

ジオタグ付きツイート分析による群集存在確率辞書の構築

千綿愛誠^{†1} 若宮翔子^{†1} 丸山直樹^{†1} 河合由起子^{†1} 川崎洋^{†2} 角谷和俊^{†3}
^{†1} 京都産業大学 ^{†2} 鹿児島大学 ^{†3} 兵庫県立大学

1 はじめに

近年、スマートフォンの普及や Twitter のような SNS の発達により、ジオタグ（緯度経度）付きツイートが大量ユーザによって投稿されている。我々はジオタグ付きツイートを活用し、実空間での群集行動を分析する研究を行ってきた [1]。しかしながら、ジオタグ情報とツイート内容に記載される場所情報間の関係性の分析には至っておらず、ユーザの移動観察は不十分であった。例えば、京都駅にいるユーザが「二条城に向かっています」というツイートを発信した場合、その後、二条城においてツイートを発信した場合は京都駅から二条城へのユーザ移動を観察できる。一方、「二条城に行く」というツイートを発信したにも関わらず、その後、二条城においてツイートを発信しなかった場合は京都駅から二条城へのユーザ移動を観察できない。

本研究では、ツイートのジオタグのみに基づく移動観察の欠損を補完するために、ツイートのジオタグに基づく場所とツイート内容のテキストに基づく場所との関係性を考慮することで群集の移動パターンを分析し、場所間の群集の移動確率を格納した辞書を構築する。具体的には、ジオタグ付きツイートからチェックインツイート（「ジオタグに基づく場所=テキストに基づく場所」）を検出して、場所名とジオタグの組を格納した場所情報データベースを作成する。次に、チェックインツイート以外のツイートと場所情報データベースをマッチングし、ジオタグに基づく場所とテキストに基づく場所が異なるツイートを抽出する。そして、抽出したツイートを発信したユーザが実際にテキストに基づく場所に移動したかを判定する。そのために、同一ユーザが一定時間内に発信したツイートのジオタグに基づく場所との近接関係を判定する。近接関係がある場合、ユーザが特定の場所についてつぶやいた後、一定時間内にユーザが実際にその場所周辺に存在したと判定する。さらに、近接関係の判定結果を集約し、特定の場所（移動元）から他の場所（移動先）への群集の

移動確率に基づく（移動先における）存在確率を算出し、群集存在確率辞書を構築する。これにより、場所 p_1 （移動元）で場所 p_2 （移動先候補）についてツイートを発信したユーザが、その後一定時間内にツイートを発信しなかったとしても、群集存在確率辞書に基づきユーザの場所 p_2 に存在する確率を推定可能となる。

2 群集存在確率辞書の構築

Twitter 上のジオタグ付きツイートには「テキストに場所名が含まれるもの」と「テキストに場所名が含まれないもの」が存在する。本研究では、前者のジオタグ情報に基づき各ユーザの居場所とテキストにおける場所名との関係性を分析する。

条件 1: ユーザの居場所と場所名が同一（「ジオタグに基づく場所=テキストに基づく場所」）

条件 2: ユーザの居場所と場所名が異なる（「ジオタグに基づく場所 \neq テキストに基づく場所」）

このうち、条件 1 を満たすチェックインツイートを検出する。チェックインツイートとは、位置情報サービスを通して発信されるツイートである。代表的なものとして、Swarm¹ におけるチェックインなどがある。それらは定型文となるため、文字列マッチングで抽出可能である [2]。例えば、「I'm at 場所名」というパターンのツイート（「I'm at 京都駅」）や、「@ 場所名」というパターンのツイート（「@ 京都駅」）がある。抽出した場所名をジオタグとともに場所情報データベースに格納する。

チェックインツイート以外のツイートは文字列マッチングで検出されないため、条件 1 と条件 2 のツイートが混在する。例えば、二条城における「二条城に到着」というツイートは条件 1 のツイートであるが、京都駅で発信された「二条城に行く」というツイートは条件 2 のツイートである。そこで、場所情報データベースを用いて、チェックインツイート以外のツイートから条件 1 を満たすツイートと条件 2 を満たすツイートをそれぞれ抽出する。そのために、チェックインツイート以外のツイートのテキストに基づく場所とジオタグに基づく場所の近接関係を判定する。

¹Swarm: <https://www.swarmapp.com>

A Proposal of crowd Existence Probability Dictionary by Analyzing Geo-tagged Tweets

^{†1} Manase CHIWATA ^{†1} Shoko WAKAMIYA ^{†1} Naoki MARUYAMA ^{†1} Yukiko KAWAI ^{†2} Hiroshi KAWASAKI ^{†3} Kazutoshi SUMIYA

