Vol.2019-DBS-169 No.2
Vol.2019-IFAT-136 No.2
2019/9/10

位置情報付きツイートを用いた観光地周辺の迷いやすいスポットの発見

鈴木 亮平[1,a)]　廣田 雅春[2,b)]　荒木 徹也[3,c)]　遠藤 雅樹[4,d)]　石川 博[1,e)]

概要：観光地や慣れていない土地での徒歩移動において，目的地までの経路探索の手段として，Google Maps などの地図アプリが主流であるが，ウェブサイトやパンフレットにあるルートガイドや，駅前などにある観光案内板を頼りに移動する場合がいまだに多く存在する．その場合，ルートは作成者や閲覧する訪問者の主観に基づき決定され，訪問者にとって本当にわかりやすく，迷いにくいルートとは限らない．本研究では，Twitter 上の位置情報付きツイートを対象として道に迷っている時のツイートを抽出し，入り組んだ交差点，類似したランドマークがある道などの迷いやすいスポットの発見を試みる．具体的には，Word2Vec を用いてツイートを学習し，教師あり学習を用いて道に迷っているときに発信されたと考えられる「迷子ツイート」を抽出する．さらに，投稿場所ごとに抽出結果を分析することで，観光地周辺で迷う人が多い場所を発見することを試みる．

キーワード：Twitter, 位置情報,Word2Vec

1. はじめに

近年，観光地や慣れていない土地へ訪問した時の徒歩移動において，Google Maps[*1]をはじめとした地図アプリの利用及び，地図アプリの経路探索機能を用いて目的地まで移動するのが一般的である．しかし，近距離にある複数の場所を巡る場合やわかりやすいランドマークがある場合などでは，観光案内板や観光パンフレットに載っている周辺地図の使用や，目的地の方向を確認して進むなど，地図アプリを使用せずに感覚を頼りに移動をする場合も多くみられる．どちらの場合でも地図に対して進行方向がわからなくなる，現在地の認識がずれるなどの理由から，道に迷う，想定よりも到着に時間がかかるといった状況に陥ることがある．地図アプリにおいては現在地や向いている方向を表示する機能が搭載されているものもあるが，精度が通信状況に左右されるなど，これらの状況にならないとは言い切れない．

[1] 首都大学東京大学院システムデザイン研究科
[2] 岡山理科大学総合情報学部
[3] 群馬大学理工学府情報科学コース
[4] 職業能力開発総合大学校基盤ものづくり系
[a)] suzuki-ryouhei2@ed.tmu.ac.jp
[b)] hirota@mis.ous.ac.jp
[c)] tetsuya.araki@gunma-u.ac.jp
[d)] endou@uitec.ac.jp
[e)] ishikawa-hiroshi@tmu.ac.jp
[*1] https://www.google.com/maps/

また，スマートフォンやタブレット端末の普及により，マイクロブログなどのソーシャルメディアが普及し，観光地を訪れた多くのユーザーは，その場の雰囲気や感想，実際の体験などを共有する投稿をしている．マイクロブログのひとつである Twitter[*2]では，投稿一つにつき 140 字までという字数制限がついており，時間や場所を問わず，気軽に投稿することができるという特徴がある．ツイートには緯度経度情報から所在地を取得できるジオタグと呼ばれるデータを付与することができる．したがって，Twitter 上にはユーザーの体験由来の地域特有の情報が蓄積されていると捉えることができる．

本研究では，「道に迷う」はユーザーのその地域における体験であると仮定し，位置情報付きツイートから道に迷っている時にツイートされたと考えられるツイート (以下，迷子ツイート) を教師あり機械学習を用いて抽出する．抽出した迷子ツイートの分布から迷いやすいスポットを発見する．迷いやすいスポットを発見することにより，案内図の作成や経路推薦を行う際に，よりユーザーが迷いにくい情報の提供に貢献する．

本論文の構成は次の通りである．2 章では，関連研究について述べる．3 章では，迷子スポット発見のための提案手法について述べる．4 章では，提案手法に基づいた実験と結果を，5 章では，実験によって得られた結果を元に考

[*2] https://twitter.com/

察を行う．6 章では，本研究のまとめを述べる．

2. 関連研究

本章では，関連研究について述べる．観光分野に限らず，マイクロブログに投稿されたデータから有益な情報を抽出する研究 [1–5] が盛んに行われている．観光分野においては，行動パターンを抽出し，観光ルートや経路の推薦に利用する研究や，地域の情報を抽出する研究が行われている．

