ソーシャルメディアにおけるユーザ位置推定に関する研究

A Study on User Location Inference in Social Media

山口 祐人 ♥

Yuto YAMAGUCHI

1. はじめに

ソーシャルメディアの台頭は，ユーザ個人に情報発信の場を提供した．このような状況で，ユーザ個人が発信した情報を分析し，新しい知見を得るソーシャルメディアマイニングに関する研究が広く行われている．中でも，ソーシャルメディアを通じて実世界を知ろうとする研究が注目を集めている．例えば，地震や台風などの実世界イベントの検知 [15]，インフルエンザなどの感染症の分析 [13]，災害の分析 [11] などはその代表例である．実世界のどこでこのような現象が起こっているのかを知ることは重要であり，それを知るためにはソーシャルメディアユーザの居住地情報が必要不可欠となる．

しかし，多くの研究で指摘されているように，ソーシャルメディアユーザは自らの居住地を公開していないことが多い [7, 2, 5]．これは，居住地公開のメリットの欠如，プライバシの問題などが原因であると考えられる．そこで本研究ではソーシャルメディアユーザの居住地の推定を行う．

ソーシャルメディアユーザの居住地推定を行う研究はその重要性から多く行われている．これらの研究の大部分はグラフベース手法 [2, 10, 8] かコンテンツベース手法 [5, 6, 7] のどちらかに分類される．グラフベース手法はソーシャルグラフを用いる手法であり，コンテンツベース手法はユーザが投稿したコンテンツを用いる手法である．前者はグラフ分析がベースとなっており，後者は自然言語処理がベースとなっている．本研究ではそれぞれの既存手法における研究課題にアプローチすることで精度向上等の拡張を行い，またそれらを既存研究と比較することにより実験的に評価する．

本研究では以下の三つの研究課題を扱う．一つはグラフベース手法に関する課題であり，続く二つはコンテンツベース手法に関する課題である．

研究課題 1：ソーシャルグラフにおけるグラフランドマークを用いた居住地推定　グラフベース手法の大部分は *closeness assumption* を基にしてユーザの居住地を推定している．Closeness assumption とは，ソーシャルグラフ上で接続されているユーザ同士（友人等）はその居住地が互いに近いという仮定である．しかし，ソーシャルグラフの性質によっては closeness assumption は必ずしも有効ではない [12]．例えば，最大のマイクロブログである Twitter ではツイートと呼ばれる記事のユーザ間の閲覧関係（フォロー）によってソーシャルグラフが定義されている．そこでは，例えばある一般ユーザのアカウントと，そのユーザが興味を持っている企業などのアカウントがソーシャルグラフ上で接続されることがある．このような環境では closeness assumption はあまり有効でないことが多い．

そこで本研究では closeness assumption とは異なる *concentration assumption* を導入する．Concentration assumption とは，ソーシャルグラフ上にはある地域から集中して注目を集めるユーザ（グラフランドマーク）が存在するという仮定である．例えば，Twitter においてつくば市天気予報アカウントはつくばに住む多くのユーザからフォローされる傾向にあるため、グラフランドマークであると言える．グラフランドマークを用いると，「グラフランドマークの隣接ノード群の居住地は互いに近い」という推定が可能になる．第 2. 節ではグラフランドマークを用いた居住地推定手法を提案し，その結果を既存のグラフベース手法と比較することにより実験的に評価する．

研究課題 2：実世界ローカルイベントを用いた居住地推定　コンテンツベース手法は，ユーザが投稿したコンテンツに含まれる単語などの地理的な局所性を用いてユーザの居住地を推定している．例えば，つくば市の地名を頻繁に投稿するユーザはつくば市に住んでいる可能性が高い．また，地名でなくとも，その地域の名物や方言などの地理的な局所性を用いることも出来る．例えば，アメリカで “rockets” と頻繁に投稿するユーザはヒューストンに住んでいる可能性が高いと考えられる [5]．これは，ヒューストンには NASA の宇宙センターや NBA のバスケットボールチームがあるためである．このような単語を**ローカルワード**と呼ぶ．

一方，本研究ではこのような定常的な局所性ではなく，実世界におけるローカルイベントによって発生する一時的な局所性に着目する．例えば，ある地域で地震が発生したときに地震に関する投稿をしたユーザはその地域に住んでいる可能性が高い．第 3. 節では，ローカルイベントを用いた居住地推定手法を提案し，その結果を既存のコンテンツベース手法と比較することにより実験的に評価する．

