DEIM Forum 2022 J43-2

ユーザの位置情報プライバシを考慮した SNS データからのイベント情報抽出・検索手法の検討

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，問合せユーザの位置情報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多く研究されている．しかし，ユーザにとって，位置情報など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．そこで我々は，ユーザの位置情報を地域メッシュ情報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することで，SNS の一種である Twitter からイベント情報を取得し，ユーザの位置情報プライバシを考慮して地理的な制約条件を満たしながら，大量の SNS データからイベント情報を検索する手法を検討した．

キーワード　SNS，情報抽出，位置情報プライバシ

Event Information Extraction and Search Method from SNS Data considering User's Location Information Privacy

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，情報収集を目的とする SNS を利用者がコミュニケーション目的の利用者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 [2] も盛んである．

本研究では，SNS の一種である Twitter [3] からイベント情報を取得し，ユーザの位置情報プライバシを考慮する処理を施すことで，地理的な制約条件を満たしながら，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．

本論文の構成は以下の通りである．2 章で関連研究について述べ，3 章で先行研究について述べる．4 章で提案システムの概要を述べ，5 章で地域メッシュを用いたイベント情報の問い

合わせ結果と考察を述べる． 最後に，6 章で本稿をまとめる．

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．

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 [4] 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 [5] も盛んである．本研究では局所差分プライバシの概念を参考に，ユーザの位置情報にノイズを加えるプライバシ処理を施し，サーバへの問い合わせを行う．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．

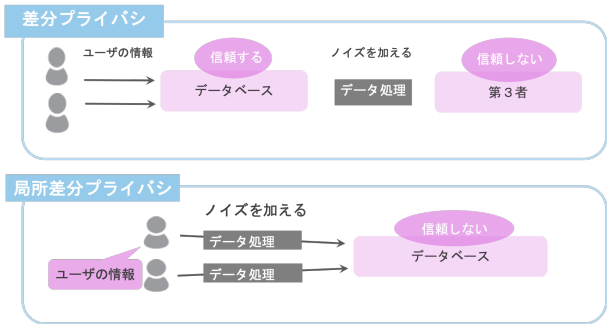

図 1 差分プライバシ・局所差分プライバシの概要

Andrés ら [2] は，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基準 Geo-indistinguishability(Geo-I) を提案した．この Geo-I もまた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．また，高木ら [6] 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り，現在も様々な議論，研究が進められている．

3. 先 行 研 究

本章では，先行研究について紹介する．同研究室工藤ら [7] 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 [8] 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 [9] 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．

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．

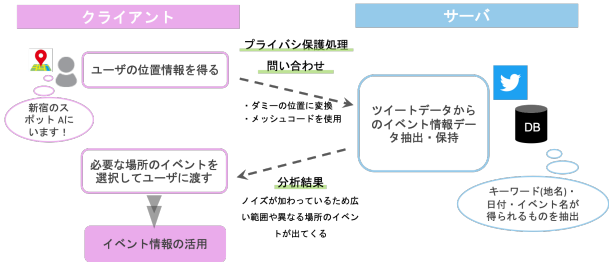

図 2 提案システムの概要

サーバ側は先行研究で紹介した工藤ら [7] 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．本研究では 4.2 節で後述する地域メッシュ [10] 情報を用いて，ユーザ位置をダミーの位置に変換する処理を行い，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 [8] 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 [10] 情報を用いて，ユーザの位置情報のプライバシ保護処理と，範囲を広げたイベント情報検索を試みる．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 [13] 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

(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 [11] とナビタイム [12] 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

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

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する．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る．本研究では，ユーザ側に現在地から短距離順の上位 k 個以上のイベント情報を得たいというような条件があると仮定し，1 回目の問い合わせで得られたイベント数が k 個より少なかった場合範囲を広げて 2 回目の問い合わせを行う．このように，得られたイベント数に応じて範囲を段階的に広げる問い合わせを繰り返すことを想定し，今回の実験では地域メッシュを利用した 3 回の問い合わせ範囲を設定した．図 3 に，問い合わせで用いたダミー位置と問い合わせ範囲の広げ方の概要を示す．1 回目の問い合わせでは，第 3 次メッシュ区画の範囲で行われるイベント情報を検索する．2 回目以降の問い合わせでは，前回の問い合わせ範囲から，四方外側の分割地域メッシュ一区画分範囲を広げてイベント情報を検索する．

