信内容だけでなく、タイミングの向上も目指している。それ故、曜日や時間帯といった時間情報を活用した習慣抽出モデルを提案する。習慣度算出の際に、例えば休日に会社近くの商業施設へ家族で行った時など、よく行く場所における非習慣状態を検出することが可能になる。Gao らは曜日・時間帯をそれぞれ考慮した習慣の積算を行ったが、提案手法のような習慣・非習慣検出に利用する習慣抽出モデルとしては、「月曜 8 時台」など曜日と時間帯の両方を考慮した習慣が抽出できず、不十分な部分があった。提案手法では、曜日と時間帯の両方を考慮した習慣と、曜日・時間帯を考慮しない習慣を加えたモデル化により、柔

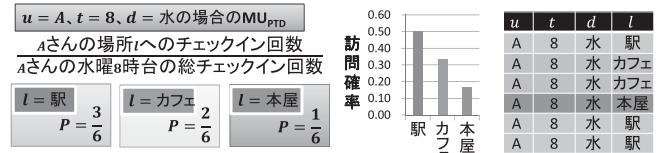


図 3: 周期的習慣の算出例

軟な習慣度算出を実現する。ここで、時間帯は「何時台」といった 1 時間刻みの単位で区切った時間区間を示す。将来的に、毎時 0 分を区切りとした単位だけでなく、他の粒度での時間情報の活用も検討しているが、本稿ではまず何時台であるかを単位として後述する方法で平滑化を行った上で実験を進める。

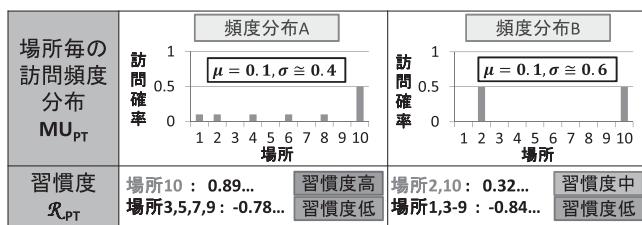
提案モデルでは、個人 (P :Personal) の周期的習慣を検出する。曜日・時間帯ともに考慮なしの習慣、曜日・1 時間刻みの時間帯両方を $24 \times 7 = 168$ 次元の 1 週間単位で考慮した習慣 (TD :Time, Day of the week), 曜日のみを考慮した習慣 (D), 時間帯のみを考慮した習慣 (T) の 4 種類を、それぞれ多項分布 (MU :Multinomial) を用いて以下のように表す。

$$\begin{aligned} MU_P: P(l|u) &= C_{l,u} / \sum_l C_{l,u} \\ MU_{PTD}: P(l|t,d,u) &= C_{l,t,d,u} / \sum_l C_{l,t,d,u} \\ MU_{PD}: P(l|d,u) &= C_{l,d,u} / \sum_l C_{l,d,u} \\ MU_{PT}: P(l|t,u) &= C_{l,t,u} / \sum_l C_{l,t,u} \end{aligned}$$

C_l は場所 l におけるチェックイン回数 (Count), すなわちログ数を示す。 t, d, u は表 2 の通りである。例えば $C_{l,t,d,u}$ では、ユーザ、時間帯、曜日、場所を全て限定してチェックイン回数を数え、 $C_{l,d,u}$ ではユーザ、曜日、場所を限定し、時間帯は限定せずにチェックイン回数を数える。従って、 $C_{l,d,u} = \sum_t C_{l,t,d,u}$ である。周期的習慣 MU_{PTD} の算出例を図 3 に示す。実際の l の値は Location ID であるが、ここでは例として駅、カフェ、本屋とした。 $u = A, t = 8, d = \text{水}$ のログを抽出した上で、各場所への訪問確率を算出する。時間帯を考慮する MU_{PTD} , MU_{PT} では、 $\sigma=0.3$ とした 1 次元の Gaussian フィルターを用いて平滑化を行った。実際の記録時間帯に近いほどチェックイン回数が大きくなるが、前後の時間帯にも値が入る。 σ の値は、4.3 節で示す習慣抽出モデルの検証で決定した。

3.3 Step2: 習慣度の算出

Step1 の習慣抽出モデルを利用して、各行動が習慣的な行動なのか、非習慣的な行動なのかを調べる習慣度算出手法を提案する。Step1 で定義した習慣のうち、 MU_P では駅やコンビニなど時間規則性はないがよく行く場所が抽出される。 MU_{PD} では出張の多い

図 4: 習慣度算出の例 (時間帯のみ考慮習慣度 R_{PT} の算出例)

会社員にとっての会社, MU_{PT} では営業時間帯の決まっている飲食店, MU_{PTD} では出張が少ない会社員にとっての会社などが抽出される。このように、習慣モデルにより、抽出されやすい場所が異なるため、本手法ではこれらの 4 つの習慣から算出された習慣度を組み合わせて、各行動の習慣度を算出する。組み合わせにより、訪問総数はあまり多くないが、月に 1 度位の頻度で日曜日に訪れる商業施設などへの訪問についても、やや習慣的であることを検出することができる。

