3I2-OS-20-5

階層的クラスタリングによる移動シーケンスの匿名化法

Anonymization of Moving Sequence via Hierarchical Clustering

石井 祐多[*1] 納 竜也[*2] 佐久間 淳[*3*4]
Yuta Ishii Tatsuya Osame Jun Sakuma

[*1]筑波大学情報学群情報科学類
College of Information Science, University of Tsukuba

[*2]筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻
Dept. of Computer Science, Graduate school of SIE, Univ. of Tsukuba

[*3]筑波大学システム情報系
Faculty of Engineering, Information and Systems

[*4]科学技術振興機構さきがけ
Japan Science and Technology Agency

We study anonymization of travel sequences. k-anonymity is known as a standard anonymization model, while k-anonymization of sequences is intractable because the size of sequences can be infinitely large. We propose a novel anonymity definition which is an extension of LKC-privacy. Most of travel sequences are provided in the form of (latitude, longitude,date-time). Our model can take such numerical sequences via discretization using clustering. Furthermore, we relaxed ordering constraints of LKC-Privacy to decrease the number of suppressed sequences. The efficiency of our anonymization model is examined by travel prediction problem using real travel sequences collected in China. Experiments show that our anonymization does not necessarily degrade the prediction accuracy.

1. はじめに

近年、個人情報を扱うサービスの利用が盛んになりつつあり、個人の精細な位置情報や行動履歴を利用した広告モデルやSNS が登場している。それと同時に、個人情報保護も重要な課題であり、個人情報の利用と保護のバランスをとる技術の模索が続いている。本稿では、個人情報の一例である位置情報の匿名化について検討する。

位置情報の匿名化を行う際、連続値で表されている位置情報の抽象化と、シーケンス匿名化の 2 つの問題がある。位置情報の抽象化の代表例としてグリッド分割が挙げられるが、データ分布を無視した分割を行うため位置情報が本来持つ疎密が失われる。そこで、本稿ではデータ分布に適応した分割を得られる手法である、クラスタリングによる抽象化 [1] を用いる。

シーケンスで表される位置情報の匿名性定義には、k-anonymity[2] や l-diversity[3] よりも匿名性条件を緩めた、LKC-Privacy[4] が提案されている。しかし既存の枠組みでは依然、匿名性条件の制約が厳しいため、有用性の高い匿名化データを生成することが困難である。そこで本稿ではデータの有用性の維持と匿名性域の緩和を両立させるために、既存の匿名性定義の拡張と順序制約の緩和を提案する。実験では、既存手法との比較と推薦精度の差から提案手法の評価を行い、有用性の高い匿名化データが生成されることを示す。

2. 移動シーケンスの匿名化法

Aggarwal ら [1] は、クラスタリングによる抽象化を用いた、連続値で表されている位置情報の匿名化法を示した。本稿では、この手法の発展として、再帰的なクラスタリングによって抽象化を行う。クラスタ分割後にある閾値未満のサイズのクラスタが生成された場合、分割を中止し一つ上の階層に戻すことで、全てのクラスタが閾値以上のデータを保持していることを保証する。各位置をクラスタの中央点に抽象化することで、閾値に応じた匿名性が達成される。

また Fung ら [4] は、k-anonymity などよりも匿名性条件の弱い LKC-Privacy を提案している。k-anonymity は全ての属性値の組合せの一意性を許容しないため、移動シーケンスのようなログデータでは組み合わせ爆発をおこし、匿名化データが達成できない。しかし現実問題では、攻撃者は被害者の全ての移動履歴を知ることは困難である。LKC-Privacy では攻撃者のターゲットに関する背景知識はたかがか L 点であることを仮定する。つまり、L 点以下の存在しうる全ての属性値の組み合わせに対して、データベース内に K 個以上存在することを匿名性定義とし、これを満たすようにデータを変更する。また、特定されたくないセンシティブ情報集合 S 内のセンシティブ情報 s を含むシーケンスにおける、L 点以内の全ての組合せにおいて、s と推定される確率が C% 以下になるようにデータを変更する。このように既存手法の制約を緩めることにより、有用性の高いかつ現実的な匿名性を満たすデータを生成する。

