

アノニマイズされた行動履歴に基づく 行動パターン検索方式の提案

川田 正明^{†1} 小川 克彦^{†2}

今日、ユビキタスコンピューティング環境の整備は急速に進んでいる。中でも特に位置情報の取得の容易性は、GPS 付き携帯電話に代表されるようなデバイスの普及によって非常に高くなっている。また、UGC (User-generated Content) と呼ばれる新しい種類のサービスの登場により、緯度・経度などのデータが含まれた位置情報付きのコンテンツが増加の傾向にある。このような状況の中、行動履歴を用いた新しいコンテンツを検索するシステムは、ユーザの好みを反映したサービスを提供することにより利便性の向上が期待できる。しかしながら、位置情報付きのコンテンツは個人情報となるためにプライバシー保護の対象となる。本論文では、緯度・経度などの行動履歴をプライバシー保護された行動パターン「アノニマイズド行動パターン」に変換し、この行動パターンを用いて似た行動パターンのユーザを探す行動パターンの検索方式を提案する。本方式の評価にあたっては 36 名の被験者の行動履歴 (位置・時間・性別・年齢・行動の状態) を用いて匿名性と情報検索精度に関して考察した。その結果、本方式が有効に機能するための条件を明らかにした。

Proposal of an Anonymized User Behavior-based Information Retrieval System

MASAAKI KAWATA^{†1} and KATSUHIKO OGAWA^{†2}

Recently, it has become easy to obtain location data, including latitude and longitude information, using a mobile phone equipped with a global positioning system (GPS) receiver. In addition, the amount of readily available user-generated content (UGC) that includes location data continues to grow. There will be great benefits to the end-user if this new type of online content can be effectively searched. However, the availability of location data may present privacy problems. The present paper proposes an anonymized user-behavior-based information retrieval system. The proposed system provides a method of creating a user-behavior-based feature pattern for each user through privacy protection from user behavior history. In addition, in the present paper, we evaluate whether the proposed system will perform effectively with respect

to anonymity and information retrieval precision with user behavior histories. These user behavior histories consist of 9,324 records gathered from 36 people over 13 days. As a result, this study investigates the most effective conditions for the operation of this system working and demonstrates with factors are location and time for creating anonymized user behavior-based feature patterns.

1. はじめに

今日、ユビキタスコンピューティングを取り巻く環境の整備は急速に進んでいる。特に、GPS (Global Positioning System) 付きの携帯電話や Wi-Fi のアクセスポイントの電波強度を用いて位置情報を取得する「PlaceEngine¹⁾」などに代表されるように、誰でも容易に位置情報を取得可能な環境が整いつつある。

また、UGC (User-generated Content) と呼ばれる新しい種類のサービスの登場により、ユーザ自らコンテンツを投稿する量が増加している。その中で、今自分自身が行っていることを一言で投稿する Twitter²⁾ や、写真投稿の Web サイトである Flickr³⁾ のようなサービスでは、ユーザ自ら位置情報を積極的に登録している。

さらに、Yahoo! では商用サービスとして Fire Eagle⁴⁾ (現在プライベート β テスト中) というサービスを開始しようとしている。これは位置情報付きの Twitter と呼ばれているもので最初から位置情報付きで現在行っていることを投稿していくサービスである。

このような状況の中、ユーザごとの位置情報を含む行動履歴を取得し行動パターンを分析することで、ユーザの趣味嗜好に合った情報提供を行うことが可能な新しいタイプの情報検索サービスを構築することが、ユーザにとって有益なサービスになるものと考えている。

しかしながらその半面、緯度・経度などの位置情報は個人情報となりうるために、公開されることやそのまま送信すると個人が特定される恐れがある。特に、位置情報の連続データは、人の行動履歴となるために、単一の位置情報より個人が特定しやすくなるプライバシーの問題が発生し、プライバシー保護が必要となる。

本論文では、位置情報を含む行動履歴から匿名性の高いアノニマイズされた行動パターンを生成し、その行動パターンに近い他の行動パターンを検索する方式 (アノニマイズド行動

^{†1} 慶應義塾大学大学院政策・メディア研究科
Graduate School of Media and Governance, Keio University

^{†2} 慶應義塾大学環境情報学部
Faculty of Environment and Information Studies, Keio University

パターン検索方式と呼ぶ)を提案する。また、本方式のシステムの実装を行い、36名の被験者によって集められた9,324件の実データである行動履歴を基に、行動パターンの匿名性と行動パターンによる行動履歴の検索精度の評価を行った。

なお、プライバシーとは、個人、グループまたは組織が自己に関する情報を、いつどのように、またどの程度他人に伝えるかを自ら決定できる権利と定義されている⁵⁾。そこで本論文では、このプライバシーの定義の中で他人に伝える権利の度合いに関して、本論文では行動履歴の2次利用を考慮し、他人に伝えてもその行動履歴単体では、特定の個人と結び付かないこととする。

2. 関連研究

関連研究として、データマイニングを基にした関連情報を検索するための研究も行われているが、ここでは位置情報を用いたContext-awareアプリケーションとして、Cyberguide、C-MAP、プライバシー保護を目的とした関連研究としてpawS、Fixerについて述べる。

2.1 Cyberguide

Cyberguide⁶⁾は、Context-awareなモバイルアプリケーションを提案したAbowdらによる携帯情報端末向けのユーザの位置情報に応じて個人に最適化された情報提供を行うためのシステムである。位置情報は、野外ではGPS、屋内では赤外線ビーコンを用いて取得している。

このCyberguideは、行動履歴は用いてはいないが、位置情報などの個人情報によるユーザ趣向のサービスを提供しようとしている。しかし、その半面プライバシー保護に対する対応は不十分になっている。