各ログが習慣的なのか、非習慣的なのかを示す指標 R を以下のように定義する。

$$R = w_P R_P + w_{PTD} R_{PTD} + w_{PD} R_{PD} + w_{PT} R_{PT} \quad (1)$$

第 1 項は曜日時間帯考慮なし習慣度、第 2 項は曜日時間帯考慮習慣度、第 3 項は曜日のみ考慮習慣度、第 4 項は時間帯のみ考慮習慣度を表す。各習慣度の算出方法は以下の通りである。

$$\begin{aligned} R_P &= \frac{MU_P - \mu}{\sigma_P}, \quad R_{PTD} = \frac{MU_{PTD} - \mu}{\sigma_{PTD}}, \\ R_{PD} &= \frac{MU_{PD} - \mu}{\sigma_{PD}}, \quad R_{PT} = \frac{MU_{PT} - \mu}{\sigma_{PT}} \end{aligned} \quad (2)$$

外れ値検定の際に一般的に使用される尺度を参照した。ここで、式(2)で算出される各習慣度の値について、習慣度が負の値を取る際にも、値が大きいほど習慣度が高くなるようにするために、 $R \leq 0$ の場合のみ以下のような補正を実施する。

$$R_{PX} \leftarrow -R_{PX} - 1.0 \quad (\text{if } R_{PX} \leq 0) \quad (3)$$

R_{PX} は式(2)で算出される 4 種類の習慣度をまとめて表現したものである。補正を行う理由の詳細は後述する。 μ は頻度分布の平均を示す。 $\sigma_P, \sigma_{PTD}, \sigma_{PD}, \sigma_{PT}$ は、それぞれ頻度分布 $MU_P, MU_{PTD}, MU_{PD}, MU_{PT}$ の標準偏差を示す。式(2)の各式を MU_{PX} (4 種類の周期的習慣をまとめた表現)について解くと $MU_{PX} = R_{PX}\sigma_{PX} + \mu$ となる。習慣度 R_{PX} は、平均 μ から MU_{PX} が標準偏差 σ_{PX} の R_{PX} 倍離れていることを示し、分布内でどれくらい突出しているかを表す。習慣度が大きいほど R, R_{PX} は大きい値を取る。

本手法では学習データ中に含まれる全場所に対する頻度分布を考える。該当時間帯に本人のログが 1 つも存在しない場所への訪問確率も定義され、その確率は全て 0 となる。全場所への訪問確率の和が 1 であることから、頻度分布の平均 μ は、全場所数を $|L|$ とするとユーザ・時間帯に関わらず全て $1/|L|$ となる。標準偏差 σ の値や習慣度の正負を決める μ を、全ユーザに共通な全場所への頻度分布から算出したのは、習慣度の大小を複数ユーザ間で比較することも想定したためである。 MU_{PX} の確率が非常に小さく $1/|L|$ 未満の場合、各習慣度 R_{PX} は負の値を取る。これは習慣度が特に低い状態と捉えられる。

習慣度の算出例

習慣度算出の例を図 4 にしめす。ユーザの特定時間帯における頻度分布 MU_{PT} を表す 2 つの頻度分布 A,B から、時間帯のみ考慮習慣度 R_{PT} を算出する例を考える。全場所数は 10 である。訪問確率 > 0 の場所にいた場合、習慣度算出では、そのログで訪問さ

れている場所へのその時間帯における訪問確率が突出しているかを調べる。従って、訪問確率がともに 0.5 である場所 10 にいた場合、同確率で場所 2 にいる可能性もある頻度分布 B よりも、頻度分布 A の方が習慣度は高くなる。

一方、訪問確率が 0 の場所にいた場合、習慣度は負の値を取る。 $|L|$ は全員分のログに記載されている全場所数の合計となるため、手元のデータでも 100,000 など非常に大きい値となることが多く、 $\sum_l C_{l,u} << |L|$ であるためである。本人の訪問が 1 度でもあれば式(2)で算出される R_{PT} が $R_{PT} > 0$ となる。一方、1 度も訪問がない場合には、 $R_{PT} \leq 0$ となり、その時間帯における頻度分布のばらつきが B のように大きいほど習慣度は小さくなる。分布 B ではその時間帯にいる可能性がある場所が少数であるのに対し、分布 A ではその時間にいる可能性がある場所が 6箇所あり曖昧なため、未訪問の場所にいる場合の習慣度は頻度分布 A の方が若干高くなる。

各習慣度の補正について

式(2)の各習慣度の最小値は -1.0 程度であるのに対し、最大値は場所数に応じて変化する。場所数が 100,000 個程度の実験で、最大値はユーザによっては 1,000 を超えた。式(2)で算出される値は、習慣度が正の場合は値が大きいほど習慣度が高く、負の場合は値が大きいほど習慣度が低くなり、値の大小関係が正負により異なる。そこで、式(3)により、値が負の場合のみ大小関係を反転させることで、値の正負に関わらず、値が大きいほど習慣度が高くなるよう補正を行う。提案する習慣度指標ではその時間帯におけるログが存在しない非習慣状態に対しても、頻度分布の分散に応じて異なる習慣度を出力できる。

各習慣度を足し合わせる重みについて

それぞれの習慣度を足し合わせる重みは、該当時間刻みにおけるログ数を考慮して以下のように求める。

$$\begin{aligned} w_P &= 1.0, \quad w_{PTD} = \frac{C_{t,d,u}}{\max_{t,d} C_{t,d,u}}, \\ w_{PD} &= \frac{C_{d,u}}{\max_d C_{d,u}}, \quad w_{PT} = \frac{C_{t,u}}{\max_t C_{t,u}} \end{aligned} \quad (4)$$

