DEIM Forum 2016 F8-5

ユーザの嗜好の保存度合いを考慮した 位置プライバシ保護のためのダミー生成手法とその評価

水野　聖也† 　原　 隆浩† 　Xing　Xie††

† 大阪大学情報科学研究科
†† Microsoft Research Asia
E-mail: †{mizuno.seiya,hara}@ist.osaka-u.ac.jp, ††xing.xie@microsoft.com

あらまし　位置情報サービスでは，ユーザの位置情報をサービスプロバイダに送信するという性質から，プライバシに関する問題が指摘されている．その一方で，ユーザの嗜好情報に基づいてサービスのパーソナライズを行うなど，ユーザの要求するプライバシ基準を満たす範囲内で嗜好情報を活用するための研究も盛んに行われている．そこで，筆者らは，先行研究において，位置情報サービスにおけるユーザのプライバシを保護しつつ，嗜好情報の活用を可能とすることを目的として，ユーザの嗜好情報の保存度合いをユーザの要求に応じて制御可能なダミー生成手法を提案した．本稿では，FourSquare のデータセットを用いて，東京 23 区の地図上でこの提案手法をシミュレートし，嗜好の保存度合いの制御可能性およびプライバシ保護性の評価を行った．

キーワード　位置プライバシ，嗜好の保存

1. はじめに

GPS を搭載した端末の普及に伴い，位置情報サービスが数多く展開されるようになった．位置情報サービスでは，自身の位置情報をサービスプロバイダに送信することにより，送信した位置情報に対応したサービスが受けられる．しかし，その位置情報の管理はサービスプロバイダに委ねられるため，サービスプロバイダが攻撃を受けたり，サービスプロバイダ自身による営利目的のデータの売買によって第三者に位置情報が露見し，ユーザの位置プライバシが侵害される危険性が指摘されている．

一方で，ユーザの嗜好に基づくサービスのパーソナライズ等，情報活用も重要視されるようになりつつある．それを受けてプライバシ保護の分野でも，ユーザの要求するプライバシ水準を満たす範囲内で可能な限り多くの情報を公開し，サービスの質や，データ分析への利用可能性の向上につなげようとする動きが活発になった [2] [3]．位置情報サービスにおいても，サービス利用履歴に基づくスポットの推薦などの研究が盛んに行われており，情報活用も考慮した位置プライバシ保護手法が望まれる．

位置プライバシ保護を目的とした研究の一つとして，ダミーを用いた位置曖昧化手法がある．この手法では，ユーザがサービスプロバイダに位置情報を送信する際に，ユーザの実際の位置情報とともに，偽の位置情報であるダミーの位置情報を複数送信する．これにより，サービスプロバイダは，送信された情報の中から実際のユーザの位置情報を一意に特定することが困難になり，ユーザの位置が曖昧化される．しかし，従来のダミーを用いた位置曖昧化手法の多くでは，ダミー生成の際にユーザの嗜好を考慮していないため，ダミーがユーザと異なる訪問傾向を示す．そのため，サービスプロバイダが観測可能な情報であるユーザとダミーが混在した位置情報系列から抽出される嗜好 (可観測な嗜好) は，本来のユーザの嗜好 (真の嗜好) と異なる可能性が高い．これは，プライバシの観点から嗜好情報も保護したいユーザには望ましい性質であるが，嗜好情報を公開することでパーソナライズされたサービスを受けたいユーザには，サービスの質を低下させる要因となる．

筆者らは，先行研究において，訪問場所の意味情報の系列であるカテゴリシーケンスのサポートをユーザの嗜好と定義し，この嗜好の保存度をユーザの要求に応じて制御可能なダミー生成手法を提案した [11]．本稿では，この提案手法の有効性を確認するため，シミュレーションによる評価実験を行った．

以下では，2 章で関連研究を説明し，3 章で文献 [11] で提案した手法の概要について述べる．4 章で本稿で行った評価について述べ，最後に 5 章でまとめと今後の課題を述べる．

2. 関連研究

2.1 ダミーを用いた位置曖昧化手法

文献 [5] [9] [10] では，ダミーを用いた位置曖昧化手法が提案されている．この手法では，ユーザはダミーの位置情報を複数生成し，自身の位置情報とともにサービスプロバイダに送信する．ユーザは，送信した全ての位置情報に対する応答を受信するが，その中から自身の位置に対する応答のみを選択することにより，自身の位置が特定されることを防止しつつ，サービスを受けることが可能となる．しかし，文献 [5] [9] で提案されている手法では，ダミーを生成する際にユーザの嗜好を考慮していないため，サービスプロバイダから観測可能なダミーとユーザが混在した位置情報系列から計算できる嗜好情報は，ユーザ本来のものとは異なってしまう．これは，嗜好情報もプライバシの観点から保護したいユーザには有用な性質であるが，嗜好情報を公開してパーソナライズされたサービスを受けたいユーザには，サービスの質を低下させる要因となる．

文献 [10] では，訪問場所のカテゴリ情報を考慮し，ユーザの

訪問履歴中で頻出するパターンに従ってダミーを遷移させることでユーザとダミーの判別を困難にし，ユーザの訪問場所の傾向を知っている攻撃者に対して効果的にユーザの位置情報を曖昧化する手法を提案している．この手法では，頻出パターンに従ってダミーを動作させるため，嗜好情報はある程度保存されると考えられるが，その度合いを制御することは難しい．

そこで筆者らは，文献 [11] において，嗜好の保存度をユーザの要求に応じて制御するためのダミー生成手法を提案した．この手法では，ユーザの嗜好を保存するダミーとユーザの嗜好と異なるダミーを，ユーザの要求する嗜好の保存度に応じた割合で混合することにより，嗜好の保存度の制御を試みる．

2.2 位置情報を用いたパーソナライズに関する研究

近年，ユーザの位置情報を用いたサービスのパーソナライズに関する研究が盛んに行われている [1] [4] [8]．これらの研究では，各訪問場所への訪問回数に関するユーザ間の類似度に基づいた協調フィルタリングが用いられている．この際，各訪問場所に対する訪問回数は，ユーザによって偏りが大きく，ユーザ間での共通部分が小さくなってしまうため，訪問場所の意味情報であるカテゴリに訪問回数をマッピングしている．例えば，文献 [1] では，各訪問場所の訪問回数をカテゴリ毎に集計したものをユーザの嗜好としている．文献 [8] では，ある時点のサービス利用が，それ以前のサービス利用に基づき決定されるという想定のもと，訪問場所のカテゴリの遷移シーケンスをユーザの嗜好としている．

