位置情報サービスにおける 属性の可観測性を考慮したプライバシ保護手法

眞野　将徳† 　石川　佳治††,†,†††

† 名古屋大学大学院情報科学研究科
†† 名古屋大学情報基盤センター
††† 国立情報学研究所
E-mail: †mano@db.itc.nagoya-u.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp

あらまし　近年普及しつつある位置に基づくサービスにはユーザのプライバシを侵害する危険が存在しうる．既存手法はユーザの位置情報のみを考慮しているため，属性情報を併用するサービス (広告配信サービスなど) には対応することができない．本研究は，このような属性情報を併用する位置に基づくサービスにおけるプライバシ保護を実現するために，属性の新たな基準として可観測性を導入する．そして可観測性から定まる一致度という指標を利用する匿名化手法を提案する．

キーワード　空間データベース，プライバシ，匿名化

1. はじめに

1.1 背　景

近年，GPS 機能を有するモバイル端末の普及や無線通信網の発展により，測位された位置情報に基づいて近隣の店舗の情報などのユーザに有益な情報を提供する，**位置に基づくサービス** (location-based services) が普及している．位置に基づくサービスは便利だが，プライバシに関わる問題が存在する．サービスを利用するためにはユーザの位置を送信する必要があるが，詳細な位置情報を送信すると，悪意を持った攻撃者である可能性があるサービス提供者にユーザがどこにいるか知られてしまう．たとえば自宅の場所が知られてしまえば，住所録と照合することでユーザを特定することもできてしまう．これを防ぐため，**位置の匿名化** (location anonymization) と呼ばれる位置情報を曖昧にする手法を用いて，ユーザのプライバシを保護しようとする研究が多くなされている [6]．位置情報を曖昧化しすぎるとサービスの質が落ちる可能性があるため，適度な匿名化が求められる．

1.2 属性情報を用いる位置情報サービス

位置に基づくサービスのうち，位置情報のみを用いるものについては，位置の匿名化がユーザのプライバシ保護に有効である．しかし，位置に基づくサービスには位置情報以外にも，ユーザの性別・年齢といった**属性情報**を併用するようなサービスも考えられる．属性情報を用いる位置に基づくサービスの一つの例として，**モバイル広告配信サービス**を考える．

本研究では以下のようなモバイル広告配信サービスを想定する．モバイルユーザが広告の要求をすると，そのエリアを対象としている広告主にその要求が伝えられ，広告主の判断に応じて広告が配信される．すなわち，プル型のサービスである．ユーザの属性情報を配信する広告を選ぶために利用する．広告の配信には，その配信数に応じて広告料がかかるため，広告主はできるだけ少ない費用で高い効果を得たいと考えている．各広告には想定しているターゲットユーザがあり (たとえば化粧品のセール広告は，特定の年代の女性ユーザなど)，属性を参照すれば，対象ユーザに絞って配信することができる．

本論文で想定するモバイル広告配信サービスの構成を図 1 に示す．モバイルユーザと広告主の間に信頼できるサードパーティの**仲介業者**が入り，各ユーザの属性情報を**プロファイル** (profile) として管理する．後述のように，仲介業者は匿名化処理も担当する．モバイルユーザが仲介業者に広告の要求をすると，仲介業者はユーザの情報を匿名化して広告主に提示し，広告主が配信を希望した広告の中から適切なものをユーザに配信する．これによりユーザはクーポンなどの特典を得ることができ，広告主は想定に合ったユーザに広告配信できる．

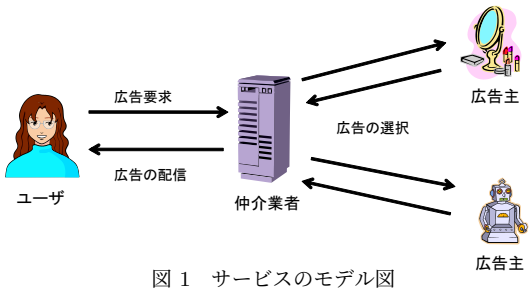
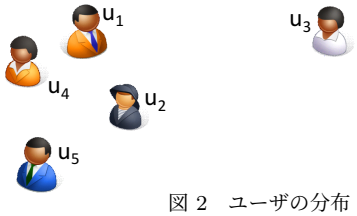


図 1　サービスのモデル図

1.3 プライバシの問題

このサービスで問題となるのは，広告主は必ずしも信頼できず，**攻撃者** (adversary) となりうるおそれがあることである．ユーザの正確な位置が広告主に通知されると，攻撃者である広告主は自身の広告エリアを観測することでユーザを特定することができる．これに対しては，既存の位置の匿名化に関する手法が適用できるが，ユーザの属性も考慮する場合には次のような問題が発生する．あるサービスにおいてユーザが図 2 のような位置・属性情報で u_1 から順番に匿名化およびサービスの要求をしたとする．

仲介サーバは，自身のプライバシを保護したいというユーザの要求とサービスの質のためできるだけ詳しくユーザの属性情報を知りたいという広告主の要求を両立させるため，さまざま



uid	年齢
u_1	23
u_2	26
u_3	22
u_4	38
u_5	24

図 2　ユーザの分布

なグループ化を検討し匿名化を試みる．k 人のユーザを位置情報に基づいてグループ化する k **匿名化** (k-anonymization) のアプローチを用いると位置情報のみが利用されるため，近傍のユーザとグループ化できるとき，たとえば u_1 から u_4 までがサービスの要求をした時点で u_1, u_2, u_4 ですぐに匿名化される (ここでは $k = 3$ とする)．ここでユーザの属性情報をそのまま提供すると，広告主にはこのエリアで 23 歳，26 歳，38 歳のユーザが広告配信を希望していることが通知される．攻撃者はこのエリアを観察することにより，38 歳のユーザをほぼ確実に特定でき，23 歳と 26 歳のユーザも高い確率で見分けることができる．

そこでより良い匿名化のため，仲介サーバは別のユーザがさらにサービスの要求をするのを待つ．そして，u_5 のサービスの要求により u_1, u_2, u_5 のグループ化を検討する．この場合には 23 歳，26 歳，24 歳と年齢が似ているため 3 人を見分けることが難しい．このため，先程の匿名化より適切な匿名化になっているといえる．