定義 1 (LKC-Privacy) 背景知識の最大長を L、秘匿したいセンシティブ情報の集合を S とする。*Trajectory* データ集合 T が *LKC-Privacy* を満たすとき、L 点未満の全てのシーケンス q が以下の条件に合致している。

1. 匿名化閾値 $K > 0$ が与えられたとき、$|T(q)| \geq K$
2. 全ての $s \in S$ において閾値 $0 < C \leq 1$ が与えられたとき、s が推定される確率 $P(s|q) \leq C$

しかしながらそれでもなお、LKC-Privacy で定義している匿名性は現実のユースケースと照らし合わせると制限が強く、位置情報のようなサイズの大きいシーケンスデータは、匿名化によってデータの有用性が失われるという懸念がある。そこで本稿では、既存の匿名性定義の拡張と順序制約の緩和による匿名化データ生成手法を提案し、連続値で表されている位置情報から生成された匿名化データの有用性の向上を目指す。

連絡先: 石井祐多, 東京大学, qq126304@iii.u-tokyo.ac.jp

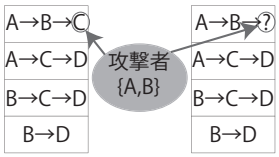

図 1: 識別推定による不利益の例。左右の表はそれぞれシーケンスのデータベースを表している。

2.1 匿名性定義の拡張

匿名化とは、個人が特定されてしまう識別推定やセンシティブ情報が特定されてしまう属性推定といった情報漏洩を防ぐものである。位置情報データベースには、シーケンスに年齢や職業などのセンシティブ情報が付加されている場合もあるが、本稿ではセンシティブ情報が付加されていないシーケンスについて扱い、属性推定については今回は考慮しない。識別推定による不利益について、攻撃者がシーケンスに含まれている背景知識以外の情報を得ることが挙げられる。(図 1 左、背景知識 A、B に対し、追加的に位置情報 C が漏れている) しかし、シーケンスに背景知識以外の情報が含まれていない場合、攻撃者に追加的に漏れる情報はない。(図 1 右、背景知識 A、B に対し、追加的に漏れる情報はない) これより、識別推定された場合でも攻撃者の背景知識以外の情報が含まれていなければ、匿名性を満たしているといえる。そこで、LKC-Privacy の枠組みを拡張し、新たな匿名性定義を提案する。

定義 2 (提案手法の定義) 背景知識の最大長を L とする。*Trajectory* データ集合 T が匿名性を満たすとき、L 点未満の全てのシーケンス q が以下のどちらかの条件に合致している。

1. 匿名化閾値 $K > 0$ が与えられたとき、$|T(q)| \geq K$
2. $|T(q)| < K$ のとき、q を含む全てのシーケンス r について $q = r$ が成立

2.2 順序制約の緩和

既存の匿名化手法では通常、位置情報だけでなく時刻も考慮に入れた上で匿名化を行っている。しかしデータの利用場面によっては、位置情報の順序だけが必要になる場合や、順序すら関係なくユーザと位置情報の結びつきのみ必要な場合も存在する。また、ある時刻では外れ値であったシーケンスも別時刻では外れ値でない場合が考えられる。このため、解析手法が時刻情報を必要としないならば、時刻を考慮しない位置情報の匿名化は、時刻を考慮する場合と比べて情報量の損失が少ないと予想することができ、有用性の高い匿名化データを生成できることが期待される。図 2 に順序制約の緩和例を示す。(図 2 中央、順序だけを考慮した場合、時刻を無視するため 2 と 3 は同一である。図 2 右、順序すら考慮しない場合、場所に訪れる順番は考慮しないので 2、3、4 は同一である。)