2.2 C-MAP

C-MAP (Context-aware Mobile Assistant Project)⁷⁾は、角らによって構築された、展示会場内で位置情報および個人情報に基づき個人の趣味趣向に応じた情報提供を行うためのガイドシステムである。このガイドシステムは、Cyberguideと似ているが、位置情報のほかに事前登録を行った趣味などの個人情報を用いて情報提供を行う点が異なる。なお、C-MAPが用いている個人情報とは、IDなどのユーザ情報と事前に登録された興味に関する情報である。また、位置情報はActive Badge System⁸⁾による赤外線ビーコンを用いて取得している。

このC-MAPは、Cyberguideと同様に、個人情報によるユーザ趣向のサービスを提供しようとしているが、プライバシー保護に関して解決方法は提示しておらず非対応のままになっている。

2.3 pawS

pawS⁹⁾はLangheinrichが提案した、ユビキタス環境においてPrivacy Policyに基づく位置情報の保護方法である。ユーザが設定したPrivacy PolicyのPrivacy Preferenceの条件を満たす場合のみ位置情報を公開する。この方法は、ユーザごとにPrivacy PolicyのPrivacy Preferenceを設定する必要がある。また、行動情報の検索を行う際、非公開の位置情報があるために検索できないデータが存在することになる。

2.4 Fixer

Fixer¹⁰⁾は、中西らが提案した位置情報保護のためのフレームワークである。公開する位置情報の粒度を動的に変更することにより、ある条件下では非公開が公開に変化する柔軟性を持つ。

Fixerは、pawSのような設定の複雑さは回避している。しかし行動情報を検索する際は、pawSと同様に非公開のデータが存在するために検索できないデータが存在することになり、有用なデータを取り逃す問題が残る。

以上のように、ユーザ趣向のサービスを行う場合は、位置情報、時間、ユーザIDなどの個人情報が必要となる。CyberguideやC-MAPのようなユーザ趣向のサービスは、個人情報(位置情報、時間、ユーザID)といった情報から誰がどこにいるのか推測されやすい。そのためプライバシー保護を行う必要があるが、pawSやFixerのようなプライバシー保護を行うと、情報提供を行う際、扱える位置情報などのデータが制限されるのでユーザ趣向のサービスに制限が発生するという問題がある。

本方式では、あらかじめ蓄えてある位置情報と時間を含めた時空間データの行動履歴を用いて、匿名性の高い行動パターンを生成する。この生成した行動パターンを用いることで、ユーザ趣向のサービスを行ううえで発生するプライバシーの問題と、扱えるデータの制限に関する問題を解決する。

3. アノニマイズド行動パターン検索方式の提案

3.1 コンセプト

まず、最初にアノニマイズド行動パターン検索方式のコンセプトを述べる。

本システムでは、GPSなどにより取得される緯度・経度・時間などのプライバシーで保護されていないデータが含まれるユーザの行動履歴から、それぞれのユーザの行動の特徴を代表する「特徴点」を抽出する。このユーザの行動の特徴をとらえた「特徴点」を使い、特徴点どうしの距離を求める。この距離を用いることによって、人それぞれの行動パターンの比較を行う。

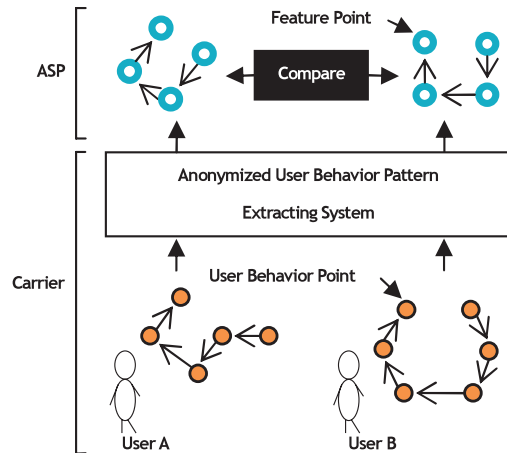


図 1 基本コンセプト
Fig. 1 Basic concept.

プライバシー保護された「特徴点」の集合である行動パターンを「アノニマイズド行動パターン」と定義する(図 1 参照)。

プライバシー保護の観点からは、ケンブリッジ大学の Alastair Beresford¹¹⁾ は、「どこでどのくらい時間をすごしたか」を集計することでユーザを識別することができる可能性を述べている。

このような危惧を回避するためにも元のデータからユーザの行動の特徴を表す「特徴点」へ置き換えることにより、元のデータを推測されにくくする。

さらに、この「特徴点」の集合を作ることにより、プライバシーの保護を行いつつユーザの行動の特徴をとらえた行動パターンを作る。これにより、オンライン上のショッピングサイトである Amazon の購買履歴に応じてそれぞれの人に合った商品をレコメンドするシステムのように、ユーザの行動履歴を用いて安全で確かな情報を提供することができると期待される。

なお、このアノニマイズド行動パターンの利用については、センサによりリアルタイムにデータを取得し、デバイスのアクセス制御によるプライバシー保護を行う The Context Toolkit¹²⁾ のような方式ではなく、ユーザはあらかじめ蓄えられた過去の行動履歴から生成するアノニマイズド行動パターンを選択することで情報提供を行うことを想定している。

実行履歴ではなく、このプライバシー保護された行動パターンを用いた情報検索は、ユー

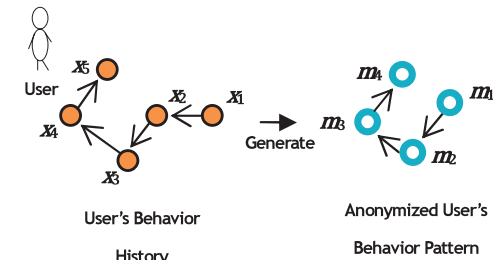


図 2 アノニマイズド行動パターンの生成
Fig. 2 Generating an anonymized user's behavior pattern.