1.4　研究の目的

本研究では，上で述べたような位置情報のみでなく属性情報も考慮する匿名化の実現を目的とする．このために，その属性が観測によりどの程度推測できるかということが重要なポイントになる．そのためそれを数値化する一致度という新たな指標を導入する．また，上の例では述べていないが，利用者によってプライバシポリシーは異なるため，そのようなユーザの嗜好を反映する匿名化を実現する．

2.　関 連 研 究

2.1　位置に基づくサービスにおける匿名化処理

位置に基づくサービスにおけるプライバシに関する問題について，これまで多くの研究がされてきた．既存手法のアプローチとして主流であるのが，**空間クローキング** (spatial cloaking) に基づく手法である．空間クローキングでは，ユーザの位置を包含する**秘匿領域** (cloaked region) を構築し匿名化をおこなう. [4] では，秘匿領域の生成に，一般のデータベース出版 (database publishing) に関する研究でしばしば用いられる k **匿名化** (k-anonymization) [8] の考え方が用いられている．多くの研究でこの考え方は用いられ，グラフ構造を利用するもの [3] やセル分割を利用するもの [1], [7] などさまざまな手法があり，本研究の手法でもこの考え方を発展させて用いる．他にも空間クローキングにはさまざまなアプローチがあるが，*location anonymizer* と呼ばれる，モバイルユーザとサービス提供者の間に位置し，全ユーザの位置情報を把握し，匿名化処理を実行する信頼できるサードパーティを想定することが一般的である．本研究では，仲介業者がこの役割を果たす．

これまで提案されてきた位置情報の匿名化のアプローチは，本研究で対象とする位置情報に加え属性情報も考慮した匿名化処理には対応することができず，新たな技術開発が必要である．この種の匿名化手法に関する提案としては [9] がある．この手法では，ユーザの属性の値に応じて属性のベクトルを作成する．匿名化処理では，近傍にいるユーザからできるだけこのベクトルの距離が近いユーザを探索し秘匿領域を生成することで匿名化をおこなうアプローチをとる．ただし，この手法では，属性の種類によって異なる攻撃者からの観測されやすさを考慮しておらず，必要以上に属性を一般化することにより，サービスの質に悪影響を与えるおそれがある．

2.2　属性の観点

静的なデータベース出版における k-匿名性 (k-anonymity) におけるプライバシ保護 [8] では，ユーザの属性を識別子，準識別子，機密情報の 3 つに分類する．

- **機密情報** (sensitive attribute)：患者と病気の関係における病気データのような個人との結びつきを秘匿したい情報
- **識別子** (identifier)：氏名や住所など個人と 1 対 1 に結びつく情報．識別子が出版されるデータにあるとプライバシが侵害される．
- **準識別子** (quasi-identifier)：年齢，性別など個人と 1 対 1 には結びつかない情報．しかしその組合せがユニークなものであった場合，別のデータ (選挙人名簿) などと組み合わせることでプライバシが侵害されるおそれがある．

データベース出版では，プライバシを保護するために属性の扱いを変えて処理するが，本研究が対象とする位置に基づくサービスで用いる属性については，そのとらえ方が異なる．位置に基づくサービスでは，サービスの対象エリアに訪ずれるユーザを想定することは困難であり，別データと準識別子を照らしあわせて個人を特定する攻撃を受ける可能性は小さい．それよりも問題であるのは，実際にそのサービスの対象エリアを観測されて，位置情報と属性情報をもとにユーザが特定されてしまうことである．そこで本研究では属性の扱いを見直す必要がある．

属性のとらえ方に関する関連研究の一つとして，ソーシャルネットワークにおけるプライバシ保護のための属性の扱いに着目した [5] がある．この論文では，プライバシに関する属性のとらえ方の基準として，以下の二つを考えている．

- **機密性** (sensitivity)：その属性がどれだけプライバシ侵害につながりやすいか，という基準である．たとえば「住所」は自宅の特定につながり非常に機密性が高いが,「出身地」はそれほど個人の特定にはつながらず，機密性は「住所」ほどは高くない. [5] では，各属性の機密度 (機密性の度合い) はユーザには依存せず，属性ごとに個別の値をとる，としている．これは，Facebook は実名登録であるため「本名」の機密度は低いが，他のハンドルネームを利用する SNS では「本名」の機密度が高くなるなど，システムによって異なる値を持つ．
- **可視性** (visibility)：ユーザがどれだけ詳細な属性値を公開するかという基準であり，ユーザごとに異なる値を持つ．たとえば「生年月日」という属性を考えてみると，年を含め全て公開，月と日だけ公開，非公開などユーザの公開ポリシーは異なる．この場合，全て公開しているユーザは可視性を高く設定し，非公開のユーザは低く設定しているといえる．

これら二つの基準を，位置に基づくサービスでも用いるのには検討が必要である．位置に基づくサービスでは，利用者がある属性値を隠したいと望もうと望むまいと，属性によっては攻

撃者がユーザを観察すればその値が明らかになってしまうものがあり制御することができない．すなわち，位置に基づくサービスにおいては，ユーザがどれだけ情報を秘匿したいか，公開したいかということよりも，外から観察したときにどれだけその属性値を推測できるかが重要となる．本研究では，これを**可観測性** (observability) と呼ぶ．これについては後ほど詳しく述べる．

2.3 個人の嗜好を反映した匿名化

本研究が想定する状況においては，どの属性をどの程度公開してもよいかは個人に依存する部分が大きいため，個人の嗜好を反映したプライバシ保護が必要となる．しかし，一般の匿名化ではユーザの嗜好を反映させることができない．たとえば，病気のデータにおいて，ガンの人はそれを隠しておきたいが，風邪の人は知られても構わない，と考えていたとする．[10] はこのような利用者ごとに異なるプライバシ保護の要求を満す静的なデータベースにおける手法である．この手法では各属性について**階層構造** (タキソノミ) を構築しておく．各ユーザはその階層構造のどのレベルまで詳細化してよいかを指定することで，プライバシ保護のレベルに個々の嗜好を反映させることができる．本研究では，個人の嗜好を反映するこのようなアプローチを発展させて用いる．