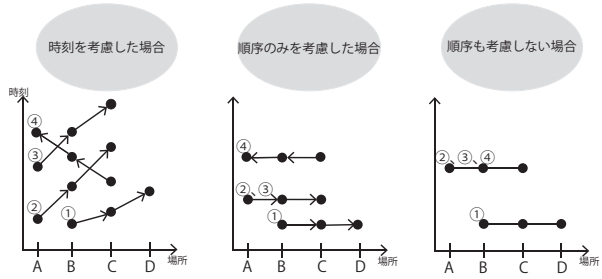

図 2: 順序制約の緩和例。直線 1〜4 がシーケンスを表す。

3. 実験

本稿では、提案手法が有用性の高い匿名化データを生成できるかどうかを確認するために既存手法との比較実験を行い、それぞれの結果に対する考察を行う。さらに、推薦アルゴリズムを用いて匿名化データの有用性を計測する。

3.1 実験データ

Microsoft Research Asia が提供している、個人の位置情報を GPS で取得した Trajectory データ [5] を使用する。本稿では、各個人のデータを 1 時間間隔で利用し、8 時から 20 時のデータを対象とした。さらに、実験に使用するデータは北京中心部約 10km 四方に存在するデータに絞った。

このデータではユーザ数が 167 人しか存在せず匿名化対象データとしては人数が少ない一方、期間が約 3 年間と長いため、同一ユーザでも日付が違えば別ユーザとして扱い、日付を無視して時刻だけを考慮するものとした。さらに、4 点以上データ点を含む Trajectory を対象に実験を行った。上記の条件を満たすユーザ数は 947 人、データ点数は 5225 点であった。

3.2 実験 I 匿名性定義の拡張による匿名化データの変化

3.2.1 実験設定

位置情報の抽象化には、各時刻のデータ点に対してクラスタリングを行う手法と、全時刻のデータ点に対してクラスタリングを行う手法の 2 通りが考えられる。各時刻のデータ点に対してクラスタリングを行う場合、時刻によってデータ点の傾向が違った場合にもそれに適応したクラスタリングができるというメリットがある。しかし各時刻を対象にするクラスタリングでは、同じクラスタ番号でも別時刻だと別の地点を示しているので、推薦アルゴリズムなど過去の事例から未来の傾向を予測するタスクには向かない場合が多い。そこで、両方の場合について匿名化を施し評価する。

匿名化のパラメータは $L = 2$、$K = 2$ を用い、結果の精度は 20 回の平均を取った。クラスタリングの際の閾値を Kh で定義し、h を $1, 2, \ldots, 30$ と変化させて実験を行った。特に断りがない場合、今後も上記の設定を用いる。

3.2.2 評価の方法

実験 I、II では匿名化によるデータ損失を、匿名化前後のデータ数の差と匿名化によるデータ点の移動距離の平均という 2 つの観点から評価する。匿名化データに求められる有用性は、応用場面によって変化する。データ分布を大きく変化させてでも匿名化によるデータの削除点数を小さくしたい場合や、削除点数の増大は許容するが、匿名化データ内のデータ点の変更を最小限に留めたい場合が考えられる。そこで前者を Suppression Loss、後者を Distortion Loss と定義し、それぞれについて評価を行う。Distortion Loss の尺度には度数のユークリッド距離を用いる。Distortion Loss に関して、匿名化によって削除された点は考慮されていないことに注意されたい。

3.2.3 実験結果

各時刻を対象にするクラスタリングと全時刻を対象にするクラスタリングそれぞれに対し、既存手法と提案手法を用いて匿名化を行い、Loss を比較した実験の結果を図 3 に示す。実験結果より、それぞれのクラスタリング手法で提案手法の Suppression Loss は既存手法よりも小さく、Distortion Loss