ザだけでなくサービスの運用者から見ても利点がある。それは、ユーザ名と行動パターン、そして行動パターンに関連付けられたコンテンツのみで情報検索を行えるようになるという点である。これは、1 度プライバシー保護された行動パターンを生成すれば、個人情報である実際の行動履歴は必要なくなるためである。

また、行動パターンを生成するための運用者 (Carrier) と、行動パターンに応じたコンテンツを配信する ASP (Application Service Provider) という役割ごとにサービスを分割することもできる。このような体制にすることにより、ユーザ名などの個人情報管理をサービス提供者 (ASP) で行えるため、Carrier にユーザ名の認証を得る必要がなくなる。

3.2 検索アルゴリズム

次に、ユーザの行動履歴からアノニマイズド行動パターンを抽出するアルゴリズムを述べる。匿名性の高い行動の特徴点の集合であるアノニマイズド行動パターンを抽出するアルゴリズムとして、本研究では教師なし学習である SOM (Self-Organizing Maps)¹³⁾ を使用した。

SOM を使用した理由として、教師なし学習を行う際、学習ステップ数を任意に変更可能な点があげられる。この学習ステップ数を大きくすれば精度が上がるがプロセッサの使用量が多くなり、逆に小さくすれば精度が下がるがプロセッサの使用量を少なくすることができる。これにより、ASP や Carrier の使用環境に合わせた環境構築が可能になる。

本アルゴリズムではまず、実際の人の行動履歴の 1 点 x_i から、アノニマイズド行動パターンを構成している特徴点である m_j を生成する。このとき、 i は行動履歴の数で、 j は特徴点の数になる(図 2 参照)

SOM で定義されている Kohonen の学習則は、式 (1) で表されるものと同様のものを使用し、学習ステップ数は、10,000 ステップとした。

$$m_j(t+1) = m_j(t) + h_{ci}(t)[x(t) - m_j(t)] \quad (1)$$

このとき、 x は行動履歴 x_i から Winner-take-all に基づき選択された入力ベクトルで、 m_j は出力ベクトルである。

この入力ベクトル群を基に Kohonen の学習則を用いて学習することで出力ベクトル m_j の配列がアノニマイズド行動パターンとなる。なお、Kohonen の学習則を用いる前に入力ベクトルの配列を、時間 t をキーとしてソートすることにより時系列順に揃える。

h_{ci} は式 (2) の近傍関数と呼ばれるガウス関数で定義した。

$$h_{ci} = \alpha(t) \cdot \exp\left(-\frac{\|r_c - r_i\|}{2\sigma^2(t)}\right) \quad (2)$$

近傍関数のパラメータ $\alpha(t)$ はスカラ値の学習率係数、パラメータ $\sigma(t)$ はカーネルの幅を定義する。 $\alpha(t)$ と $\sigma(t)$ の両パラメータは時間の単調減少関数である。 r_c は m_i の位置であり、 r_i は m_i の周辺の位置を表す。

アノニマイズド行動パターンを生成するにあたり、緯度 (lat)・経度 ($long$)・時間 (t)・ユーザの性別 (g)・年齢 (a) の 5 つのデータを、次の式 (3) に示すように SOM の学習に適用させる際の入力ベクトルとした。

$$x_i = \begin{bmatrix} lat_i \\ long_i \\ t_i \\ g_i \\ a_i \end{bmatrix} \quad (3)$$

$$lat_i = \frac{(latitude_i - \min latitude)}{(\max latitude - \min latitude)} \quad (4)$$

$$long_i = \frac{(longitude_i - \min longitude)}{(\max longitude - \min longitude)} \quad (5)$$

入力ベクトルの要素 lat は緯度を表す。特徴点を効果的に出力するよう lat の値を、最小値 0 から最大値 1 の間に収まるよう正規化した。

式 (4) の $latitude$ は緯度、 $\min latitude$ は最小の緯度の値、 $\max latitude$ は最大の緯度の値である。

入力ベクトルの要素 $long$ は経度を表す。要素 $long$ も要素 lat の場合と同じく、効果的に最小値 0 から最大値 1 の間に収まるようにした。

式 (5) の $longitude$ は経度の値、 $\min longitude$ は最小の経度の値、 $\max longitude$ は最大の経度の値である。

また、入力ベクトルの要素 t は時間を表す。要素 t は、時と分だけを入力値とした。緯度・経度のとく同じく特徴点を効果的に出力するため式 (6) の示すように、最小値 0 から最大値 1 の間に収まるようにした。なお、 $hours$ が時、 $minutes$ が分の値である。

$$t_i = \frac{(hours_i \cdot 60 + minutes_i)}{24 \cdot 60} \quad (6)$$

さらに、入力ベクトルの要素 g は性別を表す。式 (7) に示すように男性を 0、女性を 1 とした。

$$\begin{cases} 0, & \text{if gender is male} \\ 1, & \text{otherwise} \end{cases} \quad (7)$$