式(4)の分母について、例えば $\max_{t,d} C_{t,d,u}$ は最もチェックイン回数が多い曜日・時間帯(月曜 8 時台など)のチェックイン回数を示す。重み w を導入することにより、例えば MU_{PTD} で、学習データ中の予測したい曜日・時間帯のログが他の曜日・時間帯に比べて少ない場合には、 MU_{PTD} の考慮比率を下げる事が可能になる。 $\max_{t,d} C_{t,d,u} = 120$ のユーザの w_{PTD} を考える。[Case1:該当曜日・時間帯に場所 A へのログ 1 回のみがある場合]、場所 A の訪問確率が 100%で、それ以外の場所への訪問確率は 0%の頻度分布が作成される。一方、[Case2:該当曜日・時間帯に 100 回分のログがあり、そのうち 100 回全てが場所 A への訪問の場合]も確率頻度分布は同様になる。場所 A への訪問確率は同じだが、式(4)を用いると、[Case1] では $w_{PTD} = 1/120$ 、[Case2] では $w_{PTD} = 100/120$ となり、[Case1] の方が MU_{PTD} の考慮比重が低くなる。母集団となるログ数が多い[Case2]の方が、習慣として信頼度が高いと考えられるため、習慣抽出時に母集団となるログが少なかった場合は、その周期的習慣の考慮比率を下げて習慣度の足し合わせを実施する。4 種類の多項分布の重み付き線形和により、行動パターンが必ずしも一定ではないユーザの習慣も表現できる。

4. 評価と考察

習慣度算出手法及び習慣抽出モデルの評価を行う。3.1 節で示したデータを用いた。実験設定として、100 ログ以上あるユーザについて、各ユーザのログのうち 80% を学習データ、残り 20% をテストデータとした。100 ログ未満のユーザのログは全て学習データとした。

	Gowalla	Brightkite
習慣	田町(258), 品川(365), 大崎(169), 青山(19), ミッドタウンオフィス(168)	オフィス 区役所(83), 品川(81)
	鶴見(205), 大宮(184), 戸塚(72), 茅ヶ崎(124), 新習志野(53)	
非習慣	新木場(10), 國際展示場(1), 幕張(11)	イベント 幕張(12)
	羽田空港(48), 成田空港(68), 新横浜駅(39)	
	舞浜(13), 京都(67), 中華街(7), 六本木ヒルズ(56), 竹下通り(8)	空港、新幹線 成田空港ゲート(66), 成田空港ビル(27), 新横浜駅(42)
		観光地 京都(16), 富士山(13), お台場(101), 六本木ヒルズ(18), 科学未来館(6), ミッドタウン広場(3)

図 5: 場所毎に全ユーザで平均した習慣度

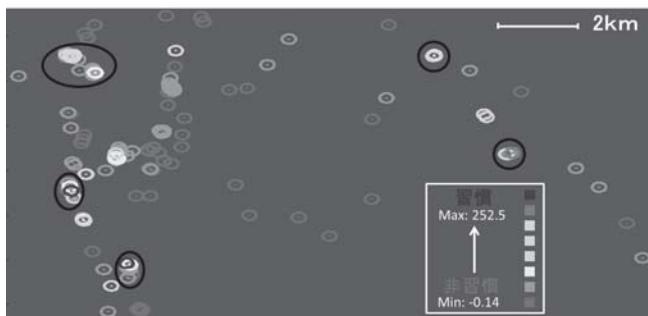


図 6: 少数ユーザ（3人分）の習慣度算出結果（場所毎平均）

4.1 習慣度算出手法の定性的評価

テストデータ中の各ログの習慣度を算出し、場所毎に全ユーザによる習慣度を平均した結果、習慣度が特に高かった場所と低かった場所を図 5 に示す。ここでは、Gowalla, Brightkite それぞれについて、Location ID が割り振られている緯度経度の組を場所として扱った。習慣性が高いと予想されるオフィスや住宅、低いと予想されるイベント会場や観光地がそれぞれ抽出された。図中の括弧内の数字は、学習データにおけるログ数を示す。必ずしも訪問回数の大小だけで習慣度が決定されていないことが分かる。

少数ユーザで場所毎に平均した習慣度

図 6 は Gowalla データのうち、少数ユーザのテストログに対して、場所毎に平均した習慣度を、緯度経度情報を参照して配置したものである。ログ数が多いユーザのうち、 R_P と R_{PTD} の差が大きいユーザを 3 ユーザ選出した。個人情報の保護のために、背景に地図を載せることは避けた。赤色に近いほど非習慣状態、紫色に近いほど習慣状態を示す。赤色は習慣度が負の場合、残りの 8 色は習慣度が正のものについて、値の大小で 8 等分した。図中の丸で囲まれた箇所のように習慣度の高い紫色が集中している場所の周辺にも、習慣度が低いオレンジ色などが見られる。頻繁に訪問する場所からの距離が遠いほど非習慣とする従来の方法では検出できない非習慣が検出できた。