研究課題 3：ソーシャルストリームを用いたオンライン居住地推定　ソーシャルメディアでは，ユーザが時々刻々と大量の投稿をしており，ソーシャルストリームと呼ばれる情報源を構成している．そのため，居住地推定の手がかりとなる情報もリアルタイムに増え続けており，インクリメンタルな推定手法が求められる．. しかし，既存のコンテンツベース手法は新たな投稿が到着したとき，過去の投稿も全て用いて推定を初めからやり直さなければならない．従って，計算コスト，記憶コストが大きいという問題点がある．

♥ 正会員　筑波大学大学院システム情報工学研究科
yuto_ymgc@kde.cs.tsukuba.ac.jp

そこで，本研究ではソーシャルストリームから次々に得られるコンテンツを基に居住地を逐次推定する手法を提案する．また，本手法ではローカルイベントに基づく手法から得られた知見を基に，*temporally-local word* という新しいローカルワードを導入する．Temporally-local word とは，従来の定常的な局所性を持つローカルワード（*statically-local word*）とは異なり，一時的な局所性を持つ単語のことである．Statically-local word と temporally-local word を併用することで，ローカルイベントを用いた手法では問題であった再現率の低下を抑えつつ精度を向上させる．第 4. 節では，ソーシャルストリームを用いたオンライン居住地推定手法を提案し，実際に時間を追って推定結果が改善されていくことの検証や既存手法との比較を通じて本手法を実験的に評価する．

2. ソーシャルグラフにおけるグラフランドマークを用いた居住地推定

先に述べたように，これまでに提案されているグラフベース手法は closeness assumption に従っている [2, 1, 8, 14, 10]．これらの手法とは異なり，本研究で提案する手法は新しく導入した concentration assumption に基づいている．Concentration assumption とは，ソーシャルグラフ上に自らのフォロワー群の居住地がある地域に集中しているユーザ（グラフランドマーク）が存在するという仮定である．フォロワーとは，Twitter において用いられる用語であるが，ソーシャルグラフが有向グラフで定義されるソーシャルメディアにおいて，あるノードに対してエッジを張っているノードのことである．Concentration assumption により，多くの既存手法で採用されている友人同士の居住地は互いに近いという仮定（closeness assumption）を前提とせずに居住地を推定できるようになった．

提案手法の概要について示す．提案手法（landmark mixture model; LMM）は，グラフランドマークをフォローするユーザの居住地は互いに近いという仮定にもとづき，ユーザの居住地を確率分布でモデル化する手法である．まず全てのユーザに対して，そのユーザのフォロワー群の居住地の分布（dominance distribution）を計算し，割り当てる．そして，あるユーザの居住地の分布を，そのユーザがフォローするユーザ群の dominance distribution を混合することにより得る．得られた居住地の分布において確率密度が最大になる点を推定した居住地とする．

2.1 評価実験

Li ら [10] によって提供されている Twitter データセットを用いて提案手法と既存手法との比較実験を実施した．本データセットはアメリカに住んでいるとされる Twitter ユーザ 3,122,842 と，ユーザ間のフォローエッジ 284,884,514 によって構成されている．いくつかの関連研究 [5, 10] で採用されている方法に従い，2010 census U.S. gazetteer[1] を用いてユーザのロケーションプロファイル（テキスト情報）を緯度経度情報に変換した．結果として，464,794 ユーザ（約 15%）が正しく緯度経度に変換された．この緯度経度情報をユーザのロケーションの正解データとして用いた．

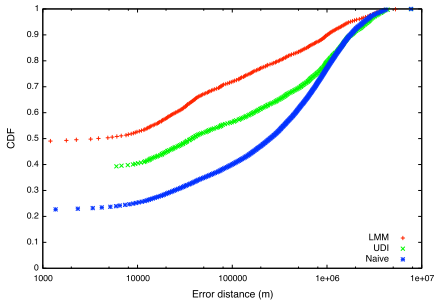

図 1 グラフランドマークを用いた提案手法とグラフベースの既存手法との比較結果

本評価実験では，提案手法（LMM），Li ら. [10] による手法（UDI），単純にユーザの隣接ユーザ（follower 及び followee）の居住地のメドイド[2]を計算し，推定した居住地とする手法（Naive）の三つの手法の精度を比較した．UDI は closeness assumption に基づいた手法であるが，有名人など，世界中の至る所に follower を持つユーザは居住地推定の役に立たないとして小さな重みを与えて評価している．我々の知る限りでは最も高い精度を実現する手法である．

結果を図 1 に示す．図 1 の横軸は推定誤差（単位:m）の値を示し，縦軸は推定誤差が対応する値以下であるユーザの割合，すなわち精度を示している．図によると，提案手法が Li らの手法及び Naive の精度を上回っていることが分かる．

