DEIM Forum 2023 5c-6-3

モバイル端末からのイベント情報検索における Geo-indistinguishability を用いたユーザ位置匿名化の評価

石神　京佳† 　榎　美紀†† 　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚２丁目１－１
†† 日本アイ・ビー・エム株式会社　〒 103–8510 東京都中央区日本橋箱崎町 １９－２１
E-mail: †{kyoka,oguchi}@ogl.is.ocha.ac.jp, ††enomiki@jp.ibm.com

あらまし　近年，ソーシャルネットワーキングサービス（以下 SNS とする）の普及に伴い，SNS 上にはローカルイベントや開催中のイベントをはじめ，大小様々な規模のイベントに関する情報が投稿されるようになった．それらの膨大なイベント情報をサーバ側に収集し，ユーザがモバイル端末の位置情報を提供して，現在地付近のイベント情報を得るようなサービスは近年多く存在するが，位置情報などの個人情報をそのままサーバに送信して情報分析に使用されることはプライバシ上の懸念がある．本研究では，ユーザの位置情報を地域メッシュと人流データを利用し，差分プライバシを位置情報匿名化に拡張した技術 Geo-Indistinguishability（以下 Geo-I とする）を満たすようなダミー位置を算出し，サーバへの問い合わせに使用することで，ユーザの位置情報を曖昧化しプライバシを考慮した．地理的な制約条件を満たしつつ，大量の SNS データからイベント情報を検索する手法を提案する．
キーワード　位置情報プライバシ，情報推薦，SNS データ

Evaluation of User Location Anonymization Using Geo-indistinguishability for Event Information Retrieval from Mobile Devices

Kyoka ISHIGAMI†, Miki ENOKI††, and Masato OGUCHI†

† Ochanomizu University
2–1–1 Otsuka, Bunkyou-ku, Tokyo 112–8610, Japan
†† IBM Japan, Ltd.
19–21 Hakozaki, Nihonbashi, Chuo-ku, Tokyo 103–8510, Japan
E-mail: †{kyoka,oguchi}@ogl.is.ocha.ac.jp, ††enomiki@jp.ibm.com

1. はじめに

近年，SNS の利用者は継続的に増加傾向で，総務省の調査 [1] によると，現在国内全体では約７割の人が SNS を利用している．また，SNS の利用目的としては，情報収集目的がコミュニケーション目的に次いで二番目に多い．これは，SNS 特有の情報発信のしやすさから，SNS 上には固定的なメディアに掲載されていないような様々な有益な情報が存在することが理由として考えられる．そこで我々は，SNS 上のローカルイベントや開催中のイベントをはじめ，大小様々な規模のイベントに関する情報が投稿されるようになったことに注目した．それらの膨大なイベント情報をサーバ側に収集し，問合せユーザの位置情報や SNS データなどを利用することで，ユーザに対し特定の場所や時間で開催されるイベントを推薦するシステムは既に研究されている．しかし，問い合わせユーザにとって，位置情報をはじめとする個人情報をサーバに送信して情報分析に使用されることはプライバシ上の懸念がある．この問題を解決するため，サーバを信頼しないようなユーザデータのプライバシ保護に関する研究が行われており，保護対象データを位置情報データに拡張する動きも盛んである．

本研究では，SNS の一種である Twiiter [2] 上からイベント情報を取得し，ユーザの必要とするイベント情報の個数的な条件，地理的な制約条件を満たすイベント情報検索システムの実装を行う．ユーザの位置情報を地域メッシュ情報 [3] を利用したダミー位置に変換し，サーバへの問い合わせに利用した実験

では，サーバに付与されるユーザ位置情報は 1km 四方に曖昧化されていた [4]．本稿では人流データ [5] を利用し，差分プライバシ [6] を位置情報匿名化に拡張した技術 Geo-I [7] を満たすようなユーザ位置のダミー位置を算出し，サーバへの問い合わせに使用することで，よりプライバシ保護の指標に則った度合いで，ユーザの位置情報を曖昧化する手法を提案する．

2. 関 連 研 究

本章では，関連研究として，近年盛んに行われているユーザのプライバシを保護しながらデータを活用するためのプライバシ保護技術と，位置情報データ保護に関する研究について紹介する．

