Vol.2015-DBS-162 No.15
2015/11/26

日本での滞在期間による 長期滞在外国人の東京での訪問先の差異の分析

佐伯 圭介[1,a)]　遠藤 雅樹[1,2,b)]　江原 遥[1,c)]　廣田 雅春[3,d)]　横山 昌平[4,e)]　石川 博[1,f)]

概要：近年，Web 上の情報から，観光情報を抽出し，分析する研究が活発である．ユーザの国籍や，滞在期間などのユーザの属性によって，日本での訪問先や，その感想には差異があることが予想される．また，長期滞在外国人については，日本での滞在期間の長さによっても，抽出される観光情報に差異があることが予想される．そこで，本研究では，ツイートの投稿時間や，付与された位置情報に着目することで，外国語を用いるユーザの属性が短期滞在外国人か，長期滞在外国人かを分類し，長期滞在外国人が日本での滞在を開始した日を推定する手法を提案した．さらに，日本での滞在期間ごとに，長期滞在外国人ユーザの訪問先を分析し，考察を行った．結果として，日本での滞在期間によって，長期滞在外国人の日本での訪問先には差異があることを確認した．

1. はじめに

旅行を計画している人に対して，観光スポットや観光ルートなどの観光に関連する情報（以下，観光情報）を適切に推薦することは，観光客を増加させるために重要である．観光客が求めている情報を正確に予測するため，観光情報を適切に提供するためには，観光客の目的地や旅行ルート，また，そこでの行動や感想を知る必要がある．

それらについて分析するための Web 上の情報源として，SNS（ソーシャル・ネットワーキング・サービス）がある．特に，ユーザ数の多い SNS として，Twitter[*1] があげられる．マイクロブログサービスのひとつである Twitter では，ユーザは，ツイートと呼ばれる短い文章を投稿する．一度に投稿可能な文字数が少なく，携帯端末での利用が容易であることが特徴である．そのため，旅行の際に，行動やその感想をその場で投稿するユーザが多い．また，その際に，一部のユーザは，位置情報をツイートに付与して投稿するため，ユーザの滞在地点を把握することが可能である．

そのため，Twitter などの Web 上の情報を用いて，旅行者の感想や，観光地のイメージなどの観光情報を抽出する研究が盛んである．例えば，Web 上の情報から観光地のイメージを分析する研究 [1], [2] や，旅行者の感想を分析する研究 [3], [4]，位置情報が付与されたツイートや写真から観光ルートを推薦する研究 [5], [6] などがある．その際に，年齢や，居住地などの観光客の属性によって，訪問先や，その感想には差異があることが予想される．例えば，ユーザの居住地や環境によって，関心を示す対象や，観光地での感想などに差異があることが予想される．そのため，より正確な分析を行うためには，観光客の国籍や，滞在期間などの属性を考慮することが重要であると考えられる．そこで筆者らは，使用言語や，短期滞在外国人，長期滞在外国人などの属性ごとに外国人ユーザの観光訪問先を分析する研究 [7], [8] を行った．ここで，短期滞在外国人とは，旅行や仕事で短期的に日本に滞在した外国人である．また，長期滞在外国人とは，仕事などで長期的に日本に滞在している外国人であり，日本に永住している在日外国人を含む．

分析の結果として，位置情報が付与されたツイートを日本で投稿している外国人ユーザについて，それらの属性によって観光情報には差異があることを確認した．さらに，長期滞在外国人については，日本での滞在期間など，時間が経過するにつれて，日本での行動に変化があることが予想される．例えば，日本での滞在を開始してからしばらくは，日本の有名な観光地を訪れるが，時間が経過するにつれて，そのような地点を訪れなくなることがあげられる．時間の経過によって長期滞在外国人の日本での行動に変化がある場合，時間の経過を考慮して長期滞在外国人に推薦

[1] 首都大学東京
[2] 職業能力開発総合大学校
[3] 大分工業高等専門学校
[4] 静岡大学
[a)] saeki-keisuke@ed.tmu.ac.jp
[b)] endou@uitec.ac.jp
[c)] ehara@tmu.ac.jp
[d)] m-hirota@oita-ct.ac.jp
[e)] yokoyama@inf.shizuoka.ac.jp
[f)] ishikawa-hiroshi@tmu.ac.jp
[*1] https://twitter.com/

するべき適切な観光情報を分析する必要がある．そこで，本研究では，主に外国語を用いるユーザを，短期滞在外国人と長期滞在外国人に分類し，長期滞在外国人ユーザについて，日本での滞在開始日を推定する手法を提案する．ここで，本研究では，外国語を日本語以外の言語とする．さらに，来日からの時間の経過によって，長期滞在外国人の訪問先に差異があることを確認するため，長期滞在外国人の東京での訪問先を来日からの時間の経過ごとに分析し，考察を行う．