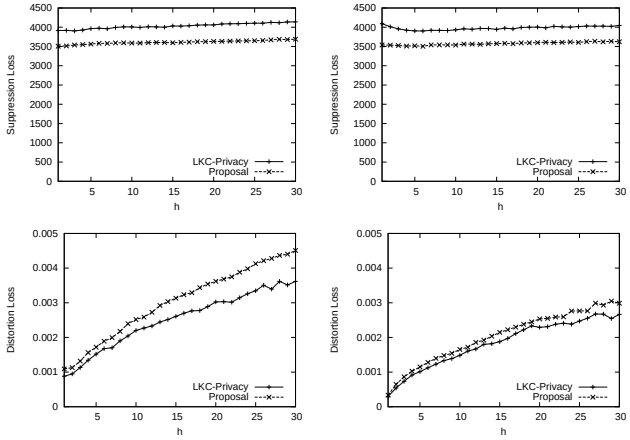

図 3: 拡張された定義による匿名化データの Loss の変化。それぞれ実線が既存手法、破線が提案手法の Loss である。上側が Suppression Loss、下側が Distortion Loss を表している。左側は各時刻のデータ点を対象にしたクラスタリング、右側は全時刻のデータ点を対象にしたクラスタリングを用いている。

は若干だが既存手法の方が小さいことがいえる。既存手法の Distortion Loss が小さい理由として、移動距離が大きい外れ値の多くが匿名化によって削除されるからと考えられる。既存手法と提案手法を比較すると、Distortion Loss の差はさほど大きくなく、Suppression Loss の差が顕著である。提案手法に比べ、既存手法は匿名化前後のデータサイズの差が大きく、データの有用性が大きく低下していると考えられる。これより、提案手法の方が有用性の損失が少ないといえる。

また、各時刻のデータ点を対象にしたクラスタリングと全時刻のデータ点を対象にしたクラスタリングを比較すると、適切なパラメータを用いた場合、Loss の差は小さかった。各時刻のデータ点を対象にしたクラスタリングの場合、それぞれの時刻のデータ点が少なく、過剰適合によりデータ分布に適応しないクラスタリングを行ってしまうからだと考えられる。このことから、以降は全時刻のデータ点を対象にしたクラスタリングについてのみ扱う。

3.3 実験 II 順序制約の緩和による匿名化データの変化

3.3.1 実験設定

前節では時刻を考慮した場合における匿名化を行ったが、データ点の順序のみ必要な場合やユーザとデータ点の結びつきのみ必要な場合も存在する。時刻を考慮する必要がない場合においては、時刻を考慮する場合と比べ削除されるデータが少ないと予想でき、有用性の高い匿名化データを生成されると期待できる。そこで、時刻を考慮した匿名化と順序のみを考慮した匿名化、順序も考慮しない匿名化の 3 つの場合について、匿名化データの Loss を評価する。

3.3.2 実験結果

順序制約の緩和による匿名化データの Loss の変化を図 4 に示す。実験結果より、順序のみを考慮した匿名化データの Suppression Loss は時刻を考慮した場合と比べ小さいことがいえる。これは、ある時刻では外れ値であったシーケンスが別時刻では外れ値ではない場合、時刻を考慮しない匿名化の際には削除されないためと考えられる。また、それぞれの場合の Distortion Loss にあまり差は見られない。時刻を考慮する匿名化によって削除される値の多くは、別時刻では外れ値でないと考えられ、これらの点が大きく Loss に影響することはない。そのため、Distortion Loss の変化が小さいと考えられる。

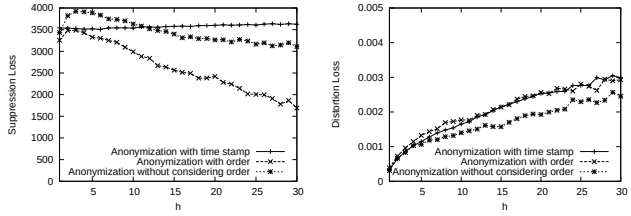

図 4: 順序制約の緩和による匿名化データの Loss の変化。凡例の上から順に時刻を考慮した匿名化 (再掲)、順序のみを考慮した匿名化、順序も考慮しない匿名化を表している。左側が Suppression Loss、右側が Distortion Loss である。