最後に、入力ベクトルの要素 a は年齢を表す。特徴点を効果的に出力するために緯度・経度の場合と同じく最小値 0 から最大値 1 の間に収まるよう式 (8) とした。 age は年齢、 $\min age$ は最小の年齢、 $\max age$ が最大の年齢である。

$$a_i = \frac{(age_i - \min age)}{(\max age - \min age)} \quad (8)$$

Kohonen の学習則を適用して出力されたアノニマイズド行動パターン M と別のアノニマイズド行動パターン N を比較するときには、式 (9) を用いて距離を計量する。

$$dist = \sum_{i=1}^n \sqrt{\sum_{j=1}^m (M_{ij} - N_{ij})^2} \quad (9)$$

本方式では式 (9) を用いて複数ユーザのアノニマイズド行動パターン間の近さを計量する。

4. システムの構築

4.1 システムの概要

実際に、アノニマイズド行動パターンを生成、行動パターンどうしを計量し比較するシステムの構築を行った。本システムでは、主にアノニマイズド行動パターンを用いた情報検索を行うためのフレームワークとして機能するよう設計されている。また、ASP や Carrier で用いられることを想定して、3 階層の MVC (Model-View-Controller) アーキテクチャで述べられている Controller と Model を主に提供する。この MVC アーキテクチャを利用することにより、用途に合わせて OS 依存の GUI アプリケーションや Web アプリケーショ

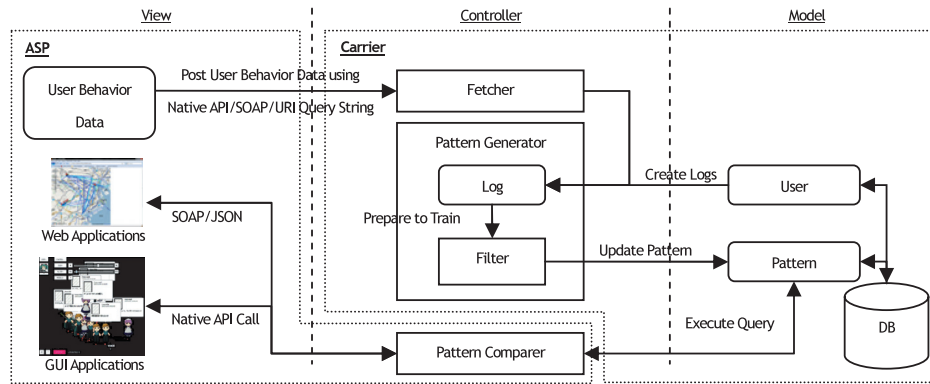


図 3 システム概要
Fig. 3 System overview.

ンを柔軟に選択し構築することができる (図 3 参照)。

本フレームワークは主に 3 つのビルディングブロックから構成される。まず View から行動履歴 (緯度・経度・時間・性別・年齢) を取得する「Fetcher」、実際に行動履歴からアノニマイズド行動パターンを生成し DB にアノニマイズド行動パターンを格納する「Pattern Generator」、そして生成されたアノニマイズド行動パターンどうしを比較し検索を行い、検索結果を View に伝える「Pattern Comparer」である。

4.2 動作概要

この 3 つのビルディングブロックの動作は、アノニマイズド行動パターンを生成する際は、「Fetcher」と「Pattern Generator」、アノニマイズド行動パターンを用いて検索する際は「Pattern Comparer」が使用される。

なお、Fetcher、Pattern Generator と Pattern Comparer は C# と C++ 言語を用いて実装を行った。DB は Microsoft SQL Server 2005 を用いた。

最初に行動履歴から行動パターンの生成を行う際には、まずユーザ側の View から Controller の「Fetcher」に対して行動履歴の送信を行う。このとき、行動履歴を送信する手段として、ローカル用に Native API のほか、Web 上で広く使われている HTTP 上で通信する SOAP (Simple Object Access Protocol)¹⁴⁾ と JSON (JavaScript Object Notation)¹⁵⁾ フォーマットをサポートする。Fetcher は行動履歴取得後、DB からユーザ情報 (ID) を照合し、「Pattern Generator」が読み取り可能な行動履歴のデータ形式である Log を作成する。

次に「Pattern Generator」は、この Log を用いて実際にアノニマイズド行動パターンを行動履歴から抽出し、その結果であるデータ Pattern を生成する。この生成されたアノニマイズド行動パターンである Pattern は DB に格納され、行動履歴 Log は Pattern を DB に格納後はプライバシー保護のため削除する。

アノニマイズド行動パターンを用いて検索を行う際には、ユーザ側 View からクエリとしてのアノニマイズド行動パターンを指定し、「Pattern Comparer」に送信する。なお、この View から Controller の「Pattern Generator」への通信は、View と「Fetcher」の通信のときと同じく、Native API、SOAP、JSON をサポートする。クエリを受け取ると「Pattern Comparer」は、クエリと DB に格納されているアノニマイズド行動パターンとの距離を求め比較を行う。この距離をキーにソートしたものを近いものから順にユーザ側の View に返却する。つまり、クエリとして発行された行動パターンと似ている別の行動パターンをユーザに対して返すことになる。また、このような似た行動パターンを検索する以外に、逆に似ていない行動パターンを探すことも可能である。

5. 評価

本システムの評価にあたって、実際の人の行動履歴を用いて定性評価と定量評価の両側面から行った。この行動履歴とは、ユーザの行動である緯度、経度、時間、性別、年齢、行動の状態を時系列順に記したデータである。

定量評価では、アノニマイズド行動パターンの匿名性と、アノニマイズド行動パターンを用いた際の検索精度の検証を行った。匿名性とは、アノニマイズド行動パターンから元の実際の行動履歴を推定できるか否かを検証するものである。検索精度の検証とは、アノニマイズド行動パターンどうしを比較して得た似た行動パターンが、生成元の実際の行動履歴に則しているか検証するものである。

定性評価ではシステムに入力したクエリとシステムが出力した検索結果を実際のユーザの行動と比較して構築したシステムが出力した結果が正しいものであるかの検証を行った。

5.1 評価データ

アノニマイズド行動パターンを生成する際に入力する人の行動履歴には、情報大航海プロジェクト¹⁶⁾ で収集されたデータを使用した^{17),18)}。

このデータは、被験者 36 名から集められた 9,324 件のデータからなる。なお、このデータはエリクソン¹⁹⁾ のライフステージ/ライフサイクル論に基づき年代・性別を均等にするよう被験者を選定し収集されたものである (表 1 参照)。

表 1 被験者の概要
Table 1 Overview of test users.