このような推薦を行うには，ユーザ間の類似度を計算するためにカテゴリレベルでの情報のみを公開すれば十分である．また，各カテゴリの訪問回数は，そのカテゴリを含むシーケンスの発生回数の和として再現可能であるため，シーケンスを用いた嗜好モデルの方が汎用性が高い．

そこで筆者らは，先行研究 [11] において，訪問場所のカテゴリシーケンスの発生率であるサポート (支持度) をユーザの嗜好と定義し，嗜好の保存度をユーザの要求に応じて制御するためのダミー生成手法を提案した．

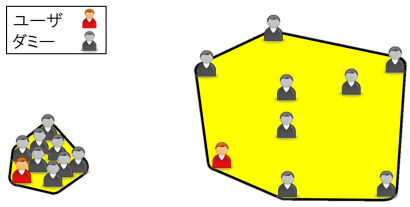
3. 先行研究 [11] で提案した手法の概要

本章では，先行研究 [11] における想定環境とダミー生成時に考慮している制約，プライバシ要求およびユーザの嗜好モデルについて説明し，提案したダミー生成手法の概要を述べる．

3.1 想 定 環 境

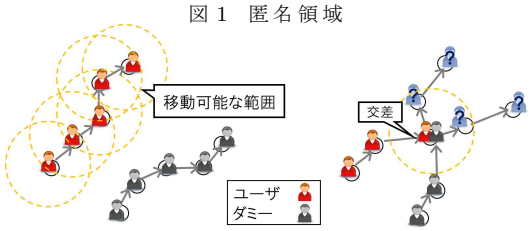
先行研究 [11] では，チェックイン型の位置情報サービスを想定している．そのためユーザの行動には，いくつかの訪問場所を訪問しながら移動する行動モデルを想定する．ユーザは，目的の訪問場所に到着した際に，先行研究 [11] で提案したダミー生成手法を用いて自身の位置情報を保護しながらサービス利用を行う．その際，訪問場所で T_m 秒以上滞在し，その後また次の訪問場所への移動を開始する．ユーザの移動は，歩行によるものとし，公共交通機関等の利用は想定しない．

ユーザの端末は地図情報や，ユーザのサービス利用履歴を十分保持しており，道路，訪問場所，訪問場所のカテゴリ，自身の嗜好情報をすべて把握しているものとする．



(a). 位置曖昧度: 小　　(b). 位置曖昧度: 大

図 1　匿 名 領 域



(a). 追跡可能性: 高　　(b). 追跡可能性: 低

図 2　追跡可能性

3.2 実環境でダミーを生成する際に考慮している制約

連続的なサービス利用では，サービス利用間のダミーの位置関係において，その時間における地理的な到達可能性を考慮する必要がある．例えば，あるユーザが一度サービスを利用してから三分後に再度サービスを利用する場合に，直前のサービスにおけるどのエンティティ(ダミーおよびユーザを表す) の位置からも三分以内に到達できない場所にダミーを配置しても，それがダミーであると容易に特定されてしまう．そのため先行研究 [11] では，各サービス利用において，直前のサービス利用時の位置から到達可能な位置にダミーを配置している．

3.3 位置プライバシに関する要求

3.3.1 匿 名 領 域

ユーザの位置プライバシを保護するためには，ダミーを識別できないことだけではなく，どの程度の大きさの領域に位置情報が曖昧化されているかも重要である．例えば図 1(a) のように，ダミーをユーザの周りに密集して配置した場合，ユーザが存在する可能性のある領域が絞り込めてしまう．ユーザの位置を十分に曖昧化するには，図 1(b) のように広範囲に分散してダミーを配置する必要がある．そこで先行研究 [11] では，Lu ら [5] の定義に基づき，ユーザとダミーで形成される凸包領域を匿名領域と定義し，この大きさがユーザの要求を達成できるようにダミーを配置している．

3.3.2 追跡可能性

短時間で連続的にサービス利用を行う場合，追跡可能性に対する配慮も必要である．例えば図 2(a) のように，あるサービス利用時刻におけるエンティティの位置から次のサービス利用時刻において到達可能な範囲に対応する位置情報が 1 つしかない場合，特定のエンティティのサービス利用が追跡できてしまう．この性質を追跡可能性と呼ぶ．追跡可能性が高い場合，目撃等によりユーザが一度特定された場合に，前後のサービス利用もユーザのものであると特定されてしまう．そのため先行研究 [11] では，図 2(b) のように同時刻に同じ場所を訪問させる (共有地点を設定する) ことにより交差を発生させ，連続的な

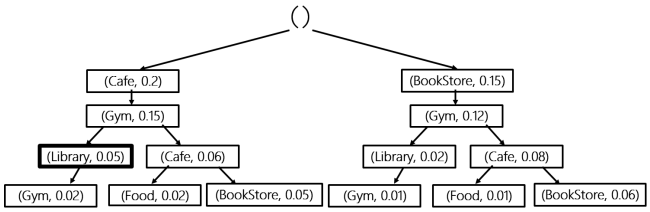


図 3　ユーザの嗜好を表す木

サービス利用の追跡を防止している．

3.4　嗜好の定義

ユーザの嗜好には，訪問場所の意味情報の系列であるカテゴリのシーケンスのサポート (支持度) を用い，それを図 3 のように木構造で表現する．この木構造において，各ノードは，根からそのノードに到達するまでに辿ったノードが保持するカテゴリを順に並べたカテゴリシーケンスに対応する．各ノードに付与された数値は，そのシーケンスのサポート値を表し，その値はユーザのサービス利用履歴に含まれる行動スケジュールの総数 N_{tj} と，その履歴中でのシーケンス S の発生度数 $n(S)$ を用いて次式で計算される．

$$Sup(S) = \frac{n(S)}{N_{tj}} \tag{1}$$

ユーザはサービス利用時に，ダミーを含めた位置情報をサービスプロバイダに送信するため，サービスプロバイダが観測可能な情報は，ユーザとダミーの位置情報系列が混在したものとなる．サービスプロバイダは，この情報を観測することでユーザの嗜好を計算する．以降では，ユーザとダミーの混在する位置情報系列の履歴から計算される嗜好を可観測な嗜好と呼び，T_o で表す．これに対し，ユーザのみの位置情報系列の履歴から計算される嗜好を真の嗜好と呼び，T_u と表記する．

