

マイクロブログにおけるジオタグのクラスタリングを用いた マイナー観光地抽出手法の改良

平久江知樹[†] 早川 智一^{††} 疋田 輝雄^{††}

[†] 明治大学大学院 理工学研究科 〒214-8571 神奈川県川崎市多摩区東三田 1-1-1

^{††} 明治大学 理工学部 〒214-8571 神奈川県川崎市多摩区東三田 1-1-1

E-mail: [†]{hirakue,t_haya,hikita}@cs.meiji.ac.jp

あらまし 現在の観光ガイドブックや観光サイトは有名な観光地の紹介が中心となっているが、有名観光地はリピータにとっては物足りなく、混雑等で十分満足できない可能性がある。そこで本稿では既存のメディアではあまり取り上げられないマイナー観光地の発見と推薦を試みる。提案手法として、マイクロブログの一種である Twitter に着目し、位置情報が付加されたツイートを収集し、ユーザを在住者と観光客に分離する。次に観光客のツイートをクラスタリングし、得られた座標に対して周辺の施設情報をタギングすることにより観光スポットの発見を行う。得られた観光スポットについて2つの手法を用いてメジャー・マイナー判定を行いマイナー観光地の抽出を行う。また、マイナー観光地の抽出例として別府エリアのマイナー観光地の抽出を行った。

キーワード Twitter, マイクロブログ, 情報推薦, 位置情報, 観光地推薦

1. はじめに

1.1 背景

現在 2020 年の東京オリンピックを控え、インバウンド観光客と呼ばれる外国人観光客が 2011 年度の 622 万人から 2016 年度には 2,404 万人と増加傾向にある [1]。また、2017 年 7-9 月期のインバウンド観光客の内 56.3 % がリピーターであり、特に香港や台湾、シンガポールなどの日本に近い東アジア・東南アジアからのインバウンド観光客のリピーター率の高さが目立つ [2]。インバウンド観光客は初来日ではゴールデンルートと呼ばれる東京大阪間の主要な観光地を巡ることが多いが、リピーターであればメジャーな観光地では飽きたらず、地方のマイナー観光地に足を運ぶようになる。このようにインバウンド観光客の多様化が進んでいるのに対し、既存の観光サイトや観光ガイドなどでは主要な観光地の情報が中心であり、マイナーな観光地の情報は少ない。そのため、マイナーな観光地を発見することが課題となっている。

そこで、我々はマイクロブログの一種である Twitter^(注1) に注目した。Twitter にはジオタグと呼ばれる位置情報をタグとしてツイートに付加できるサービスがあり、携帯端末上から Twitter にジオタグを投稿できる。日本のジオタグ付きツイートの割合はツイート全体の 0.5% に過ぎないが [3]、ツイートの総数が多いため、ジオタグ付きツイートの絶対数も統計的に有意な量を収集することができる。Twitter の月間アクティブユーザは 2016 年 6 月時点で 3 億 1000 万人であり、モバイルユーザの割合は 82% である [4]。事前調査では東京駅周辺 50km 圏内において、1 ヶ月に約 75 万件のジオタグ付きツイートを取得できた。また、

新井ら [5] の研究により位置情報ゲームである Foursquare^(注2) や画像共有サービスの一種である Instagram^(注3) といったサービスを利用して、観光客が観光地で位置情報付きツイートをしていることが示されている。

1.2 提案手法概要

本稿では、観光サイトや観光ガイドブックに載っていないマイナー観光地を観光客の位置情報付きツイートより求める手法を提案する。観光客の位置情報付きツイートをクラスタリングによって集約し、求められた座標に対し Google Place API^(注4) を用いてタグ付けを行う。求められた観光地に対し、二つの手法を用いてマイナー観光地の判定を行う。第一の手法として各クラスタの元となったツイートのユニークユーザ数よりマイナー観光地の判定を行う手法を提案する。第二の手法として著名観光サイトである TripAdvisor^(注5) を用いて観光地候補の名称を入力として与え、該当施設の口コミ数からマイナー観光地の判定を行う手法を提案する。