^{†1} Kyoto Sangyo University

^{†2} Kagoshima University

^{†3} University of Hyogo

表 1: 群集存在確率辞書

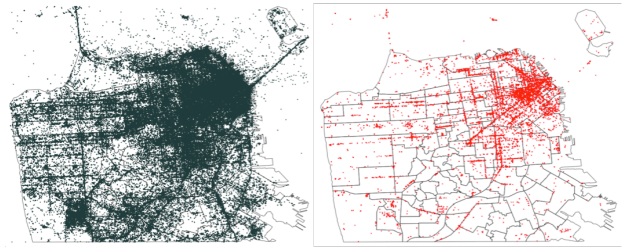
移動元			移動先			存在確率
場所名	緯度	経度	場所名	緯度	経度	
京都駅	34.98	135.75	二条城	35.01	135.74	30%
京都駅	34.98	135.75	清水寺	34.99	135.78	25%

さらに、近接関係の判定結果から、特定の場所についてつぶやいた後、一定時間内にユーザがその場所周辺に存在したかを調べる。まず、テキストに基づく場所とジオタグに基づく場所の近接関係がないツイートを発信したユーザを発見する。次に、同一ユーザがその後一定時間内に発信したツイートのジオタグに基づく場所と元のツイートのテキストに基づく場所との近接関係を判定する。これらに近接関係がある場合、ユーザが特定の場所についてつぶやいた後、一定時間内にユーザが実際にその場所周辺に存在したと判定する。各ユーザの移動先候補における存在の有無を抽出して集約することにより、場所間の群集の移動確率を算出する。表 1 に群集存在確率辞書の例を示す。辞書は、「移動元」と「移動先」の場所情報および「(移動先における) 存在確率」で構成される。

3 実装

群集の移動確率を算出し、辞書構築を行った。Twitter Streaming API を用いて、San Francisco, CA, USA (経度 = [-122.51490, -122.3570], 緯度 = [37.70809, 37.8324]) において約4ヶ月 (2013年9月25日から2014年1月17日) にわたり発信された 583,536 件のジオタグ付きツイートを取得した。ユーザ数は 54,793 人であった。このうち、ツイートを 100 件以上発信したヘビー Twitter ユーザは 845 名 (1.5%) であった。なお、ローカル情報や広告を発信するユーザアカウントは省いた。また、本実験では、Swarm を通して発信されたツイートをチェックインツイートとした。図 1(a) に取得したジオタグ付きツイートの地理分布を示す。このうち、Swarm のチェックインツイートは 40,600 件 (取得したツイートの約 7.0%) であった。図 1(b) にチェックインツイートの地理分布を示す。

2 章で述べた Swarm のチェックインツイートのジオタグに基づく場所とテキストに基づく場所は同一であるという仮定から、場所名とジオタグを抽出した。抽出結果より場所情報データベースに登録された場所情報は 8,491 箇所 (重複を省く) であった。登録された場所情報の場所名を含むチェックインツイート以外のツイート数のうち 40,217 件 (取得したツイートの約



(a) 全ツイート (b) チェックインツイートのみ

図 1: 収集したツイートの地理分布

6.9%) がジオタグの近接判定内であった。ただし、他のチェックインツイート (Instagram など) が含まれていたため、今後それらに対応することでより精度の高い辞書構築を目指す。

4 おわりに

本研究では、ツイートのジオタグ情報のみに基づく移動観察の欠損を補完しユーザの移動推定を行うために、特定の場所から他の場所への群集の移動確率に基づく群集存在確率辞書の構築手法を提案した。提案手法は、ツイートのジオタグに基づく場所とテキストに基づく場所との関係性を考慮しており、群集の移動パターンの分析につながる。

今後の課題として、時間帯により移動する確率は異なると考えられるため、時間帯を考慮した場所間の群集の移動パターンを算出して辞書を改良する計画である。また、1つのツイートのテキストに複数の地名が含まれている場合や、ツイートのテキストには場所名以外にも、その場所を示したり頻繁に用いられるキーワード (ローカル語) が含まれていると考えられるため、これら複数語、ならびにローカル語の抽出手法を検討する。

謝辞

本研究の一部は、総務省戦略的情報通信研究開発推進事業 (SCOPE) および JSPS 科研費 26280042 の助成を受けたものである。ここに記して謝意を表す。

参考文献

- [1] 若宮 翔子, 李 龍, 角谷 和俊, “位置ベース SNS を通じた群集の移動経験に基づく都市空間の近接性分析”, 情報処理学会論文誌データベース (TOD58), Vol. 6, No. 3, pp. 159-176, 2013.
- [2] 服部 哲, 速水 治夫, “位置情報を含むツイートを効率的に発掘するための基本方式の検討”, マルチメディア, 分散, 協調とモバイル (DICOMO2011) シンポジウム講演論文集, pp. 1526-1530, 2011.