2.1 差分プライバシと局所差分プライバシ

代表的なプライバシ保護保証の評価手法の一種に，Dwork らにより発表された差分プライバシ [6] が挙げられる．これは，あるユーザのデータが含まれるデータベースと含まないデータベースの差分を少なくすることで，攻撃者が統計的結果を得ても，結果がどちらのデータベースから得たものなのか見分けを付きにくくするという概念である．差分プライバシの概念では，データ収集者であるサーバを信頼し，データ解析結果を第三者に提供する際のプライバシ保護を想定している．近年ではユーザがデータ収集者を信頼せず，ユーザ自身が各々データに差分プライバシを満たすようなノイズを加える，局所差分プライバシに関する研究 [8] も盛んである．図 1 に差分プライバシと局所差分プライバシの概要を示す．

2.2 Geo-Indistinguishability（Geo-I）

Geo-I は，Andrés ら [7] によって提案された, 差分プライバシにおけるデータを位置情報，ハミング距離をユークリッド距離に置き換え，位置情報データの保護に応用したプライバシ保護基準である．局所差分プライバシと同様に，データ収集者を信頼しないようなモデルで，攻撃者に大まかな位置が予測されても，具体的な位置は予測されないようにするという概念である．X をユーザの位置の集合，Z をメカニズム K によって曖昧化した結果得られる位置の集合としたとき，任意の二つの点 $x; x^0 \ 2 \ X$ において以下が成り立てば，位置 $z \ 2 \ Z$ を確率 k_{xz} で出力するメカニズム K は -Geo-I を満たすという．

$$k_{xz} \leqq e^{\ d_X(x;x')} k_{x'z}$$

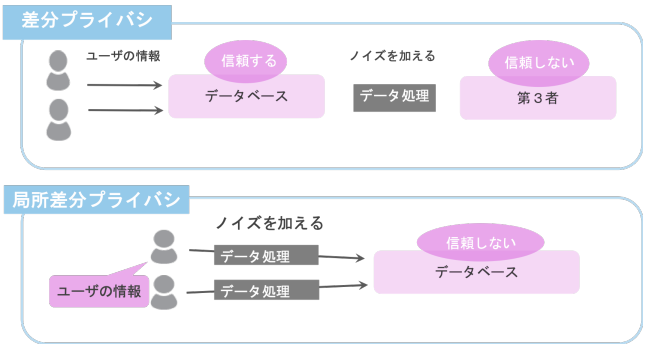

図 1: 差分プライバシ・局所差分プライバシの概要

また，高木ら [9] は Geo-I の弱点として，攻撃者が道路ネットワークを考慮した場合，ユーザの予測の精度が上がってしまうことを挙げ，道路ネットワークにおける位置情報プライバシという題で道路ネットワークを考慮した基準 Geo-Graph-Indistinguishability(GGI) を提案した．

このように，有効性のあるユーザデータのプライバシ保護基準が研究されている．プライバシ侵害の判定基準はユーザによって異なる主観的な概念であるため，明確な基準を定めるのは非常に困難であるが，これらの数式で示された保護基準を用いることで，プライバシ保護度合いを値　で表現でき，統計的なデータの保護に応用することが可能となった．

3. 先 行 研 究

本章では，本研究の先行研究について紹介する．同研究室工藤ら [10] が，Twitter のツイートデータからのイベントに関する情報の抽出に関する研究を行っている．工藤らの提案システムは以下に示すような，ツイートの抽出パート，イベントのカテゴリ分類パート，ユーザへの情報配信準備パートの大きく分けて 3 つのパートで構成されている．

I　ツイートの抽出
- i　Twitter API [11] のキーワード検索で地名をキーワードに設定しツイートを収集
- ii　取得したツイートの情報を整理
- iii　さらに日付と時間，イベントの開催地がツイート本文に含まれるものを抽出
- iv　正規表現を用いてイベント名を取得，外部情報を利用しイベント情報を補完

II　イベントのカテゴリ分類
- i　ランダムフォレストを用いてイベントを Music Event,Comedy などのカテゴリに分類

III　ユーザへの情報配信準備
- i　イベント名，開催日時，カテゴリ等のユーザに提供する情報を取得し整理
- ii　提供する情報を複数言語に分類
- iii　ユーザの位置を取得
- iv　ユーザ位置をもとに提供する情報の順位付け