我々の先行研究 [6] においては TripAdvisor を用いた手法を使いマイナー観光地を発見する手法の提案を行ったが、本稿ではマイナー観光地の抽出精度の改善を行う。また、TripAdvisor を用いた手法に加え、ユニークユーザ数を用いてマイナー観光地を抽出する手法を提案する。観光地候補に対し人手でメジャー観光地かマイナー観光地かの判定を行い、両手法の結果と比較をし、得られた比較結果に対して考察を行う。

提案手法は以下の通り行う。(図 1)

(1) 著名観光エリアで投稿されたジオタグ付きツイートを収集する。同様に著名観光エリアでジオタグ付きツイートを投

(注1) : Twitter, <https://twitter.com/>

(注2) : Foursquare, <https://ja.foursquare.com/>

(注3) : Instagram, <https://www.instagram.com/>

(注4) : Google Place API, <https://developers.google.com/places/>

(注5) : TripAdvisor, <https://www.tripadvisor.jp/>

表 1 ツイート収集地点一覧

エリア名	中心地点	中心緯度	中心経度	取得半径
浅草エリア	隅田公園	35.712963	139.804724	1.5 km
鎌倉エリア	鎌倉駅	35.319001	139.550733	4.0 km
	江ノ島駅	35.311046	139.487378	2.5 km
京都エリア	京都駅	34.985656	135.766577	4.0 km
	御霊神社	35.036496	135.761533	4.0 km
奈良エリア	奈良駅	34.68105	135.819375	4.0 km
広島エリア	広島駅	34.396165	132.475433	4.0 km
宮島エリア	厳島神社	34.289942	132.316729	2.0 km
札幌エリア	札幌駅	43.068661	141.350755	4.0 km
	羊ヶ丘展望台	42.999047	141.394512	5.0 km
	藻岩山	43.021799	141.322393	5.0 km
別府エリア	別府駅	33.279688	131.500308	3.0 km
	別府地獄めぐり	33.315795	131.469763	3.0 km
	おさるの湯	33.23537	131.53367	3.0 km
日光エリア	日光駅	36.747448	139.62204	4.0 km
仙台エリア	仙台駅	38.260123	140.882429	4.0 km

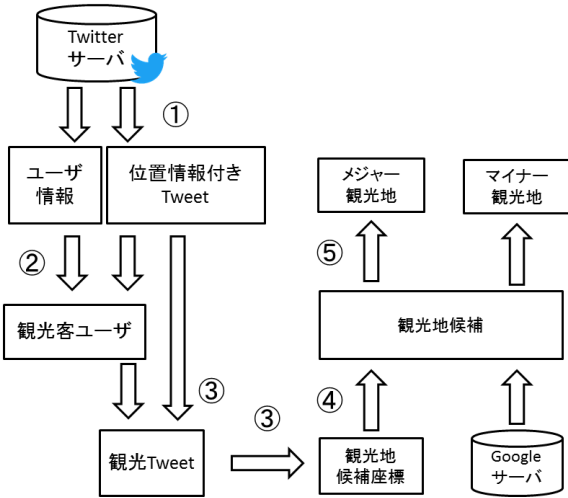


図 1 提案手法のフローチャート

稿したユーザ情報も収集する。(3.1 節)

(2) 収集したツイートとユーザ情報を元に観光客を分離し、観光客をユーザ名と観光エリアのペアとして保存する。(3.2 節)

(3) 各観光エリアごとに、3.2 節で求めた観光客と位置情報付きツイートを元に観光ツイートを抽出する。抽出された観光ツイートをクラスタリングすることにより観光地候補座標を求める。(3.3 節)

(4) 観光地候補座標を入力クエリとして Google Place API を用い、求められた座標と観光地を対応させ、観光地候補を発見する。(3.4 節)

(5) 2 種類の手法を用いて観光地候補をマイナー観光地とメジャー観光地の分離を行う。第一の手法では観光地候補の元となるツイートのユニークユーザ数に注目し、ユニークユーザ数が閾値以上の場合にはメジャー観光地、閾値以下の場合にはマイナー観光地と判断する。第二の手法では Trip Advisor を用い、入力として観光地候補の名称を与え、得られた結果の口コミ数に注目し、口コミ数が閾値以上の場合にはメジャー観光地、閾値以下の場合にはマイナー観光地と判断する。(3.5 節)