2.1 マイクロブログの情報抽出に関する研究

亘理ら [1] は，電車の混雑具合に関する情報を含んでいるツイートを抽出することを目的として，“混雑”，“混む” などの混雑ワードや “比較的”，“予想以上” などの比較ワードと駅名からなる駅名ワードを定義し，混雑表現辞書を作成，混雑ワードや比較ワードの自動抽出手法を提案している．従来研究が位置情報を元に混雑状況のリアルタイム取得を目標しているのに対して，本研究では，抽出した位置情報付きツイートを場所ごとに集計・可視化することによって地域属性の取得を目標にしている．

三浦ら [2] は，ユーザーの属性に起因する訪問地の違いに着目して，位置情報付きツイートからユーザーごとの特徴量を作成し，属性を推定，実際にユーザーの男女の推定を行い，場所によってツイートの投稿者に男女の割合に違いが出ることをあげている．

2.2 地域の情報の抽出に関する研究

長谷川ら [6] は，Twitter 上に投稿されたコンテンツの中から，地域の特徴を表す特徴語を抽出し，地域特徴語辞書を構築する手法および，構築された地域特徴語辞書を利用して Twitter からユーザーの観光体験を検索する手法を提案している．本研究では，迷子体験に限定し，汎用的な単語辞書を用いて迷いやすい特徴をもつ地域の発見を提案する．

高木ら [7] はランドマークを目印とした経路推薦システムにおいて，ランドマークを視認性だけでなく，位置ベースソーシャルメディアである Foursquare の情報を用いて話題性をもとに抽出する手法を提案している．ユーザーの移動において有益な情報を抽出するという点で共通しているが，本研究では，地理的情報ではなく，ユーザーの体験をもとに抽出する．

堂前ら [8] は，ツイートの中のトピックには，地域に偏りがあるものと共通で現れるものがあるという仮定のもと，文章の確率的な生成モデルである LDA(Latent Dirichlet Allocation) から得たトピックを利用し，Twitter ユーザの生活に関わる地域を推定している．本研究でも，ツイート中のトピックに地域の特性が現れるとして，メッシュごとに可視化することで迷子スポットの発見に取り組む．

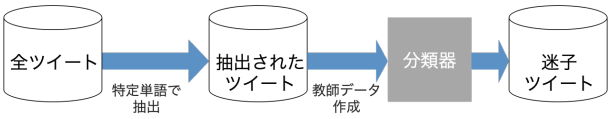

図 1 迷子ツイート抽出の流れ

2.3 ユーザーの行動分析に関する研究

新井ら [9] は，観光スポットでの観光客のツイートを収集し，てがかり語や品詞の特徴から観光ツイートを “食事”，“景観”，“行動”，“土産” に分類し，観光ツイートの時間帯分布，観光ルート内における観光スポットの共起頻度から，観光ルート推薦手法を提案している．投稿内容からツイート内容を推定している点が共通しているが，本研究では，従来研究においてのてがかり語となる，迷子の時につぶやかられる可能性の高い単語を含むツイートを抽出し，教師あり機械学習で分類することにより迷子ツイートの高精度での抽出を提案する．

3. 提案手法

本章では，本研究の提案手法について述べる．はじめに，Twitter ユーザーの投稿した位置情報付きツイートの投稿内容に基づいた特徴ベクトルを作成する．次に，作成した特徴ベクトルと機械学習手法を用いて，迷子ツイートの抽出を行う．続いて，地図をグリッドで区切り，グリッドで区切られたメッシュ内の迷子ツイート数に応じて色分けを行い，可視化する．

3.1 投稿内容に基づいた特徴ベクトルの作成

3.1.1 前処理

本節では，特徴ベクトル作成の前処理について述べる．はじめに，本研究で使用する位置情報付きツイートのうち，Instagram や Foursquare など他のソーシャルメディアと連携しているツイート及びリプライ，リツイートを除去する．次に，投稿内容から，URL 部分や記号，絵文字を取り除く．

3.1.2 特徴ベクトルの作成

本節では，位置情報付きツイートの投稿内容を用いて特徴ベクトルを作成する方法について述べる．

はじめに，形態素解析を行い単語を分割し，基本型に直す処理を行う．その後，品詞判定を行う．ユーザーの発信している情報を適切に特徴ベクトル化するため，名詞，動詞，形容詞及び助動詞と判定された単語を抽出する．続いて，抽出した各単語に対して Word2Vec [10] を用いて単語ベクトルを作成する．ツイート内の各単語ベクトルの和を，そのツイートの特徴ベクトルとする．