3. 提案手法の概要

3.1 匿名化の目標

匿名化においてモバイルユーザのプライバシ保護とサービスの質の両立のため，次の方針で匿名化をおこなう．

- 識別確率：ユーザとプロファイルが結びつく確率はユーザにとっては低いほうが良いが，サービス提供者にとっては高いほうが良い．そこで後述のプロファイルでユーザはその閾値を設定する．匿名化処理では閾値以下でできるだけ高い識別確率となるよう匿名化する．
- 属性の汎化：ユーザの属性についても適度に一般化する必要がある．しかし，属性の汎化はサービスの質を落とすことにつながり，ユーザごとにどこまで汎化してもよいかの嗜好は異なる．そこでユーザは属性の開示レベルを設定し，匿名化の過程では，この開示レベル以下でできるだけ被覆領域内の属性が似るように汎化する．
- 領域サイズ：被覆領域が大きすぎるとサービスの質の低下につながる．本研究では，被覆領域の大きさをシステムによって定められた最大サイズ以下にする．

3.2 属性のタキソノミ

匿名化過程における汎化処理で用いる属性のタキソノミについて述べる．各属性について，あらかじめシステムによって定められた階層的なタキソノミが設定されているとする．例として，年齢のタキソノミを図 3 に示す．レベル 0 のノード any は根ノードであり，全ての値を含む．そして，根ノードから葉ノードにレベルが進むにつれて値が詳細化されていく．図 3 では [20-39] ノードの子孫しか示していないが，[-19] ノードや [40-] ノードについてもやはり同様に子孫ノードは定義されているものとする．タキソノミは，この年齢属性以外についても，性別，ZIP コード，職業などあらゆる属性について作成されるが，本論文ではその作り方については踏みこまない．

また，各ユーザは属性の**開示レベル** (disclosure level) を指定できるものとする．たとえば，あるユーザの年齢が 23 歳であるとする．そのユーザが 20 代であることまで情報をぼかしてよいなら，ノード [20-29] を開示ノードとして指定する．開示レベルが葉ノードに近づけば近づくほど，より個人に特化された広告等のサービスを受けることができることを意味する．しかし，葉ノードに近いほど，匿名性は低くなりやすいことから，開示レベルの設定は，プライバシ保護とサービスの質にどれだけ重みを置くかのトレードオフとなる．

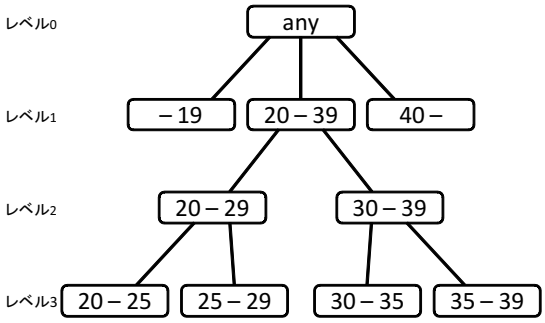


図 3　年齢のタキソノミ

3.3 プロファイル

各モバイルユーザは，自身のサービスや匿名化への嗜好を明確にするためプロファイルを利用する．プロファイルは信頼できる第三者である仲介サーバで管理される．プロファイルの例を図 4 に示す．プロファイルには，サービスを得るために用いる属性情報と，匿名化の条件の 2 つがあり，匿名化の条件は，さらに属性の開示レベルと識別確率閾値の 2 つに分かれる．これらの詳細は次の通りである．

- 属性値：各属性についてそのユーザの持つ値を示す．
- 属性の開示レベル：各属性を匿名化の際どこまでならばぼかしてよいかを示す．より強い汎化 (より曖昧になる) を許せば，匿名化の成功率の上昇が期待できるが，自身の属性に合わないものも対象となるため，サービスの質の低下につながるおそれがある．
- 識別確率閾値：提供者サーバに送信される匿名化されたプロファイルが自身のものであると識別されてしまう確率をこの閾値以下にして欲しいということを示す．既存研究では匿名化によって生成される被覆領域に含まれるユーザ数 k が用いられることが多いが，識別確率の閾値を設定することでユーザ数 k よりもわかりやすいプライバシ保護を実現することができる．

ID	年齢	年齢汎化許容	性別	性別汎化許容	識別確率閾値
u_1	23	[20-29]	男	[Any]	0.4
u_2	26	[20-39]	男	[男]	0.5
u_3	22	[20-24]	女	[女]	0.6
u_4	38	[30-39]	女	[Any]	0.5
u_5	24	[20-24]	男	[男]	0.5

図 4　プロファイルの例

3.4 属性の基準

位置に基づくサービスでは，ユーザが対象とするエリアにまさしく存在するため，攻撃者がその地点を観測すれば，サービスを要求したユーザの候補について一部の情報がわかってしまう．そこで，本研究では位置に基づくサービスにおけるプライバシ保護のための属性の新たな基準として，**可観測性** (observability) を導入する．

- **可観測性** (observability)：ユーザを外部から観察したと

きに，その属性をどれだけ推測しやすいかを示す．たとえば，「性別」は観察すればその推測は容易であるのに対し，「出身地」は観測のみで判別することは困難である．

本研究では，プライバシに関わる属性の性質について，以下のようなアプローチをとる．

- 機密性について：先に述べたように，属性の種類によってはユーザが望まなくても，外部からの観察によってその属性値が攻撃者に知られてしまう．ただし，属性には開示レベルを指定でき，より一般化されたノードまで汎化を許可することで，より送信されるプロファイルの属性値を曖昧にすることができるため，これにより機密性に関するユーザの要求を反映することができる．
- 可視性について：可視性についても，ユーザの指定する開示レベルによって制御されるものとする．開示レベルを葉ノードに近づければ可視性は高くなり，根ノードに近づければ可視性は低くなる．
- 可観測性について：各属性についての可観測性はシステムにおいて固定された値をとるものとする．「性別」といった属性が最大となり，「出身地」は小さい値をとる．