1.3 本稿の構成

2 節では位置情報を活用して観光地を発見する研究と観光客と在住者を分類する研究について述べる。3 節では提案手法と実装について述べる。4 節では提案手法により抽出できた観光地例を示す。5 節では提案手法の評価と結果を報告する。6 節では得られた評価結果に対しての考察を述べる。7 節では本稿のまとめと今後の展望について述べる。

2. 関連研究

マイクロブログの位置情報を観光に活かす研究は盛んに行われている。Pat ら [7] は、ソーシャルメディア上のジオタグ付き投稿の本文が位置情報との関連性を持つことに注目し、この性質を利用して検索クエリから関連する単語や地域を抽出する手法を提案した。櫻川ら [8] は、ソーシャルメディアサイトにアップロードされた位置情報付きの写真を用いて、多数の写真が投稿されたスポットをホットスポットと定義し、観光地の発見を

行った。更に写真の撮影者を在住者と観光客に分類し、観光客は在住者に人気の飲食店や娯楽施設を知りたいと仮定し、在住者に注視し、在住者のジオタグをクラスタリングすることによりホットスポットの発見を行った。当手法では在住者と観光客の分類アルゴリズムやクラスタリングアルゴリズムが本稿とは異なり、また本稿では観光客の観光地訪問傾向から観光地のメジャー・マイナーを判定したいため在住者でなく観光客に注視している点で異なる。

佐伯ら [9] は、ツイートの投稿時間や位置情報を分析することで、ユーザの使用言語に依存せず日本国内でツイートを投稿したユーザを日本人、訪日外国人、在日外国人に分類する手法を提案した。また、各グループごとの観光スポットの訪問傾向を分析し、属性ごとに観光スポットの訪問傾向が異なることを確認した。また日本国内でツイートを投稿したユーザを日本人、訪日外国人、在日外国人に分類し、それぞれの観光スポットの訪問傾向を分析した。本稿とはユーザ属性の分類を目的としている点で異なり、観光スポットの発見に関しても訪問傾向を分析するためにユーザが多く訪れたスポットを抽出しているため、マイナー観光地の発見を目的とした本稿とは異なる。

3. 提案手法

3.1 ツイートとユーザの収集

観光客のツイートを収集するにあたり、解析を行う観光エリアを選別した。評価結果を比較できるように、都市の中に観光地が点在しているエリアとして京都・仙台・札幌・広島を、観光地に特化している地域であるエリアとして日光・宮島・別府を、両方の属性を兼ね備えたエリアとして浅草・奈良・鎌倉の全 10 エリアを観光ツイートの収集エリアとして選定した。また、10 エリアに対し、表 1 で示す通り 15 地点の取得半径範囲内で投稿されたジオタグ付きツイート（以下、観光エリアツイートと

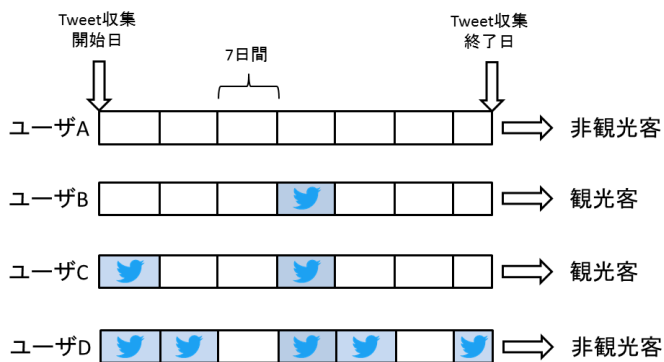


図2 観光客と在住者の分類方法

呼ぶ)を収集した。ツイートの収集はTwitter API^(注6)の内、GETsearch/tweets メソッドを用いて観光エリアツイートを集める。Twitter API を呼び出す言語として Java, ラッパークラスとして Twitter4j^(注7)を用いる。

3.2 観光客ユーザの抽出

前節で取得した観光エリアツイートを元に、取得したユーザから観光客ユーザだけを抽出する。観光客ユーザは観光地とユーザのペアで取得する。これは京都在住のユーザが旅行で浅草を訪れた場合など、ある地方では在住者であるが他の地方では観光客となる場合が考えられるからである。観光客ユーザの判定には、佐伯ら[9]の在日外国人と訪日外国人を分類する手法を改良したものを用い、以下のように処理を行う。