	男性	女性	合計
成年層(18-29 才)	6	7	13
壮年層(30-44 才)	8	6	14
老年層(45-60 才)	4	5	9
合計	18	18	36

表 2 1 日ごとの行動履歴の概要
Table 2 Overview of user behavior histories for a day.

MIN	AVG	MAX
2	10.7	34

具体的には年齢層を，成年層（18-29 歳），壮年層（30-44 歳），老年層（45-60 歳）と 3 つに分け，成年層が 13 名，壮年層が 14 名，老年層が 9 名という構成で収集された。

また，調査期間については平日と休日を含む 2008/2/5（火）から 2008/2/17（日）までの計 13 日間にわたって行われたものである。

なお，1 日ごとの行動履歴は被験者辺りの行動履歴の数は，最小が 2 件，最大が 34 件，そして平均は 10.7 件であった（表 2 参照）。

さらに，下記のようなこの行動履歴には何をしていたかを示す 15 の行動内容のいずれかが関連付けされている。

- | | |
|---------------|--------------|
| 1 運動 | 2 仕事 |
| 3 家事/育児/介護 | 4 団らん中 |
| 5 勉強 | 6 移動 |
| 7 食事 | 8 買物 |
| 9 趣味・娯楽 | 10 暇つぶし・ぶらぶら |
| 11 人や何かを待っている | 12 おしゃべり |
| 13 休憩中・リフレッシュ | 14 準備・身支度 |
| 15 その他 | |

この 9,324 件の実際の人の行動履歴から 1 日ごとに時間を区切りアノニマイズド行動パ

ターンを生成した。その結果，生成した行動パターンの数は 445 パターンとなった。

5.2 アノニマイズド行動パターンの匿名性の評価

アノニマイズド行動パターンを実際の行動履歴と比較することによる匿名性の検証を行った。

具体的には，出力する特徴点の数（Number of Feature Points）を変化させながら，445 件ある生成したアノニマイズド行動パターンと実際の行動履歴の距離である実空間上での 2 点間の距離および時間差を比較した。ここで距離とは，アノニマイズド行動パターンの特徴点と実際の行動履歴の 1 地点が最も近い距離の平均である。この平均の距離が長ければ実際の行動履歴との誤差は大きくなり匿名性は高くなる。逆に平均の距離が短ければ実際の行動履歴との誤差は小さくなり匿名性は低くなる。

なお，アノニマイズド行動パターンを生成する際，実空間上での位置（緯度・経度）と時間のほかに，性別と年齢を考慮して生成したが，つねに同じ値となる性別や年齢のようなものは検索アルゴリズムで用いた SOM の特性上，入力と出力が同じになり匿名性がなくなる結果となった。よって，今後は実空間上の位置と時間の匿名性のみ配慮する。

その結果，出力する特徴点の数が 5 点以上の場合には，実際の位置より 106 m 以上の誤差があることが分かった。

また，特徴点の数が 6 点の場合は 62 m，7 点の場合は 39 m，8 点の場合は 26 m，9 点の場合は 15 m となり誤差は小さくなり，特徴点の数が 15 点を超えると誤差が 3 m 以下になった（図 4 参照）。

この結果から匿名性の必要性が高い都市部を考えると，都市部である渋谷や新宿の 1 ブロックのおおよその大きさである 50 m より大きな誤差が必要である。よって特徴点は 6 点以下でアノニマイズド行動パターンを生成するのが望ましいことが分かる。逆に，特徴点の数を 15 以上にすると距離が 0 に近くなり，匿名性はなくなる。

次に，実際の行動履歴とアノニマイズド行動パターンとの時間差を，特徴点の数を変化させながら測った。

その結果，どの特徴点の場合も実際の時間と比較した場合，140 分以上の誤差があることが分かった（図 5 参照）。

この結果から実空間上の位置の匿名性を高くしたときと比較すると，特徴点の数を制限することは必要ないことが分かる。また，特徴点の数が多くなれば匿名性は高くなるが，ただ単調減少するわけではない。1 日ごとの行動履歴の平均である 10.7 個あたりまでは単調減少するが，それ以降では同じ位置で違う時間に行動履歴をつけているために，時間だけが違う特徴点が生じられることになり，50 分から 60 分という振幅が発生する。

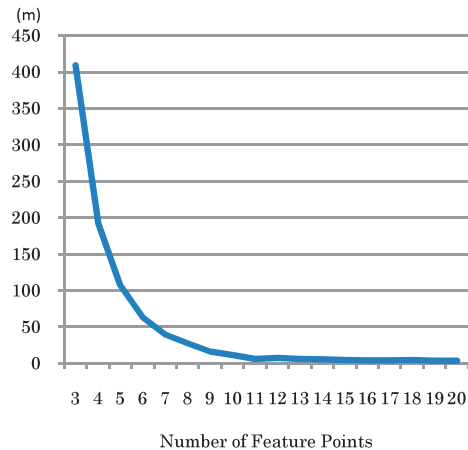


図 4 実際の行動履歴とアノニマイズド行動パターンとの距離 (緯度・経度)

Fig. 4 Distance between user behaviors and anonymized user-behavior patterns (Latitude and Longitude).

