様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　２６　日現在

機関番号：　１２６０１
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１２５４８

研究課題名（和文）現実の人の動きを元にした移動履歴解析とその解析に適したプライバシー保護手法

研究課題名（英文）Preliminary Study about Advantageous Trajectory Anonymization methods Based on Real Trajectory data

研究代表者

山口　利恵（繁富利恵）（Yamaguchi, Rie）

東京大学・大学院情報理工学系研究科・特任准教授

研究者番号：９０４４３１９２

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：近年の ICT 技術の発展に応じて,スマートフォンのアプリケーションや IoT デバイスの移動履歴を利用して個 人が特定されにくい移動履歴データの生成交通統計データを生成する ことを検討することが必要である．従来の履歴ごとにエリアを拡大しながら匿名化を行 う方法ではなく,四分木を利用してグリッドごとの人口分布のデータを生成したうえで.エリア内の人口 が一定以上になるようにエリア範囲を調節することで,人口の多いエリアでは小エリアに分割,人口の 少ないエリアでは大エリアに分割することとした．それらの匿名化手法を良し悪しの検討と匿名化結果に関する評価の提 案を行う

研究成果の概要（英文）：Growing mobile networks and widely spread of Global Positing System (GPS) devices enables to collect large scale location and trajectory data. Using trajectory data to succeed, privacy and data characteristics are essential. Most anonymization methods are losing characteristic for service providers, like adding noise to trajectories.
In this research, firstly we present adaptive quadtree grid population calculation to determine grid size of trajectories. Our anonymization method dynamically adjust cluster size to maintain trajectory data characteristics, small mesh to the dense populated area, large mesh to sparse populated area, base on quadtree geospatial data structure. Proposed method is satisfied that adaptive size scaling and more efficient to maintain characteristic. Our experiment suggest that proposed method correctly anonymize Tokyo Urban flow data of 1.3M Trajectories.

研究分野：　情報セキュリティ

キーワード：　プライバシー保護　情報セキュリティ　位置情報　匿名化

１．研究開始当初の背景

現在、IoT 社会と言われ、人々がたくさんの端末を持ち歩くようになり、データの利活用が叫ばれている。この端末を利用することにより、位置情報をより簡単に利用できるようになってきた。

特に位置情報は、（人の動き）を洗い出すことによって、ビジネス、交通、健康や医療に役立つ情報を提供することを通して社会、産業に寄与するポテンシャルを持っている。一方で、このような位置情報を利活用するためには、プライバシーに問題への配慮がかかせない。

●利活用における有用性

利活用をする人々にとって、意味のあるまとまりを持つような技術をどのように作れば良いのであろうか。すべての分野に通用する技術を作ることは困難であるので、この研究では上で述べたように実用上高い価値をもつ位置情報の時系列、すなわち移動履歴情報に焦点を絞る。現状のビックデータの研究への流行りから鑑みて、集められた情報から応用上有用な知識を取り出すことは移動履歴に限らず重要な研究テーマである。しかし、生の移動履歴情報からプライバシーが保護され、かつ応用上有用な情報を抽出する手法は未だ不十分な状況である。

●プライバシー保護

現状既に提案されているプライバシー保護手法の多くは、情報の利活用という観点よりもプライバシーの配慮だけにフォーカスしている傾向にあり、利活用するには情報量が少なすぎるまで削る、もしくは情報にノイズを混ぜる等、実際に利用する人々にとって意味をなさない手法が多いという状況にある。例えば、携帯電話やスマホから発信される個人の位置情報は概略、

[氏名、携帯番号、住所、年齢、(位置情報の系列)]

という構造となる。ここでプライバシー保護のために、氏名、携帯番号を削除し、住所、年齢を「文京区本郷、30 代」などと肌理の粗い情報にして、同じ住所、年齢の人がデータベース中に k 人以上いるように変換する方法が良く知られている k-匿名化である。