観測可能な情報から T_o を計算するには，観測した位置情報の系列から発生したカテゴリシーケンスを計上し，(1) 式によりサポートを計算すればよい．ただしこの際，交差によって，各エンティティのカテゴリシーケンスが一意に決定できない場合が生じる．その場合，発生した可能性のある全てのシーケンスに対して発生度数の期待値を計算し，その期待値を発生度数として計上することで嗜好情報を更新する．

3.5　嗜好の保存度の定義

嗜好の保存度 $Pres(T_u, T_o)$ は，可観測な嗜好 T_o が真の嗜好 T_u の情報をどの程度保存しているかを表す指標として，次式で定義される．

$$Pres(T_u, T_o) = 1 - \frac{1}{2} \sum_{\{S \in SSet_u \cup SSet_t\}} |Sup_o(S) - Sup_u(S)| \tag{2}$$

ここで，$SSet_u$ は真の嗜好 T_u に含まれるすべてのシーケンスの集合，$SSet_o$ は可観測な嗜好 T_o に含まれる全てのシーケンスの集合，$Sup_u(S)$ は真の嗜好 T_u におけるシーケンス S のサポート値，$Sup_o(S)$ は可観測な嗜好 T_o におけるシーケンス S のサポート値である．この式において第二項は，二つの嗜好の木に含まれるシーケンスのサポート値の累積誤差を表し，この値が小さいほど $Pres(T_u, T_o)$ の値は 1 に近づく．すなわち，$Pres(T_u, T_o)$ が 1 に近いほど嗜好が保存されていることを表す．逆に累積誤差が大きいとこの値は 0 に近づき，嗜好のプライバシが保護されている状態であることを表す．

3.6　文献 [11] の提案手法の概要

文献 [11] の提案手法では，ユーザの行動プラン tj_0，要求ダミー数 n，要求匿名領域 A_r，ユーザの真の嗜好 T_u，可観測な嗜好 T_o，および嗜好の要求保存度 $\alpha = [0, 1]$ に基づき，n 個分のダミーの行動スケジュールを生成する．ここで，ユーザの行動プランおよびダミーの行動スケジュールは，i 回目のサービス利用における訪問場所 v_i とその訪問時刻 t_i の組 $< v_i, t_i >$ の系列である．この手法では，ユーザの嗜好の保存度を制御するため，生成する n 個のダミーを要求保存度 α に応じて，$\lfloor \alpha n \rfloor$ 個と $\lceil (1-\alpha)n \rceil$ 個の 2 つのグループに分割する．以降では，前者を Preserving グループ，後者を Obfuscating グループと呼ぶ．

Preserving グループのダミーは，ユーザの嗜好を保存するために，真の嗜好 T_u 中のカテゴリシーケンスに従って動作させる．Obfuscating グループのダミーは，ユーザの嗜好を考慮せずに動作させる．ダミーの生成では Preserving グループ，Obfuscating グループの順で，各ダミーの行動スケジュールを 1 つずつ確定させていく．この際，$id = i$ のダミーの生成には，ユーザの行動プランおよび生成済みである $id < i$ のダミーの行動スケジュールを考慮する．

3.6.1　Preserving グループのダミーの生成手順

Preserving グループのダミー生成では，(1) ユーザの嗜好の保存と，(2) ユーザの嗜好を知っている攻撃者に対する追跡可能性低下のための効果的な交差の 2 つの観点を考慮し，ユーザの真の嗜好 T_u に従うユーザらしいダミーを生成する．この際ダミーは，3.2 節で述べた到達可能性の制約を満たす必要があるため，シーケンスを一つ定めたとしても，そのシーケンスに従う経路を地図上から容易に発見できるとは限らない．これらの点を考慮し，まず，ユーザの真の嗜好 T_u 中に含まれるすべてのシーケンス S とそのシーケンス上に設定する共有地点 SP の組 $< S, SP >$ に対し，上記 2 つの観点からスコアリングを行い，各シーケンスを発生させる優先度を決定する．その後，地図情報を参照し，優先度の高いシーケンスからそのシーケンスに従う経路の生成を試行し，到達可能性の制約を満たしつつ，ユーザの要求匿名領域が可能な限り満たせる経路をダミーの移動経路として確定させる．ここで，共有地点 SP は，地点を共有させる $id = k$ の生成済みエンティティ e_k とその設定時刻 t_i の組 $< e_k, t_i >$ である．以降では，これらの手順を説明する．

a)　カテゴリシーケンスのスコアリング

ダミーに従わせるカテゴリシーケンスを決定するため，ユーザの真の嗜好 T_u 中に含まれるすべてのシーケンス S とそのシーケンス上に設定する共有地点 SP の組 $< S, SP >$ に対し，(1) ユーザの嗜好の保存と (2) ユーザの嗜好を知っている攻撃者に対する追跡可能性低下の 2 つの観点から次式でスコアリングを行う．

$$Score_s(< S, SP >) = \beta Score_{pref}(S) + (1-\beta) Score_{cross}(< S, SP >) \tag{3}$$

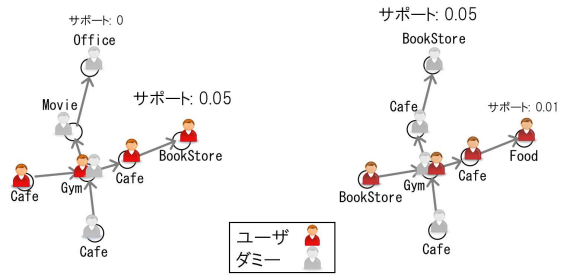


図 4　嗜好を知っている攻撃者に対する交差の効果

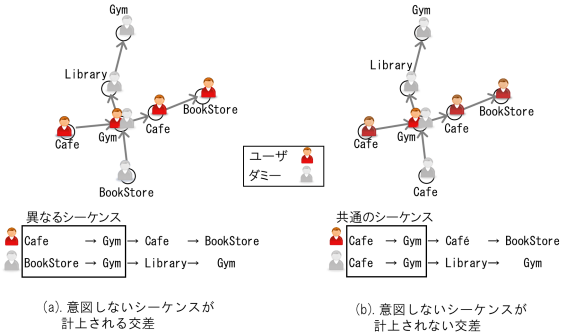


図 5　嗜好の保存が阻害される交差と阻害されない交差

ここで，$Score_{pref}$ は嗜好の保存のためのスコア，$Score_{cross}$ は追跡可能性低下のためのスコアであり，$\beta = [0, 1]$ は，両スコアの重みを決定するパラメータである．本稿では，事前実験に基づき，この値を $\beta = 0.25$ と定めた．