3.5 一 致 度

上で述べた可観測性を匿名化アルゴリズムで利用するために，属性の可観測性を数値化する必要がある．可観測性とはあるユーザの属性値が構築された属性のタキソノミにおけるあるノードにどれだけ結びつきやすいか (観察されたときにどのように推測されるか)，ということである．たとえば「年齢」の 21 という属性値はノード [20-24] に強く結びつき，ノード [15-19] には少しだけ結びつくが，ノード [30-34] にはほとんど結びつかない．このユーザ u_i とタキソノミのノード n_k が一致する度合いを**一致度** (matching degree) と呼び，$match(u_i \to n_k)$ で表す．一致度を件付確率として

$$match(u_i \to n_k) = \Pr(n_k \mid u_i) \tag{1}$$

と定義する．タキソノミのあるレベルのノードが K 個のとき一致度は，

$$\sum_{k=1}^{K} match(u_i \to n_k) = \sum_{k=1}^{K} \Pr(n_k \mid u_i) = 1 \tag{2}$$

が成り立つ．仲介サーバはあらゆる属性値とタキソノミの各ノードの一致度を前もって用意しており，u_1 は 23 歳などモバイルユーザの属性値に応じて必要な値を取り出す．一致度の例を図 5 に示す．ここではレベル 1 のノード [Any] が省略されていたり，レベル 2 のノードが 20 代から 30 代のもの ([20-39]) しか示していないが他のノード (40 代から 60 代を示す [40-59] など) も存在するものとする．

	レベル 2	レベル 3		レベル 4			
uid	[20-39]	[20-29]	[30-39]	[20-24]	[25-29]	[30-34]	[35-39]
u_1	0.88	0.88	0.00	0.54	0.34	0.00	0.00
u_2	1.00	0.90	0.10	0.38	0.52	0.10	0.00
u_3	0.79	0.79	0.00	0.56	0.23	0.00	0.00
u_4	0.64	0.00	0.64	0.00	0.00	0.11	0.53
u_5	0.97	0.95	0.02	0.51	0.44	0.02	0.00

図 5　属性の一致度

ユーザのプロファイルは，複数の属性のタキソノミのノードの集合といえる．本研究ではプロファイルの各属性は独立であると想定する．そこで一致度は確率のように扱うことができるので，プロファイルの一致度は各属性の一致度の積で表わすことができる．

3.6 識 別 確 率

匿名化処理が成功したか否かの判定では，匿名化済のプロファイルとその被覆領域内のユーザとが結びつき，どのプロファイルが誰のものであるか識別されてしまう確率 (**識別確率**) を用いる．この識別確率がユーザがプロファイルで指定した閾値以下であればユーザの指定より強く保護できていると判定することができる．識別確率は先に述べた一致度を用いて計算できる．

3.6.1 確率の計算方式：ユーザが 2 名の場合

ユーザ (u_1, u_2) と匿名化されたプロファイル (p_1, p_2)(図 6) を考える．ユーザとプロファイルの対応付けは明らかにされていないものとする．このため，攻撃者は $(u_1 : p_1, u_2 : p_2)$ または $(u_1 : p_2, u_2 : p_1)$ の 2 通りの対応付けを考えることになる．明らかに，

$$\Pr(u_1 : p_1, u_2 : p_2) + \Pr(u_1 : p_2, u_2 : p_1) = 1 \tag{3}$$

が成立する．

pid	タキソノミのノード
p_1	[20-24]
p_2	[25-29]

図 6　プロファイル

確率の計算には次のようなアイディアを用いる．各ユーザ u_i について，サイコロを振る．サイコロにはタキソノミのノードごとに面があり，その面が出る確率は一致度にしたがう．この例では，u_1, u_2 のサイコロを同時に振る．このときサイコロの目の出かたには，$(u_1 : p_1, u_2 : p_1), (u_1 : p_1, u_2 : p_2), (u_1 : p_2, u_2 : p_1), (u_1 : p_2, u_2 : p_2)$ の 4 パターンがある．$(u_1 : p_1, u_2 : p_2)$ の出現確率は，

$$\Pr(p_1 \mid u_1) \times \Pr(p_2 \mid u_2) = 0.54 \times 0.52 = 0.281 \tag{4}$$

であり，$(u_1 : p_2, u_2 : p_1)$ の出現確率は，

$$\Pr(p_2 \mid u_1) \times \Pr(p_1 \mid u_2) = 0.34 \times 0.38 = 0.129 \tag{5}$$

となる．ただし，$(u_1 : p_1, u_2 : p_1), (u_1 : p_2, u_2 : p_2)$ は禁止されている組合せ (一つのプロファイルが複数のユーザには対応しない) なので，これが出た場合はサイコロの試行自体がなかったものとなる．よって，有りうる場合のみを考慮すると，

$$\Pr(u_1 : p_1, u_2 : p_2) = 0.281/(0.281 + 0.129) = 0.69 \tag{6}$$

$$\Pr(u_1 : p_2, u_2 : p_1) = 0.129/(0.281 + 0.129) = 0.31 \tag{7}$$

となる．

3.6.2 確率の計算方式：一般の場合

基本的には 2 人の場合と同じ考え方で進める．たとえばユーザが 3 人の場合，6 通りの対応付けのそれぞれについてまず確率を計算して，それらの和によって各確率を割ればよい．

また，匿名化において大事なのはユーザごとの識別確率である．たとえば，ユーザ (u_1, u_2, u_3) とプロファイル (p_1, p_2, p_3) があるとき，ユーザ u_1 にとっては自身が各プロファイル p_1, p_2, p_3 と結びつく確率が閾値以下になるかどうかが重要であり，ユーザ u_2, u_3 の識別確率については気にしない．このような個別の

確率については，あるプロファイルと結びつく組合せの確率の和で求めることができる．この例では，ユーザ u_1 とプロファイル p_1 が結びつくと識別される確率は

$$\Pr(u_1 : p_1) = \Pr(u_1 : p_1, u_2 : p_2, u_3 : p_3) + \Pr(u_1 : p_1, u_2 : p_3, u_3 : p_2) \quad (8)$$

で求められる．今後の議論では識別確率とは，この個別の識別確率を指す．

4. 匿名化処理のアルゴリズム

アルゴリズムで使用する記号を表 1 にまとめる．匿名化処理のアルゴリズムは，プロファイルの汎化処理と被覆領域生成処理の二つから成る．

表 1　アルゴリズム中で使用する記号

記号	意味
u_i	モバイルユーザ
p_j	プロファイル
n_k	タキソノミのノード
u_q	サービスを要求して，匿名化処理を発生させたユーザ
$u_q.t$	u_q がサービスを要求した時間
$u_q.e_t$	u_q が匿名化を諦めて失敗とする時間
$u_q.th$	u_q の識別確率の閾値
U_R	ある被覆領域に含まれるユーザの集合，匿名化の候補
$\mathcal{U}_C$	U_R の候補の集合
H_U	サービスを要求したユーザを管理するヒープ
P_R	U_R におけるプロファイルの集合