特定ユーザのログ毎の習慣度変遷

図 7 はある特定ユーザの 21 日分のテストデータの習慣度変遷を時系列に沿って可視化したものである。横軸の点線は日付を示し、1 日の中を 1 時間ずつ 24 分割して描画位置を決定した。縦軸は場所を表す。テストデータに登場した順に、下から表示を割り当てた。なお、個人情報の保護のために、当該ユーザのログがある場所のうち、一部のみを抜粋した上で、平日の一部曜日を置換した。桃色の矩形で囲まれた場所 A を見ると、ほぼ毎日訪問しているものの、その習慣度には微妙な変化があることが分かる。場所 A における周期性習慣の抽出結果と、各訪問における習慣度変

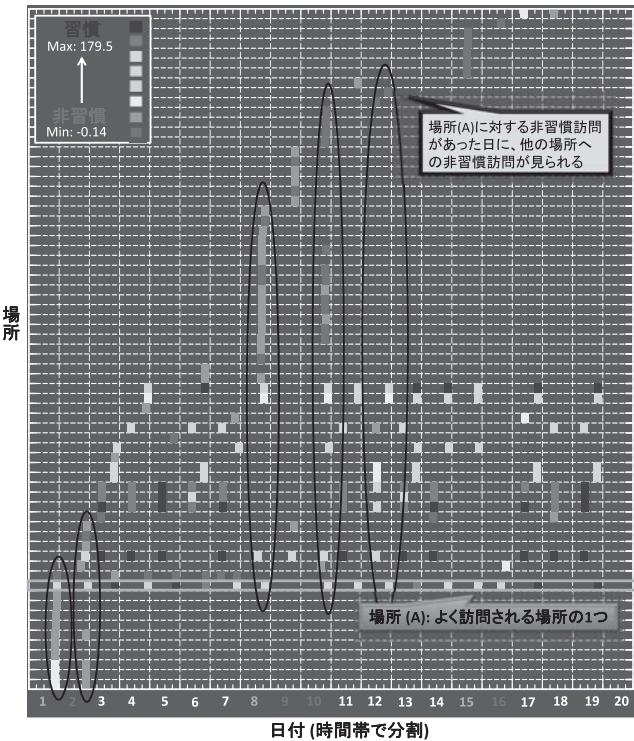


図 7: 特定ユーザの習慣度算出結果の変遷

遷の詳細を図 8 にまとめた。曜日 ID のうち、1-5 は平日、6-7 は休日である。1 日目 (No.1), 2 日目 (No.2), 4 日目 (No.4), 12 日目 (No.11), 16 日目 (No.15) は訪問時間帯や曜日が習慣と異なるため、習慣度が比較的低いことが分かる。例えば 12 日目の訪問は従来よりも遅い時間に場所 A にいた事例であると考えられる。この後の訪問を調べると、非習慣的にいくつかの場所を訪問していることが分かった。帰りが遅れたため、普段あまり利用しない飲食店で食事をしたなどの可能性が考えられる。12 日目の訪問だけでなく、図 7 で桃色で囲まれているように、いつも訪れている場所への訪問にも関わらず訪問時間帯が非習慣的だった場合、その前後にも非習慣行動が続きやすいことが可視化結果から分かる。このような非習慣行動を続いているタイミングで、本人履歴がない、新規性・意外性の高い飲食店を推薦するなどといった情報配信への活用が考えられる。

4.2 習慣度算出手法の定量的評価

連続ログ間の習慣度持続性に関する分析

1 章において、(1) 配信直前の行動における習慣度が、直後の行動でも持続しやすく、直前の習慣度により配信タイミングを決定できるという習慣度の要件を設定した。遷移時間が 2 時間以内で連続している 2 ログ間の習慣度差を算出し、閾値以下であった割合を求めたところ、Gowalla で 73.2%, Brightkite で 81.2% であった。全ユーザの習慣度の (最大値 - 最小値) × 0.1 を閾値とした。同一ユーザの任意のログ間に同じく習慣度差を算出した場合、閾値以下であった割合は Gowalla で 55.8%, Brightkite で 52.9% であり、2 時間以内に連続するログでは、任意のログ間に比べ、習慣度差が小さいことが分かった。直前の行動ログの習慣度による配信タイミングの決定が有効であると考えられる。

連続ログ間における習慣度変化の相関分析

習慣度と配信内容の相関を詳細に調べるために、2 時間以内に連続する 2 ログ間の習慣度変化について相関分析を行った。 $n-1$ 番目のログにおける習慣度 R_{n-1} と n 番目のログにおける習慣度 R_n の関係を