(1) 収集した各ユーザに対し、位置情報の誤差を考慮し、観光エリアツイート数が10に満たないユーザは全観光地において判定対象から除外する。

(2) 各ユーザのロケーション情報を参照し、判定対象の観光地と対になるキーワードが含まれている場合には在住者と判定し判定対象から除外する。例えばあるユーザのロケーション情報が「京都府 京都市」であれば京都在住と判定し京都においては次の処理を行わずに在住者と判定する。

(3) 上記の二条件に該当しない場合、図2の通り、ツイートの取得開始日から終了日までを全解析区間とし、全解析区間を7日ごとに区切る。この時最後の区間に関しては7日以下になった場合でも1つの区間としてみなす。次に各区間でユーザが対象の観光エリアでツイートをを行っているかを確認する。ユーザがツイートをを行っている区間が閾値以下の場合には観光客と判定する。閾値について、解析している観光エリアがツイート数が多い浅草・京都・札幌の場合には2区間、その他の場合には3区間とする。

3.3 観光ツイートによる観光地候補座標の抽出

前節で取得した観光客ユーザのユーザIDと観光地の組に合致する観光エリアツイートを観光ツイートとして抽出する。抽出した観光ツイートに対しクラスタリングを行うことでデータ量を削減しつつ観光地の候補となる座標（以下、観光地候補座

標と呼ぶ)を抽出する。本稿では、クラスタ間距離が最も近いクラスタ同士を階層的に結合する階層型クラスタリングを用い、クラスタ間距離を求める方法としてウォード法を用いた。ウォード法によるクラスタ間距離 Δ は要素数 m のクラスタ M と要素数 n のクラスタ N を結合し、要素数 p のクラスタ P を求める時、クラスタの M, N, P の各要素 i と重心の距離を L_M^i, L_N^i, L_P^i とすると式1により求めることができる。求めたクラスタ間距離 Δ の小さいものから順にクラスタの結合を進める。

$$\Delta = \sum_{i=1}^p (L_P^i)^2 - (L_M^i)^2 - (L_N^i)^2 \quad (1)$$

ウォード法はクラスタリング精度が高くなる反面、計算量が多く処理に時間がかかるという欠点がある。そのためアルゴリズムに改良を加え、処理速度の高速化を図った。クラスタリングを開始する際の各クラスタは原則として1つの観光ツイートあたり1クラスタとしているが、全く同じ座標で投稿された観光ツイートを複数の場合には同一座標で投稿された複数の観光ツイートを1つのクラスタとみなした。また、クラスタリングは最も近いクラスタ間を順次統合するため遠く離れたクラスタ間の距離を求める必要はないと考えられる。このためクラスタリング開始時に全クラスタ間の距離を予め求め、クラスタ間距離リストとして保存した。この際にクラスタリング終了距離の1.5倍以上の距離が離れているクラスタ間距離に関してはリストに追加しないようにした。クラスタの統合時のクラスタ間距離リストに関しても差分だけを修正することにより計算時間の圧縮を行った。以上のアルゴリズムの改良による計算時間の圧縮で、30倍以上の処理時間の向上が見られた。このようにクラスタリング処理を行い、クラスタ間距離が100mに達した時点で処理を終了し、処理終了時に残っていたクラスタを観光地候補座標として保存した。

3.4 観光地候補座標のタグ付けによる観光地候補の発見

抽出した観光地候補座標に対し、Google Place APIを用いてタグ付けを行った。観光地候補座標はそのままの形で緯度経度といった座標情報しか示していないためその位置がどのような観光地であるかを示す必要がある。本稿ではGoogle Place APIを用い、観光地候補座標を入力として与え、観光地候補座標の周辺50mに存在する施設をすべて取得し、施設リストとして保存した。この施設リストの中から最も適切な観光地を以下の手法で観光地候補座標に対して割当を行う。

(1) 各施設に対し、施設スコア S_i を式2によって求める。Google Place APIによって取得した施設情報にはPlace Type^(注8)と呼ばれる施設情報が一施設に対し複数設定されており、ここでは各施設のPlace Typeの数が n 、 i 番目のPlace Typeのスコアを表2に対応させて P_i としている。

