

マイクロブログのメッセージを用いた発信場所推定

伊川 洋平[†] 榎 美紀[†] 立堀 道昭[†]

[†] 日本アイ・ビー・エム株式会社 東京基礎研究所 〒242-8502 神奈川県大和市下鶴間 1623 番 14 号

E-mail: [†]{yikawa,enomiki,mich}@jp.ibm.com

あらまし 本研究では、マイクロブログ上で発信されるメッセージの内容を元に、あるメッセージが発信された場所を推定することを試みる。マイクロブログは他のソーシャルメディアと比較してリアルタイム性が極めて高いという特徴を持ち、情報が発信された場所を特定できると、場所に関連した情報推薦や、位置情報と結びつけたイベント検出など様々な応用が考えられる。しかしながら、メッセージ発信時に携帯端末の GPS 情報を付与するなど、直接位置情報をメッセージに付与するユーザーは非常に少数であるため、一般的にはメッセージに座標データなどの位置情報が付いていることは期待できない。そこで本研究では、ユーザーの過去のメッセージから位置情報とキーワードの関連付けを行うことで、メッセージの発信場所を推定することを試みた。

キーワード マイクロブログと地理情報、テキスト分析

Location Inference using Microblogging Messages

Yohei IKAWA[†], Miki ENOKI[†], and Michiaki TATSUBORI[†]

[†] IBM Research-Tokyo, IBM Japan Ltd.

1623-14, Shimotsuruma, Yamato, Kanagawa, 242-8502 Japan

E-mail: [†]{yikawa,enomiki,mich}@jp.ibm.com

Abstract In this paper, we attempt to discover geolocation information from microblogging messages. Since microblogging services are more timely compared to other social media, understanding the geolocation information of each microblogging message is useful for location-based services such as recommendation based on user's current location, and event detection associated with geolocation information. Some microblogging services provide a function for adding geolocation information to messages from mobile device equipped with GPS detectors. However, few users use these functionality, so most messages do not have geolocation information. Therefore, we attempt to discover the location where a message was generated by using its textual content.

Key words Microblogging, Geographical Information, Text Analysis

1. はじめに

近年、ソーシャルメディアの発展と普及を背景にマイクロブログの利用者数が増大している。マイクロブログとは、200 字程度の短いメッセージで情報発信が行えるサービスのことで、従来のブログやソーシャルネットワーキングサービスと比較して、手軽に情報発信ができることを特徴としたソーシャルメディアである。マイクロブログの手軽さと、シンプルで使いやすい操作性を提供するスマートフォンの普及とが相まって、自宅やオフィスだけでなく、外出先や移動中など屋外においても携帯端末を用いた情報発信が活発に行われるようになった。このような背景から、マイクロブログは他のソーシャルメディアと比較してリアルタイム性が極めて高いという特徴を持つ。

本研究では、マイクロブログ上で発信されるメッセージの内

容を元に、あるメッセージが発信された場所を推定することを試みる。情報が発信された場所を特定できると、ユーザーの行動予測や、場所に関連したリアルタイム性の高い情報推薦、位置情報と結びつけたイベントや災害情報の検出など様々な応用が考えられる。マイクロブログには、スマートフォンの GPS 機能を利用して現在地の緯度経度の情報を位置情報タグとしてメッセージに付与する機能がある。しかしながら、現在位置の情報をメッセージに直接付与しているユーザーは非常に少数であり、一般的にはメッセージに座標データなどの位置情報が付いていることは期待できない。

そこで本研究では、位置情報タグを用いずに、メッセージのテキスト情報のみを用いて、あるメッセージが発信された場所を推定することを試みた。位置情報サービスからマイクロブログに発信されたメッセージを用いて発信場所を特定し、それを

元に時間的に近接して発信されたメッセージ中の表現と場所との関係性を学習する．そして、新たなメッセージに対して、含まれる表現の類似性が高いメッセージに紐付けられた場所を新メッセージの発信場所として推定する．ユーザー毎に学習を行った場合とユーザーを区別せずに学習を行った場合とで比較実験を行い、前者の方が良好な結果を示すことを確認した．