しかし、位置情報の場合、夜間に長時間滞在する場所は自宅であるとか、昼間に長時間滞在する場所は勤務先であるというようなプライバシー情報が含まれる。そのため、短期間の移動に限定するような時間的切り出しが必要である。その上で、なお勤務先と交渉相手会社の間の移動履歴には個人に辿り着く手がかりが残るため、細かい領域に分割してその内部に一定の時間幅において k-人以上存在するような処理を施す。これによってはじめてプライバシーの保護が完璧ではないにしても達成できる。ただし、k の値は２や３などという小さな数では個人特定の危険性があるため、例えば 10 以上のようにできるだけ大きくしたい。

特に位置情報の活用に関しては、既にヒューリスティックスな手法が多数存在しているが、その手法が匿名化手法に十分に適用されているとはいえない。本研究では、位置情報、特に人の動きに関する移動履歴情報について、既存のヒューリスティックスや、地図に記載されている街路、鉄道路線などの情報を組み合わせ、これらに機械学習のアルゴリズムを適用することで、有用な人々の行動に関する有用な情報が抽出でき、同時に有効な匿名化が達成できる手法の提案を行う。

２．研究の目的

IoT 社会において端末から位置情報を容易に手に入れられるようになってきた。このような端末から手に入れることができる人々の移動履歴情報は、より有用な利活用をしたい反面、プライバシー保護が必須である。しかし、移動歴情報は、人々の社会的に培われた歴史と経験を踏まえることで移動の特徴を捉え、その特徴を生かしたプライバシー保護手法の提案を行う。これらの手法は、現状の移動履歴情報の機械学習等のアルゴリズムによる解析によって特徴をとらえ、その特徴を生かした上のプライバシー保護手法を提案することで、従来の手法の単純な適用では NP 困難であった問題に対して､解決策を提案することとなる。

３．研究の方法

ｋ匿名性の単純な適用では NP 困難となる問題に対して、従来社会的に培われた人々の動きにおける特徴を手に入れデータ有用性を保つための数理モデルを検討する。このために、現状の人の動きを機械学習等のアルゴリズムを活用して、データの有用性について検討を行う。次に、数理モデルを基にしたプライバシー保護手法を提案する。

４．研究成果

従来の履歴ごとにエリアを拡大しながら匿名化を行う方法ではなく，四分木を利用してグリッドごとの人分布のデータを生成したうえで，エリア内の人口が一定以上になるようにエリア範囲を調節した．このようにすることで，人口の多いエリアでは小エリアに分割,人口の少ないエリアでは大エリアに分割することができる．また，それらのエリア分割をもとにし, 匿名化手法を適用した. また, この匿名化に関する評価を行った.

ジオコーディングとしての四分木による領域分割について説明を行った上で,実際のデータを用いた四分木グリッドの人口分布について説明を行う.更にこの人口分布データを用いた匿名化手法の提案を行う.

４．１領域分割手法と四分木(Quadtree)

本節では地理空間情報において利用される領域分割手法を紹介しながら,本稿で利用する四分木に関する説明を行う.位置座標は X,Y または緯度経度で表される二次元空間の

座標であるがこの位置座標を計算機で扱いやすい情報に変換する必要がある.このような変換はジオコーディングと呼ばれており,階層化や符号化の方法によって複数の手法が使われている.
階層化手法で有　　　　名な符号化方法がGeohash[15]で位置情報を符号化するジオコーディングの一種であり,階層構造を持ちつつも位置座標を空間分割する機能を持つ.
また米軍で採用されている MRGS と呼ばれる手法ももあり,この手法は UTM 手法として国土地理院のコーディングでも採用されている.MRGS は全球をメルカトル図法で分割しており,単なる XY の分割よりも制度が高い特徴があるが XY を分離して保有するため,階層化する際の計算量が増えるという難点がある.
Quadtree(四分木)は二分木のような木構造による格納手法であり,2 分割する二分木と異なり,図 2 のように枝を 0,1,2,3 の四分割して格納する.木構造の四分木は非平衡木であるが位置情報を取り扱う場合には先頭からエンコードができることにメリットが有る.
今回はこの四分木の手法を位置情報の符号化に用いて整理したものを利用し,この符号化を Quadtree と記すことにする

４．２実データを用いた四分木グリッドの人口分布

実際の移動履歴データを利用して,Quadtree の手法を適用した場合にグリッドあたりの分布がどのようになるのかを検討することにする.