(2) 各施設に対し、距離スコア S_d を式3によって求める。この時 P_{lat}, P_{lng} を各施設の緯度、経度、 C_{lat}, C_{lng} を観光地候補座標の緯度、経度とする。

(注6) : Twitter API, <https://developer.twitter.com/>

(注7) : Twitter4j, <http://twitter4j.org/>

(注8) : Place Type, https://developers.google.com/places/supported_types/

表 2 PlaceType に対する得点の対応表

name	score	name	score
amusement_park	20	jewelry_store	5
aquarium	20	library	5
art_gallery	20	lodging	5
casino	20	night_club	5
museum	20	subway_station	5
spa	20	train_station	5
zoo	20	university	5
airport	10	bar	3
campground	10	cafe	3
hindu_temple	10	restaurant	3
mosque	10	convenience_store	1
movie_theater	10	department_store	1
park	10	food	1
place_of_worship	10	funeral_home	1
stadium	10	liquor_store	1
bowling_alley	5	meal_delivery	1
cemetery	5	meal_takeaway	1
church	5	shopping_mall	1

(3) (1)(2) で求めた施設スコア, 距離スコアを用い, 総合スコア S_a を式 4 によって求める.

$$S_t = \sum_{i=1}^n P_i \quad (2)$$

$$S_d = \sqrt{(P_{lat} - C_{lat})^2 + (P_{lng} - C_{lng})^2} \quad (3)$$

$$S_a = S_t - S_d \times 10000 \quad (4)$$

以上により S_a が最も大きくなる施設を観光地候補情報により割り当て, 観光地候補を求めた. S_a が 0 以上となる施設が存在しない場合には対応する観光地が存在しないと判定した.

3.5 観光地候補のメジャー・マイナー分類

前節で求めた観光地候補をメジャー観光地とマイナー観光地への分類を行う. 分類方法としてユニークユーザ数を用いて分類する方法 (以下, ユニークユーザ法とする.) と TripAdvisor を用いて分類する方法 (以下, TripAdvisor を用いた手法とする.) の二手法を提案する.

3.5.1 ユニークユーザ法

ユニークユーザ法では各クラスタのクラスタ元となったツイートを参照し, ツイートを投稿したユニークユーザ数から分類を行う. 判定を行う閾値を D とし, ユニークユーザ数 U が $U \leq D$ の時にはマイナー観光地, $U > D$ の時にはメジャー観光地と判定し分類を行う. D の値は 5 節における評価の適合率と再現率の F 値の値が大きくなるように選択した.

3.5.2 TripAdvisor を用いた手法

TripAdvisor を用いた手法では評価方法として提案手法により求められた観光地候補の名称を TripAdvisor に与え, 該当した施設の口コミ数を求める. 判定を行う閾値を D とし, 該当した施設が存在しない場合や, 該当した施設の口コミ数 U が $U \leq D$ であればマイナー観光地, $U > D$ であればメジャー観

