系列データの匿名化に関する研究
A Study on Data Anonymization for Series Data

高橋 翼 ♥
Tsubasa TAKAHASHI

本稿では，学位論文として取り組んだ系列データの匿名化に関する研究の概要を紹介する．この研究では，系列データに対する連続的な匿名化と，系列中の属性値間の値の特定を困難にする匿名化 **(**関係多様化**)** について新たな手法の提案と評価を行った．

This paper presents a summary of a study on data anonymization for series data. This study is focus on a continuous anonymization method for trajectory stream and an anonymization which ensures relation diversity among attributes.

1. はじめに

近年，パーソナルデータが大量に蓄積され，第三者提供や二次活用による活用が期待されている．特に，パーソナルデータのシーケンスである系列データからは，系列中のパターンや因果関係といった高度な分析を実現できる．このような系列データをデータ保有者以外も利活用可能になれば，様々な分野で新たな発見やサービスの発展が期待できる．系列データの例として，移動軌跡，医療データをそれぞれ図 1，図**??**に示す．しかし，パーソナルデータの第三者提供や二次活用に際しては，プライバシへの配慮が必要となる．

1.1 データ匿名化

従来，氏名や会員番号などの直接識別子 (明示的識別子, Explicit Identifier) を削除するといった処理が匿名化として認知されてきた．k-匿名化 [1] に代表されるデータ匿名化では，1 つ以上の属性の組合せでレコードを識別し得る属性である準識別子 (Quasi-identifier, QI) を加工することで，レコードの識別を困難にする．さらに，レコードの特定だけでなく，他人に知られたくない属性であるセンシティブ属性 (Sensitive Attribute, SA) の特定を困難にする (ℓ-多様化 [2]) など，攻撃のモデルに合わせた匿名化手段が研究されている．

1.2 系列データの匿名化

系列データに対する既存の匿名化技術は，蓄積された複数の属性値の系列に対して静的に匿名化を行う技術であり，連続的に蓄積されていくデータを逐次匿名化することができなかった [3][4]．また，個人が特定されずともある属性の値が他の属性群から特定されてしまう場合がある．このようなときに，系列データの属性間の関係を多様化するような加工には，属性値や属性間の関係を大きく曖昧化してしまう問題があった．

表 1 系列データ:移動軌跡

u	l	t
Alice	(10, 5)	1
Alice	(15, 5)	2
Alice	(18, 8)	3
Alice	(18, 10)	4
Bob	(10, 5)	1
Bob	(9, 4)	2
Bob	(8, 3)	3
Bob	(8, 3)	4

表 2 系列データ:医療データ

ID	傷病	薬剤
1	a	x
2	a	y
3	b	x
4	b	y
5	a	x
6	a	z
7	c	x
8	c	w

表 3 匿名化の実行例

(a) 元テーブル

ID	年齢	性別	病名
1	22	男	かぜ
2	28	男	HIV
3	33	男	HIV
4	36	男	かぜ
5	42	女	ガン
6	48	女	ガン

(b) ID を削除したテーブル

年齢	性別	病名
22	男	かぜ
28	男	HIV
33	男	HIV
36	男	かぜ
42	女	ガン
48	女	ガン

(c) 2-匿名化したテーブル

年齢	性別	病名
[20, 29]	男	かぜ
[20, 29]	男	HIV
[30, 39]	男	HIV
[30, 39]	男	かぜ
[40, 49]	女	ガン
[40, 49]	女	ガン

(d) 2-多様化したテーブル

年齢	性別	病名
[20, 29]	男	かぜ
[20, 29]	男	HIV
[30, 49]	ANY	HIV
[30, 49]	ANY	かぜ
[30, 49]	ANY	ガン
[30, 49]	ANY	ガン

1.3 本研究の貢献

本研究では，系列データの連続的な利活用をプライバシを考慮しながら実現するために，主に移動軌跡ストリームを対象とした連続的匿名化手法を提案する．また，センシティブ属性間の多様性の保証を，属性値を加工せずに実現する関係多様化を導入し，さらに関係の曖昧性を抑止しながら関係多様化を実現する手法を提案する．