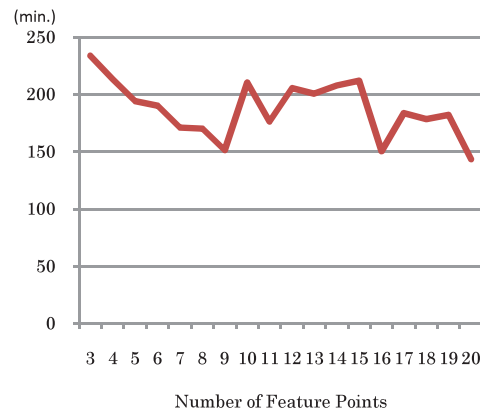


図 5 実際の行動履歴とアノニマイズド行動パターンとの距離 (時間)

Fig. 5 Distance between user behaviors and anonymized user behavior patterns (Time).

これらのことから、特徴点が 6 点以下であればいくつでも高い匿名性を維持できると考えられる。

5.3 アノニマイズド行動パターンを用いた場合の検索精度の評価

次に、アノニマイズド行動パターンを用いた場合の検索精度を検証した。これはアノニマイズド行動パターンを用いても、実際の行動履歴に則した正しい検索ができるのかを検証するものである。

この検索精度の検証では、あらかじめ実際の行動履歴から生成された 445 パターンあるアノニマイズド行動パターンの中から検索キーとなるパターン 1 つとそれに対応する検索結果として正しいパターンの集合を用意する。このようなテストパターンを用いて検索された検索結果から、適合率 (Precision) と再現率 (Recall) を求める。この 2 つを検索精度の指標とする。

なお、検索キーに含まれる行動パターンと同じ地域が含まれる行動パターンの検索結果を正しいものとした。

評価では、新宿から渋谷へ移動する行動範囲が狭いタイプと、東京都から神奈川県へ移動する行動範囲が広いタイプの 2 つの検索精度を検証した。この 2 つのタイプの行動は、使用した実際の行動履歴に多くみられたものである。この 2 つのタイプの行動に対する適合率と再現率は、それぞれ検索クエリは 5 回発行を行い、その平均を結果とした。

最初に行動範囲が狭い場合の検索精度であるが、アノニマイズド行動パターンの特徴点の数が多くなるに従い、検索精度は向上することが分かった (図 6 参照)。

高い匿名性が得られる特徴点が 6 点以下の場合の検索精度としては、特徴点が 5 点の場合は適合率が 80% で再現率が 62%、6 点の場合は適合率が 84% で再現率が 65% となった。特徴点が 3 点、または 4 点の場合は再現率が 70% 台まで下がり適合率も 50% 台まで下がる結果となった。

この結果から特徴点が 5 点、または 6 点の場合に有効な匿名性と検索精度を得られることが分かった。また、行動範囲が狭い場合は、スケールの狭い同じ地域を行動しているため特徴点の数が少ない場合でも特徴を得やすい。このため検索精度は高くなるが、逆に特徴点を多くしても同じ地域を行動しているため特徴を得ることはできないことも分かる。

次に、行動範囲が広いタイプの検索精度の結果であるが、アノニマイズド行動パターンの特徴点の数が多くなるに従い検索精度は向上する (図 7 参照)。

高い匿名性が得られる 6 点以下の場合の検索精度は、5 点の場合は適合率が 82% で再現率が 59%、6 点の場合も同じく適合率が 82% で再現率が 59% となった。特徴点が 3 点、ま

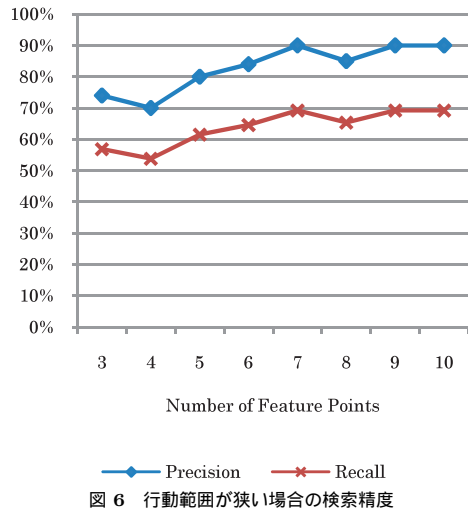


Fig. 6 Precision and Recall: User activities of the small scale case.

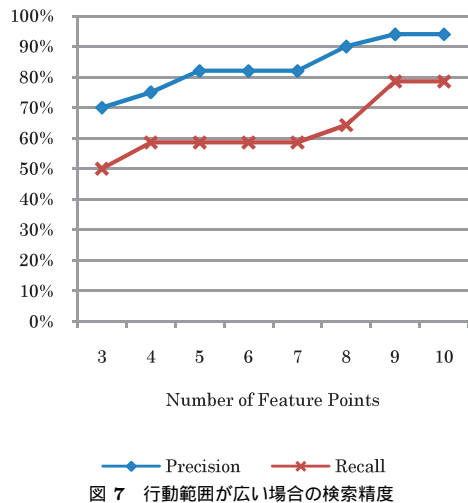


Fig. 7 Precision and Recall: User activities of the large scale case.