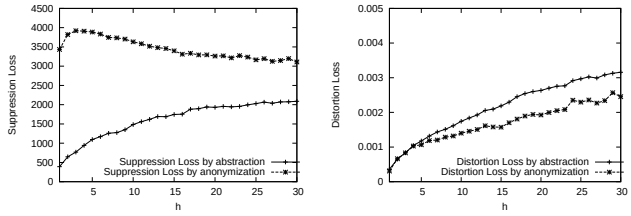

図 5: 順序も考慮しない場合における、抽象化のみと、抽象化および匿名化の Loss の変化。グラフの下側が抽象化のみ、上側が抽象化および匿名化の Loss である。左側が Suppression Loss、右側が Distortion Loss を表している。Suppression Loss に関して、抽象化および匿名化の Loss と抽象化のみの Loss との差が、匿名化による Loss である。

直感的には、順序制約の緩和によって Suppression Loss が減ると考えられるが、順序も考慮しない匿名化の Suppression Loss と順序のみを考慮した匿名化の Suppression Loss を比較すると、順序のみを考慮した匿名化の Suppression Loss の方が小さい。順序を考慮しない場合の匿名化は、抽象化を行った後にシーケンス内の同一クラスタに属するデータ点は 1 点を残して全て削除することになる。そこで Suppression Loss は抽象化によるものと匿名化によるものに分けて考察する。順序も考慮しない場合における抽象化のみの Loss と抽象化および匿名化後の Loss を、図 5 に示す。図 5 と図 4 より、抽象化による Suppression Loss を考慮せずに匿名化による Loss のみを比較すると、順序も考慮しない匿名化の Loss の方が小さいことがいえる。また Distortion Loss は抽象化によって生まれるが、匿名化によって大きく値が変化しないことがいえる。このことから、順序を考慮しない場合も匿名化前後で有用性の損失が少ないと考えられ、順序制約の緩和の妥当性が確認できる。

実験結果より、Suppression Loss と Distortion Loss はトレードオフの関係であることがいえる。そのため匿名性定義は、解析対象に応じて選択する必要がある。

3.4 実験 III 推薦を用いた有用性の評価

前節までは、匿名化データの有用性を Suppression Loss と Distortion Loss を用いて評価した。しかし実際にデータを用いて機械学習を行った場合に、データの匿名化前後で有用性に大きな差があっては意味がない。そこで本稿では、遷移確率を利用した推薦と、行列分解を利用した推薦 [6] を用いて、匿名化前後のデータを用いた推薦結果の差から有用性を評価した。

3.4.1 実験設定

実験 II より、順序制約を緩和した場合は Suppression Loss と Distortion Loss がトレードオフの関係になることを確認し

表 1: 遷移確率を利用した推薦の精度比較

	h=1		h=30	
	匿名化前	匿名化後	匿名化前	匿名化後
時刻を考慮	0.039	0.011	0.247	0.176
順序のみを考慮	0.033	0.032	0.248	0.255
順序も考慮しない				

表 2: 行列分解を利用した推薦の精度比較

	h=1		h=30	
	匿名化前	匿名化後	匿名化前	匿名化後
時刻を考慮	0.157	0.127	0.072	0.173
順序のみを考慮	0.154	0.072	0.075	0.112
順序も考慮しない	0.127	0.096	0.154	0.234

た。これより、それぞれの Loss が小さい場合の匿名化前後のデータを用いて推薦を行う。本稿では、時刻を考慮した場合、順序のみを考慮した場合、順序も考慮しない場合それぞれについて、$h = 1$ と $h = 30$ のときに得られる匿名化前後のデータを用いた。実験では 2 点以上含んでいる Trajectory の最後のデータ点をテストデータ、残りのデータ点を訓練データとした。

3.4.2 評価の方法

遷移確率を利用した推薦は、訓練データを用いて場所間の遷移確率を求め、降順に並べ替えたスコアの上位を推薦結果として提示する。行列分解を利用した推薦は、訓練データを用いて各ユーザに対してすべての未訪問地に対するスコアを行列分解により求め、降順に並べ替えたスコアの上位を推薦結果として提示する。推薦結果として上位 N% 提示された中に正解が含まれている確率を Top-N% Precision とし、精度とする。本稿では、匿名化前後のデータによる推薦結果の精度を比較し、有用性を評価する。