3.2 迷子ツイートの抽出

本節では，3.1.2 節で作成したツイートごとの特徴ベクト

ルを用いて教師あり機械学習で迷子ツイートを抽出する方法について述べる．

図 1 に流れを示す．はじめに，迷子ツイートに含まれている可能性が高い単語または単語の組み合わせを人手で決定する．迷子ツイートに含まれる可能性が高い単語とは “迷う” や “迷子” など，単語の組み合わせは「“道” と “わかる” と “ない”」や「“ここ” と “どこ”」などの一般的に道に迷っている時に呟くと考えられる単語である．“わかる”,“ない” に関しては “わからない” が 3.1.2 節の処理により分割，基本型になっていることを想定している．次に，教師あり機械学習を用いて，抽出した単語ごとに分類器を作成し，迷子ツイートを抽出する．本研究では，教師あり機械学習手法として，Support Vector Machine (SVM) と Random Forest を用いた．各分類器において，グリッドサーチを行い各パラメータを決定した．また，Stratified K-Fold 法により，交差検証を行い，F 値の平均を推定精度とした．

3.3 地図上に可視化

本節では，迷子ツイートの地図上への可視化方法について述べる．

ユーザーがどのような場所で迷子ツイートを投稿しているかを可視化するため，地図を一定距離ごとのグリッドで区切り，グリッドで区切られたメッシュ内で投稿されている迷子ツイートの数に応じてメッシュを色分けする．

4. 実験

本章では，実際に Twitter から収集したツイートを用いて，3 章で提案した手法により迷子ツイートを抽出，迷子スポットを可視化する．

4.1 データセット

本節では，本研究で用いたデータセットについて述べる．2016 年 1 月 1 日から 2018 年 12 月 31 日の 3 年間の間に投稿された位置情報付きツイートをランダムに収集した．3.1.1 節の処理の結果，15,298,856 件が収集できた．

4.2 実装

3.1.2 節の形態素解析及び品詞判定には Mecab[*3]を用いた．Mecab の辞書データには新語・固有語表現に強く，語彙数も多い mecab-ipadic-NEologd [11]

Word2Vec の実装には Python のライブラリである gensim[*4]を用い，事前学習には Wikipedia 日本語版[*5]のデータを用いて事前学習している japanese-word2vec-model-builder[*6]にて公開されているデータを使用した．

SVM および Random Forest の実装には scikit-learn [12] の SVC，Random Forest を使用した．

表 1 分類器性能

機械学習手法		条件 1	条件 2
SVM	正答率	0.859	0.750
	F 値	0.857	0.769
Random Forest	正答率	0.846	0.541
	F 値	0.838	0.521

4.3 迷子ツイートの抽出

4.3.1 分類器作成

本節では，3.2 節にもとづく迷子ツイートの抽出結果について述べる．本論文では，迷子ツイートに含まれている可能性が高い単語または単語の組み合わせとして「“迷う” または “迷子”」を含むツイートを条件 1，「“道” と “わかる” と “ない”」を含むツイートを条件 2 としての 2 通りで迷子ツイートの抽出を行なった．

教師データとして，条件 1 では正解不正解ツイートを各 300 件，条件 2 では正解不正解ツイートを各 150 件抽出し，これらの特徴ベクトルを特徴量として，機械学習に入力した．また，$K = 5$ と設定して，正解不正解のラベルのツイート数が等しくなるようにデータを 5 分割し，1 つをテストデータ，残りの 4 つを教師データとして，交差検証を行なった．

4.3.2 分類・抽出結果

SVM, Random Forest で分類を行った結果を**表 1** に示す．正答率と F 値は，Stratified K-Fold を用いて交差検証を行った結果を平均した値を示している．

分類を行った結果，条件 1 では正答率と F 値が，SVM において，それぞれ 0.859，0.838 となり，Random Forest において，それぞれ 0.846, 0.838 となった．条件 2 では，SVM において，それぞれ 0.750，0.769，となり，Random Forest において，それぞれ 0.541，0.521 となった．条件 1 ではわずかではあるが，条件 1，条件 2 双方において，Random Forest よりも SVM が正答率，F 値共に高く，分類性能が高いと考えられるため，迷子ツイートの抽出には SVM の分類器を用いた．条件 1 と比べて，条件 2 の分類性能が低くなってしまっているのは，“道” が “迷う” や “迷子” と比べてより一般的に使われる語であること，教師データが少ないことが原因と考えられる．作成した分類器を用いて抽出を行なった結果，迷子ツイートと思われるツイートが 84,192 件抽出された．

*3 http://taku910.github.io/mecab/
*4 urlhttps://radimrehurek.com/gensim/
*5 https://ja.wikipedia.org/wiki/