2. 関連研究

マイクロブログに関する研究は、[3] を始めとして、これまでに多くの研究が行われている．ここでは、その中でも位置情報に関連した研究について述べる．

Cheng ら [1] は、ユーザーが過去に発信したメッセージを分析することで、ユーザーの居住地を都市レベルで推定することを試みている．コンテンツベースで場所推定をしようとしている点で本研究と共通しているが、本研究ではメッセージが発信された時点でのユーザーの居場所を推定しようとしている点で目的が異なっている．

Kinsella ら [4] は、位置情報タグが付与されたメッセージを教師データとしてユーザーの区別をせずに学習を行い、メッセージが発信された時点でのユーザーの現在位置の推定を試みている．本研究と目的は同じだが、位置情報タグを学習に利用している点で本研究と異なっている．

また近年、マイクロブログをソーシャルセンサーと捉えて、災害時に発信されたメッセージを分析する試みが行われている [6], [7]．災害時におけるマイクロブログの活用は本研究における主要な応用例の一つであり、本研究によりメッセージと位置情報が紐付けられることで、災害発生場所の特定等に役立てることができる．

3. マイクロブログと位置情報サービス

本節では、マイクロブログサービスの中でも最も有名なものの一つであり、本稿でも実験データとして利用している Twitter^(注1)と、メッセージからユーザーの現在位置を特定する手掛かりとなる位置情報サービスについて説明する．

3.1 Twitter

Twitter は、1 億人以上のアクティブユーザーを擁する世界最大規模のマイクロブログサービスである．140 字を上限としてメッセージを一般に広く発信できるほか、特定のユーザーに向けてメッセージを発信するリプライ機能や、他のユーザーが発信した情報を再発信するリツイートと呼ばれる機能が備わっている [5]．短い文章で手軽に情報発信ができ、スマートフォンとの親和性も高いことから、従来のブログやソーシャルネットワークサービスと比較して、リアルタイム性が極めて高いという特徴を持っている．

また、Twitter には、携帯端末の GPS 機能を利用して現在の位置の緯度経度を位置情報タグとして付加したメッセージを発信する機能が備わっている．しかし、明示的にこの機能をオンにしなければ位置情報タグの付加は行われず、自分の位置情報

表 1 位置情報サービスから発信されるメッセージ形式

位置情報サービス	メッセージ形式
Foursquare	I'm at [場所] ([住所]) [コメント] (@ [場所])
ロケタッチ	[場所] にタッチ！ [コメント] @ [場所] にタッチ！

が常に公開されることに抵抗を持つユーザーも多いことから、位置情報タグが付加されたメッセージは非常に少ないのが現状である．

3.2 位置情報サービス

位置情報サービスとは、一般的には携帯端末の GPS 機能を利用したユーザーの現在位置に基づくサービスの総称であるが、本稿では、その中でも特に自分の位置情報を外部に発信することを目的としたサービスのことを位置情報サービスと呼ぶ．主要なものとして、Foursquare^(注2)やロケタッチ^(注3)等があり、その中でも Foursquare は、1,000 万人以上のユーザーから利用されているようなポピュラーなサービスである^(注4)．

位置情報サービスの多くは Twitter と連動しており、サービス上で自分の現在位置を通知することによって、その情報が Twitter にも同時に発信されるようになっている．Twitter で常に現在位置の情報が発信されることについては抵抗を持つユーザーが多い一方で、位置情報サービスを利用するユーザーが増加していることから、あるタイミングで今自分が特定の場所にいることを他のユーザーに知らせることについては許容できるユーザーが多いことが伺える．

位置情報サービスから Twitter に発信されたメッセージは、サービスによって特定の形式を持っており、それらの特徴は [2] にまとめられている．ここでは、主要な位置情報サービスのメッセージ形式を表 1 に示す．位置情報サービスからは、このような定型化されたメッセージが発信されるため、メッセージ本文に対して文字列マッチングを行うことで、ユーザーの現在位置を特定することができる．