（1）　嗜好の保存のためのスコア $Score_{pref}$

シーケンス S のシーケンス長を l，$SSet_{u,l}$ をユーザの真の嗜好 T_u 中に含まれる長さ l のシーケンスの集合としたとき，シーケンス S の嗜好の保存に対する効果を表すスコア $Score_{pref}(S)$ は次式で表される．

$$Score_{pref}(S) = \frac{Sup_u(S) - Sup_o(S)}{\max\limits_{S_j \in SSet_{u,l}} \{Sup_u(S_j) - Sup_o(S_j)\}} \quad (4)$$

この式において，分子は真の嗜好に対する可観測な嗜好におけるサポート値の不足を表し，分母はそれを正規化するための項である．これにより，2 つの嗜好の差分を埋められるシーケンスが優先的に選択されるため，嗜好の保存度の向上が見込まれる．

（2）　追跡可能性低下のためのスコア $Score_{cross}$

ユーザの真の嗜好 T_u を知っている攻撃者に対しては，単純に交差を多く発生させるだけでは追跡可能性を十分に低下させられない．例えば，ユーザの真の嗜好 T_u を図 3 のものとしたとき，図 4(a) のようにダミーをユーザの真の嗜好 T_u 中に含まれないサポート値が 0 であるシーケンスで交差させたとしても，サポート値が 0.05 であるシーケンスに従って移動している位置情報がユーザのものであると容易に特定されてしまう．そのため，嗜好を知っている攻撃者に対して効果的に追跡可能性を低下させるには，ダミーを図 4(b) のように，サポート値の高いシーケンスに従わせた上で交差させる必要がある．

ここで，あるエンティティの位置情報がユーザのものである確率をユーザ確率と定義し，交差発生時に期待されるダミーに対するユーザ確率の配分 $distro(< S, SP >)$ を考え，交差の効果を表すスコアの一部として導入する．S_{e_k} を交差相手のエンティティ e_k が従うカテゴリシーケンス，l をシーケンス S および S_{e_k} のシーケンス長，$S^{(i,j)}$ をシーケンス S の i 番目の要素から j 番目の要素までのサブシーケンス，$\{S_1, S_2\}$ を シーケンス S_1 の末尾にシーケンス S_2 を結合したシーケンスを表すものとしたとき，$distro < S, SP >$ は次のように計算できる．

$$distro(< S, SP >) = \frac{Sup_u(S) + Sup_u(\{S_{e_k}^{(1,i)}, S^{(i+1,l)}\})}{Sup_{sum}},$$
$$Sup_{sum} = Sup_u(S) + Sup_u(\{S_{e_k}^{(1,i)}, S^{(i+1,l)}\}) + Sup_u(S_{e_k}) + Sup_u(\{S^{(1,i)}, S_{e_k}^{(i+1,l)}\}) \quad (5)$$

交差を設定する際には，交差回数が時刻間で均一になることが望ましい．これは，同時刻に集中して交差が発生した場合に，ダミーとユーザが密集することによって，匿名領域が小さくなるためである．また，ユーザに集中して交差を設定した場合，交差回数が他のエンティティと比べて多いものがユーザであると容易に特定されてしまう．そのため，特定のエンティティに交差が集中することを防ぎ，エンティティ間の交差回数についても均一化を図る必要がある．

交差が発生した場合，サービスプロバイダはそれらの発生した期待度数を計上することによって，可観測な嗜好 T_o を計算する．そのため，上記だけでは，ユーザの嗜好の保存を阻害する交差を発生させてしまう可能性がある．例えば，図 5(a) のように，共有地点までのシーケンスが異なるような共有地点を設定した場合，サービスプロバイダによってシーケンスは，$Cafe \to Gym \to Cafe \to BookStore$，$Cafe \to Gym \to Library \to Gym$，$BookStore \to Gym \to Cafe \to BookStore$，$BookStore \to Gym \to Library \to Gym$ が，それぞれ 0.5 回ずつ発生したものと計上され，本来のダミーによって生成されたシーケンスである $BookStore \to Gym \to Library \to Gym$ というシーケンスの発生度数が，$BookStore \to Gym \to Cafe \to BookStore$ という意図しないシーケンスに分散してしまう．これを防ぐには，図 5(b) のように，共有地点を設定する時刻までのシーケンスが同一となるようにする必要がある．これにより $Cafe \to Gym \to Library \to Gym$，$Cafe \to Gym \to Cafe \to BookStore$ というダミーとユーザによって本来生成されていたシーケンスが 1 度ずつ発生したと計上されるようになり，嗜好の保存度の制御可能性の向上が見込まれる．

以上を考慮して，嗜好を知っている攻撃者に対する交差の効果を表すスコア $Score_{cross}$ を次式で定義する．

$$Score_{cross}(< S, SP >) = \frac{\delta \cdot distro(< S, SP >)}{1 + n_{sh}^3(e_k) + n_{sh}^2(t_i)}, \quad (6)$$

$$\delta = \begin{cases} 1 & (S^{(1,i)} = S_{e_k}^{(1,i)}) \\ 0 & (otherwise) \end{cases}$$

$n_{sh}(e_k)$ はエンティティ e_k に設定済みの共有地点数，$n_{sh}(t_i)$ は時刻 t_i に共有地点を設定済みのエンティティ数を表す．

b）　経路の決定方法

スコアリングによって決定したシーケンス S と共有地点 SP を基に，地図情報を参照し，到達可能性の制約を満たすダミー

表 1　シミュレーションにおけるパラメータ

パラメータ	値
領域 $[m^2]$	4000 × 14000
交差点数	325946
訪問場所数	278153
訪問場所のカテゴリ数	10
サービス利用間隔 [s]	[480, 720]
歩行速度 [km/h]	[4.0, 6.0]

の移動経路を生成する．具体的には，決定した SP を起点にそれ以前とそれ以降のサービス利用時刻におけるダミーの訪問場所を順に決定する．ある時刻 t_i におけるダミーの訪問場所を v_{t_i} としたとき，時刻 t_{i+1} におけるダミーの訪問場所 $v_{t_{i+1}}$ を次の手順で決定する．まず初めに，v_{t_i} から時刻 t_{i+1} までの間に到達可能な範囲にある訪問場所のうち，カテゴリシーケンスに従ったカテゴリに該当する訪問場所を全て取得する．$v_{t_{i+1}}$ は，この中から選択するが，この際，これらの訪問場所に対し次式でスコアリングを行い，スコアが最も高くなるものを次の訪問場所として決定する

$$Score_v(v_{t_i}) = \begin{cases} \frac{SimA_r - CA_r}{(1+n_{reach}(v_{t_i}))^{2\alpha-1}} & (CA_r \leq A_r) \\ \frac{1}{|SimA_r - A_r|(1+n_{reach}(v_{t_i}))^{2\alpha-1}} & (otherwise) \end{cases} \tag{7}$$