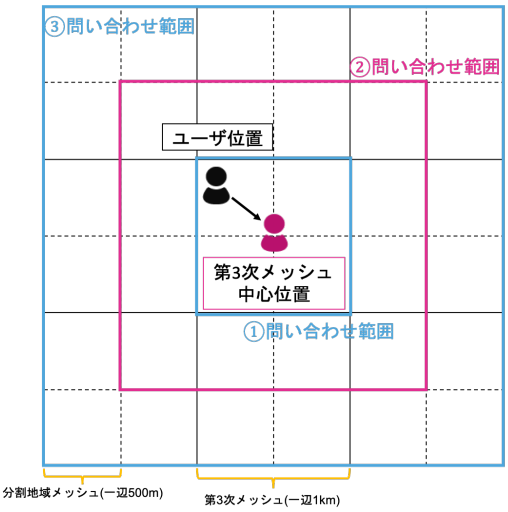

図 3　ダミー位置と問い合わせ範囲の広げ方

5. 地域メッシュを用いたイベント問い合わせ

5.1 実験データ

地域メッシュを用いたイベントの問合せの実験を行うために，2021 年の 11/20 から 11/24，12/10 から 12/14 のそれぞれ 5 日間，計 10 種類のキーワードを含むツイートを取得した．表 2 に設定したキーワード (地名) と抽出できたイベント情報数 (イベント名の重複を含む) の関係を示す．

山手線沿線のターミナル駅である新宿や渋谷をキーワードとしてツイートに含んでいたイベント情報数が比較的多いことが分かる．

5.2 問い合わせ結果

例として，都内の駅やスポット 6 種類のユーザ位置を設定し，それぞれに対し第 3 次メッシュの中心位置を算出しダミーの位

表 2　キーワードと取得したイベント数の関係

キーワード	取得ツイート数(個)				
	11/20(土)	11/21(日)	11/22(月)	11/23(火)	11/24(水)
原宿	16	16	11	18	11
渋谷	85	116	116	118	92
新宿	101	96	126	108	90
赤坂	6	6	10	13	7
代々木	21	29	18	28	18
池袋	51	76	69	56	48
六本木	9	12	25	13	12
	12/10(金)	12/11(土)	12/12(日)	12/13(月)	12/14(火)
立川	6	16	8	10	8
足立	12	7	7	4	5
八王子	35	34	49	38	47

置とした．ユーザ位置と第 3 次メッシュ中心位置それぞれの緯度経度と 2 点間の距離を表 3 に示す．

表 3　ユーザ位置と第 3 次メッシュ中心位置の緯度経度と距離

ユーザ位置名	ユーザ位置 緯度，経度	第 3 次メッシュ中心位置 緯度，経度	ユーザ位置と 中心位置の距離(m)
新宿三丁目駅	35.6908823, 139.7048821	35.6875, 139.70625	395.17542
東京都庁	35.6896342, 139.6921007	35.6875, 139.69375	279.92834
東京芸術劇場	35.729842, 139.7080751	35.729167, 139.70625	181.30916
東京ドーム	35.7056396, 139.75189	35.704167, 139.75625	427.06464
六本木ヒルズ	35.6602384, 139.7300767	35.6625, 139.73125	272.49411
八王子駅	35.6556157, 139.338853	35.654167, 139.34375	471.67985

ダミーの位置から範囲を広げる問い合わせを行い，問い合わせ範囲内で開催されるイベント情報を取得した．

今回の実験は，各回の問い合わせで得られたイベント情報数に関わらず，図 3 に示した①1 回目問い合わせ範囲,②2 回目問い合わせ範囲,③3 回目問い合わせ範囲で計 3 回の問い合わせを行った．段階的に問い合わせ範囲を広げた計 3 回の問い合わせで得られたユーザ位置とイベント取得数の関係を表 4 に示す．

表 4　第 3 次メッシュ中心位置から範囲を広げた問い合わせ結果

ユーザ位置	問い合わせで得られたイベント数(個)		
	1 回目問い合わせ	2 回目問い合わせ	3 回目問い合わせ
新宿三丁目駅	44	144	180
東京都庁	2	122	178
東京芸術劇場	29	39	45
東京ドーム	1	1	10
六本木ヒルズ	17	19	30
八王子駅	7	7	7

5.3　考　　察

ユーザ位置によって，範囲を広げるサーバへの問い合わせをしても，得られるイベント数があまり増加せず，ユーザが必要とする k 個のイベント情報を取得するまでに必要な問い合わせ回数が多くなる．また，何度か繰り返して範囲を広げた問い合わせをしても k 個のイベント情報が得られずユーザの待機時間が長くなってしまう可能性があるため，問い合わせ回数の上限を設定する必要がある．これは，東京ドームなど，ターミナル駅から離れた位置にいた場合に顕著で，イベント開催地の分布は一様では無く，ターミナル駅付近に集中していることが理由として考えられる．例えば，池袋駅付近のイベント開催地の分布は以下の図 4 のようになっていて，池袋駅付近に集中して分布していることが見て取れる．