たは4点の場合は適合率が70%台まで下がり、再現率は50%と適合率ほど下がることはなかった。

この結果から、特徴点が多い方が検索精度は上がるものの特徴点が5点以上でも適合率が82%以上、再現率が59%以上であることから、有効な匿名性と行動範囲が狭い場合と同じような検索精度を維持できる。

このようにアノニマイズド行動パターンを使った検索で、行動範囲の違いに関係なく有効な検索精度を得ることができた。また、行動範囲の違いに関係なく匿名性を維持しつつ有効な検索精度を得るためには、特徴点の数を、5点または6点に設定するのが望ましいことが分かった。

5.4 アノニマイズド行動パターンの生成におけるパフォーマンスの評価

このアノニマイズド行動パターン生成に関して、匿名性、検索精度のほかに、生成にかかる処理時間の評価を行った。この処理時間の評価では、445パターンの生成にかかる処理時間を特徴点数ごとに計測した。そこから、1つのアノニマイズド行動パターンを生成するためにかかる処理時間の平均を出した。なお、処理時間を計測する際に使用した環境は、CPUがクロック周波数の2.2GHzであるIntel Core 2 Duo T7500、メモリは2GBのものを使用した。

その結果、処理時間は特徴点の数に比例して長くなり、匿名性を維持しつつ有効な検索精度を維持できる特徴点の数である、5点の処理時間は18.25ミリ秒、6点の処理時間は22.82ミリ秒となった。この処理時間は高いリアルタイム性を保持しつつ生成可能な処理能力を持っていることから、システムを運用するうえで十分な処理能力を持っていることが分かる。

このように、アノニマイズド行動パターンの生成におけるパフォーマンスは、十分に実用に耐えるものであることが分かった。また、今回の生成処理で用いているSOMは、特徴点の数を増加させる場合はシーケンシャルにデータを読み書きする必要があることから処理時間は特徴点の数に比例して計算コストが大きくなる。しかし、入力する行動履歴はSOMの特性上、シーケンシャルに呼び出す必要がないのでスケーラビリティは保証できる。さらに1つのパターンを生成する際、他のデータを共有することなく生成可能な特徴を持っているため、並列処理に向いておりスケール可能なアルゴリズムともいえる。

5.5 ユーザの実際の行動内容の比較

最後に、検索クエリとしてのアノニマイズド行動パターンに関連付けられている行動内容と、検索結果の行動パターンに関連付けられている行動内容を比較する。この比較により、似ている行動内容を取得できているか検証を行う。

表 3 行動範囲が狭い場合の検索クエリと検索結果
Table 3 Search query and results of small scale activities.

行動範囲	クエリ	結果 1	結果 2	結果 3
が狭い場 合(新宿か ら渋谷の 範囲)	3. 家事/育 児/介護	3. 家事/育 児/介護	6. 移動	3. 家事/育 児/介護
	6. 移動	6. 移動	7. 食事	6. 移動
	7. 食事	7. 食事	9. 趣味・娛 楽	7. 食事
	9. 趣味・娛 楽	9. 趣味・娛 楽	13. 休憩 中・リフレ ツシュ	8. 買物
	11. 人や何 かを待って いる	11. 人や何 かを待って いる		9. 趣味・娛 楽

この検証では、新宿から渋谷へ移動する狭い範囲を行動するタイプと、神奈川県から東京都へ移動する広い範囲を行動するタイプの 2 種類を用いた。この 2 種類のいずれの場合でも、それぞれ検索クエリは 5 回発行し、その妥当である結果の中から検索クエリに対し特徴的な行動内容を抜き出し、検索クエリと検索結果を比較した。

最初に行動範囲が狭いタイプの結果であるが、行動範囲が広いタイプと同様に移動や食事がクエリから同じ行動として取得できていることが分かった。また、趣味・娯楽といった行動の状態まで取得できていることが分かった。これは、渋谷や新宿といった街が多くの人には趣味や娯楽目的で利用されることが多いことが起因していると考えられる。つまり、場所と人の目的は深い関係にあるので、行動履歴はユーザの行動内容を表現していることが多いといえることが結果から分かる。

しかし、家事/育児/介護といった今回アノニマイズド行動パターンの際に使用した位置・時間・年齢・性別といった情報以外の行動履歴が含まれる場合は、当然ながら検索結果に含まれない(表 3 参照)。

次に、行動範囲が広いタイプの結果としては、移動や食事などについては、同じ行動の状

表 4 行動範囲が広い場合の検索クエリと検索結果の行動内容
Table 4 Search query and results of large scale activities.

行動範囲	クエリ	結果 1	結果 2	結果 3
が広い場 合(神奈川 県から東 京都の範 囲)	2. 仕事	2. 仕事	7. 食事	7. 食事
	6. 移動	3. 家事/育 児/介護	6. 移動	6. 移動
	7. 食事	6. 移動	10. 暇つぶ し・ぶらぶ ら	9. 趣味・娛 楽
		7. 食事	13. 休憩 中・リフレ ツシュ	

態がクエリから取得できていることが分かった。しかし、仕事の状態が入っているクエリに対して、趣味・娯楽といったような望ましくない項目が検索結果として得られてしまうことも分かった。これは現在、平日と休日を分けてランキングを行うといったような処理をしていないため起きているものであると考えられる(表 4 参照)。

これらのことから、アノニマイズド行動パターンによる行動パターン検索は、アノニマイズド行動パターンを生成する際に使用する個人情報が行動内容を示すものであれば、有効に機能することが分かった。また時間については、平日と休日の区別をするなど、より考慮しなければならないことも分かった。

6. まとめと今後の課題

本論文では、ユーザの行動履歴から、プライバシー保護を行いつつ検索精度の高い行動パターン検索を行うために、アノニマイズされたユーザの特徴を示す行動パターンを用いた検索方式を提案した。

この方式では、実データである緯度・経度、時間、性別、年齢が含まれたユーザの個人情報から、アノニマイズされた行動パターンを生成する。この生成した行動パターンを検索するシステムの構築を行い、実際の行動履歴と比較しながらアノニマイズされた行動パターンの匿名性と検索精度について評価を行った。

その結果、今回提案したアノニマイズド行動パターン検索方式の有効性は以下のような条

件下で有効であることが分かった．

- 高い匿名性と検索精度を保つにはアノニマイズド行動パターンを構成する特徴点の数が5点, または6点とする．
- アノニマイズド行動パターンを生成する際の行動履歴の各要素の値は, 性別や年齢のような一定の要素でなくリアルタイムに変化する位置や時間を表すデータを使用する．
- アノニマイズド行動パターンを生成する際の行動履歴の要素は, 行動内容に密接に関連する位置を用いる．