ここで，CA_r は生成済みのダミーとユーザから形成される匿名領域の大きさ，$SimA_r$ は訪問場所 v_{t_i} を次の訪問地点として決定した場合に形成される匿名領域の大きさ $n_{reach}(v_{t_i})$ は，時刻 t_i に訪問場所 v_{t_i} に到達可能な生成済みエンティティの数を表す．すなわち，生成済みのダミーとユーザから形成される匿名領域の大きさと要求匿名領域の大きさを比較し，それが要求匿名領域より小さい場合，匿名領域の増分が大きくなるように次の訪問場所を選択し，逆に，既にこれが要求匿名領域より大きい場合，匿名領域の大きさが要求匿名領域 A_r に最も近くなる地点を次の訪問場所として決定する．その際，要求保存度が高い場合には他の生成済みエンティティから到達しにくい訪問場所を選択することで意図しない交差を抑制し，嗜好の保存度の向上を図る．

3.7　Obfuscating グループのダミーの生成手順

Obfuscating グループのダミーは，ユーザの嗜好を考慮せずに動作させることで，ユーザの嗜好を保護する．さらに，設定されている共有地点数が少ないエンティティとの間に共有地点を設定することで，嗜好を知らない攻撃者に対する追跡可能性の低下を図る．具体的にはまず，設定されている共有地点数が最小のエンティティを取得する．さらに，共有地点を設定されているエンティティ数が最小の時刻を検索し，この時刻において，取得したエンティティとの共有地点を設定する．この共有地点を基準に，カテゴリシーケンスの制約を除き，Preserving グループと同様の方法で経路を生成することでダミーの行動スケジュールを決定する．

4. 評 価 実 験

先行研究 [11] の提案手法の有効性を検証するため，FourSquare のデータセットを用いて，東京 23 区の地図上で実際のユーザのサービス利用を再現し，評価実験を行った．シミュレーションにおける各パラメータを表 1 に示す．訪問場所には，FourSquare API(注1) を用いて取得した Venue を用いて実際の訪問場所の分布を再現し，訪問場所のカテゴリには FourSquare で定められた第 1 階層のカテゴリ 10 種類を割り当てた (注2)．ユーザの行動スケジュールの生成には，文献 [7] で用いられているチェックインのデータセットを利用した．ただし，このデータセットには，公共交通機関や自動車を用いた移動が含まれているため，次の手順で歩行のみによる行動スケジュールに変換し，それを基にユーザの嗜好モデルを構築した．

（1） チェックイン時刻の間隔が 2 時間以上となっている部分でチェックイン系列を分割し，それぞれを 1 つのユーザの行動スケジュールとした．

（2） 各行動スケジュールで訪問場所をカテゴリ情報に変換し，(1) 式によりサポート値を計算することで，ユーザの真の嗜好 T_u を構築した．

（3） 各行動スケジュールにおけるカテゴリシーケンスが保存されるようにしつつ，それぞれ前のサービス利用の時刻から最短経路を通って $[8, 12]$ 分の移動時間で到達できる経路となるように訪問場所を変換した．

評価では，上記の手順により生成されたユーザの行動スケジュール系列，および真の嗜好 T_u をダミー生成の際の入力とする．具体的には，行動スケジュール系列の先頭から各行動スケジュールをユーザの行動プランとみなしてダミーを生成し，その結果に対して 4.1 節で定義する各評価指標を計算した．

4.1　評 価 指 標

- **嗜好の保存度**

生成されたダミーの行動スケジュールに基づき可観測な嗜好 T_o を構築し，式 (2) を用いて実際の嗜好の保存度を計算する．これが嗜好の要求保存度 $\alpha = [0, 1]$ に対し，どのように変化するかを調べることで，提案した手法における嗜好の保存度の制御可能性を検証する．

- **AR-Count / AR-Size**

サービス利用の総数に対し，要求された匿名領域のサイズをダミーの配置によって実際に達成できた回数の割合を AR-Count，達成できた匿名領域の平均面積の要求匿名領域に対する割合を AR-Size と定義する．これらが高い場合，ユーザの要求に対して，十分に位置曖昧度を確保できていることを表す．

- **MTC**

文献 [6] の定義に基づき，追跡可能性を評価するための指標である MTC を次のように定義する．攻撃者から見たある位置情報がユーザのものである確率をユーザ確率と呼ぶ．ある時点において，目撃等によりユーザが特定された場合，ユーザ自身の

(注1)：https://developer.foursquare.com/docs/venues/search
(注2)：https://developer.foursquare.com/categorytree

ユーザ確率は 1 となり，他のダミーのユーザ確率は 0 となる．ここで，エンティティ間で図 2(b) のように交差が発生した場合，交差前後の遷移の対応関係が一意に定まらなくなるため，それらのエンティティ間でユーザ確率が配分される．この条件下で MTC は，ユーザが特定された時点から各エンティティ e_k のユーザ確率 $p(k)$ のエントロピー $H = -\sum p(k)\log_2 p(k)$ が閾値 T を超えるまでにかかる平均時間と定義される．なお，本稿では $T = 1$ とした．すなわち，ユーザの各行動スケジュールを tj_k，ユーザの行動スケジュールの総数を N_{tj}，tj_k におけるサービス利用の総数を $n(tj_k)$，tj_k において時刻 t_i でユーザが特定されてから $H > T$ となるまでにかかる時間を $TC(tj_k, t_i)$ とすると，MTC は次式で計算される．

$$MTC = \frac{1}{N_{tj}} \sum_{k=1}^{N_{tj}} \frac{1}{n_{achieve}(tj_k)} \sum_{i=1}^{n(tj_k)} TC(tj_k, t_i) \qquad (8)$$

ただし，$TC(tj_k, t_i)$ は，時刻 $t_{n(tj_k)}$ までに $H > T$ が満たせない場合は 0 と見なし，MTC の計算には考慮しない．$n_{achieve}(tj_k)$ は，tj_k において，最後のサービス利用時刻 $t_{n(tj_k)}$ までに $H > T$ を達成できたユーザ特定時刻 t_i の数である．この指標は，ユーザが特定されてから再び曖昧化されるまでにかかる平均時間であるため，この値が小さければ追跡可能性が低いことを表す．