*6 https://github.com/shiroyagicorp/japanese-word2vec-model-builder
Copyright (c) 2013-2017 Shiroyagi Corporation. https://shiroyagi.co.jp

Vol.2019-DBS-169 No.2
Vol.2019-IFAT-136 No.2
2019/9/10

表 2　色分けされたグリッド数

投稿数 (件)	メッシュ数	メッシュの色
1〜2	2,797	白
3〜4	59	橙
5〜9	48	緑
10 以上	22	赤

4.4　地図上に可視化

本節では，3.3 に基づいて地図を一定距離四方のグリッドに区切り，メッシュ内での投稿された迷子ツイートの数に応じて色分けを行なった．今回は抽出した迷子ツイートのうち 5,000 件をサンプリングし，グリッドを 250m 間隔とした．色分けされたメッシュ数および，色の内訳を**表 2** に示す．本論文では，メッシュ内に迷子ツイートが確認できたもののうち，迷子ツイート件数が 1〜2 件のものを白，3〜4 件のものを橙，5〜9 件のものを緑，10 件以上のものを赤で色分けを行なった．

5.　考察

本章では，4.4 節の結果をもとに，各地域の迷子ツイートが多かったスポットについて考察を行う．

迷子ツイートの投稿数に応じて地図上に色分けしたものを，**図 2**, **図 3**, **図 4** に示す．

まず，図 2 の東京タワーを含む A で囲んだメッシュに注目する．東京タワー周辺にはアミューズメント施設などが併設されており，展望台や施設などの屋内での移動において迷子ツイートが投稿されている可能性が考えられる．

次に B で囲んだメッシュに注目する．このメッシュは六本木ヒルズを含むメッシュとなっている．六本木ヒルズは森美術館など観光目的で訪れるスポットも含むが，有数のオフィス街でもあり，ビジネス目的の来訪者も多く，迷子スポットである可能性は考えられるが，観光目的の訪問者にとって迷いやすい場所であるかはさらなる検証が必要である．

続いて，C で囲んだ銀座周辺のメッシュ群に注目する．二つのメッシュにまたがって迷子ツイートを確認できる．メッシュ内および近辺は，直結していない地下鉄駅が複数存在するエリアである．地下鉄の乗り換えや，目的地へ向かう際に，複数存在する出口を間違える，自分の利用した駅を他の地下鉄の駅と間違えるなど，複数の迷う理由が推測できる．六本木ヒルズと同じく，観光目的の訪問者にとって迷いやすい場所であるかはさらなる検証が必要である．

次に，図 3 の京都駅を含む A で囲んだメッシュに注目する．メッシュ内はほぼ京都駅構内となっており，施設内部での迷子であることが推測される．京都駅構内は構造が複雑であることが知られており，観光客にとって迷いやすい場所である可能性が高い．

続いて，B で囲んだメッシュに注目する．メッシュ内には駅や名が知られている観光地は確認できない．この場所はあくまで経由地点であると考えられ，さらなる検証が必要であるが，なんらかの要因が存在しており，迷いやすい場所だと考えることができる．