最後の位置情報(緯度・経度)である特定の場所と人の行動目的は合致する場合が多い．たとえば, 渋谷などの街の特徴として, 今回の評価に用いた行動履歴からは趣味や娯楽目的で訪れる場合が多い．よって, 渋谷など街自体が趣味や娯楽で訪れるといった目的のはっきりした特徴を持つことが判断できる．このことから, 本方式を用いてユーザに情報をリコメンドするサービスに有効であると考えている．よって, 本論文では, ユーザが安心して行動履歴という個人情報を投稿する際, プライバシ保護を行い, かつ, 高い検索精度を持つ, 相反する2つの問題の解を示したことに意義がある．

今後の課題として, 匿名性を高く保ちつつ, 行動内容としての高い精度の検索ができるように改善していくことがあげられる．よりユーザの目的と合致した検索結果を得るために, 平日や休日といった時間によってランキングを変更することや, アノニマイズド行動パターンを生成する際の最適な要素を検証することがあげられる．

また, 今回はSOMによりアノニマイズド行動パターンを生成したが, 別のアルゴリズムやSOMと別のアルゴリズムを組み合わせたアルゴリズムによる評価も必要である．その理由として, SOMを用いた場合の問題点は, 実際の行動履歴の数より少ない特徴点を生成することでプライバシー保護ができる半面, 複数ある場所と目的が合致している特徴的な街がそれぞれ近い場合は, 1つにまとめられる可能性が高いことがあげられる．よって, このような複数ある特徴的な街がそれぞれ近い場合に, その場所を検出し1つにまとめられないようにするためのSOMとは異なるアルゴリズムやSOMと別のアルゴリズムを組み合わせたアルゴリズムを評価することが必要である．

さらに, ショッピングモールなどのより狭い範囲の行動履歴や, 購買履歴や視聴履歴などの行動履歴にも応用できるよう汎用性を高めていく考えである．

参 考 文 献

1) PlaceEngine. <http://www.placeengine.com/>

- 2) Twitter. <http://www.twitter.com/>
- 3) Flickr. <http://www.flickr.com/>
- 4) Fire Eagle. <http://fireeagle.yahoo.net/>
- 5) 青柳武彦: 情報化時代のプライバシー研究: 「個の尊厳」と「公共性」の調和に向けて, NTT出版(2008).
- 6) Abowd, G.D., Atkeson, C.G., Hong, J., Long, S., Kooper, R. and Pinkerton, M.: Cyberguide: A mobile context-aware tour guide, *Wireless Networks*, Vol.3, No.5, pp.421-433 (1997).
- 7) 角 康之, 江谷為之, フェルス・シドニー, シモネ・ニコラ, 小林 薫, 間瀬健二: C-MAP: Context-aware な展示ガイドシステムの試作, 情報処理学会論文誌, Vol.39, No.10, pp.2866-2878 (1998).
- 8) Want, R., Hopper, A., Falcão, V. and Gibbons, J.: The active badge location system, *ACM Trans. Inf. Syst.*, Vol.10, No.1, pp.91-102 (1992).
- 9) Langheinrich, M.: A Privacy Awareness System for Ubiquitous Computing Environment, *Proc. Ubicomp 2002*, Lecture Notes in Computer Science, Vol.2498, pp.237-245, Springer-Verlag (2002).
- 10) 中西健一, 高汐一紀, 徳田英幸: 粒度の動的変更による位置匿名性についての考察, 情報処理学会論文誌, Vol.46, No.9, pp.2260-2268 (2005).
- 11) Beresford, A.R. and Stajano, F.: Location privacy in pervasive computing, *IEEE Pervasive Computing*, Vol.2, No.1 (2003).
- 12) Salber, D., Dey, K.A. and Abowd, D.G.: The Context Toolkit: Aiding the Development of Context-Enabled Applications, *Proc. CHI'99*, pp.434-441 (1999).
- 13) Kohonen, T.: *Self-Organizing Maps*, Springer-Verlag, Berlin, Germany (1995).
- 14) SOAP Version 1.2. <http://www.w3.org/TR/soap12-part0/>
- 15) JSON. <http://www.json.org/>
- 16) 情報大航海プロジェクト. <http://www2.igvpj.jp/>
- 17) 天笠邦一, 加藤文俊, 岡部大介: シチュエーション情報の収集とその活用可能性に関する一考察, 第18回情報処理学会ユビキタスコンピューティングシステム研究会(2008).
- 18) 川田正明, 小川克彦: アノニマイズした行動履歴に基づく行動情報検索システムの提案, 第18回情報処理学会ユビキタスコンピューティングシステム研究会(2008).
- 19) エリクソン, E.H., エリクソン, J.M.: ライフサイクル, その完結, みすず書房(2001).
(平成20年7月4日受付)
(平成21年1月7日採録)



川田 正明

2004年慶應義塾大学環境情報学部卒業．2006年同大学院政策・メディア研究科修士課程修了．現在，同大学院政策・メディア研究科後期博士課程に在学中．ユーザの行動履歴を用いた情報検索等の研究に従事．



小川 克彦（正会員）

1978年慶應義塾大学大学院工学研究科修士課程修了．同年NTT入社．NTTサイバーソリューション研究所長を経て，2007年より慶應義塾大学環境情報学部教授．工学博士．画像通信システム，ヒューマンインタフェース，マルチメディアサービスの研究開発を行い，現在は場所やケータイの未来サービスに関する研究と教育に従事．主な著書に『デジタルな生活 ITがデザインする空間と意識』（NTT出版）がある．電子情報通信学会，HFES等の会員．