4.1 プロファイルの汎化

各モバイルユーザの識別確率を下げるために，匿名化ではある被覆領域にいるユーザのプロファイルの汎化をおこなう．プロファイルは先に述べたようにタキソノミのノードの集合であり，各属性は独立であると想定するため，プロファイルを汎化するには各属性ごとにそのノードを汎化すればよい．識別確率は汎化をすればするほど小さくなるわけではなく，できるだけ全てのノードが近いものであるほうが良い．本研究では，全てのノードの共通の先祖ノードのうち最もレベルの高いノード (最小上界) に近づけることにする．最小上界までは各ノードは汎化すればするほど識別確率は小さくなるが，それ以上は汎化しても識別確率は小さくはならず，サービスの質を悪化させるだけである．そして，すべてのノードが最小上界に汎化されたとき，最良の識別確率 (N 人のとき $1/N$) を得る．

被覆領域内にユーザが N 人いるときの，ある属性のノードの汎化をアルゴリズム 1 に示す．$getLUB(n_1, n_2, ..., n_N)$ はある属性のタキソノミのノード n_1 から n_N の最小上界を返す．たとえば図 3 のタキソノミにおいて，$getLUB([20-25], [25-29]) = [20-29]$，$getLUB([20-25], [30-39], [40-]) = [any]$，$getLUB([20-29], [20-25]) = [20-29]$ となる．$generalize$ はノードの汎化をおこなう関数で，n_i を指定されたレベルに汎化する．ここでは最小上界もしくはそのユーザの開示レベルのどちらかレベルが高いものに汎化する．

4.2 被覆領域の生成

あるユーザがサービスの要求をしたときにおこなわれる匿名化の基本的な流れをアルゴリズム 2 に示す．2 行目では，サービス要求があった時間と anonymizer で保持する時間 (処理の締切時間) を優先度付きキュー H_U に挿入する．5 行目ではグループ化したユーザによる MBR がシステムで定められた最大サイズより小さいか確認する．6 行目の $generalizeProfile$ はプロファイルの汎化を行なう関数であり，プロファイルの各属性について，それぞれ先述した $generalizeNode$ によってノードの汎化をおこなう．7 行目から 12 行目では，汎化後のプロファイルの識別確率が閾値以下か確認され，成功ならその U_R に含まれるユーザを含む集合が全て $\mathcal{U}_C$ から取り除かれる．17 行目からの関数は締切時間を越えたサービス要求を匿名化失敗として処理するもので，適切なタイミングで呼ばれる．

アルゴリズム 1 ノードの汎化

```
1: procedure GENERALIZENODE
2:     n_lub ← leastUpperBound(n_1, n_2, ..., n_N)
3:     for all i such that 1 ≦ i ≦ N do
4:         n'_i ← generalize(n_i, max(u_i.disclosure_level, n_lub.level))
5:     end forreturn {n'_1, n'_2, ..., n'_N}
6: end procedure
```

アルゴリズム 2 被覆領域生成処理

```
1: procedure ANONYMIZATION(u_q)
2:     Add {u_q, u_q.t + u_q.e_t} into H_U order by u_q.t + u_q.e_t
3:     for all U_R such that U_R ∈ U_C do
4:         U_R ← U_R ∪ u_q
5:         if getMBRSize(U_R) ≦ MAX_RECT_SIZE then
6:             P_R ← generalizeProfile(U_R)
7:             if ∀u_i ∈ U_R, ∀p_j ∈ P_R, Pr(u_i : p_j) ≦ u_i.th then
8:                 ∀S ∈ U_R, removeSfromU_C
9:                 return U_R, P_R
10:            else
11:                U_C ← U_C ∪ U_R
12:            end if
13:        end if
14:    end for
15: end procedure

16: procedure CHECKEXPIRATION
17:    while true do
18:        {u, deadline} ← Pop first item in H_U
19:        if deadline > now then
20:            Remove AllSetsContain(u) from U_C
21:        else
22:            break
23:        end if
24:    end while
25: end procedure
```

図 2 のユーザを用いて実際の流れの例を以下に示す．サービス要求は u_1, u_2, u_3, u_4, u_5 の順でおこなわれたとする．仲介サーバにおける候補の管理の様子を図 7 に示す．この図は初期状態から $u_1, u_2, ...$ と順番にユーザが問合せをおこなうことで仲介サーバで管理される被覆領域の候補が増えていく様子を示している．被覆領域の候補はユーザとそのプロファイル，およびその被覆領域におけるそのユーザの識別確率からなる．

初期状態では候補 $\mathcal{U}_C = \phi$ であるが，ユーザが次々とサービスの要求により候補が増えていく．そして先述したアルゴリズムに従い，プロファイルの汎化および識別確率の計算がおこな

問合せユーザ	候補
(初期状態)	{ϕ}
u_1	{ϕ, **{u_1[20-24]:1.0}**}
u_2	{ϕ, {u_1[20-24]:1.0}, **{u_2[25-29]:1.0}, {u_1[20-29]:0.5, u_2[20-29]:0.5}**}
u_3	{ϕ, {u_1[20-24]:1.0}, {u_2[25-29]:1.0}, {u_1[20-29]:0.5, u_2[20-29]:0.5}, **{u_3[20-24]:1.0}**}
u_4	{ϕ, {u_1[20-24]:1.0}, {u_2[25-29]:1.0}, {u_1[20-29]:0.5, u_2[20-29]:0.5}, {u_3[20-24]:1.0}, **{u_4[30-34]:1.0}, {u_1[20-29]:1.0, u_4[30-39]:1.0}, {u_2[20-39]:0.91, u_4[30-39]:0.91}, {u_1[20-29]:0.55, u_2[20-39]:0.5, u_4[30-39]:0.95}**}
u_5	{ϕ, {u_1[20-24]:1.0}, {u_2[25-29]:1.0}, {u_1[20-29]:0.5, u_2[20-29]:0.5}, {u_3[20-24]:1.0}, {u_4[30-34]:1.0}, {u_1[20-29]:1.0, u_4[30-39]:1.0}, {u_2[20-39]:0.91, u_4[30-39]:0.91}, {u_1[20-29]:0.55, u_2[20-39]:0.5, u_4[30-39]:0.95}, **{u_5[20-24]:1.0}, {u_1[20-24]:0.5, u_5[20-24]:0.5}, {u_2[20-29]:0.56, u_5[20-24]:0.56}, {u_1[20-29]:0.4, u_2[20-29]:0.37, u_5[20-24]:0.34}**}
({u_1,u_2,u_5}の出力後)	{ϕ, {u_3[20-24]:1.0}, {u_4[30-34]:1.0}}