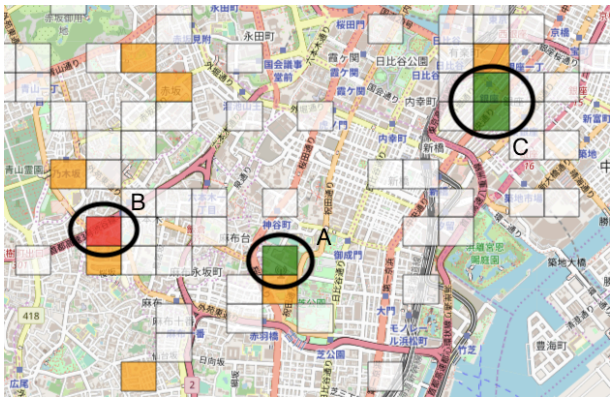

図 2　東京タワー周辺の迷子ツイート投稿数により色分けを行なった可視化結果

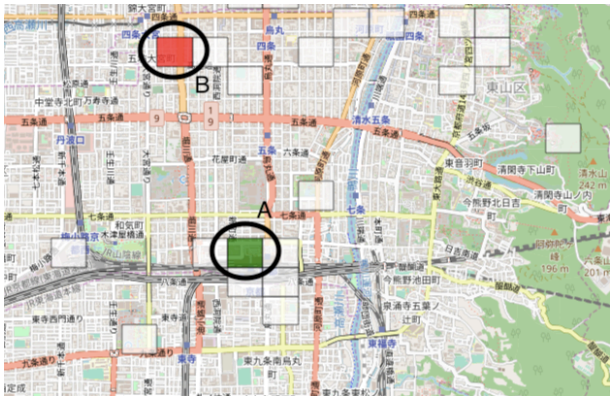

図 3　京都駅周辺の迷子ツイート投稿数により色分けを行なった可視化結果

図 4 の USJ を含む A で囲んだメッシュに注目する．赤い部分は USJ 内部となっており，アトラクションの場所がわからないなど迷っている可能性も考えられるが，混雑に起因する同伴者とはぐれた場合の迷子を多く含んでいる可能性が高い．右上の緑のメッシュについても，駅から USJ までのルートを含んでおり，ほぼ一本道であることから USJ に向かおうとしている最中に迷っているとは考えにくい，しかし，周辺施設は入り組んでおり，迷いやすいスポットである可能性があり，さらなる検証が必要である．

6.　まとめ

本研究では，教師あり機械学習を用いて，迷子の時に投稿したと思われる迷子ツイートを抽出し，道に迷いやすい迷子スポットを発見する手法を提案した．そして，提案手法を用いて，場所ごとに集計，可視化することで，迷子スポットの可能性がある場所を複数発見した．課題として，総投稿数の多いメッシュは迷子になりやすいかどうかに関わらず，一定数の迷子ツイートが抽出されてしまう点がある．迷子ツイートの数だけでなく，総ツイートにおける迷

Vol.2019-DBS-169 No.2
Vol.2019-IFAT-136 No.2
2019/9/10

図 4 USJ 周辺の迷子ツイート投稿数により色分けを行なった可視化結果

子ツイートの割合を考慮してメッシュの分類，色分けを行い，分析する必要があると考える．また，迷子ツイートの抽出に任意の単語を選択する必要があるため，想定していない単語を用いている迷子ツイートを取得できない．単語を任意に選択するのではなく，迷子に関連する語を判定する方法が必要だと考えられる．または，単語に依存せず，全ツイートに対して利用できる分類器の作成の必要がある．

今後の展望として，本研究では形態素解析および品詞判定の段階で日本語以外のツイートを除外していたが，外国語ツイートを対象とすることで訪日外国人にとっての迷いやすいスポットを発見できる可能性がある．また，発見した迷子スポットを考慮した経路推薦システムの構築などがあげられる．

謝辞 本研究は，JSPS 科研費 19K20418 による．

参考文献

[1] 亘理湧, 豊田哲也, 大原剛三. 鉄道の混雑検出センサとして機能する twitter ユーザの推定. 第 79 回全国大会講演論文集, 2017.

[2] 三浦理緒, 廣田雅春, 加藤大受, 荒木徹也, 遠藤雅樹, 石川博. マイクロブログのジオタグを用いた訪問地の違いに着目したユーザ性別推定手法の提案. 第 10 回データ工学と情報マネジメントに関するフォーラム, 2018.

[3] 古賀裕之, 谷口忠大. 潜在トピックに着目した twitter 上のユーザ推薦システムの構築. ヒューマンインタフェースシンポジウム, pp. 867–872, 2010.

[4] 落合涼, 伊與田光宏. 投稿場所に着目したソーシャルメディア上の情報拡散の分析. 第 80 回全国大会講演論文集, 2018.

[5] Tatsuya Fujisaka, Ryong Lee, and Kazutoshi Sumiya. Discovery of user behavior patterns from geo-tagged micro-blogs. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*. ACM, 2010.

[6] 長谷川馨亮, 馬強, 吉川正俊. Twitter からの地域特徴語辞書の構築とその観光情報検索への応用. 第 6 回データ工学と情報マネジメントに関するフォーラム, 2014.

[7] 森永寛紀, 若宮翔子, 谷山友規, 赤木康宏, 小野智司, 河合由起子, 川崎洋. 点と線と面のランドマークによる道に迷いにくいナビゲーション・システムとその評価. 情報処理学会論文誌, Vol. 57, No. 4, pp. 1227–1238, 2016.

[8] 堂前友貴, 関洋平. 地域に偏りのあるトピックを用いた twitter ユーザの生活に関わる地域推定. 研究報告データベースシステム (DBS), 2013.

[9] 新井晃平, 新妻弘崇, 太田学. Twitter を利用した観光ルート推薦の一手法. 第 7 回データ工学と情報マネジメントに関するフォーラム, 2015.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013.

[11] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き用辞書生成システム neologd の運用-文書分類を例にして. 研究報告自然言語処理 (NL), 2016.

[12] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in

python. *Journal of machine learning research*, Vol. 12, No. 10, pp. 2825–2830, 2011.