また，同研究室今井ら [12] が，ユーザの過去のツイートデータから趣味趣向を判別し，工藤らのシステムを用て収集したイベント情報のデータを，取得したユーザ趣向に基づいて順位付けし推薦するシステムを提案している．

本研究では，サーバ側が行う Twitter からのイベント情報抽出に，先行研究の手法を使用する．先行研究ではユーザの位置情報やユーザの過去のツイートをサーバに付与しているため，プライバシ上の課題があった．本研究ではユーザの位置情報プライバシ保護のためのデータ処理について検討し，先行研究のプライバシを考慮したモデルへの拡張を目指す．

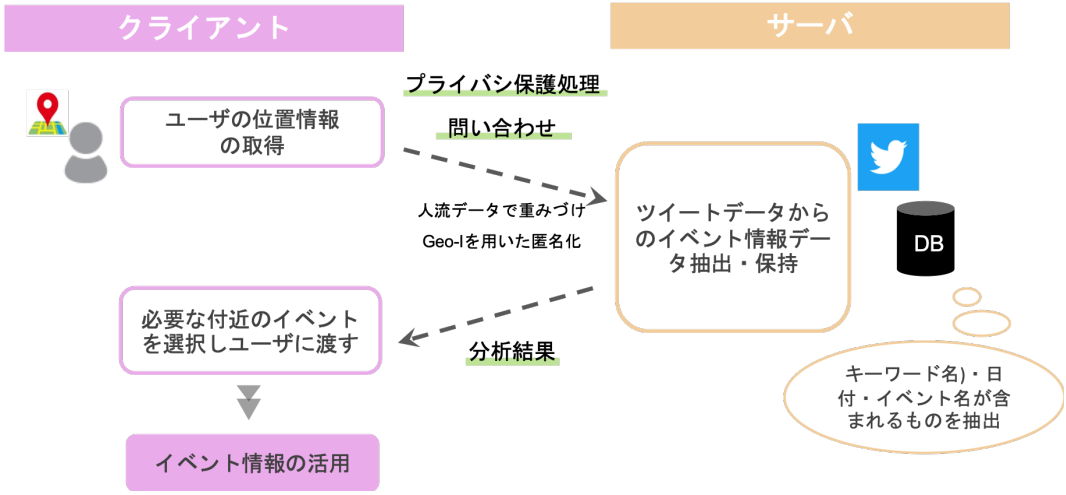

図 2: 提案システムの概要

4. 提案システム

本研究ではユーザの位置情報プライバシを保護しつつ，膨大な SNS 上のイベント情報から，ユーザに適したイベント情報を推薦するシステムの構築を目指す．提案するシステムの概要を図 2 に示す．

サーバ側は先行研究で紹介した工藤ら [10] が提案する手法を参考にし，パブリックなツイートデータからのイベント情報抽出・分析処理を行う．クライアント側はユーザのプライバシに関わるような情報の保護処理をしてからサーバへイベント情報の問い合わせをする．本提案手法ではメッシュ単位の人流データを使用し算出した確率分布を満たすダミー位置を問い合わせに使用する．クライアントは問い合わせ結果で得られたイベント情報から，ユーザが必要としている範囲や個数などの制約を満たすイベント情報を抽出しユーザへのイベント情報推薦を行う．

4.1 イベント情報抽出

Twitter API [11] を使用しキーワード (地名) を含むツイートを収集する．尚，同一内容のツイートの重複した収集を防ぐため，リツイート機能を用いて再投稿されたツイートは除いて収集する．続いて，取得したツイートデータから，正規表現を用いて日付とイベント名が含まれるツイートを抽出し，それらをイベント情報としてイベント名が重複するものを除きデータベースに格納する．表 1 にデータベースとして格納されるイベント情報の一例を示す．

表 1: データベースに格納されているイベント情報 (一部抜粋)

Event Name	Location	Start date	End date
D 王 GRAND PRIX 2021 Ⅱ in Korakuen Hall	後楽園ホール	2021/12/26	2021/12/26
15TH ANNIVERSARY LIVE KAT-TUN	国立代々木劇場	2021/03/22	2021/03/22
『炎炎ノ消防隊』POP-UP ショップ	新宿アルタ	2021/11/20	2021/11/20
ぐるぐる魂！	新宿バッシュ	2021/11/22	2021/11/22
Fantasy Passport	渋谷公会堂	2021/12/30	2021/12/30
GO!GO!トレジャーロード!!	渋谷マルイ	2021/11/24	2021/11/24