4. メッセージを用いた発信場所推定

本研究では、マイクロブログに付与されている位置情報タグを使わずに、メッセージのテキスト情報のみを用いて、あるメッセージが発信された場所を推定することを試みる．ユーザーが現在の状況について発信したメッセージには、位置情報サービスから発信されたメッセージなど、特定の場所が記載されており、ユーザーの現在位置が分かるものもあるが、場所が直接記載されていないものも多数存在する．そこで、場所が記載されているメッセージと時間的に近接して発信されたメッセージに含まれる語を、その場所と紐付けることによって、場所が直接記載されていないメッセージの発信場所を推定する．

本稿で検討した以下の推定手法は、学習と場所推定の 2 つのステップに分けて実施される．学習ステップで過去のメッセー

(注2): <https://foursquare.com/>

(注3): <http://tou.ch/>

(注4): <http://blog.foursquare.com/2011/06/20/holysmokes10millionpeople/>

(注1): <http://twitter.com/>

ジから語と場所の関係性を学習し、場所推定ステップで学習結果に基づき新メッセージの場所推定を行う。

4.1 学 習

最初に、過去のメッセージから学習に用いるデータの選定を行う。マイクロブログでは、特定の友人と会話したり、他人が発信した情報を再発信することがよく行われるが、これらは必ずしもその時点での自分の状況を表しているものではない。例えば、あるイベントに参加しているユーザーが発信したメッセージが、別のユーザーによって有益と判断され再発信された場合がこれにあたる。このようなメッセージは、メッセージの内容と場所に関係性がないと考えられるため、学習対象から除外する。Twitter においては、メッセージ本文に対して文字列マッチを行うことでリプライとリツイートのメッセージを特定し、学習対象から除外する。

続いて、上記で選定された学習に用いるメッセージを、ユーザーの現在の位置情報について述べている location、位置情報は直接記載されていないが、ユーザーの現在の状況について述べている expression の 2 つのタイプに分類する。location の判定は、位置情報サービスから発信されたメッセージかどうかを判定することによって行う。各メッセージに対して表 1 に示されているメッセージ形式かどうかをチェックすることで、ほぼ確実に location の判定を行うことができる [2]。また、地名辞書を用いて「地名 + デ格」やマイクロブログに特徴的な「地名 + なう」といったパターンを適用して判定を行うことも可能である。しかし、一般に地名には曖昧性があり、人名など他の固有名詞や一般名詞との区別がつかない場合があるため、位置情報サービスのメッセージ形式かどうかを判定する方法と比較して精度は低くなる。また、location と判定されなかったメッセージを expression とする。

最後に、expression と判定された各メッセージに対して、前後一定時間 t 内に location のメッセージが存在すれば、その場所情報と expression を紐付ける。複数の location が存在する場合は、時間的に近い方をひとつ選択する。この expression のメッセージ本文に対して形態素解析を行い、キーワード（名詞句）のリストを生成して、場所情報と紐付けて学習結果として保持しておく。

以上の手順により、各 expression のキーワードリストに対して、ひとつの場所情報が紐づいているデータが学習結果として生成される。

4.2 場 所 推 定

学習結果を元に、新メッセージに対して場所推定を行う。場所推定を行いたいメッセージに対して形態素解析を行い、キーワードのリストを生成する。このキーワードリストと、学習結果として保持している各 expression のキーワードリストとのコサイン類似度を算出し、最も類似度が高い expression に紐付けられている場所情報を、新メッセージの発信場所の推定結果とする。

5. 実 験

位置情報サービスを利用しているユーザーに対して、4. 節で

述べた発信場所推定の有効性を評価する。ここで比較のために、各ユーザーが過去に最も多くメッセージを発信した場所を推定結果として返すナイーブな手法をベースラインとして採用し、ユーザー毎に学習を行った場合、全ユーザーで学習を行った場合、ベースラインの 3 つの推定手法について、推定精度の評価を行った。なお、実験に使用したデータは、全て Twitter 社が提供する API から取得しており、広く一般に公開されているものである。