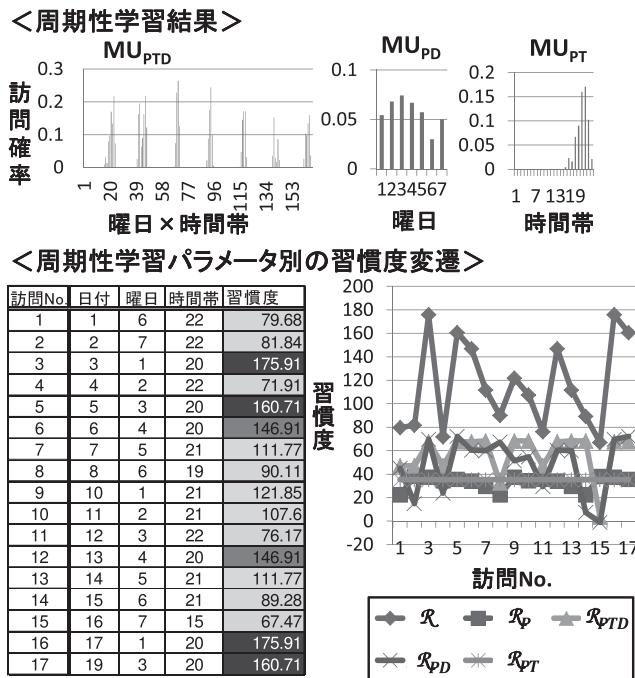


図 8: 特定ユーザの特定場所への習慣度変遷

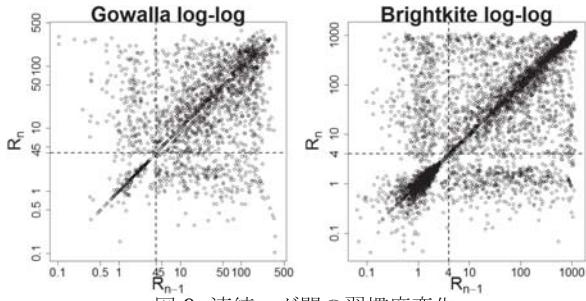


図 9: 連続ログ間の習慣度変化

分析した。相関分析では、2つの確率変数 x, y 間の相互依存性を分析し、2変数間にどの程度直線的な関係があるかを示す相関係数の値を求める。相関係数は $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}$ で求められ、-1から1の間の実数値を取り。1に近い場合には2変数間に正の相関があることを、-1に近い場合には負の相関があることを示す。ここで、 \bar{x}, \bar{y} はそれぞれ、 $x = \{x_i\}, y = \{y_i\}$ の相加平均である。 R の値が大きい場合、小さい場合両方について分析を行うため、 $R + \epsilon$ の常用対数を取った値を用いた。習慣度は最小値が-1.0である4種類の習慣度の線形和であったことから、 $\epsilon = 4.0$ とした。 R_{n-1} と R_n の関係を図9に示す。Gowalla, Brightkite の相関係数はそれぞれ 0.60, 0.72 であり、ともに正の相関が見られた。直前、直後の習慣度がともに負となった領域に限定して相関係数を算出すると、Gowalla で 0.87, Brightkite で 0.72 となり、特に Gowalla で強い相関関係があることが分かった。偶然このような相関関係が観測される確率を示すP値は、すべて $2.2e-16$ 未満と十分に小さい値であった。提案手法では、習慣度が負になる場合についても、その時間帯の分布のばらつきから習慣度を変化させているが、その定義が効果的であることが示された。

連続ログ間の習慣度変化の平均値による分析

直前の行動の習慣度と直後の行動の習慣度の関係について、4.1節において、1人のユーザの習慣度変遷の定性的評価により、特に非習慣行動を行っている際に、続けて非習慣行動を行う場合がある

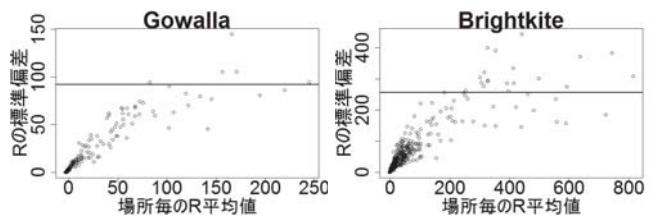


図 10: 場所毎の平均習慣度と習慣度のばらつきの関係

ことが示唆された。これについて、統計的な検証を実施する。移動時間が2時間以内で連続する2ログ間の習慣度の差を平均したところ、Gowalla で-3.35, Brightkite で-8.11 であった。習慣度差の平均が負の値を取ることは、非習慣状態（習慣度は小さい）から習慣状態（習慣度は大きい）への変化が少ないことを示す。特に習慣度が低い時に、本人履歴にない場所を推薦することが有効であると考えられる。

以上の分析から、直前の行動の習慣度と、次の行動の習慣度には一定の相関があり、特に非習慣行動時には、非習慣行動を続ける可能性が高いため、他人の履歴を利用した情報配信を行うことが有効な可能性があることが示された。直前の行動の習慣度を、情報配信における配信タイミングとその内容決定に活用できると期待される。

場所毎の習慣度のばらつきについて

1章において、(2) 同一場所は複数ユーザから類似した習慣度で訪問されやすいという習慣度の要件を設定した。特に非習慣状態にあるユーザに対して、近い習慣度を持つ他のログを考慮することが考えられる。ここでは、あるユーザに非習慣的に訪問される場所が、他のユーザからも類似した習慣度で訪問されやすいかどうかを検証する。3ユーザ以上から訪問されている場所への訪問を抽出した上で、各場所の習慣度平均と習慣度のばらつき（標準偏差）の関係を分析した（図10）。図中の横軸に平行な太線は、抽出したログ全体の標準偏差を示す。Gowalla, Brightkite とともに、全ログの標準偏差に比べて、同一の場所ではばらつきが少ない。また、特に平均習慣度が低い場所について、標準偏差も小さい傾向があることが分かり、非習慣状態において、類似した習慣度で訪問されやすい場所を提示することが有効であると考えられる。