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，ツイートに付与された位置情報を用いて，外国語を用いるユーザを短期滞在外国人と長期滞在外国人に分類する手法と，長期滞在外国人の滞在開始日を推定する手法について述べる．4 章では，3 章で述べた手法の適切なパラメータの探索と，長期滞在外国人の滞在開始日の推定手法の評価結果について述べる．5 章では，滞在期間ごとの，長期滞在外国人の訪問先の分析結果を示し，考察を述べる．6 章では，本研究のまとめを述べる．

2. 関連研究

Web 上の情報を用いて，観光情報を抽出する研究は，近年盛んに行われている．例えば，Web 上の情報から観光イメージを分析する研究 [1], [2] や，旅行者の感想を分析する研究 [3], [4]，SNS に投稿された情報から観光ルートを推薦する研究 [5], [6] などがある．Choi ら [1] は，Web 上の旅行に関連した情報源から，マカオの観光イメージを分析している．Wenger[4] は，オーストリアを旅行した人のブログを分析し，オーストリアの観光イメージを抽出している．中嶋ら [6] は，観光に関連するツイートを分析し，旅行者の好みに合わせた観光ルートを推薦している．これらの研究では，観光情報を抽出する際に，ユーザの属性を考慮していない．使用言語や滞在期間などの，ユーザの属性を考慮することで，さらに正確な観光情報の抽出が可能になると考えられる．

SNS のデータから，観光情報の分析に応用可能なユーザの属性を推定する研究は盛んに行われている．これらの研究は，以下のように分類することができる．ユーザに対して一意に決まる属性を推定する研究，一意とは限らない属性を推定する研究，および状況によって変化する属性を推定する研究である．

ユーザに対して一意に決まる属性を推定する研究として，以下の研究があげられる．Cheng ら [9] は，ユーザの投稿したツイート本文を用いることで，ユーザの居住地を都市レベルで推定している．また，Burger ら [10] は，ブログの本文とメタデータを組み合わせて用いることで，ブロガーの年齢を推定している．さらに，伊藤ら [11] は，プロフィール文やツイート集合と，会話関係を用いて，性別や年齢などのユーザの属性を推定している．

一意とは限らない属性を推定する研究として，以下の研究があげられる．Pennacchiotti ら [12] は，機械学習とソーシャルグラフを組み合わせることで，Twitter ユーザについて，政治的趣向や民族，特定のビジネスへの親近感を推定している．また，奥谷ら [13] は，ツイート本文やフォロー関係ではなく，ユーザ間でやり取りされるメンション情報を用いることで，Twitter ユーザの所属や興味といった複数の属性に対応したプロフィール情報の推定を行っている．

状況によって変化する属性を推定する研究として，以下の研究があげられる．岩井ら [14] は，ユーザの投稿したツイートの感情極性を計算することで，特定のトピックに対して，ユーザが肯定的か，否定的かを推定している．また，野呂ら [15] は，属性によって記事内容に差が表れるような特徴ベクトルを作成し，作成した特徴ベクトルに基づいて Support Vector Machine を適用することで，ブログ記事ごとに，ブロガーをイベントの主催者と参加者の 2 つの属性に分類している．

本研究は，状況によって変化する属性を推定する研究の 1 つである．これは，時間の経過によって，抽出される情報に差異があることが考えられるためである．本研究では，時間の経過によって抽出される情報に差異があることを確認するため，外国人 Twitter ユーザを短期滞在外国人と長期滞在外国人に分類し，更に長期滞在外国人の滞在開始日を推定する手法を提案する．また，滞在開始日からの時間の経過による訪問先の変化について分析する．

3. 提案手法

本章では，外国語を用いるユーザを短期滞在外国人と長期滞在外国人に分類する手法について述べる．また，この手法によって長期滞在外国人と分類された Twitter ユーザについて，日本での滞在開始日を推定する手法について述べる．

分析の手順を図 1 に示す．はじめに，日本国内でツイートを投稿したすべてのユーザの主な使用言語を判定し，外国人ユーザを抽出する．次に，外国人ユーザに対して，短期滞在外国人か，長期滞在外国人かを分類する手法を適用する．長期滞在外国人であると分類されたユーザについて，ユーザごとに滞在開始日を推定する．

3.1 外国人 Twitter ユーザの抽出

本節では，分析に用いるツイートの取得方法と，外国人ユーザの抽出方法について述べる．

Twitter からのツイートの取得には，Twitter Streaming API[*2] を用いた．その際に，ツイートに付与された位置情報に基づいて，日本国内で投稿されたツイートを取得する

