6K-07

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

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† 明治大学 理工学部‡

1.はじめに

現在，日本に訪れる外国人観光客の観光先が多様化している．従来は東京・大阪間の主要な観光地を巡るルートが主流であったが，現在は地方の穴場観光地を巡る外国人観光客の存在が目立ち始めている．このような外国人観光客の多様化を受けて，主要な観光地だけでなく穴場観光地を発見する方法が必要である．この理由として，第一に，市販の観光ガイドブックでは主要な観光地の情報が中心であり，観光地を網羅していないこと，第二に，主要な観光地はリピータにとっては物足りなく，混雑等で十分に楽しめない可能性があげられる．

そこで，我々はマイクロブログの一種である Twitter 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[1]，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 1,000 万人であり，モバイルユーザの割合は 82%である[2]．事前調査では東京駅周辺 50 ㎞圏内において，1 ヵ月に約 75 万件のジオタグ付きツイートを取得できた．

以上より本稿では，観光ガイドブックに載っていない穴場観光地を観光客によるツイートの位置情報をクラスタリングによって集約し，発見する方法を提案する．

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多く行われている．櫻川ら[3]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，写真の撮影者を在住者と観光客とに分類し，撮影者ごとのホットスポットを発見した．当手法では在住者と観光客の分類アルゴリズムが本稿とは異なる．

佐伯ら[4]は，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．しかしユーザ属性の分類を目的としている点が本稿とは異なる．

3.提案手法

本節では，有名観光地周辺のジオタグ付きツイートから穴場観光地を抽出する提案手法について説明する．提案手法は図 1 の通り，以下の手順で行う．

(1) 主要観光地周辺で投稿されたジオタグ付きツイートを収集する．（3.1 節）

(2) 収集したツイートのユーザが観光客か在住者かを判定する．（3.2 節）

(3) 収集したツイートの内，観光客と判定したユーザのツイートのみをクラスタリングする．（3.3 節）

Discovery of sightseeing spots by clustering geotags of mycroblogs
Tomoki Hirakue, Tomokazu Hayakawa, Teruo Hikita
†Graduate School of Science and Technology, Meiji University
‡School of Science and Technology, Meiji University

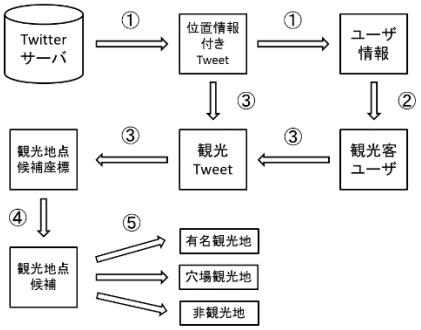

図 1 提案手法のフローチャート
Figure 1 The proposed method.

(4) 求めた座標を Google Place API を用いてタグ付けし，観光地点候補の発見を行う．（3.4 節）

(5) 発見した観光地点候補を非観光地，有名観光地，穴場観光地に分類し，穴場観光地を発見する．（3.5 節）

3.1.ツイートの収集

ツイートの収集には，Twitter API のうち，座標を入力として周囲のジオタグが付加されたツイートを返す GET search/tweets メソッドを用いた．これを用いて，表 1 で示す 10 エリア 15 地点の取得半径範囲内のジオタグ付きツイートを収集した．Twitter API を呼び出す言語として Java，ラッパークラスとして Twitter4j を用いた．収集期間は 2016 年 11 月 8 日から 2016 年 12 月 23 日までの 46 日間である．結果として，256,438 ツイートを取得した．

3.2.観光客判定

観光客の判定には，佐伯ら[4]の手法を改良したものを用いた．判定対象として，収集したユーザの内，収集したツイートの中で 10 以上の投稿が見られるユーザのみを観光客判定対象とした．また，ユーザのロケーション情報中に，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は，在住者と判定し除外した．この 2 つに該当しない場合，ツイートの取得開始日から終了日までの期間を全解析区間とし，全解析区間を 7 日ごとに区切り，各区間でユーザがツイートをしているかを調べる．

全解析区間の内，ユーザがツイートしていた区間が閾値以上の場合には在住者とし，閾値未満ならば観光客と判定する． 観光エリアごとに観光客判定を行い，ユーザ ID と観光エリアとの組で観光客判定を行う．

3.3.観光ツイートによる観光地点候補座標の抽出

収集したツイートの内，前項によって求めたユーザ ID と観光地との組に合致する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用いて，データ量を削減しつつ観光地を推定する．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．ウォード法のクラスタ間距離は結合前と結合前のクラスタの各要素と重心間の距離の二乗和の差で求めることができる．

表 1 ツイート収集地点
Table1 Tweet collection points.