４．２．１使用データについて

匿名化実験のデータとしては人為的に作成したデータと実データの二種類が考えられる .完全に人為的に生成したデータはアルゴリズムの確認ならびに匿名化のベンチマークとしては利用可能であるが,現実的なデータとして利用するにはデータの特性が異　なるため難しい.一方で実データの利用には本論文で議論したように ID と位置情報が個人を特定する情報になるため,自ら同意を取得してデータを収集するか,またはユーザの同意を得て第三者提供を受ける必要がある.現時点ではこのような実データの利用には困難が伴うと言える.
今回提案手法の実験を実施するためには現実的な移動履歴を持ち,人数が多く,なおかつユーザからの同意または個人情報保護上問題がないデータを利用する必要がある.そこで上記の条件を満たすデータとして,東京大学空間情報科学研究センターの「人の流れプロジェクト」の「2008 年東京都市圏人の流れデータセット(空間配分版)」のデータを利用することとした.今回の実験データは東京都市圏交通計画協議会が収集したパーソントリップ調査によるデータを元にしている.実験データとしては関しては元のパーソントリップ調査のデータを用いて,住所詳細を記載していないものをベースとし,以下に示す空間配分を行った空間配分版とした.空間配分とはゾーンごとにまとめられた地点情報について,個々人の位置情報をゾーン範囲内の建物の分布に合わせて詳細位置に確率的に再配分し,現実のデータに近づける処理のことである.
表 1 に,人の流れプロジェクトのデータにおける位置履歴情報定義を利用した位置履歴情報の例を引用する.このデータ定義については実際のものだが,データ自体に関しては定義に合わせて著者が作成したダミーデータとなっている.
この例では 20-25 歳の学生でかつ女性であるユーザ 12345 は,東京大学構内から徒歩で本郷三丁目駅に移動し,本郷三丁目駅から新宿駅まで移動をした後に,新宿駅から高尾山口駅まで鉄道で移動し,最後に高尾山山頂まで徒歩で移動をした,この移動の目的はレジャーであったことがわかる.
今回利用したデータセットには本データセットには特定日付の 586882 ユーザの 1316100 の移動履歴が含まれている.今回はこの移動履歴を利用することにした.

４．２．２Quadtree による人口分布の計算

まずはメッシュごとの人口計算を行う.今回はトリップの発着地点を双方用いることにする.すなわち発着地点ごとに人口を+1 しているので 1 名は 2 箇所に換算される.同一地点を往復した場合は 4 地点に対して+1 される.
この Quadtree メッシュを 2 文字ごと(1/16 ごと)に計算を行う形で人口を算出した[313200312132223031]であれば,[313200312132223031],[3132003121322230][31320031213222],[313200312132],[3132003121],[31320031]のエリアごとの人口を算出している
ここから抜き出したデータをトリップごとに整理を行う.たとえば「[本郷三丁目,高尾山口]」を元に整理する.この際にこれ以下であれば領域を拡大し,これ以上であれば領域サイズをそのままにするスレショルド値の決定が必要である.この値を仮に 250 としてこの先の議論をすすめる.
まずは[本郷三丁目]エリアのデータを使って整理する.人口が多い箇所の場合は 18 文字の Quadtree 内の人口が 400 を超えることも稀ではない.そのため本郷地区については眺めの文字例すなわち,狭めのエリアで検討が可能である.
一方,高尾山口のようなエリアには人口がそれほどいない,そのため 250 というスレショルド値を満たすためには領域を拡大する必要がある.先ほど Quadtree メッシュごとの人口を作成したので,quadtree エリアの人口をプログラムに問い合わせることで適切な領域がわかる.
本郷エリアと高尾エリアのサイズが決まっ