本研究では，他にも二つの制約によって精度，再現率，および計算コストのトレードオフが調節できることを示した．結果として，再現率を高い値に保ったまま精度を向上させることが可能であることと，再現率を高い値に保ったまま計算コストを削減できることを示した．詳細は論文を参照されたい．

2.2 本研究の貢献

本研究の貢献を以下にまとめる．

1. Concentration assumption を導入することにより，友人同士の居住地は近いという closeness assumption を前提としない居住地推定を可能にした．
2. Concentration assumption に基づいた手法 LMM を提案した．また，二つの制約の導入により，既存手法では行えなかった，精度，再現率，計算コストのトレードオフの調節を実現した．
3. Twitter データセットを用いた実験により，提案手法が現時点で最も高い精度を実現している Li らの手法より 27% 高い精度を実現したことを示した．

[1] http://www.census.gov/geo/maps-data/data/gazetteer2010.html

[2] セントロイドに最も近いデータ点

3. 実世界ローカルイベントを用いた居住地推定

コンテンツベースの居住地推定手法としては，Cheng ら [5] が先駆的である．Cheng らはローカルワードと呼ばれる，投稿される地域に偏りのある単語を用いてユーザの居住地を推定する手法を提案した．ローカルワードは地名辞書を用いず，それを投稿したユーザの地理的な偏りのみによって抽出される．他にも様々な手法が提案されているが [4, 7, 9, 3]，いずれも上記のアイデアに基いている．

これらの手法は定常的に地理的な偏りを持つローカルワードを用いているが，本研究では時間的な特徴を考慮する．実世界では，地理的な局所性を持つ**ローカルイベント**（地震や火事など）が発生する．ローカルイベントが発生すると，そのイベントが発生した付近にいるユーザは一斉にそのイベントに関する投稿を行う傾向がある．例えば，東京で地震が発生した時には，東京にいるユーザの多くが「地震だ！」などの投稿をする．本研究ではこのローカルイベントに関する投稿を用いて居住地が未知であるユーザの居住地を推定する手法を提案する．

提案手法の概要について示す．提案手法はまずソーシャルメディアの投稿からローカルイベントを抽出する．ここでのローカルイベントとは，投稿時刻が近く，内容が類似し，投稿位置が互いに近い投稿の集合のことをいう．内容の類似性は投稿内容を用いたクラスタリングによって評価し，投稿位置の近接性は投稿集合の中心点とそれぞれの投稿との距離の平均を評価する．そして，抽出されたローカルイベント（投稿集合）に含まれる投稿をしたユーザはそのローカルイベント付近に住んでいるという仮定に基づき、居住地を推定する．

3.1 評価実験

Twitter から独自に収集したツイート及びユーザを用いて提案手法と既存手法の精度を比較した．本実験では地震，天気，竜巻，緊急（救急車やサイレンなど）に関するキーワードで Twitter を検索，それぞれに合致するツイートを収集し，四つのデータセットを作成した．本評価実験では，提案手法（Proposed），UDI [10]，Cheng ら [5] による手法（Cheng），ユーザの居住地を取りうる位置の集合 L からランダムに決定する手法（Random）の四つの手法の精度を比較した．

結果を図 2 に示す．図 2 の縦軸と横軸は図 1 と同様である．図によると，推定誤差 160km での提案手法の精度を UDI，Cheng と比較するとそれぞれ約 34%，約 122% の向上を示していることが分かる．ただし，既存手法によるカバレッジ，すなわち居住地を推定できたユーザの割合はほぼ 100% であるのに対し，提案手法のそれは約 10% から 1% 程度であった．

本研究では，この他にも検出されたローカルイベントの妥当性，提案手法のパラメータを変えた時の精度及び再現率の比較，及び四つのデータセットによる精度の比較を実施している．結果として，妥当なローカルイベントが検出されており，四つのデータセットの中では天気に関するデータセットが最も精度が高いことが示された．

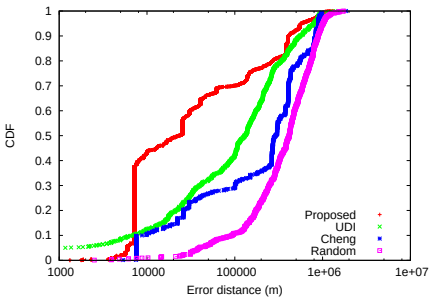

図 2　提案手法と他の手法との比較結果.