4.2 地域メッシュ情報の付与

本研究では，地域メッシュ [3] 情報を用いて，ユーザの位置情報のプライバシ保護処理を試みる．地域メッシュは，統計に利用するために緯度・経度に基づいて地域を隙間のない網の目のような区画としたものである．イベント開催地にスポット名 (注1) または住所が保存されているイベント情報に対し，Geolocation API [15] を用いて開催地の緯度経度を取得し，地域メッシュコードを算出し保持しておく．本研究では，現在日本で用いられている，昭和 48 年 7 月 12 日行政管理庁告示第 143 号に基づく「標準地域メッシュ・コード」から以下の区画を使用する．

第 1 次メッシュ (第 1 次地域区画)：
緯度の間隔 40 分，経度の間隔 1 度，一辺の長さ約 80km

第 2 次メッシュ (第 2 次地域区画)：
第 1 次メッシュを緯線方向及び経線方向に 8 等分してできる区域，一辺の長さ約 10km

第 3 次メッシュ (基準地域メッシュ)：
第 2 次メッシュを緯線方向及び経線方向に 10 等分してできる区域，一辺の長さ約 1km

分割地域メッシュ (2 分の 1 地域メッシュ)：
第 3 次メッシュを緯線方向及び経線方向に 2 等分してできる区域，一辺の長さ約 500m

4.3 プライバシを考慮した問い合わせ

[4] で提案した地域メッシュを用いたイベント情報問い合わせ手法では，ユーザの位置情報プライバシを保護するために，ユーザ位置から算出した第 3 次メッシュの中心位置をダミーの位置としてサーバへの問い合わせに使用した．この処理を行うことで，サーバに付与されるユーザの位置情報は第 3 次メッシュコードに限られ，約 1km 四方の区画に曖昧化されていたが，ユーザ位置に関わらずプライバシ保護度合いが固定的であり，攻撃者が周辺の人流やユーザの移動経路，道路ネットワー

(注1)：東京 23 区内のアミューズメント施設，ミュージアム，ショッピング施設，エンターテインメント施設，温泉，劇場，ホールを東京ウォーカー [13] とナビタイム [14] から抜粋しイベントスポット辞書を作成する．ツイートにイベントスポット辞書内の要素が存在する場合は，その要素をイベント開催地として保持

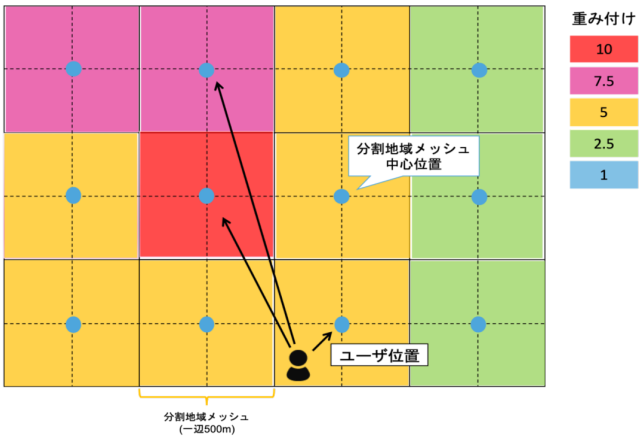

図 3: Geo-I を用いた位置情報匿名化の概要

クを考慮した予測を行うとユーザ位置の予測精度が高くなってしまうという課題があった．そこで本提案手法では，ユーザの位置情報を地域メッシュと人流データを利用し，2.2 節で紹介した Geo-I を満たすようなダミー位置を算出し，サーバへの問い合わせに使用することで，ユーザ位置の匿名化を保護度合い で行う．Geo-I を満たすメカニズム K は，データの平均にラプラスノイズを加えるラプラスメカニズムをはじめ様々なものが提案されているが，今回の実験では Bordenabe [16] らが提案した，ユーザ位置が $x\ 2\ X$ である事前確率 π_x を用いたメカニズム K を参考に実装を行った．このメカニズム K を使用したときのユーザ位置 x とノイズ付きの位置 z との距離 $d_Q(x;z)$ の期待値は以下のように表せる．

$$\sum_{x;z\,2} \pi_x k_{xz} d_Q(x;z)$$

分割地域メッシュ単位の人流データを用いた重みづけをユーザ位置の事前確率 π_x とし，ユーザ位置の匿名化を保護度合い で行う．求めた -Geo-I を満たす確率分布に従ってダミー位置を算出し，問い合わせの位置として使用する．今回の実験では図 3 に表すように，試験的に分割地域メッシュ 20 区画に対し，人流数が多い区画から少ない区画に 10 から 1 の 5 段階の重み付けを行い， の値と確率分布の関係を調査した．