図 7　候補の管理

われる．そして図 4 のプロファイルによると u_1 の識別確率の閾値は 0.4 であるので，u_1 については自身の識別確率が 0.4 以下になる被覆領域が生成できれば成功ということになる．位置属性については被覆領域の MBR の最大値がシステムによって定められているため，u_3 のように u_1 や u_2 と離れた地点にいるユーザは u_1, u_2 と一緒の候補になることはない．

計算を続けていくと，u_4 まではユーザの閾値を満たすグループ化はおこなえないが，u_5 がサービスの要求をしたときに検討される u_1, u_2, u_5 による被覆領域が，それぞれの閾値以下の識別確率となるため匿名化成功となる．仲介サーバは u_1, u_2, u_5 の情報をサービス提供者に送信し，$\mathcal{U}_C$ から u_1, u_2, u_5 を含む候補をとりのぞく．そしてまだ匿名化が成功していない u_3, u_4 については，さらに別のユーザがサービスの要求をしてくるのを待つことになる．

4.3　戦略と評価基準

4.2 節で示したアルゴリズムは基本的なアルゴリズムであり，細かい点をあまり考慮していない．たとえばアルゴリズムでは，閾値を満たすユーザのグループを作ることができたらすぐに，そのグループによる匿名化をおこなう．この naive なアルゴリズムに対して，ユーザが指定する締切時間までにはまだ時間があるため，すぐには匿名化せずに別のより良いグループ化ができないか他のユーザがサービスの要求をするのを待つという戦略が考えられる．戦略の決定には，匿名化の結果をどのように評価するかが重要になる．評価基準は以下の通りである．

- **スループット**：サービスを要求したユーザのうち，何人のユーザを匿名化することができたかという基準である．ユーザにとってもサービス提供者にとっても多いほうが良い手法であると言える．
- **被覆領域の評価**：サービスの提供者にとっては，匿名化された属性値が，より詳細なものに近いほうがより良い匿名化であるといえる．その基準として，匿名化後の属性値を表わすタキソノミのノードのレベルの平均を用いる．たとえば，二つの被覆領域があり，一つの汎化後の年齢属性は {[20-24],[20-24],[25-29]} で，もう一つは {[20-24],[20-29],[20-29]} であるとする．このとき，[20-24] と [25-29] のレベルが 3，[20-29] のレベルが 2 なので，{[20-24],[20-24],[25-29]} は 3，{[20-24],[20-29],[20-29]} は 2.33 となり，より大きい前者の方が有用性が高いと評価できる．

5.　評 価 実 験

5.1　使用した実験データ

シミュレーションによる評価実験をおこなうにあたり，ユーザの位置情報については人工データと実データを用いた．人工データは中心座標および分散の異なる複数の二次元ガウス分布によって生成した．また，実データはオルデンブルクの道路ネットワークを用いた．実データにおけるモバイルユーザの生成には Brinkhoff の生成器 [2] を用いた．これはグラフデータを元に道路上を移動するオブジェクトを生成するものである．ただし，本研究はユーザの移動を考慮していないため，各オブジェクトの出現位置のみを利用した．

また，その他のパラメータの基本設定を表 2 に示す．ユーザのサービスの要求の発生はポアソン到着に従うとし，1/100 秒ごとに $\lambda = 0.1$ の確率で新しいユーザがサービスを要求する (ただし同じ時刻に複数のユーザがサービスを要求することもあり，その場合には同時にサービスを要求した別のユーザは，他のユーザの処理が終わるまで待つことになる)．一度サービスを要求したユーザは，その後再度サービスを要求することはない．ユーザの属性としては年齢データを用いた．年齢は 20 から 39 歳までとし，その一致度は図 5 を元に全ての年齢について設定したものを用いる．この年齢のタキソノミは図 3 のものを利用し，各ユーザの属性の開示レベルは 1($[20-39]$) から 3(葉ノード) とした．

表 2　基本的なパラメータ

ユーザの人数	1000 人
サービス要求の単位時間	1/100 s
ユーザのサービス要求頻度	10 回/ s
使用した属性	年齢データ
ユーザの年齢	$[20, 39]$
開示レベル	$1, 2, 3$
識別確率の閾値	$0.3, 0.4, 0.5$
締切時間	$10\,\mathrm{s} \pm 10\%$
被覆領域の最大サイズ	1000×1000

5.2　使用した被覆領域生成のための戦略

実験では，4.3 節で示した考え方に基づき，次のような戦略をとる手法を用いた．

- **naive**: アルゴリズム 2 に示したものである．順に問合せユーザを含めた候補を調べていき，**閾値を満たす被覆領域を作ることができたらすぐに**出力する手法．
- **deadline-based**: 閾値を満たす被覆領域を作ることができてもすぐには出力せずに，その被覆領域に含まれるユーザの最も近い締切が来るまで保持する．そして，保持している間，追加のユーザの要求があれば，まず**この保持している領域に追加することができるか**確認し，できないのであればアルゴリズム 2 にしたがい，他のまだ匿名化できていないユーザとの組合せによる被覆領域生成を試みる．
- **lazy**: deadline-based と似ているが，新しくサービス要求したユーザを追加する候補として，deadline-based が既に閾値を満たしている領域から先に調べるのに対し，この手法では