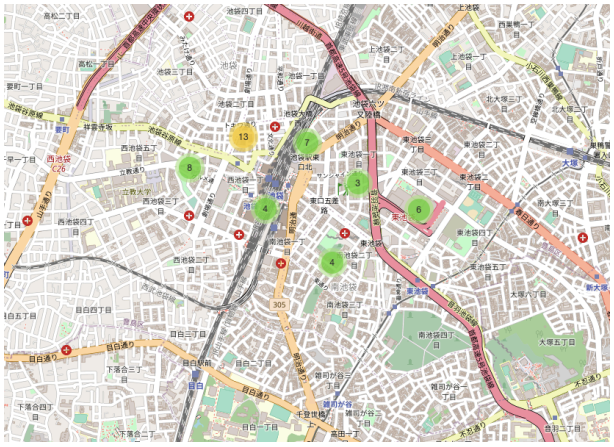

図 4　池袋駅付近で取得したイベント開催地の分布（周辺で開催されるイベントをクラスタ化し件数表示）

しかし，1 回目の問い合わせから広い範囲で問い合わせをすると，新宿三丁目駅や東京芸術劇場などターミナル駅付近の位置にいる場合，問い合わせの結果得られるイベント数が膨大となり，ユーザが必要とする k 個のイベント情報を算出するクライアント側の計算量が大きくなってしまう．よって，段階的に範囲を広げる問い合わせは有効であると考えられる．

6.　まとめと今後の課題

本研究では，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．ユーザ位置の第 3 次メッシュ中心位置をダミー位置として設定し，サーバへの問い合わせに使用することで，サーバに付与されるユーザ位置情報は第 3 次メッシュ情報に限定されることを保証した．攻撃者からユーザの位置情報の予測はされにくくなるが，問い合わせ位置をずらしたことによってデータの有効性が多少欠損してしまうことが課題として挙げられる．また今回得られたイベント情報数が，ユーザ位置によって問い合わせ回数に応じてあまり増加しないものがあったため，問い合わせ回数とクライアントの計算量のトレードオフの関係を考慮し，ユーザ位置の周辺ターミナル駅との距離に

よって問い合わせ範囲の広げ方を変更するなど，イベントの密集状況に応じたより効率的な範囲の広げ方を検討していきたい．また，本研究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．攻撃者が道路ネットワークを考慮した予測を行うとユーザ位置の予測精度は高くなると考えられる．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．

文　献

[1] 令和元年度通信利用動向調査 (総務省). https://www.soumu.go.jp/johotsusintokei/statistics/data/200529_1.pdf.

[2] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. *Proceedings of the 2013 ACM SIGSAC conference on Computer and Communications Security (CCS'13))*, pages 901–914, 2013.

[3] Twitter. http://twitter.com/.

[4] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Differential privacy. *Proc. 33rd International Colloquium on Automata, Languages and Programming (ICALP' 06)*, pages 1–12, 2006.

[5] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. *Proc. IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS' 13)*, pages 429–438, 2013.

[6] Shun Takagi, Yang Cao, Yasuhito Asano, and Masatoshi Yoshikawa. Geo-graph-indistinguishability: Location privacy on road networks based on differential privacy. *CoRR*, abs/2010.13449, 2020.

[7] Ruriko Kudo, Miki Enoki, Akihiro Nakao, Shu Yamamoto, Saneyasu Yamaguchi, and Masato Oguchi. Real-time event search corresponding to place and time using social stream. *he 6th IEEE International Congress on Big Data (BigData Congress2017)*, pages 1047–1053, 2017.

[8] Twitter search api. https://dev.twitter.com/rest/public/search.

[9] Miki Imai, Miki Enoki, Ruriko Kudo, and Masato Oguchi. Personalized local event search based on sns data analysis. *the 14th IEEE International Conference on Ubiquitous Information Management and Communication*, pages 6–14, 2020.

[10] 総務省情報局 地域メッシュ統計の概要. https://www.stat.go.jp/data/mesh/m_tuite.html.

[11] Walker+ 東京都のおでかけスポット一覧. https://www.walkerplus.com/spot_list/ar0313/.

[12] Navitime. https://www.navitime.co.jp/category/.

[13] Google maps geolocation api. https://developers.google.com/maps/documentation/geolocation/overview.