5. Geo-I を用いた位置情報匿名化

5.1 実　　験

例として，新宿駅周辺と新橋駅周辺における 500m 四方の分割地域メッシュ 20 区画の人流データを用いて重み付けを行ない，問い合わせの位置として使用される -Geo-I を満たすダミー位置として各区画が選ばれる確率を調査する．ユーザ位置を固定しプライバシ保護度合い = 3:0 としたときの結果（図 4(a)）と = 2:0 としたときの結果（図 4(b)）を示す．また， = 2:5 と固定したとき，ユーザが周辺の人流数の多い地域（新宿）にいる場合の結果（図 5(a)）と，海などに囲まれ周辺の人流数の少ない地域（新橋）にいる場合の結果（図 5(b)）を示す．各メッシュ区画が選ばれる確率を数値で記載し，各メッシュ区画の色は図 3 における，5 段階の人流数に従う重み付けを表している．

5.2 考　　察

はノイズ付きの位置から元の位置に関して漏洩する情報量を制御するパラメータであり，その値が小さいほどプライバシ保護度合いは高くなるが，ノイズは大きくなり，よりユーザ位置とは無関係な区画が選ばれる．図 4(a) = 3:0 では元のユーザ位置がそのままダミー位置として選ばれる確率が 35.0% であるのに対し，図 4(b) = 2:0 においては 25.5% となっていて低い．また，人流数が多い赤色や桃色の区画に注目すると， = 3:0 の場合より = 2:0 の場合の方がより選ばれる確率が高くなっていることが見て取れる． が大きい値のときはよりユーザ位置からの距離が近い区画が選ばれ， が小さい値のときはより人流数が多い区画が選ばれる．また，図 5(a) の (b) を比較すると，ユーザ位置や周りの区画にあまり人がいない場合は，人流数の多い区画がより選ばれやすくなっていることがわかる．このような の値とユーザ位置における人流数の大小の傾向を利用し，適切な の値を調整することで，より人のいるところ，またはユーザ位置から近い場所を問い合わせの位置として選んで欲しいというユーザの要件に柔軟に対応することが可能となる．

6. まとめと今後の課題

ユーザ位置周辺という地理的な制約条件を満たすような，SNS データからのイベント情報の検索をする際の，ユーザの位置情報プライバシを保護するためのデータ処理方法について検討した．第 3 次メッシュの中心位置をダミーの位置としてサーバへの問い合わせに使用した手法における，ユーザ位置の人流数に関わらずプライバシ保護度合いが固定的であるという課題を解決するため，本稿では地域メッシュと人流データを利用し Geo-I を満たすようなダミー位置を算出する処理を行った． の値と Geo-I を満たす確率分布に従ってダミー位置が選ばれる確率を調査した結果， の値を調整することで，より人のいるところまたはユーザ位置から近い場所を問い合わせの位置として選んで欲しいというユーザの要件に柔軟に対応でき，従来の手法よりプライバシ保護の指標に則った度合いで，論理的，定量的なプライバシ保護を行うことが可能となる．今回の実験では限定的な区画で実験を行い，年間の人流データを使用したが，今後はより広い連続する範囲での実装，時間帯別や年齢別の人流データを使用し，より細かなユーザの要件に柔軟に対応し，適切な の値の設定を行いたい．また，より適切にイベント開催地分布に応じた場合分けを行うために過去のイベントの発生情報の利用や，ユーザの要望に合うような情報の推薦を行うために他のユーザデータを利用したイベント情報の推薦順位付けを考えていきたい．また，本手法はユーザが移動しながら都度複数回の問い合わせを行い，攻撃者が問い合わせ結果を全て得られるような場合を考慮していない．今後は取得したイベント情報の精査に加え，ユーザの移動経路を保護できるようなモデルも構築していきたい．