提案手法を用いることで，系列データの提供をプライバシに配慮した形で実現し，特にリアルタイムな移動軌跡の提供や，属性間の関係に多様性 が保証された系列データの提供が，あ

♥ 正会員 筑波大学，日本電気株式会社クラウドシステム研究所 t-takahashi@nk.jp.nec.com

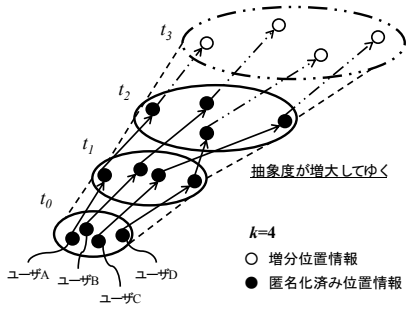


図 1 移動軌跡の連続的匿名化

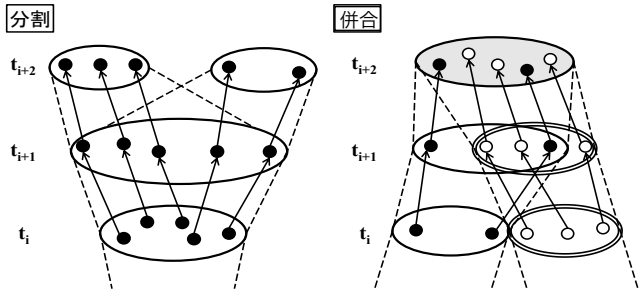


図 2 動的再構成

る一定の精度を維持しつつ匿名性が保証された形で実現する．

2. 移動軌跡ストリームの連続的匿名化

2.1 概要

位置情報の時系列である移動軌跡は，ユーザの移動経路や生活の動線を表す情報である．一方で，移動軌跡はユーザに特有の情報であるため，移動軌跡中のいくつかの位置情報から移動軌跡を識別することや，個人を特定することが可能である．識別，特定されることによって，センシティブな滞在先の漏洩が生じ得る．さらに，リアルタイムに移動軌跡を提供するような状況下では，追跡や監視といった脅威に晒される．

本研究では，移動軌跡をリアルタイムに提供するための連続的匿名化に関する問題を扱う．提案匿名化手法 CMOA (Continuous Moving Objects Anonymization) は，新たに測位された移動軌跡の増分位置情報を，既に匿名化された匿名移動軌跡を考慮しながらリアルタイムかつ連続的に匿名化を行う手法である．また，時々刻々と変化する人々の移動に合わせて匿名グループの構成を動的に再構成することで，過度な抽象化を抑制する．

2.2 移動軌跡ストリームの k-匿名性

各ユーザの位置情報 l は一定のインターバル毎に測位され，タイムスタンプ t を付加して信頼できるプラットフォームに蓄積されるものとする．l の例として，緯度・経度によって表される座標等がある．

定義 1 (移動軌跡ストリーム): データ主体 u の移動軌跡ストリームは位置情報の時系列で表される: $\tau_u = \{(u, l_0, t_0), (u, l_1, t_1), \ldots, (u, l_{last}, t_{last})\}(t_0 < t_1 < \ldots < t_{last})$．ここで，$t_{last}$ は最新のタイムスタンプ (現在時刻) とする．また，時刻 t の位置情報は $\tau_u[t_i](= l)$ と表す．

ここで汎化した位置情報を l^* とし，l^* のシーケンスから成る移動軌跡ストリームを汎化移動軌跡ストリームとする．

定義 2 (汎化移動軌跡ストリーム): $\tau^*_{tid} = \{(tid, l^*_1, t_1), (tid, l^*_2, t_2), \ldots, (tid, l^*_m, t_m)\}$ $(t_1 < t_2 < \ldots < t_m)$. tid は汎化移動軌跡ストリームの識別子であり，ランダムに割り当てた値を用いる．

移動軌跡ストリーム中の位置情報 l を汎化して，複数の移動軌跡ストリームが滞在する場所が共起するようにすることで，ある場所に滞在する移動軌跡ストリームがどのデータ主体のものであるのか，識別を困難にすることができる．本研究では，以下の移動軌跡の k-匿名性の充足を考える．