4.3 習慣抽出モデルの定量的評価

Step2の習慣度算出時に、時間帯や曜日の考慮有無に応じて、4つの習慣を重みづけて足し合わせた。この組み合わせの効果を検証する。式(4)と同様の重み付けで4つの習慣を組み合わせた頻度分布を用いた行動予測精度評価により検証する。Step2の習慣度算出時と同様に、Step1で抽出した4種類の習慣を組み合わせた MU_{PALL} を以下のように求める。重み \widehat{w} は式(4)と同様に求めるが、全場所への訪問確率の和を1にするために、 $\widehat{w}_P + \widehat{w}_{PTD} + \widehat{w}_{PD} + \widehat{w}_{PT} = 1.0$ とする。

$$\begin{aligned} MU_{PALL} &= \widehat{w}_P MUP + \widehat{w}_{PTD} MU_{PTD} \\ &\quad + \widehat{w}_{PD} MUPD + \widehat{w}_{PT} MUPT \end{aligned} \quad (5)$$

位置情報に基づく行動予測では、定食屋 A に来た人は、その後カフェ B へ行くことが多いなど、 n 番目の場所 I_n から $n+1$ 番目の場所 I_{n+1} への行きやすさを表す遷移確率が有効であると示されており、Gao らの手法でも個人ログから抽出した遷移習慣と周期的習慣を掛け合わせた確率を算出する。そこで、提案手法においても、周期的習慣に加え、遷移習慣を考慮する。ここでは、各場所から各場所への遷移確率を算出し、個人だけでなく、全ユーザ (G:Global) についてもそれぞれ遷移習慣を抽出する。これは、本人の習慣に従わない行動をしている際に、全体の遷移習慣を合わせて利用すると、行動予測精度が向上する可能性があると考えられるた

表 4: 習慣を利用した行動予測精度結果

モデル	Gowalla		Brightkite	
	p@5	p@1	p@5	p@1
(a) MU_P	29.6	10.9	40.6	22.0
(b) MU_{PTD}	28.8	12.2	35.1	21.0
(c) MU_{PD}	28.0	10.5	38.0	22.0
(d) MU_{PT}	32.4	14.0	40.1	23.0
(e) MU_{PALL}	33.2	14.3	40.8	23.8
(f) $0.3MU_{PALL} + 0.7MM_{GP}$	39.9	20.0	48.8	34.1
(g) $MU_{PD} \times MU_{PT} \times MM_P$ (Gao ら)	29.6	17.1	33.4	24.8
(h) $MU_{PALL} \times MM_P$ (参考)	30.6	17.8	34.5	26.4

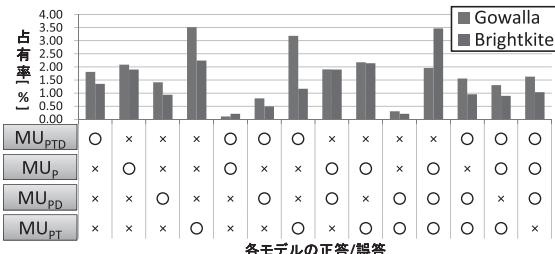


図 11: 各周期的習慣の予測性能比較

めである。遷移習慣は一次マルコフモデル (MM :Markov Model) により算出する。各遷移習慣の抽出方法を以下に示す。 MM_{GP} では、 l_n を出発する本人の遷移履歴がある場合は MM_P を、本人履歴はないが全体遷移履歴のみがある場合は MM_G を使い分ける。

$$MM_P(\text{個人遷移習慣}): P_u(l_{n+1}|l_n)$$

$$MM_G(\text{全体遷移習慣}): P(l_{n+1}|l_n)$$

$$MM_{GP} = \begin{cases} MM_P & (\text{if } P_u(l_{n+1}|l_n) > 0) \\ MM_G & (\text{else if } P(l_{n+1}|l_n) > 0) \end{cases}$$

Gao らの手法では、周期的習慣と遷移習慣を独立と捉えて掛け合わせて用いていたが、掛け合わせを行うと、周期的習慣、遷移習慣のどちらかが 0 に近い場合に、他方の確率が反映されないという問題がある。そこで、提案手法においては、 MU による訪問確率と MM による訪問確率の重み付き線形和で各場所への訪問確率を算出する。各習慣の重みは予測精度による検証から $0.3MU + 0.7MM$ とした。

予測するテストログの中には、非習慣行動のログも混在しているため、予測精度には限界があるが、習慣的行動の予測精度評価により、提案法による習慣抽出モデルがもたらす効果を定量的に評価できると期待される。テストデータ中の訪問の曜日・時間帯と、1 つ前の訪問場所を与えた上で全場所への訪問確率を算出し、そのうち上位 5 個(p@5)または 1 個(p@1)を出力することで、正答率を計算する。すなわち、 $n+1$ 番目の訪問場所 l_{n+1} を、 $P(l_{n+1}|l_n, u, t, d')$ が大きいものから k 個選ぶ p@k 評価を行う。 t' , d' はそれぞれ $n+1$ 番目のログの時間帯、曜日を示す。予測精度は正答回数/総予測回数で算出し、単位は %とする。総予測回数は Gowalla で 5,511 回、Brightkite で 32,254 回であった。