エリア名	中心地点	取得半径	エリア名	中心地点	取得半径	エリア名	中心地点	取得半径
浅草エリア	浅草駅	3km	広島エリア	広島駅	4km	別府エリア	別府駅	3km
鎌倉エリア	鎌倉駅	4km	宮島エリア	厳島神社	2km		別府 地獄めぐり	3km
	江の島駅	2.5km	札幌エリア	札幌駅	4km			
京都エリア	京都駅	4km		羊ヶ丘 展望台	5km		おさるの湯	3km
	御靈神社	4km				日光エリア	日光駅	4km
奈良エリア	奈良駅	4km		藻岩山	5km	仙台エリア	仙台駅	4km

この時，クラスタ間距離が最も近いクラスタから結合してゆき，クラスタ間距離が 100m を超えた際に処理を終了させた．処理終了時に残っていたクラスタを観光地点候補座標として抽出した．

3.4.観光地点候補座標のタグ付け

抽出した観光地点候補座標に対し，Google Place API を用いてタグ付けを行った．Google Place API を用いることで，入力地点の周囲にある施設情報の一覧を取得できる．施設情報の内 Place Type と呼ばれる施設情報属性に予め指定した観光に関するプロパティが含まれる最も近い施設情報（施設の緯度経度，施設名称）を，観光地点候補に対応させた．また，入力した座標の周囲に観光地に関するプロパティが含まれる施設情報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断し，除去した．同様に，対応する施設情報が重複する場合やクラスタリング前のツイート数が閾値以下の観光地点候補も，ノイズとして除去した．

以上により図 2 のように観光地点候補を抽出した．

3.5.観光地点の分類

前項においてタグ付けを行った観光地点候補の中には観光地でない施設や，有名な観光地，穴場な観光地が含まれている．これらを分類するために，Trip-advisor と呼ばれる旅行検索サイトを用いた．前項において得られた観光地点候補の名称を入力として与え，一致する結果が得られた場合には観光地，得られない場合には非観光地と判定した．一致する施設情報が得られた場合には該当施設の口コミ数が閾値以上の場合には有名観光地，それ未満の場合には穴場観光地と判定した．

4.評価

提案手法により抽出した観光地および穴場観光地の評価を行った． 正解データとして，提案手法により抽出した観光地点候補 1,061 ヶ所を人手で非観光地・有名観光地・穴場観光地に分類し，提案手法の結果と比較し，観光エリアごとの適合率，再現率及び F 値を求めた．適合率は提案手法により抽出されたデータの内，正解データが含まれている割合であり，再現率は正解データの内，抽出したデータが含まれている割合である．F 値は適合

図 2 得られた観光地点候補例
Figure 2. Examples of tourist point candidates.

表 2 抽出した観光地の精度
Table2 Accuracy of extracted tourist spots.

	観光地			穴場観光地		
	適合率	再現率	F 値	適合率	再現率	F 値
別府	0.94	0.84	0.89	0.64	0.75	0.69
札幌	0.99	0.67	0.80	0.93	0.46	0.61
仙台	0.96	0.72	0.83	0.88	0.65	0.75
広島	1.00	0.82	0.90	0.94	0.68	0.79
鎌倉	1.00	0.62	0.77	0.97	0.57	0.72
浅草	1.00	0.70	0.83	0.94	0.63	0.75
京都	1.00	0.72	0.84	0.92	0.58	0.71
宮島	1.00	0.81	0.90	0.86	0.50	0.63
奈良	1.00	0.76	0.86	1.00	0.57	0.73
日光	0.75	0.60	0.67	0.00	0.00	0.00
計	0.99	0.71	0.83	0.92	0.57	0.71

率（Precision）と再現率（Recall）の調和平均として式 1 により求められる．

$$F=\frac{2\times Precision\times Recall}{Precision+Recall} \quad (式\ 1)$$

以上により求められた評価結果が表 2 の通りとなる．全地域合計の観光地，穴場観光地の適合率は 0.99， 0.92 と高い一方，再現率は 0.71，0.57 とやや低い結果となった．これは Trip-advisor において個人経営の飲食店など小規模な施設を網羅していないためと考えられる．そのために，小規模な施設が多く集中する都市部の観光エリアでは再現率が低くなる傾向が見られた．また，日光エリアにおいての結果が極端に悪い理由は，日光エリアの観光地点候補の数が 7 ヶ所と少なかったためである．

5.まとめ

本稿では，Twitter の位置情報をクラスタリングすることにより穴場観光地を発見する方法を提案した．主要な観光地域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抽出し，Trip-advisor を用いて穴場観光地を分類した．結果として，抽出した観光地，穴場観光地の適合率をそれぞれ 0.99，0.92 と高い精度で求めることができた．

今後の課題として，Google Place API を用いた観光地情報のタグ付けアルゴリズムの改良が挙げられる．また，観光地点の分類に関して，Trip-advisor では小規模な施設を検出できないため，分類方法の代替手段を探すことが挙げられる．

参考文献

[1] 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
[2] Twitter.Inc.について https://about.twitter.com/ja/company
[3] 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
[4] 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