本研究では，攻撃者として，嗜好を知らない攻撃者と嗜好を知っている攻撃者の両方を想定し，それぞれに対する MTC を評価する．これらの攻撃者は背景知識が異なるため，交差が発生した際のユーザ確率の配分が異なる．以下でそれぞれの攻撃者を想定した場合のユーザ確率の配分について説明する．

a） ユーザの嗜好を知らない攻撃者の場合

まず一般的に，図 6 のように，時刻 t_i において，ユーザ確率が α で訪問場所 v_1 に存在するユーザと，ユーザ確率が β で訪問場所 v_2 に存在するダミーが時刻 t_i と t_{i+1} の間で交差し，それぞれ訪問場所 v_4, v_3 に遷移した場合のユーザ確率の配分を考える．X_i をユーザが時刻 t_i で訪問した訪問場所を表す確率変数とすると，交差後のそれぞれのユーザ確率は $Pr(X_{i+1})$ と表され，次のように計算できる．

$$Pr(X_{i+1}) = \sum_{X_i} Pr(X_{i+1}|X_i)Pr(X_i) \qquad (9)$$

すなわち，図 6 の場合，

$$\begin{aligned}
\gamma &= Pr(X_{i+1} = v_4) \\
&= \alpha Pr(X_{i+1} = v_4|X_i = v_1) + \beta Pr(X_{i+1} = v_4|X_i = v_2) \\
&= \frac{\alpha + \beta}{2}, \\
\delta &= Pr(X_{i+1} = v_3) \\
&= \alpha Pr(X_{i+1} = v_3|X_i = v_1) + \beta Pr(X_{i+1} = v_3|X_i = v_2) \\
&= \frac{\alpha + \beta}{2}
\end{aligned}$$

となる．この配分方法に基づく MTC を MTC1 とする．

b） ユーザの嗜好を知っている攻撃者の場合

ユーザの真の嗜好 T_u を把握している攻撃者に対しては，ユーザ確率の配分を T_u に含まれるサポート値に基づいて行う．真の嗜好 T_u が図 3 のものであるとしたときのユーザ確率の配分を図 7 を用いて説明する．図は，時刻 t_1 においてユーザ確率がそれぞれ α，β であるユーザとダミーが時刻 t_2 において交差した際のユーザ確率の配分を表している．式 (9) に基づき，交差後のユーザのユーザ確率 γ は，

$$\begin{aligned}
\gamma &= Pr(X_3 = Cafe2) \\
&= \alpha Pr(X_3 = Cafe2|X_2 = Gym1, X_1 = Cafe1) \\
&\quad + \beta Pr(X_3 = Cafe2|X_2 = Gym1, X_1 = BookStore1)
\end{aligned}$$

と計算できる．ここで，C_i をユーザが時刻 t_i において訪問した訪問場所のカテゴリを表す確率変数とすると，

$$\begin{aligned}
\gamma &= \alpha Pr(C_3 = Cafe|C_2 = Gym, C_1 = Cafe) \\
&\quad + \beta Pr(C_3 = Cafe|C_2 = Gym, C_1 = BookStore) \\
&= \alpha \frac{\sum_{C_4} Pr(C_4, C_3 = Cafe, C_2 = Gym, C_1 = Cafe)}{\sum_{C_4}\sum_{C_3} Pr(C_4, C_3, C_2 = Gym, C_1 = Cafe)} \\
&\quad + \beta \frac{\sum_{C_4} Pr(C_4, C_3 = Cafe, C_2 = Gym, C_1 = BookStore)}{\sum_{C_4}\sum_{C_3} Pr(C_4, C_3, C_2 = Gym, C_1 = BookStore)}
\end{aligned}$$

と置き換えられる．ここで，Pr(S) をユーザがシーケンス S に従って遷移する結合確率とすると，サポート値との関係は，

$$Pr(S) = \frac{Sup(S)}{\sum Sup(S)} \qquad (10)$$

である．この例において，ユーザがとり得たシーケンスは，$S_{u,1} = Cafe \rightarrow Gym \rightarrow Cafe \rightarrow BookStore$, $S_{u,2} = Cafe \rightarrow Gym \rightarrow Library \rightarrow Gym$, ダミーがとり得たシーケンスは，$S_{d,1} = BookStore \rightarrow Gym \rightarrow Cafe \rightarrow BookStore$, $S_{d,2} = BookStore \rightarrow Gym \rightarrow Library \rightarrow Gym$ であり，それぞれのサポート値は図 3 から，$Sup_u(S_{u,1}) = 0.05$, $Sup_u(S_{u,2}) = 0.02$, $Sup_u(S_{d,1}) = 0.06$, $Sup_u(S_{d,2}) = 0.01$ と求まる．よって，

$$\begin{aligned}
\gamma &= \alpha \frac{Sup_u(S_{u,1})}{Sup_u(S_{u,1}) + Sup_u(S_{u,2})} + \beta \frac{Sup_u(S_{d,1})}{Sup_u(S_{d,1}) + Sup_u(S_{d,2})} \\
&= \alpha \frac{0.05}{0.02 + 0.05} + \beta \frac{0.06}{0.06 + 0.01} = \frac{0.05\alpha + 0.06\beta}{0.07},
\end{aligned}$$

同様に，

$$\delta = \alpha \frac{0.02}{0.05 + 0.02} + \beta \frac{0.01}{0.06 + 0.01} = \frac{0.02\alpha + 0.01\beta}{0.07}$$

と計算できる．この配分方法に基づく MTC を MTC2 とする．

- **CR (Confusion achieving Ratio)**

式 (8) における N_{tj}, $n(tj_k)$, $n_{achieve}(tj_k)$ を用いて，追跡可能性を評価するためのもう一つの指標として CR を定義する．

$$CR = \frac{1}{N_{tj}} \sum_{k=1}^{N_{tj}} \frac{n_{achieve}(tj_k)}{n(tj_k)} \qquad (11)$$

CR は，ユーザの行動スケジュール全体のうち，最初のサービス利用からどの程度の割合のサービス利用までのユーザ特定に対し，最終サービス利用時刻までにユーザの曖昧性を回復することができたかを表す．そのため，この値が高いほど追跡可能性が低いことを意味する．嗜好を知らない攻撃者を想定した場合のユーザ確率の配分方法に基づいて計算される CR を CR1，嗜好を知っている攻撃者を想定した場合の CR を CR2 とする．