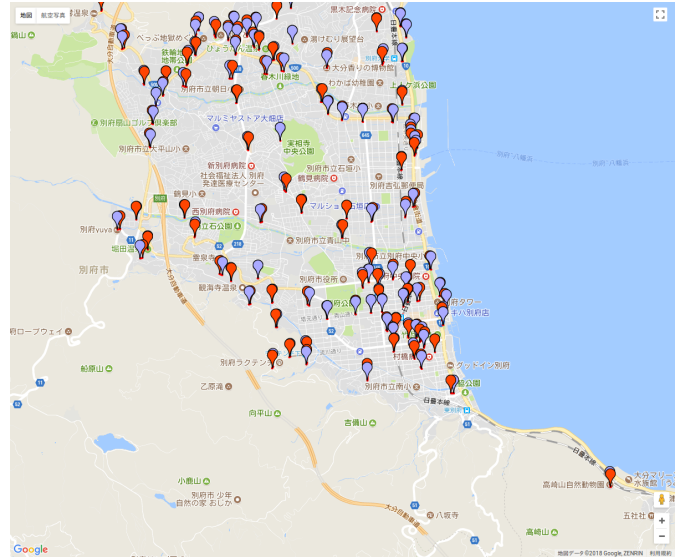


図 3 ユニークユーザ法を用いた別府全体のマイナー観光地抽出例

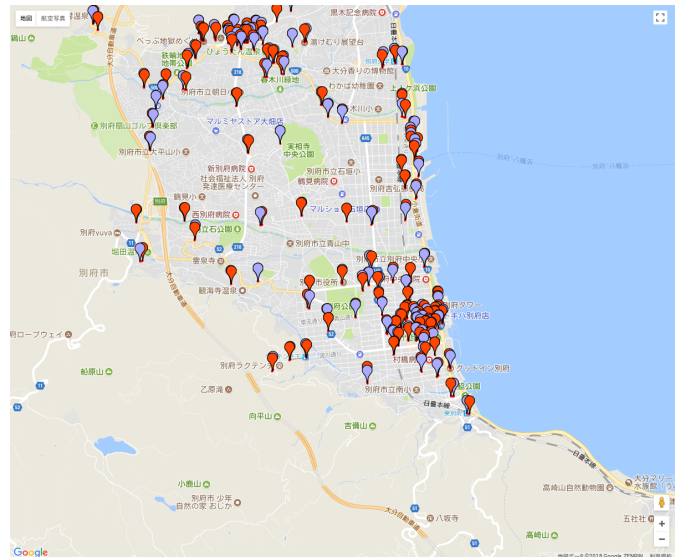


図 4 TripAdvisor を用いた手法による別府全体のマイナー観光地抽出例

光地と判定し分類を行う. D の値は 5 節における評価の適合率と再現率の F 値の値が大きくなるように選択した.

4. マイナー観光地の抽出例

前節の提案手法により抽出できたマイナー観光地の具体例を示す. 初めに提案した 2 つの手法を用いて求めた別府エリア全体のマイナー観光地抽出例を示す. ユニークユーザ法を用い, 閾値 $D = 5$ として抽出を行った際の結果は図 3 の通りとなる. 赤色のマーカーが観光地候補座標, 青色のマーカーが観光地候補を示している. また TripAdvisor を用いた手法により, 閾値 $D = 10$ として抽出を行った際の結果は図 4 の通りとなる. TripAdvisor を用いた手法では市街地に集中してマイナー観光地が抽出されたのに対し, ユニークユーザ法では比較的均一にマイナー観光地の抽出がされたことが分かる.



図 5 マイナー観光地抽出例

別府エリアにおいて抽出できたマイナー観光地の具体例として図 5 の浜脇温泉^(注9)が挙げられる。別府エリアは温泉地として名高いが、その中でも浜脇温泉は深夜まで営業しており、地域住民に密着した温泉として親しまれている。またユニークユーザ数も 4 と少なく、具体的な観光ツイートから浜脇温泉に対する投稿であることが分かる。

5. 評価実験

本章では本稿で提案したマイナー観光地発見手法に対する評価実験を行った。2016 年 11 月 8 日から 2017 年 12 月 8 日までの 396 日間に表 1 のエリア内で投稿された 1,663,563 観光エリアツイートおよび観光エリアツイートの投稿ユーザ 158,385 ユーザを実験として用いた。上記の観光エリアツイートおよび投稿ユーザを用い、提案手法により観光地候補を求める。求められた観光地候補のうち、各観光エリアごとに観光地候補を無作為に 100 箇所抽出を行い、正解データとして人手によりマイナー観光地とメジャー観光地に分類を行った。ただし、観光地候補が 100 箇所に満たない宮島エリアと日光エリアについては全観光地候補である 64 箇所と 95 箇所に対して分類を行った。正解データとユニークユーザ法、TripAdvisor を用いた手法による分類結果と比較し適合率と再現率、F 値を求めた。提案手法で用いる分類閾値をユニークユーザ法では $D = 15$ 、TripAdvisor を用いた手法では $D = 10$ とした時、実験結果は表 3 の通りとなった。