定義 3 (移動軌跡の k-匿名性) 移動軌跡ストリーム τ_u(または汎化移動軌跡ストリーム τ^*_{tid}) は，時間 $[t_i, t_j]$ に共通の場所に滞在する他の移動軌跡ストリームが $k-1$ 個存在するとき，時間 $[t_i, t_j]$ に k-匿名性を満たす．

2.3 提案手法 CMOA

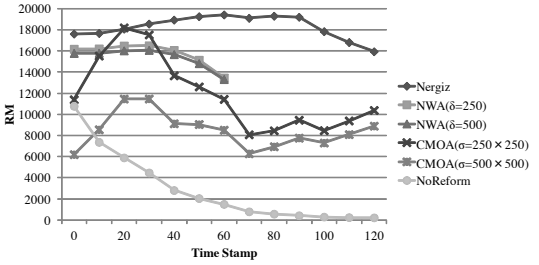
提案手法 CMOA は，移動軌跡ストリームを一定の面積 (σ) 以下に加工をして連続的に k-匿名化を行う手法である．CMOA は，移動軌跡の各位置情報を k 個以上の移動軌跡が常に包含されるように汎化する．ここで，汎化された移動軌跡が等価な移動軌跡を等価クラスと呼ぶ．k-匿名性を維持し続けるためには，既に匿名かして公開したデータと k 以上の移動軌跡が同じ等価クラスに属し続けることが必要になる．しかし，同じメンバで等価クラスを形成し続けると，メンバの移動方向が変わることで汎化の度合いを大きくなってしまう (図 1)．

そこで，移動軌跡ストリームの等価クラスの組み替え (動的再構成) を導入し，位置情報の精度を σ に保つ．

2.3.1 動的再構成

動的再構成では，分割と併合による再構成を行う (図 2)．分割では，メンバ数が $2k$ 以上かつ面積が σ を超えた等価クラスを 2 分割する．分割を繰り返すことで面積を小さく保つことができるが，一定以上の分割を行うと，k-匿名性を充足した分割が不可能になる．分割不可能な場合には，近傍の等価クラスとの併合を行い，十分なメンバ数を持った新たな等価クラスを作る．併合後に分割が行われると k-匿名性に違反してしまう場合があり，このときには TID の再割当を行う．ID の再割当が生じると，移動軌跡をトレーサビリティが毀損し，匿名化データの有用性が低下してしまう．

そこで併合を行う際には，併合後の等価クラスがどの程度トレーサビリティを維持できるかを評価して，この評価値が高い等価クラス群で併合を行う．このトレーサビリティの維持可



能度合いを以下の評価式 $E(c[t])$ で評価する．

$$E(c[t]) = \frac{\sigma(1+\log_2 \frac{|c[t]|}{k})}{S(c[t])} \quad (1)$$

ここで，$S(c[t])$ は時刻 t におけるクラス c の面積である．

2.4 評価

手法の有効性を評価するために，匿名化した移動軌跡の精度 (解像度) の評価と計算時間の計測を行った．

実験用のデータセットとして，東京大学が提供している人の流れデータ (PFLOW[5]) を利用した．PFLOW からランダムに 10 万人分の 2 時間分の移動軌跡を抽出して利用した．

図 2.4 は提案手法 CMOA と，動的再構成をしない連続的匿名化 (No Reform)，静的な手法 (NWA[3] と Nergiz 方式 [4]) と解像度の比較を行った結果を示している．RM は位置情報の解像度を表す指標であり，大きいほど詳細な位置情報であることを表す．提案手法は，動的再構成によって一定以上の解像度を維持していることがわかる．一方，動的再構成のない手法は解像度が単調減少していることがわかる．移動軌跡の全容が分かった上で匿名化する静的手法と比べて，提案手法は解像度が低い．ただし，一部静的手法の解像度に近づいており，一定の有用性は有していると言える．

以上より，移動軌跡をリアルタイムかつ連続的に k-匿名化する提案手法が，ある一定の位置情報の精度を保証しながら匿名化できることが分かった．なお，位置情報の精度だけではなく，一定のトレーサビリティを保証できるようにすることは今後の課題の一つであると考えている．