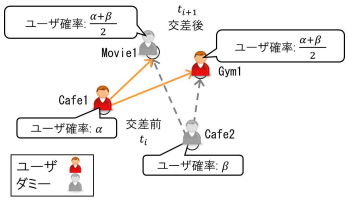


図 6　嗜好を知らない攻撃者からみたユーザ確率の配分

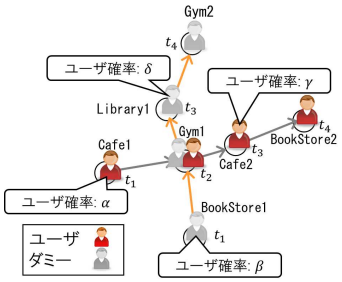


図 7　嗜好を知っている攻撃者からみたユーザ確率の配分

表 2　各ユーザのチェックインデータの特徴

パラメータ名	値
総チェックイン数	1322
行動スケジュール数 (経路数)	251
1 行動スケジュール中の平均サービス利用回数	5.1
平均サービス利用間隔 [s]	594.4

4.2　評価結果

本評価では，ユーザの嗜好情報を十分に確保するために，データセットに含まれるユーザのうち，特にチェックイン数の多かったユーザを選出し評価を行った．ユーザのチェックインデータの特徴は表 2 の通りである．

ユーザの真の嗜好のヒストグラムを図 8 に示す．このヒストグラムにおいて，横軸は真の嗜好中の各カテゴリシーケンス，縦軸がそのシーケンスのサポート値に対応し，各シーケンスをサポート値により降順にソートして表示している．一部のシーケンスのサポートが高くなっていることから，普段よく行っている行動パターンがあることがうかがえる．

本評価では，このユーザに対し，要求保存度 α を変化させながら提案したダミー生成手法を適用し，各評価指標の値の変化を調べた．なお，要求匿名領域は $A_r = 1500^2[m^2]$，ダミー数は $n = 9$ とした．まず，各ユーザの行動スケジュールの系列に提案手法を適用した際の，実際の嗜好の保存度の遷移を図 9 に示す．このグラフにおいて，横軸はユーザの行動スケジュール系列中の各行動スケジュールに対応し，縦軸は，その系列の先頭から順にダミー生成を行った結果の累積から計算される嗜好の保存度を表す．この結果から，行動スケジュール系列が進むにつれて，嗜好の保存度が一定の値に収束する傾向があることがわかる．そのため，以降では，最後の行動スケジュールに対するダミー生成が終了した時点の嗜好の保存度の値を収束値とし，この値を用いて嗜好の保存度の制御可能性について論じる．嗜好の要求保存度に対する嗜好の保存度の収束値の遷移を図 10 に示す．このグラフから，要求保存度の増加に対し実際の嗜好の保存度が概ね線形に増加していることがわかる．このことから，提案手法は，嗜好を保存するダミーと嗜好のプライバシを保護するダミーの割合を要求に応じて混合することで，嗜好の保存度を制御できていることを確認した．

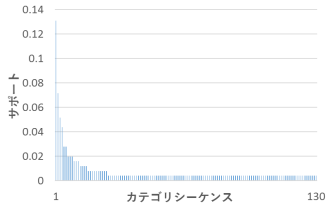


図 8　真の嗜好のヒストグラム

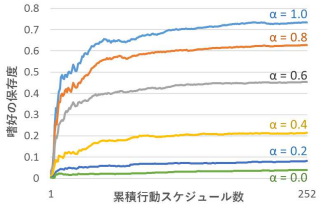


図 9　嗜好の保存度の遷移

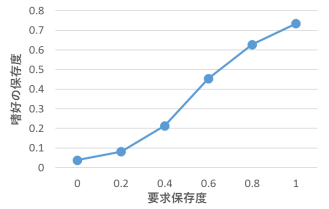


図 10　嗜好の保存度の収束値

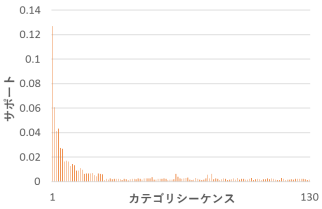


図 11　可観測な嗜好のヒストグラム (要求保存度 1.0)

次に，要求保存度を 1.0 とした際の可観測な嗜好のヒストグラムを図 11 に示す．ただし，このヒストグラムでは，ユーザの真の嗜好に含まれないシーケンスは省略している．このグラフを図 8 と比較すると，類似性が高いことがわかる．しかし，嗜好の保存度は 0.73 程度に留まっている．これは，意図しないシーケンスが計上される交差が発生していることに起因する．真の嗜好においてサポート値が 0 でないシーケンスの数が 130 個であるのに対し，可観測な嗜好では，サポート値が 0 でないシーケンスが 3601 個存在する．しかし，意図しないシーケンスのサポート値はユーザの真の嗜好に含まれるシーケンスのサポート値と比べると小さいことが確認されたため，パーソナライズ等に用いる際には，これらをフィルタリングすることにより十分なパフォーマンスを期待できると考えられる．

次に，要求保存度に対する AR-Count，AR-Size の変化を図 12，図 13 に示す．AR-Count，AR-Size ともに要求保存度 1.0 の場合と 0.0 の場合で達成率が大幅に低下している．1.0 の場合に低下するのは，要求保存度が増加すると，ユーザの真の嗜好に含まれるシーケンスに従って移動するダミーの数が増加するためである．すなわち，カテゴリシーケンスの制約を満たし，かつ匿名領域の要求を満たせる経路の発見が実環境において困難であることを意味する．一方，要求保存度が 0.0 の場合にも AR-Count，AR-Size ともに小さい値となっている．これは，提案した手法で他の生成済みエンティティからの到達可能性を考慮している影響である．要求保存度が小さい場合，他の生成済みエンティティから到達しやすい訪問場所が選択されることでエンティティが密集しやすくなるため，匿名領域が小さくなったと考えられる．

各ユーザにおける MTC と CR の評価結果を図 14，図 15 に示す．図 14 において，各プロットに付記した数値は，ユーザの曖昧化を一度でも達成できた行動スケジュールの数，つまり，

平均の計算に用いられた追跡可能時間の数を表す．

MTC1 は要求保存度の値に対する変化が小さい．提案した手法では，各ダミー生成時に 1 つしか共有地点を設定していないが，各要求保存度に対する MTC1 の平均値は，704.1 秒と十分に小さい値を維持できている．ユーザの平均サービス利用間隔はそれぞれ 594.4 秒であるので，平均で 1.2 回程度のサービス利用の間に再び曖昧性を回復できたことを表す．このことから，8 ～ 12 分程度の間隔でサービス利用が起こる場合，適度な大きさの領域にダミーを配置すれば，意図的な交差を設定しない場合でも，嗜好を知らない攻撃者に対しては容易にユーザの曖昧性を回復することが可能であることがわかる．そのため，CR1 は常に達成率が 0.8 以上の高い値を維持できている．