6. 考察

前章によって求められた評価結果より提案手法、観光エリアによる差異について考察を行う。

6.1 提案手法による差異

表 3 より全地域合計の F 値がユニークユーザ法による 80.26

表 3 評価結果

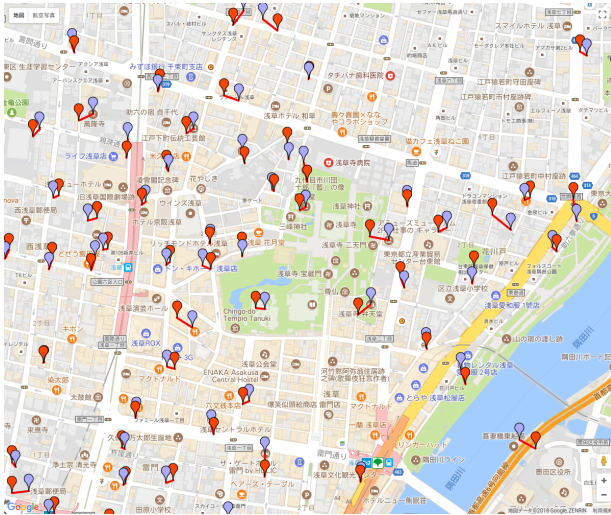
観光エリア	ユニークユーザ法			TripAdvisor を用いた手法		
	適合率	再現率	F 値	適合率	再現率	F 値
浅草エリア	84.85	70.00	76.71	85.53	81.25	83.33
鎌倉エリア	77.92	90.91	83.92	80.77	95.45	87.50
京都エリア	76.62	81.94	79.19	84.81	93.06	88.74
奈良エリア	83.54	84.62	84.08	94.52	88.46	91.39
広島エリア	78.57	75.34	76.92	90.91	95.89	93.33
宮島エリア	68.57	72.73	70.59	75.76	75.76	75.76
札幌エリア	88.24	72.29	79.47	98.46	77.11	86.49
別府エリア	77.11	91.43	83.66	83.08	77.14	80.00
日光エリア	88.41	87.14	87.77	89.06	81.43	85.07
仙台エリア	90.00	64.29	75.00	96.10	88.10	91.93
全地域合計	81.73	78.84	80.26	88.50	85.75	87.11



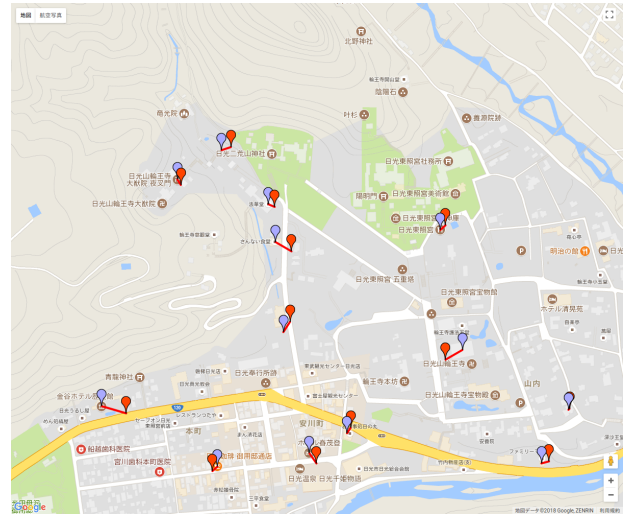
図 6 ユニークユーザ法による割り当て失敗例

に対し、TripAdvisor を用いた手法では 87.11 と TripAdvisor を用いた手法の方が高い結果であることが示された。この理由としてユニークユーザ法ではクラスタリング元となったツイートが必ずしも割り当てた観光地を示しているとは限らないため、誤差が生じるためであると考えられる。具体例として図 6 のケースが挙げられる。この時赤色のマーカーが観光地候補座標、緑色のマーカーが Google Place API により観光地候補座標を入力として得られる周囲の施設情報、青色のマーカーが 3.4 節の提案手法により割り当てられた観光地候補を示している。この場合では cacom と呼ばれるバーが観光地候補として割り当てられているが、観光地候補座標のクラスタリング元となったツイートより多くのユーザが言問橋に関する投稿をしていることが分かる。しかしながら現在の提案手法では Google Place API