3. センシティブ属性間の関係多様化

3.1 概要

複数のセンシティブ属性を持ったパーソナルデータからは，属性間の相関や変化を観察することができる．一方で，あるセンシティブ属性に関する知識から他のセンシティブ属性値が特定され得る．この特定を防ぐためには，あるセンシティブ属性から他のセンシティブ属性が一意に対応付かないように，センシティブ属性間の対応関係が多様 (多対多) になること (関係多様性) を保証する必要がある．しかしながら，属性間の対応関係の多様化を実現するためには，表 4 のように属性値が過度に加工されてしまう問題がある．

そこで，センシティブ属性をそれぞれ異なるテーブルへと分割し，対応関係を曖昧化する関係多様化という操作を導入する (表 5(a)，表 5(b))．このとき，センシティブ属性を汎化せずに関係多様性を保証することができる．本研究では，関係の曖昧化を抑制しながら関係多様化する手法を実現する．

表 4 汎化データ

傷病	薬剤
a / b	x / y
a / b	x / y
a / b	x / y
a / b	x / y
a / c	w / x
a / c	w / x
a / c	x / z
a / c	x / z

表 5 関係多様化データ

(a) 傷病

CID	傷病
1	a
1	b
1	a
1	b
2	a
2	c
3	a
3	c

(b) 薬剤

CID	薬剤
1	x
1	y
1	x
1	y
2	x
2	w
3	x
3	z

3.2 関係多様化

以降では，関係多様化されたテーブル T_i と T_j の二項関係を対象に議論する．3.1 節で述べたセンシティブ属性値の特定を防ぐためには，センシティブ属性 S_i に対応する他のセンシティブ属性 S_j の属性値が一定の種類以上存在することを保証する必要がある．

二項関係にあるセンシティブ属性に対して，一方のセンシティブ属性値から特定可能な他方のセンシティブ属性値の種類数を表す関係の多様性指標 (ℓ_1,ℓ_2)-関係多様性を定義する．

定義 4 ((ℓ_1,ℓ_2)-関係多様性) 任意のタプルについて，センシティブ属性 S_i から特定可能なセンシティブ属性 S_j の属性値の種類数が ℓ_2 以上，S_j から特定可能なセンシティブ属性 S_i の属性値の種類数が ℓ_1 以上のとき，(S_i, S_j) に二項関係を持つ T_i と T_j は (ℓ_1,ℓ_2)-関係多様性を満たす．

ここで同一の CID を持つタプルの集合をクラスと呼ぶ．

(ℓ_1,ℓ_2)-関係多様性を保証する (ℓ_1,ℓ_2)-関係多様化を行うと，S_i の値はサイズ ℓ_1 以上の集合に，S_j の値はサイズ ℓ_2 以上の集合となる．よって，(S_i, S_j) の二項関係は，集合間の二項関係へと曖昧化される．

関係多様化によって混入するオリジナルの関係には存在しない関係をノイズとし，クラス中のオリジナルの関係とノイズとの比を表す関係ノイズ比 RNR を以下のように定義する．

定義 5 (関係ノイズ比)

$$RNR(c) = \frac{|S_i(c)||S_j(c)|}{|R(c)|} \quad (2)$$

ここで，$|S_i(c)|$ は c の S_i の属性値の種類数を表す．$|R(c)|$ は c の元の関係の種類数である．$RNR(c)$ は 1 以上の値を取り，最小値 (1) のとき，クラスにノイズが混入していないことを表す．ノイズの混入のないクラスをノイズレスクラスと呼ぶ．

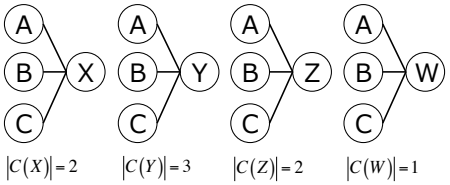


図 3　前提部の ℓ_1-多様化 ($(\ell_1, 1)$-関係多様化)

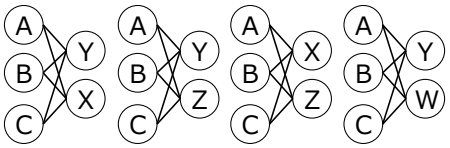


図 4　結論部の ℓ_2-多様化 ((ℓ_1, ℓ_2)-関係多様化)