一方，MTC2 は要求保存度が高くなると上昇している．これは，MTC2 要求保存度が低い場合に，曖昧化を一度も達成できなかった行動スケジュール数が多いことに起因する．MTC の計算では，その行動スケジュール終了時までに曖昧化の条件が達成されない場合，その追跡可能時間は計算に含められない．要求保存度が小さい場合，ユーザの嗜好を考慮したダミーの数も少なくなるため，曖昧化の条件を達成できた回数が減少している．そのため，平均の計算に考慮された追跡可能時間の数が著しく少なくなり，MTC2 が小さくなったと考えられる．逆に，要求保存度が高い場合は，曖昧化を達成できた回数が増加しているが，その際，要求保存度が小さい場合には考慮されなかった長い追跡可能時間が平均の計算に考慮されるようになり，MTC2 が増加したと考えられる．そのため，CR2 は，要求保存度が小さい場合にはほぼ 0 となっている．この結果は，嗜好を知っている攻撃者の追跡能力の高さを示唆しており，嗜好を考慮しないダミー生成手法では，サービス利用を追跡される危険性が極めて高いことを意味する．それに対し提案手法では，要求保存度を高く設定することで，CR2 の値に改善が見られ，その値は，要求保存度を 1 とした場合に平均で 0.44 となっている．この値は，平均すると行動スケジュール前半でのユーザの特定に対し，最後のサービス利用までの間にユーザの曖昧性を回復できたことを意味する．この結果から，ユーザの嗜好を考慮することは，嗜好の保存の観点のみでなく，位置プライバシ保護の観点からも極めて重要であるといえる．

5. おわりに

本稿では，先行研究 [11] で提案した手法の有効性を確認するために，FourSquare のデータセットを用いてシミュレーション評価を行った．評価の結果，提案手法では，ユーザの嗜好を保存するダミーとユーザの嗜好と異なるダミーをユーザの要求する保存度に応じた比率で生成することで，嗜好の保存度を制御可能であることを確認した．また，ユーザの嗜好を考慮してダミーを生成することで，従来手法よりもユーザの嗜好を知っている攻撃者に対する追跡可能性を低下できることを確認した．

今後は，公共交通機関を利用するユーザに対しても効果的にダミーが生成できるように，手法の拡張を行う予定である．

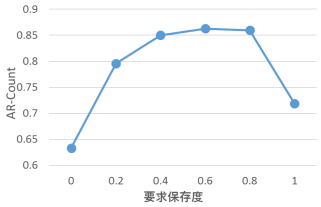


図 12　AR-Count

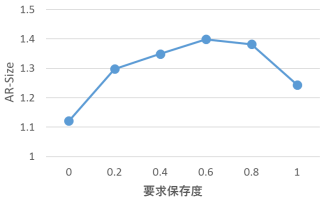


図 13　AR-Size

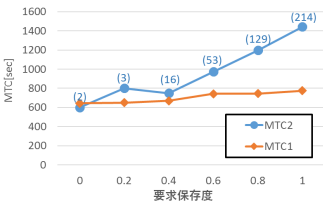


図 14　MTC

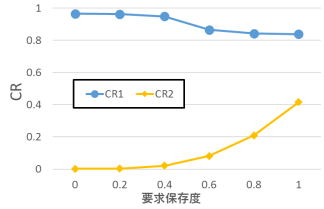


図 15　CR

謝　辞

本研究の一部は，文部科学省科学研究費補助金・基盤研究 (A)(26240013)，および日立財団研究助成「倉田奨励金」の研究助成によるものである．評価で利用した東京 23 区の地図データは一般財団法人日本デジタル道路地図協会より貸与されたものである．ここに記して謝意を示す．

文　献

[1] Bao, J., Zheng, Y. and Mokbel, M. F.: Location-based and Preference-aware Recommendation Using Sparse Geo-social Networking Data, *In Proc. GIS*, pp. 199–208 (2012).

[2] Chen, R., Acs, G. and Castelluccia, C.: Differentially Private Sequential Data Publication via Variable-length N-grams, *In Proc. CCS*, pp. 638–649 (2012).

[3] Götz, M., Nath, S. and Gehrke, J.: MaskIt: Privately Releasing User Context Streams for Personalized Mobile Applications, *In Proc. SIGMOD*, pp. 289–300 (2012).

[4] Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W. and Ma, W.-Y.: Mining User Similarity Based on Location History, *In Proc. GIS*, pp. 34:1–34:10 (2008).

[5] Lu, H., Jensen, C. S. and Yiu, M. L.: PAD: Privacy-area Aware, Dummy-based Location Privacy in Mobile Services, *In Proc. MobiDE*, pp. 16–23 (2008).

[6] Shokri, R., Freudiger, J., Jadliwala, M. and Hubaux, J.-P.: A Distortion-based Metric for Location Privacy, *In Proc. WPES*, pp. 21–30 (2009).

[7] Yang, D., Zhang, D., Zheng, V. W. and Yu, Z.: Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs, *IEEE Trans. SMC*, Vol. 45, No. 1, pp. 129–142 (2015).

[8] Ying, J. J.-C., Lee, W.-C., Weng, T.-C. and Tseng, V. S.: Semantic Trajectory Mining for Location Prediction, *In Proc. GIS*, pp. 34–43 (2011).

[9] 加藤 諒，原 隆浩，Xie, X.，岩田麻佑，西尾章治郎：ユーザの行動プランの変更を考慮したダミーによるユーザ位置曖昧化手法，第 7 回データ工学と情報マネジメントに関するフォーラム論文集 (2015).

[10] 加藤 諒，松野有弥，原 隆浩，荒瀬由紀，Xie, X.，西尾章治郎：ユーザの訪問場所の傾向を考慮したダミーによるユーザ位置曖昧化手法，情報処理学会マルチメディア，分散，協調とモバイルシンポジウム論文集，Vol. 2014, pp. 1174–1181 (2014).

[11] 水野聖也，原 隆浩，Xie, X.：位置プライバシ保護のための嗜好の保存度合いを考慮したダミー生成手法の検討，情報処理学会研究報告，Vol. 2015-DPS-165, No. 6, pp. 1–7 (2015).