5.1 データの準備

評価実験で用いる実験データの準備方法について述べる。最初に、位置情報タグ付きのメッセージを発信しているユーザーを選定する。この手法では位置情報タグを使わずに発信場所推定を行うが、推定場所と実際の発信場所の誤差を評価するために、位置情報タグが付与されたメッセージを発信しているユーザーを対象にして評価を行った。

Twitter Streaming API^(注5)を用いて取得した Spritzer と呼ばれるデータセット（サンプリング率：約 1%）から、位置情報タグ付きのメッセージを発信をしたユーザーを抽出する。これらのユーザーに対して、Twitter REST API^(注6)により過去に発信されたメッセージ（サンプリングなし、最大 3,200 件）を取得する。このメッセージ内容に対して表 1 に示すパターンを適用し、位置情報サービスを利用しているユーザーを抽出して実験対象のユーザーとする。2011 年 11 月 15 日から 11 月 23 日の間に収集したデータセットに対して実験対象のユーザーを抽出したところ、110 ユーザーが抽出された。上記 110 ユーザーが発信したメッセージのうち、現在の自分の状況について述べているものに限定するため、メッセージ本文に対してパターンマッチを行い、リプライ、リツイートを除外し、それ以外のメッセージを対象データとする。

続いて、この対象データを学習データと評価に使用する質問データに分類する。4.1 に示した手順に従って学習を行い、場所情報との紐付けが行われなかった expression のデータのうち位置情報タグが付与されているものを質問データとした。ここで、学習ステップにおいて expression と場所情報を紐付ける時間の閾値 t は 10 分とした。上記データセットに対して学習データと質問データの数それぞれ 20,535 件、12,463 件となった。

このようにして作成した質問データに対して発信場所推定を行い、推定結果と実際の発信場所の緯度経度情報から誤差を算出して評価を行った。

5.2 実験結果と考察

ユーザー毎に学習を行った場合、全ユーザーで学習を行った場合、ベースラインの 3 つの推定手法について、推定結果と実際の発信場所の距離の誤差が 0.5km, 1.0km, 3.0km, 5.0km, 10km, 30km 以内であったものを正解としたときの精度をそれぞれ算出した。誤差と精度の関係を図 1 に示す。

誤差が小さい範囲においては、ユーザー毎に学習を行った場合に高い精度で推定が行えることが分かる。また、誤差を許容

(注5): <https://dev.twitter.com/docs/streaming-api/methods>

(注6): <https://dev.twitter.com/docs/api>

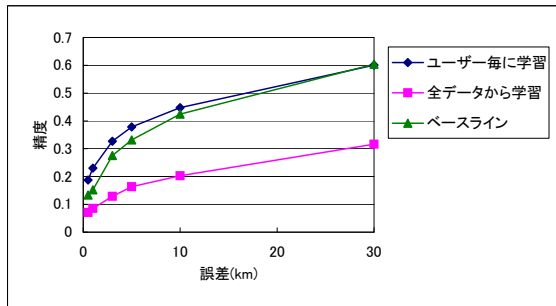


図 1 誤差と精度の推移

するに従ってベースラインの精度が追いついてくるという結果が得られた。ベースラインの手法はユーザーの行動範囲の大きな地域を特定すればよいようなケースでは一定の精度で場所推定が可能だが、より正確な場所推定を行うためには、メッセージを用いた場所推定が有効であることが分かる。また、全データから学習した場合は、ユーザー毎に学習を行った場合と比較して精度が半分程度にとどまることが確認された。

ユーザー毎に学習を行った場合と、全ユーザーから学習を行った場合の再現率の推移を表 2 に示す。ここで再現率とは、質問データ全体のうちその誤差の範囲内で正解した割合を表している。質問データ 20,535 件中、推定を行ったのは、ユーザー毎に学習を行った場合が 7,658 件、全ユーザーから学習した場合は 16,380 件と後者の方が 2 倍以上の質問データに対して推定を行っている。しかし、後者は不正解のケースが多く、ユーザー毎に学習を行った方がわずかではあるが再現率が高いという結果になっている。