表 4 に各習慣を利用した行動予測精度をしめす。4 つの習慣を組み合わせた (e) MU_{PALL} を用いることにより、各習慣を単独で利用する (a)-(d) に比べ、精度が向上することが分かった。

(a)-(d) の各周期的習慣の予測性能を比較したものを図 11 に示す。縦軸は総予測数に対する占有率を示す。左の 2 つの棒グラフを見ると、Gao らの研究でも抽出していた MU_{PD} , MU_{PT} では誤答だが、提案モデルで考慮に加えた MU_{PTD} や MU_P では正答となつたログが、全体の 3-4% 存在することが分かる。また、 MU_{PT} では正答になるが、 MU_{PD} では誤答になるログが 10% 近くある。4 つの周期的習慣を適切に組み合わせることで、従来法では誤答となるログに対して正答できると期待される。図 11 において全てのモデルで正答だったものが Gowalla で 18.38%, Brightkite で

29.03%あり、全モデルで誤答だったものが Gowalla で 57.83%, Brightkite で 52.04% であった。いずれかのモデルで正答できたものは Gowalla で 42.17%, Brightkite で 47.96% であり、4 モデルの組み合わせ方法の改良により、最大でこの精度までの向上が期待できる。4 つの習慣を組み合わせて重み付き線形和を取る際の適切な重み w を、ユーザや時間に応じて学習させるなどの改良も考えられる。

本人と全ユーザの遷移習慣を組み合わせて利用した MM_{GP} との線形和を取った (f) を用いると、さらに精度が向上した。曜日考慮習慣、時間帯考慮習慣、個人遷移習慣をそれぞれ独立と捉えて掛け合わせた Gao らのモデル (g) の精度を上回っている。参考のために、提案した MU_{PALL} を周期的習慣として用いた上で、Gao らと同様に個人遷移習慣との掛け合わせを行った (h) を見ると、曜日・時間帯を独立と考えた Gao らの (g) よりも精度が高いことが分かり、遷移習慣との組み合わせ方法改良による効果を取り除いた場合においても、 MU_{PALL} が有効であることが分かる。以上のように、4 つの習慣を組み合わせた MU_{PALL} は、単独で用いる場合、遷移習慣と組み合わせて用いる場合の両方において効果があり、習慣度算出においても、同様の組み合わせを行うことが有効であると考えられる。遷移習慣の抽出方法、周期的習慣・遷移習慣の組み合わせ方法にも改良を加えることにより、Gao らの既存手法に比べて、p@5 で 10% 以上の精度向上が見られることが示せた。

5. 提案手法を活用した情報配信の検討

4.2 節で示したように、連続する 2 ログ間の習慣度は正の相関を持つ場合がある。従って、情報配信を行う直前のログの習慣度を利用し、習慣度が高いほど本人の履歴を、習慣度が低いほど他人の履歴を強く考慮した情報配信が考えられる。本人の履歴を利用した情報配信では、4.3 節で示した行動予測結果が活用できる。習慣度が低い場合には、本人の履歴ではなくジャンル情報などを活用することにより、本人の嗜好のみを反映することも考えられる。一方、他人の履歴を利用した情報配信について、4.2 節で述べたように、情報を配信したいユーザと類似した習慣度を持つ場所を推薦することが考えられる。一例として図 12 はアメリカのサンフランシスコ付近における、全ユーザのログを用いた場所毎の習慣度平均を表す。本稿で用いた Gowalla, Brightkite の 2 つのチェックインログには、評価実験で用いた日本国内だけでなく、世界中の場所に対するログが含まれている。そこで、本章では情報配信例を検討するためにサンフランシスコ付近のログを用いた。サンフランシスコ付近のログにおいても、空港や AT&T パークなどが非習慣的に訪れるやすい場所として、大学や駅が習慣的に訪れやすい場所として抽出されるなどの結果が確認できた。このような地図を利用すれば、例えば配信対象のユーザがサンフランシスコを非習慣的に訪れている場合、そのユーザは出張や旅行で現地を訪れている可能性が高いと判断し、比較的平均習慣度が低い Golden Gate Bridge などを推薦するなどのユーザの状況に応じた情報配信が実現できる。提案手法を用いれば、このように配信対象のユーザ自身のログが少ない場所においても、既に訪れたことがある他のユーザのログから各場所の平均習慣度を算出することができ、それを配信対象のユーザの習慣度と関連付けて活用することができる。具体的な場所の推薦だけでなく、図 12 のような、平均習慣度をプロットした地図を提示し、ユーザ自身が希望に合わせて訪問場所を選ぶ、などの活用方法も考えられる。

6. まとめと今後の予定

位置情報ログを活用した情報配信の満足度向上に向け、ユーザ行動理解の指標として行動の「習慣度」を定義した。情報配信における習慣/非習慣状態推定の必要性とその要件を検討し、時間的特徴を考慮した習慣抽出モデルと、従来よりも詳細な行動習慣度算出手法を提案した。加えて、チェックインログを用いた習慣度