3.2 本研究の貢献

本研究の貢献を以下にまとめる．

1. ソーシャルメディア上では実世界におけるローカルイベントに関する多くの投稿が行われることを生かし，その投稿を用いた居住地推定手法を提案した．これにより，既存手法では用いることの出来なかった情報を基に居住地推定が可能であることを示した．
2. Twitter データセットを用いた実験により，提案手法が既存手法より 34% から 122% 高い精度を実現したことが示された．

4. ソーシャルストリームを用いたオンライン居住地推定

既存のコンテンツベース手法はソーシャルメディアにおいてインクリメンタルな推定を行うには，計算コストや記憶コストが大きいという問題点がある．そこで，本研究では時々刻々と得られるコンテンツを用いて居住地を逐次推定する手法（Online Location Inference Method; O-LIM）を提案する．また，本研究では一時的に地理的な局所性を持つローカルワード（temporally-local word）を導入する．例えば，実世界で地震が発生した時には，地震に関連する単語（地震，揺れる，震度など）が一時的に地理的な局所性を持つと考えられる．既存のコンテンツベース手法 [4, 7, 9, 3] は全て定常的なローカルワード（statically-local word）のみを用いているが，本研究では statically-及び temporally-local word の両方を用いることで精度の向上及びカバレッジの向上を実現した．

提案手法の概要を示す．まず，二種類のローカルワードをそれらが投稿された位置の分布を用いて検出する．ある単語を含む投稿をしたユーザの居住地の分布をその単語の分布とし，その分布と全ユーザの居住地の分布との乖離を計算する．計算した乖離が閾値より大きい場合は対応する単語をローカルワードとして抽出する．単語の分布を計算する期間をある一定の期間（一時間）などに限定することで，一時的な局所性を持つ temporally-local word の抽出が可能となる．そして，抽出されたローカルワードを含む投稿をしたユーザはそのローカルワードが示す場所に住ん

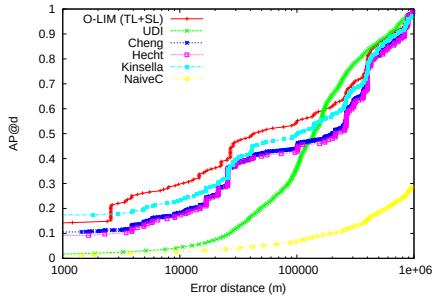

図 3　既存手法との比較結果

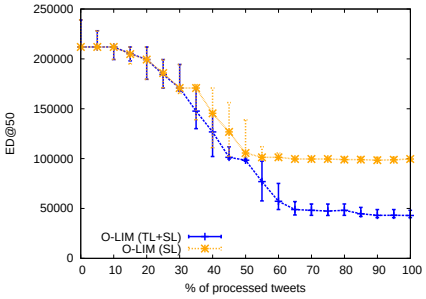

図 4　時間経過による推定誤差減の減少

でいる可能性が高いとして居住地を推定する．本研究では，推定結果は新しく得られた投稿に含まれるローカルワードのみを用いて逐次更新できることを示した．それにより，更新にかかる計算コストを削減し，また過去の投稿を保存する必要が無いため記憶コストも削減した．

4.1　評価実験

本実験では，Twitter から独自に収集したデータを用いた．本データは日本に住む 201,570 の Twitter ユーザと，それぞれのユーザの直近の 200 ツイート，ユーザ間のフォローエッジ 33,569,924 で構成されている．ユーザの居住地の正解データは，ユーザのロケーションプロファイル（テキスト情報）を Yahoo! Geocoder を用いて緯度経度情報に変換することによって得た．本評価実験では，Statically-，及び temporally-local word の両方を用いる提案手法（O-LIM (TL+SL)），statically-local word のみを用いる提案手法（O-LIM (SL)），UDI [10]，Cheng [5]，Kinsella ら [9] による手法（Kinsella），Hecht ら [7] による手法（Hecht），単純にユーザが投稿した地名のメドイドを取る手法（NaiveC）の七つの手法の精度を比較した．

図 3 は各手法の精度の比較結果を示している．図 3 の縦軸と横軸は図 1 と同様である．結果から，提案手法の精度が最も高いことが分かる．

図 4 は時間を追って提案手法が推定誤差を減少させていく結果を示している．横軸はデータセットに含まれるツイートを時系列順に処理した時の処理したツイートの割合を表し，縦軸は推定誤差の中央値を表している．結果から，推定誤差はオンライン推定により徐々に減少していくことが分かる．また，statically-local word だけでなく，temporally-local word も用いたほうが推定誤差が小さいことが分かる．