3.4.3 遷移確率を利用した推薦

時刻を考慮した場合と順序のみを考慮した場合について、匿名化前後のデータを用いた推薦精度の比較を表 1 に示す。実験結果より、$h = 30$ のときの順序のみを考慮した匿名化後の推薦精度は、匿名化前に比べ高くなったが、その他の場合は全て匿名化後の推薦精度は匿名化前に比べ低くなった。順序のみを考慮した匿名化の場合、他の時刻では外れ値のシーケンスでも削除されない。このため、順序のみを考慮した匿名化データ内に数十以上あるシーケンスが存在する確率が上がり、確率の高い遷移が生まれやすく、その分精度が上がっていると考えられる。さらに $h = 1$ のときはデータ点を細かく分割するため、それに伴い数十以上あるシーケンスが存在することがない。これより確率の高い遷移が生まれず、精度が低いと考えられる。

3.4.4 行列分解を利用した推薦

時刻を考慮した場合、順序のみを考慮した場合、順序も考慮しない場合について、匿名化前後のデータを用いた推薦精度の比較を表 2 に示す。行列分解の正則化パラメータ $\alpha = 0.01, 0.1, 1$、ランク $K = 3, 5, 10$ と変化させて推薦を行ったが、パラメータを変化させた場合でも推薦精度にあまり差がないため、それぞれの条件において一番精度の高い値を採用した。

実験結果より、$h = 1$ のとき全ての順序制約において匿名化前のデータを用いた場合の精度が匿名化後に比べ高かった。一方 $h = 30$ のとき、全ての順序制約において匿名化後のデータを用いた場合の精度が匿名化前に比べ高かった。h が大きい場合、抽象化によるデータの歪みが大きく trajectory の分布が変わってしまうため、匿名化前のデータに関する推薦の精度が低くなると考えられる。しかし h が大きくなると対象となる場所数が減り、さらに匿名化によって外れ値が削除されるため、匿名化後に推薦の精度が上がると考えられる。このことから、$h = 30$ のとき匿名化後のデータを用いた場合の精度が匿名化前に比べ高いと考えられる。

実験結果より、匿名化が必ずしも推薦の精度を悪化させるわけではなく、適切に抽象化等を設定することで、精度と匿名化を両立できることが示されたといえる。

4. 結論

本稿では、クラスタリングを用いて有用性の高い位置情報の匿名化データを作成するという目的で実験を行った。適切なパラメータを用いた場合、既存手法に比べ、提案手法を用いた匿名化によるデータ損失が小さいという結果を得ることができ、有用性の高い匿名化データが生成できることが示された。

しかし、多くの問題が残されている。本稿では最適なパラメータを実験によって求めたが、大規模なデータの匿名化を行う際、パラメータを求めるために何度も匿名化を行うのは、計算コストが高く現実的ではない。さらに位置情報は日々増加し、データ分布も変化していくため、パラメータも動的に変化させることが求められる。このことから、オンラインでの匿名化へ拡張する必要がある。

更なる課題として、高速な匿名化アルゴリズムの作成がある。実験で用いたアルゴリズムでは、MFS を求める際の処理時間のオーダーは、Trajectory 数を M、場所数を N、Trajecoty の最大シーケンス長を L とした場合、$O(MN^L)$ であるため、アルゴリズムの高速化を行う必要がある。これらは今後の検討課題である。

謝辞

本研究は最先端研究開発プログラム (FIRST)「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」の助成を受けたものである.

参考文献

[1] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 153–162. ACM, 2006.

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

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 1, No. 1, p. 3, 2007.

[4] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman and Hall/CRC, 8 2010.

[5] Microsoft Research Asia. Geolife project. http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/.

[6] A. Paterek. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD Cup and Workshop*, Vol. 2007, pp. 5–8, 2007.