たところで,この経路ペアをシステムに登 録す る .システムでは発着エリアとも同エリア・同サイズであった場合に経路が重 複 しているとみなす.すなわち履歴において個人特定性があるとは,その履歴が全体の中で単独で存在することであり,同じ履歴が複数存在すれば個人特定性はないこととなる.
今回は移動履歴を表 2 のように管理する.この例の場合,トリップ数は 5 であり,経路数は[東京]→[新宿],[新宿]→[渋谷],[代々木]→[渋谷]の 3,棄却経路は[代々木]→[渋谷]の 1 となる.この棄却経路の履歴が個人特定性のある履歴ということになる.
また棄却率=個人特定率を[棄却経路数/総トリップ数]と定義し,この例の場合は 1/5=20%となる.
このときに[代々木]→[渋谷]は区間としては[新宿]→[渋谷]に内包されるが,今回はこのような区間の内包は計上せず別 件 と し て数えることとした.
実データでは Quadtree の長さ 8 で最大 22 万のグリッドが存在したがプロットは 2500 で人口が 2500 で打ち切っている.
実際の計算では最大 5000 まで変動させたので,それ以上の領域も利用している

４．３人口比に応じた移動履歴の加工手法

本節では前説までの結果を踏まえて,人口比を利用したエンコード手法の提案を行う. 1.全データからメッシュごとの人口を計算する
2.人口スレショルド値の決定
3.履歴を取り出し,発着点の人口メッシュを検索する
4.発着点のグリッドサイズを調整して匿名加工を終了する
まず 1 のように,全データからメッシュごとの人口を計算する.今回は一日分のデータをすべて利用した.
次に 2 の人口スレショルド値を決める.今回は 100,250,500,1000,2500,5000 というスレショルド値を用意した.
ここからデータ処理に入る.3 のように履歴から発着点のみを抜き出す.その上で発地点,着地点ごとにスレショルド値よりも多い人口が存在する Quadtree メッシュを 4 発着地点のメッシュとして利用する.このようにして履歴データを加工していく.
ショルドを 100,250,500,1000,2500,5000 と適用し,提案手法による加工の効果を見ることとした.
注意が必要なのは本提案手法で加工された情報の匿名性は保証されないということである.すなわち繁華街と広大な領域との移動であったとしてもその履歴が単独でしか存在しない一意な履歴であればこれは特 定 可能ということになる.
従来の手法では k-匿名化を利用していたので加工後のデータは必ず k よりも多いことが保証されていたが本件手法ではその保 証 はない.そのため加工処理後に必要に応じてスクリーニングを行う必要がある.実際の利用時に関してはこのような特定が可能な履歴をどのようにするかの検討が必要であり,必要に応じて履歴を削除することが求 め ら れる.
ただ,本稿執筆時に経済産業省から出されているガイドラインでは必ずしも履 歴 に 関 して ,一意な履歴を削除しなければならないという立場のようであり,一意な履歴に関する取り扱いに関しては今後の検討が必要であると考えられる.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）
〔学会発表〕（計　３件）

疋田敏朗, 山口利恵, 匿名化移動履歴のアプリケーション評価, 暗号・情報セキュリティシンポジウム 2018
Toshiba Hikita, Rie Yamaguchi, Preliminary Study about Advantageous Trajectory Anonymization methods Based on Population, 2018 10th International Conference on Communication Systems & Networks (COMSNETS), 査読有り
鈴 木 宏 哉 , 山 口 利 恵 , iOS の Significant-Change Location Service による機能呼び出しの タイムスタンプ情報を用いた利用者行動パターンの分類, 暗号・情報セキュリティシンポジウム 2018

〔その他〕
http://www.yamagula.ic.i.u-tokyo.ac.jp

６．研究組織

(1)研究代表者
山口　利恵（繁富利恵）(Yamaguchi, Rie)
東京大学・大学院情報理工学系研究科・特任准教授
研究者番号：90443192

(2)研究分担者
中川 裕志 (Hiroshi Nakagawa)
東京大学・情報基盤センター・教授研究者番号： 20134893