(注9)：浜脇温泉, <https://www.city.beppu.oita.jp/sisetu/shieionsen/detail1.html>



(a) 浅草寺周辺のマイナー観光地



(b) 日光東照宮周辺のマイナー観光地

図 7 観光エリアによる観光地集中度の差異

を用いて取得できた周囲の施設情報の中に言問橋は含まれていないため正解を導くことが出来ず、割り当てに失敗していることが分かる。このようにユニークユーザ法では基準となるユニークユーザ数が実際の観光地と乖離しているケースが含まれるため、口コミ数と実際の観光地が一致している TripAdvisor を用いた手法と比較して低い結果になったと考えられる。また、分類閾値をユニークユーザ法では $D = 15$ 、TripAdvisor を用いた手法では $D = 10$ と設定したことにより適切にマイナー観光地の抽出ができたと考えられる。

6.2 観光エリアによる差異

表 3 よりユニークユーザ法を用いた場合の F 値が宮島エリアの 70.59 や浅草エリアの 76.71 と低い値であるのに対し、別府エリアでは 83.66、日光エリアでは 87.77 と高い値を示している。観光エリアごとに評価結果に大きな差異が見られる理由として観光地の集中度合いが異なるためであると考えられる。具体例としてユニークユーザ法を用いて閾値 $D = 10$ とした際の浅草エリアと日光エリアのマイナー観光地は同縮尺で図 7 の通りになった。浅草エリアの場合では小規模な飲食店や観光地が密集しているため、本来複数個に分割すべき観光地に対しても過剰なクラスタリングを行い、一箇所にまとまってしまっているために誤差が生じていると考えられる。一方の日光エリアでは観光地が分散しており、クラスタリングが比較的高精度で行われているために高い評価が得られたと考えられる。

7. おわりに

本稿ではマイクロブログのジオタグ付きツイートより、マイナー観光地を発見する手法を提案した。主要な観光地の周囲で投稿されたジオタグ付きツイートの内観光客のツイートだけを抽出・クラスタリングすることにより観光地候補を発見した。発見した観光地候補をマイナー観光地とメジャー観光地に分類する手法として、ユニークユーザ数を用いた手法と TripAdvisor を用いる手法の 2 種類を提案し、それぞれの F 値は 80.26, 87.11 という結果を出すことができた。現状におい

ては TripAdvisor を用いる手法の方が高い結果となったが、ユニークユーザ法では外部サービスに依存せず、TripAdvisor では網羅していないマイナーな観光地に対しても尺度を図ることができる利点がある。

今後はクラスタリング元となったツイート本文を考慮することによる観光地候補座標に対するタグ付け精度の向上を図ることが考えられる。また、現状においては全観光エリアに対してもクラスタリング終了距離を一律にしているため、都市部などの観光地が密集しているエリアにおいて過剰クラスタリングが発生しており、観光地発見精度の低下を招いている。このためクラスタリングの段階においてエリアごとに柔軟なクラスタリング終了距離を変えることや、k-means 法などの非階層的手法を用いること、クラスタリング距離に現行の物理的な距離に加え、ツイート本文の類似度をレーベンシュタイン距離等により求め、別ベクトルとして加えることが考えられる。

文 献

- [1] 観光白書, 平成 29 年度版, 国土交通省観光庁
- [2] 訪日外国人の消費動向 平成 29 年 7-9 月期報告書, 国土交通省観光庁
- [3] 中澤昌美, 池田和史, 服部元, 小野智弘, “位置情報付きツイートからのイベント検出手法の提案”, 情報処理学会全国大会論文講演集, Vol.74, No.1, pp.503-504, 2012
- [4] Twitter, Inc. について <https://about.twitter.com/ja/company>
- [5] 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", DEIM Forum 2015 G7-6
- [6] 平久江知樹, 早川 智一, 疋田 輝雄, "マイクロブログにおけるジオタグのクラスタリングを用いた観光地発見", 情報処理学会 第 79 回全国大会講演論文集, pp.523-524, 2017
- [7] Barak Pat, Yaron Kanza, Mor Naaman, "Geosocial Search: Finding Places based on Geotagged Social-Media Posts", International World Wide Web Conference Committee, pp.231-234, (2015)
- [8] 櫻川直洋, 廣田雅春, 石川博, 横山昌平, “ジオタグ付き写真の在住者と観光者に分類することによるホットスポットの発見”, DEIM Forum 2015
- [9] 佐伯圭介, 遠藤雅樹, 廣田雅春, 倉田陽平, 横山昌平, 石川博, “外国人 Twitter ユーザの観光訪問先の属性別分析”, DEIM Forum 2015