まだ閾値を満たしていない領域を先に調べていく．すぐに出力せずに締切まで待ち，匿名化できないユーザを取り込む naive とも言える．

- **many-first**: 閾値を満たすあらゆる被覆領域を保持しつつ，あるユーザの締切が来たら，そのユーザを含む被覆領域の中で**被覆領域内のユーザ数が最も大きいもの**を出力する．
- **next-deadline-based**: 閾値を満たすあらゆる被覆領域を保持しつつ，あるユーザの締切が来たら，そのユーザを含む被覆領域の中で**今出力するユーザの次に締切が近いユーザの締切** (被覆領域の中で 2 番目に近い締切) が最も早いものを出力する．
- **avg-deadline-based**: 閾値を満たすあらゆる被覆領域を保持しつつ，あるユーザの締切が来たら，そのユーザを含む被覆領域の中で**平均の締切時間**が最も早いものを出力する．
- **threshold-based**: 閾値を満たすあらゆる被覆領域を保持しつつ，あるユーザの締切が来たら，そのユーザを含む被覆領域の中で**閾値が最も低く，その中で閾値と識別確率の比が最も大きいもの**を出力する．

すなわち，deadline-based と lazy は，新しい問合せユーザと組合せる候補の順番が異なる手法であり，many-first，next-deadline-based，avg-deadline-based，threshold-based は基本的には同じ手法だが，あるユーザの締切が来たときに，いくつかあるそのユーザを含む被覆領域の中からどれを選択して出力するかが異なる．

5.3 実験 1：ユーザのサービス要求頻度

まずユーザがサービスを要求する頻度を変化させ，各手法の匿名化が成功したユーザ数の変化を調べた．ユーザ数のサービス要求頻度が高くなれば，あるユーザと同じ被覆領域に含まれるユーザ数が多くなるため，スループットが向上することが期待できる．ユーザのサービス要求頻度は 1 秒間に，5 回，10 回，50 回，100 回の 4 種類とした．またこの実験は人工データのみでおこない，サービス要求頻度以外のパラメータは表 2 のものを利用した．実験の結果を図 8 に示す．naive，deadline-based，lazy の 3 手法は，ユーザのサービス要求頻度が高くなるのにしたがい出力されるユーザ数を伸ばしている．それに対して，many-first，next-deadline-based，avg-deadline-based，threshold-based の 4 手法は，1 秒に 50 人以上がサービスを要求をするときに出力されるユーザ数が著しく少ない．その原因は，この 4 手法が閾値を満たすあらゆる被覆領域の候補を保持するため，候補となるユーザの増加によりその候補数が非常に多くなることである．図 9 に，naive，deadline-based，lazy，many-first の 4 手法における，サービスを要求しても，前のユーザの処理が終了しないために自身の処理が遅延して始まるユーザの人数と，その中で自身の締切になってもまだ処理が始まらないユーザの人数を示す．ただし遅延自体は多くの場合起こる (同時に複数のユーザがサービスの要求をすれば当然一人を除けば遅延することになる) ため，0.1 秒以上の遅延のみを対象とした．この結果から deadline-based と many-first で遅延が発生し，特に many-first ではそのほとんどが締切を越えて遅延していることがわかる．なお，next-deadline-based，avg-deadline-based，threshold-based の 3 手法は many-first と同じ値であった．many-first，next-deadline-based，avg-deadline-based，threshold-based は閾値を満たすあらゆる被覆領域を出力候補として保持しているため，極端にその候補の数が増えることでユーザのサービス要求に対し処理が追いつかなくなることが欠点であることがわかった．

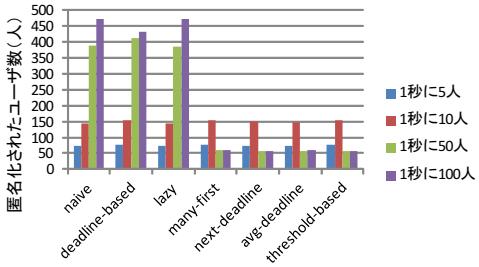


図 8　サービス要求頻度と出力人数

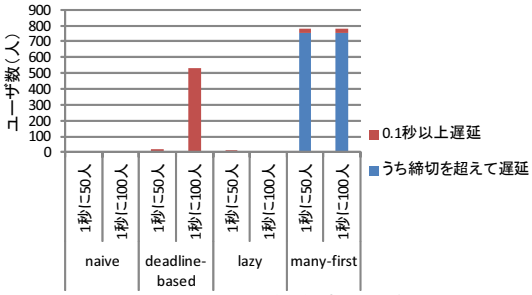


図 9　サービス要求頻度と遅延

5.4 実験 2：最大領域サイズの変化

次に被覆領域の最大サイズ (アルゴリズム 2 における定数 *MAX_RECT_SIZE*) を 500 × 500 から 2000 × 2000 まで変化させて実験をおこなった．

人工の位置情報と一様分布な属性分布における，各手法の匿名化できたユーザの人数を図 10 に示す．最大領域サイズが 2000 × 2000 のとき，avg-deadline-based のみに実験 2 でも発生した締切を越える遅延が発生し，出力されるユーザ数に影響が出た．出力される人数は many-first と deadline-based と threshold-based の手法が多い．また，出力されたユーザの属性がどのレベルまで汎化されたかを図 11 に示す．これによると，naive と lazy では最大サイズが大きくなったときでも，汎化されすぎず詳細なものに近い属性値を送信するため，サービスの提供者にとって良い匿名化といえる．