4.2　本研究の貢献

本研究の貢献を以下にまとめる．

1. 既存手法ではコストが大きかったオンライン推定を可能にした．提案手法により，時々刻々と得られるソーシャルメディアの投稿を用いて居住地を逐次推定することが可能になった．
2. 既存のコンテンツベース手法で用いられていた，定常的に地理的な局所性を持つローカルワード（statically-local word）に加えて，一時的に地理的な局所性を持つローカルワード（temporally-local word）を導入し，それら二種類のローカルワードを合わせて用いることで精度の向上及びカバレッジの向上を実現した．

5.　まとめと今後の方針

本研究では，ユーザの居住地推定に着目し，グラフベース手法とコンテンツベース手法それぞれの課題について取り組んだ．グラフベース手法については，ソーシャルグラフ上において友人同士の居住地は近いという closeness assumption とは異なる concentration assumption を導入し，高い精度を実現した．コンテンツベース手法については，実世界ローカルイベントを用いる居住地推定手法を提案した．さらに，時々刻々とコンテンツが発信されるソーシャルストリームを用いたオンライン居住地推定手法を提案した．提案手法では定常的に地理的な局所性を持つ statically-local word に加えて，一時的に地理的な局所性をもつ temporally-local word を新たに導入し，それらを併用することで高い精度とカバレッジを実現した．

今後の課題としては以下の二つが挙げられる．一つ目はグラフランドマークの応用である．グラフランドマークは地域に密着した有用な情報発信をする傾向にあり，その地域のユーザから多くの関心を集めていた．そのため，新しくその地域に移住したユーザ，その地域を旅行しているユーザなどにグラフランドマークの投稿や，グラフランドマーク自体を推薦すると有益であると考えられる．

二つ目はユーザの他の属性への，本研究の適用である．本研究で提案した三つの推定手法は居住地を推定するものだったが，他の属性についても適用可能であると考えられる．

【文献】

[1] Abrol, S. and Khan, L.: Tweethood: Agglomerative Clustering on Fuzzy k-Closest Friends with Variable Depth for Location Mining, *SocialCom/PASSAT*, pp. 153–160 (2010).

[2] Backstrom, L., Sun, E. and Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity, *WWW*, pp. 61–70 (2010).

[3] Chandra, S., Khan, L. and Muhaya, F. B.: Estimating Twitter User Location Using Social Interactions-A Content Based Approach, *SocialCom/PASSAT*, pp. 838–843 (2011).
[4] Chang, H.-W., Lee, D., Eltaher, M. and Lee, J.: @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage, *ASONAM*, pp. 111–118 (2012).
[5] Cheng, Z., Caverlee, J. and Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users, *CIKM*, pp. 759–768 (2010).
[6] Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P.: A Latent Variable Model for Geographic Lexical Variation, *EMNLP*, pp. 1277–1287 (2010).
[7] Hecht, B., Hong, L., Suh, B. and Chi, E. H.: Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles, *CHI*, pp. 237–246 (2011).
[8] Jurgens, D.: That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships, *ICWSM* (2013).
[9] Kinsella, S., Murdock, V. and O'Hare, N.: "I'm eating a sandwich in Glasgow": modeling locations with tweets, *SMUC*, pp. 61–68 (2011).
[10] Li, R., Wang, S., Deng, H., Wang, R. and Chang, K. C.-C.: Towards social user profiling: unified and discriminative influence model for inferring home locations, *KDD*, pp. 1023–1031 (2012).
[11] Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B. and Rodrigue, J.: A demographic analysis of online sentiment during hurricane irene, *Proceedings of the Second Workshop on Language in Social Media*, Association for Computational Linguistics, pp. 27–36 (2012).
[12] McGee, J., Caverlee, J. A. and Cheng, Z.: A geographic study of tie strength in social media, *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, pp. 2333–2336 (2011).
[13] Paul, M. J. and Dredze, M.: You Are What You Tweet: Analyzing Twitter for Public Health., *ICWSM* (2011).
[14] Rout, D. P., Bontcheva, K., Preotiuc-Pietro, D. and Cohn, T.: Where's @wally?: a classification approach to geolocating users based on their social ties, *HT*, pp. 11–20 (2013).
[15] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, *WWW*, pp. 851–860 (2010).

山口 祐人 Yuto YAMAGUCHI
2014 年筑波大学大学院システム情報工学研究科博士後期課程修了．博士（工学）．日本学術振興会特別研究員（PD）．現在，筑波大学大学院システム情報工学研究科博士研究員，米カーネギーメロン大学客員研究員．データマイニング等の研究に従事．情報処理学会，日本データベース学会各会員．