3.3　提案手法：関係ノイズ比を抑制した関係多様化

関係多様化データの有用性を高くするために，関係ノイズ比を抑制した関係多様化を実現したい．提案手法では，ノイズレスクラスを多数生成することで関係ノイズ比を抑制する．

S_j に同じ値を持ち，S_i に異なる値を持つ ℓ_1 個のタプルからは (ℓ_1,1)-関係多様なノイズレスクラスを生成できる．さらに，S_i の集合が同一かつ，S_j の値が異なる (ℓ_1,1)-関係多様なノイズレスクラスの集合から，(ℓ_1,ℓ_2)-関係多様なノイズレスクラスが生成できる．ここで，S_i の値が同じタプル群毎に S_j の出現パターンを考える．S_j の出現パターンが類似するこのタプル群からは，ノイズレスクラスが多数生成できる．

そこで，以下のステップでノイズレスクラスの生成を行う．

1. 類似度の高いタプル群を選択
2. 前提部の多様化 ($(\ell_1, 1)$-関係多様化)（図 3）
3. 結論部の多様化 ((ℓ_1, ℓ_2)-関係多様化)（図 4）
4. ノイズレスクラスが生成可能なら 1. へ

残りのタプルに対しては，RNR が小さくなるようにクラスタリングをすることで，関係多様性を充足させる．

3.4　評価

提案手法の有効性を評価するために関係ノイズ比と関係多様化の計算時間を評価した．提案手法を NLC とし，比較手法として関係ノイズ比を考慮したクラスタリングによる手法 DGRL と，関係ノイズ比を考慮しないナイーブな手法 DG を実装して比較した．データセットには，人工データを用いた．

評価尺度には，テーブル全体の関係ノイズ比 RNR を用いた．図 5 に (2, 2), (3,3)-関係多様化した 3 手法の RNR を示す．提案手法は，(2,2)，(3,3)-関係多様化において，RNR が 1 に近く，非常に小さい．

また，他の手法より 10 倍以上高速であることも確認できた．

以上より，提案手法は系列データの関係多様化を関係の曖昧化を抑止しながら効率よく実現できると言える．

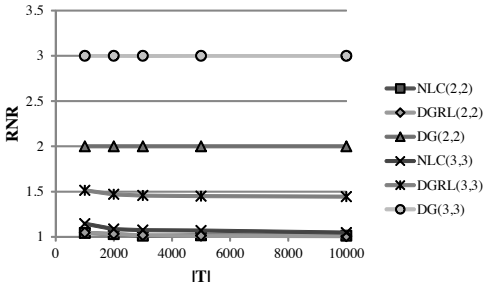


図 5　関係ノイズ比

4.　おわりに

本稿では，系列データの匿名化に関する研究の概要を紹介した．特に，移動軌跡ストリームに対して一定の精度を保ちながら連続的に k-匿名化する手法と，センシティブ属性間の値の特定の回避を属性値や属性間の関係を維持して実現する関係多様化を提案した．これらの研究成果によって，系列データの第三者提供や二次活用における有用性を維持しながらプライバシ侵害の懸念を軽減できたと考えられる．

謝辞

筑波大学の北川博之教授を初めとする諸先生方やご指導くださった皆様に深く感謝致します．

[文献]

[1] L. Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No. 5, pp. 557–570, 2002.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. TKDD, Vol. 1, No. 1, p. 3, 2007.

[3] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. ICDE 2008, pp. 376 – 385, 2008.

[4] M. E. Nergiz, M. Atzori, Y. Saygin and B. Güç. Towards Trajectory Anonymization: a Generalization-Based Approach. Transactions on Data Privacy, Vol. 2, No. 1, pp. 47–75, 2009.

[5] People Flow Project (pflow). `http://pflow.csis.u-tokyo.ac.jp/index.html`

高橋 翼　Tsubasa TAKAHASHI

2010 年筑波大学大学院システム情報工学研究科博士前期課程修了．同年，日本電気株式会社入社．2014 年筑波大学大学院システム情報工学研究科博士後期課程修了．博士（工学）．情報処理学会正会員．日本データベース学会正会員．