また，実データや属性の分布に偏りがあるものについても実験をおこなったところ，同様の結果が得られた．

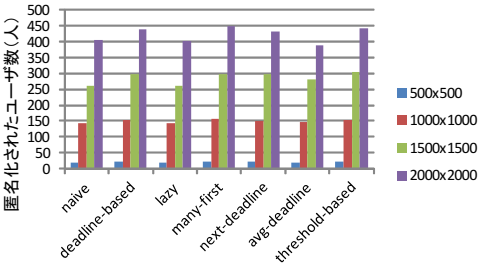


図 10　最大サイズとスループット

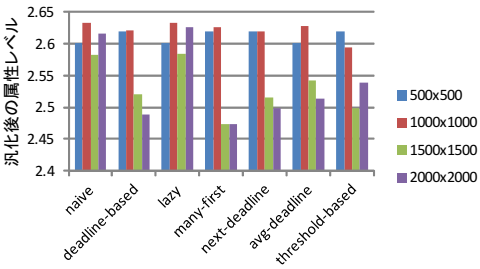


図 11　属性の平均汎化レベル

5.5 実験 3：ユーザの条件の変更

この実験では，表 2 のパラメータについて締切時間と識別確率の閾値を変更したときの振舞いを人工データを用いて調べた．

まず，締切時間を 10 ± 50% に変更して実験をおこなった．締切時間ごとに出力されたユーザの人数を図 12 に示す．next-deadline-based，avg-deadline-based の結果が良くなることを期待したが，締切時間を気にせずにユーザを取り込む deadline-based や many-first のほうが良い結果を出した．詳しく結果を調べたところ，確かに締切時間を考慮する手法は，many-first が出力しなかった締切時間の近いユーザを出力できていたが，その結果，より多くのユーザを含む被覆領域 (その中に締切時間がやや近いユーザも含まれていた) を出力することができていなかったことがわかった．

次に，締切時間は 10 ± 10% に戻し，確率の閾値に 0.2 を加えて実験をおこなった．閾値ごとに出力されたユーザの人数を図 13 に示す．こちらは締切時間と異なり，予想通り threshold-based が閾値の低いものを優先して出力しようとして，閾値が 0.3 のユーザの出力数で良い結果を示した．しかし，これも閾値を気にせずとにかくユーザ数の多い被覆領域を出力しようとする deadline-based や many-first よりも良い結果を出すことはできなかった．さらにどの手法であっても閾値が 0.2 のユーザの匿名化を成功させることはできなかった．

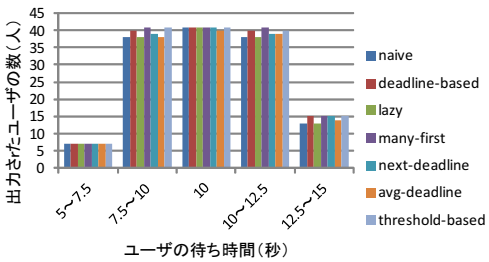


図 12　締切時間ごとの出力ユーザ数

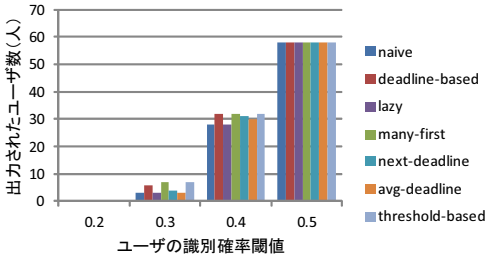


図 13　確率の閾値ごとの出力ユーザ数

5.6 考　　察

スループットという観点では many-first がとても良い性能を示した．ユーザの締切や確率の閾値を考慮する avg-deadline-based，next-deadline-based および threshold-based という手法と比べてみても，それらの違いをすべて吸収するような被覆領域を生成し出力することができる．ただし，これら 4 つの手法は次々とユーザがサービスの要求をするようなあまりにも作ることができる候補の数が膨大になるときに，その処理に時間がかかるため次にサービスを要求したユーザに対応できなくなってしまう欠点がある．このような膨大な候補ができるときには，naive な手法などでも十分なスループットを出すことができるため，適宜手法を切り替えて，できるだけ遅延がおこらないような仕組みを用意しておくことが実際の運用では必要になるだろう．

被覆領域の有用性という観点では lazy が良い性能を示した．この手法による匿名化では，ユーザの属性が汎化がされる度合いが低いため，サービス提供者にとって非常に有用性が高く，柔軟なサービスの提供が可能になる．また，一度にサービス要求をおこなうユーザ数が膨大になったときにも深刻な遅延を発生させずに処理を続けることができる．

いずれの手法も，他のユーザに比べて閾値が非常に低いような条件の厳しいユーザを匿名化することができなかった．匿名化が成功せずに締切が来たら，仲介サーバは匿名化が失敗したことをユーザに通知するが，その際失敗した原因 (そもそも周辺に他のユーザがいないのか，条件が厳しいためできなかったのかなど) を通知するような，匿名化が失敗したユーザをフォローする仕組みが必要であろう．

6. ま と め

本論文では，位置に基づくサービスにおけるユーザのプライバシの問題について，これまでの位置のみを考慮する匿名化手法に対し，それに加え属性情報も考慮する手法を提案した．そのために，観察によりどれだけ属性値を推測されるかという新たな属性の基準である可観測性と，それを確率として扱うために一致度を導入した．また，複数の手法の実装をおこない，評価実験により各手法の評価および考察をおこなった．

今後の課題は，スループットが高くロバストなものや，今回の手法では匿名化できなかった条件の厳しいユーザも匿名化することができるもののような，より優れたアルゴリズムの考案が挙げられる．

謝　　辞

本研究の一部は，内閣府最先端研究開発プロジェクト (FIRST) および科学研究費 (22300034) の助成による．

文　　献

[1] B. Bamba, L. Liu, P. Pesti, and T. Wang. Supporting anonymous location queries in mobile environments with PrivacyGrid. In *Proc. of WWW*, pp. 237–246, 2008.

[2] T. Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6:153–180, 2002.

[3] B. Gedik and L. Liu. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing*, 7(1):1–18, 2008.

[4] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proc. MobiSys*, pp. 31–42, 2003.

[5] K. Liu and E. Terzi. A framework for computing the privacy scores of users in online social networks. In *Proc. ICDM*, pp. 288–297, 2009.

[6] L. Liu. Privacy and location anonymization in location-based services. *SIGSPATIAL Special*, 1(2):15–22, 2009.

[7] M. F. Mokbel, C.-Y. Chow, and W. G. Aref. The New Casper: Query processing for location services without compromising privacy. In *Proc. VLDB*, pp. 763–774, 2006.

[8] P. Samarati. Protecting respondents' identities in microdata release. *IEEE TKDE*, 13(6):1010–1027, 2001.

[9] H. Shin, V. Atluri, and J. Vaidya. A profile anonymization model for privacy in a personalized location based service environment. In *Proc. MDM*, pp. 73–80, 2008.

[10] X. Xiao and Y. Tao. Personalized privacy preservation. In *Proc. ACM SIGMOD*, pp. 229–240, 2006.