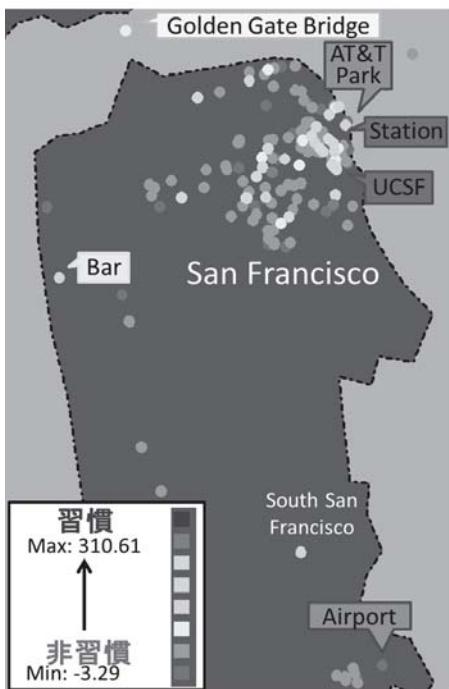


図 12: サンフランシスコ付近の習慣度平均

算出結果の定量的・定性的評価と情報配信への活用例を示し、提案法の有用性を示した。

今後は、習慣抽出において、曜日・時間帯に対する 1 週間毎の周期的習慣だけでなく、1 週間という単位を取り扱ったその他の周期的習慣や、習慣の変容を考慮していきたい。現在の手法では学習データに含まれる非習慣行動のログが大量に存在した場合、それらも習慣として学習してしまう可能性がある。習慣を 1 週間周期ではなく連続的に捉えることにより、短期間にしか存在しなかったログを非習慣として扱い、習慣抽出段階で無視することが可能になると期待される。また習慣の変容を考慮することにより、例えば A さんがある場所に 3 か月に一度訪問するという習慣がある場合に、1 年以上来なければ習慣が変容したとみなして習慣を更新するとともに、クーポン配布など何らかのアクションを行う、などの応用が考えられる。さらに、習慣度変化の状態遷移をモデル化することにより、人や場所に応じた確実性・意外性の考慮比率をより詳細に推定することも検討している。

【文献】

- [1] D. Ashbrook and T. Starner. Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing*, 7(5):275–286, 2003.
- [2] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, 2012.
- [3] E. Cho, S. A. Myers, and J. 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*, pages 1082–1090. ACM, 2011.
- [4] E. Davami and G. Sukthankar. Online learning of user-specific destination prediction models. *HUMAN JOURNAL*, 1(3):144–151, 2012.
- [5] H. Gao, J. Tang, and H. Liu. Mobile location prediction in spatio-temporal context. In *Nokia Mobile Data Challenge 2012 Workshop. p. Dedicated task*, volume 2, 2012.
- [6] T. Kurashima, T. Iwata, T. Hoshida, N. Takaya, and K. 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*, pages 375–384. ACM, 2013.
- [7] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 579–588. ACM, 2010.
- [8] K. W.-T. Leung, D. L. Lee, and W.-C. Lee. Clr: a collaborative location recommendation framework based on co-clustering. In *Proceedings of SIGIR*, pages 305–314, 2011.
- [9] E. Malmi, T. M. T. Do, and D. Gatica-Perez. From foursquare to my square: Learning check-in behavior from multiple sources. 2013.
- [10] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 637–646. ACM, 2009.
- [11] B. Shaw, J. Shea, S. Sinha, and A. Hogue. Learning to rank for spatiotemporal search. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 717–726. ACM, 2013.
- [12] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 363–372. ACM, 2013.
- [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*, pages 791–800. ACM, 2009.
- [14] D. Zhou, B. Wang, S. M. Rahimi, and X. Wang. A study of recommending locations on location-based social network by collaborative filtering. In *Advances in Artificial Intelligence*, pages 255–266. Springer, 2012.

林 亜紀 Aki HAYASHI

2010 年、お茶の水女子大学理学部情報科学科卒業。2012 年、同博士前期課程修了。同年、日本電信電話株式会社入社。以来、NTT サービスエボリューション研究所にて、可視化の研究開発に従事。お茶の水女子大学大学院人間文化創成科学研究科理学専攻博士後期課程在学中。日本データベース学会会員。

松林 達史 Tatsushi MATSUBAYASHI

2000 年、京都大学理学部物理学科卒業。2002 年 10 月より 2 年半、理化学研究所非常勤研究員。2005 年、東京工業大学大学院理工学研究科地球惑星科学専攻博士課程修了。同年、日本電信電話株式会社入社。以来、NTT コミュニケーション科学基礎研究所、NTT サービスエボリューション研究所にて、可視化の研究開発に従事。博士（理学）。

澤田 宏 Hiroshi SAWADA

1991 年、京都大学工学部情報工学科卒業。1993 年、同修士課程修了。同年、日本電信電話株式会社入社。以来、NTT コミュニケーション科学基礎研究所にて、VLSI 向け CAD および計算機アーキテクチャの研究に従事。2000 年より、信号処理、特にブラインド音源分離の研究に従事。2009 年より、知能創発環境研究グループリーダ。2013 年より、サービスエボリューション研究所ヒューマンアナリシスプロジェクトグループリーダ、現在に至る。2001 年、京都大学博士（情報学）。電子情報通信学会、日本音響学会、IEEE 各会員。