全ユーザーから学習を行った場合に、精度が半減してしまう原因として、学習データが不足していることが挙げられるものの、両手法の精度には大きな差があり、これはメッセージの発信場所推定というタスクにおいてユーザー毎に学習を行うことの重要性を示唆しているものと考えられる。例えば「会社で仕事」というメッセージが発信されたとき、推定すべき場所は、そのユーザーにとっての会社の場所だが、全てのユーザーを対象に学習を行うことで、ユーザー固有の特定の場所と結びつきのある表現が正しく学習されなくなってしまう。一方で、ユーザー毎の学習データのみを利用する場合は再現率の向上が難しいことが予想されるため、精度、再現率を共に向上させていくためにはハイブリッドな手法を開発する必要がある。

最後に、ユーザー毎に学習を行った場合の誤差 10km 以内の推定結果を正解数が多いユーザーから順番に並べたものを図 2 に示す。一部のユーザーの場所推定に大きく失敗していることが精度低下の要因になっており、これらのケースに対応してい

表 2 誤差と再現率の推移

学習方法	誤差					
	0.5km	1.0km	3.0km	5.0km	10km	30km
ユーザー毎に学習	0.07	0.09	0.12	0.14	0.17	0.22
全データから学習	0.06	0.07	0.10	0.13	0.16	0.25

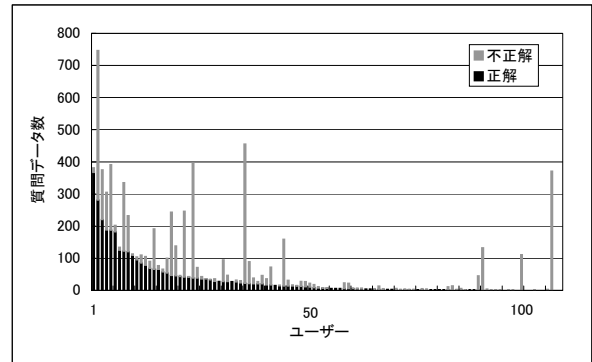


図 2 ユーザー毎に学習を行った場合の誤差 10km 以内の推定結果

くことで今後精度を改善できる可能性がある。また、ユーザーの現在位置についてのコンテキストが全く存在しないようなメッセージについては推定を行わないことも、精度向上のために必要である。

6. ま と め

本稿では、マイクロブログの位置情報タグを用いずにメッセージ本文から発信場所推定を行う方法について検討した。発信場所の特定が容易なメッセージと時間的に近接して発信されたメッセージ中の表現を学習して、新たなメッセージに含まれる表現と類似性の高いものを特定することで場所推定を行った。また、位置情報サービスを利用しているユーザーを対象として、ユーザー毎に学習を行う場合と全ユーザーから学習を行う場合で比較実験を行い、前者の方が精度、再現率ともに良好な結果を示すことを確認した。

今後の課題としては、特定のアプリケーションを想定してその上で実用レベルの精度が出るように、あらゆるアプローチに取り組むことが挙げられる。また、より大規模なデータを用いた実験を行い、ユーザー毎の学習と全データでの学習を組み合わせたハイブリッドな推定手法の開発が必要である。

文 献

- [1] Z. Cheng, J. Caverlee, and K. Lee. "You are where you tweet: A content-based approach to geo-locationg Twitter users". In *Proceedings of CIKM*, 2010.
- [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". In *Proceedings of DICO2011*, 2011.
- [3] A. Java, X. Song, T. Finin, and B. Tseng. "Why We Twitter: Understanding Microblogging Usage and Communities". In *Proceedings of WebKDD/SNAKDD*, 2007.
- [4] S. Kinsella, V. Murdock, and N. O'Hare. "I'm Eating a Sandwich in Glasgow: Modeling Locations with Tweets". In *Proceedings of SMUC*, 2011.
- [5] H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?". In *Proceedings of WWW*, 2010.
- [6] T. Sakaki, M. Okazaki, and Y. Matsuo. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In *Proceedings of WWW*, 2010.
- [7] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness". In *Proceedings of CHI*, 2010.