[*2] https://dev.twitter.com/streaming/overview

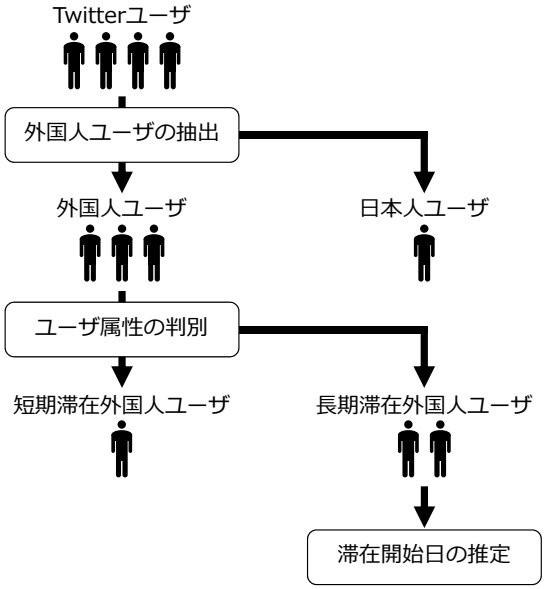

図 1 分析の手順

ように設定した．また，取得したツイート本文の URL 部分は削除した．

取得した位置情報が付与されたツイートを用いて，日本国内でツイートを行ったユーザの主な使用言語の判定を行う．本研究では，ユーザが Twitter のプロフィール情報に登録している使用言語情報と，ツイートの本文に基づいて使用言語を判定する．プロフィール情報に加えて，ツイートの本文を用いるのは，ほかのサービスを経由して投稿したツイートには，自動生成された文字列がツイートに挿入されることや，ツイートの文字数の少なさによって，言語判定が有効に機能しない場合を考慮したためである．ツイートの言語判定は， Nakatani の Language-Detection[*3] をそれぞれのツイートに対して適用し，それぞれのツイートの言語の判定結果を集計し，最もツイート数の多い言語を算出する．その言語がユーザの全ツイートの半数以上を占め，かつユーザの使用言語情報が一致していれば，そのユーザの使用言語とする．使用言語が日本語以外であると判定されたユーザを外国人ユーザとする．

本研究では，外国人ユーザを短期滞在外国人と長期滞在外国人に分類する 2 つの手法を提案する．1 つは，外国人ユーザが日本国内でツイートを投稿した期間が一時的な滞在かを推定することで分類する手法である．もう 1 つは，分析する期間内の日本国内でのツイートの割合に基づいて分類する手法である．

3.2 投稿期間に基づいた分類手法（分類手法 1）

分類手法 1 では，ユーザが日本国内でツイートを投稿した期間が，一時的かどうかを推定することによって，外国人ユーザを短期滞在外国人と長期滞在外国人に分類する．

[*3] https://github.com/shuyo/language-detection

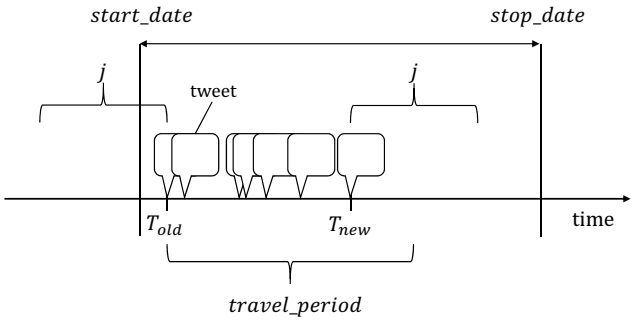

図 2 分類手法 1

分類手法 1 の概要を図 2 に示す．まず，ユーザが日本国内で投稿したツイートを投稿時間の古い順に並び替えて，i 番目に投稿されたツイートを t_i とする．また，分析する期間の開始日 $start\_date$ と終了日 $stop\_date$ を設定する．次に，$start\_date$ から $stop\_date$ までの期間で，ユーザが投稿した投稿時間の最も古いツイートの投稿時間 T_{old} を選択する．短期滞在外国人の日本での滞在期間の最長日数 $travel\_period$ を設定し，T_{old} から $travel\_period$ 日後までの期間の中で，投稿時間の最も新しいツイートの投稿時間 T_{new} を選択する．日本に到着する前であること，帰国後であることを判定するための日数 j を設定し，T_{old} の j 日前までの期間と，T_{new} の j 日後までの期間において，ユーザがツイートを投稿したかどうかを判定する．これらの期間にて，ツイートが投稿されておらず，T_{old} から T_{new} までの期間に投稿されたツイート数が，T_{min} 以上である場合，T_{old} から T_{new} の期間を短期滞在外国人の一時滞在期間とする．ここで，T_{min} は，ユーザの投稿するツイート数が少ないことによる，誤判定を防ぐために設定する最低ツイート数である．そして，T_{new} 以降のツイートに対して，同様の処理を繰り返す．

分析期間内のすべてのツイートが一時滞在期間における投稿であると判定された場合，ユーザは短期滞在外国人ユーザであると判定する．また，短期滞在外国人ではないと判定されたユーザの中で，ツイート数が T_{min} 以上であったユーザを長期滞在外国人ユーザであると判定する．短期滞在外国人ユーザ，長期滞在外国人ユーザのどちらにも判定されなかったユーザは，分類不能ユーザであるとする．これまでの研究で確認した結果，この手法は短期滞在外国人と長期滞在外国人の約 9 割をそれぞれ正しく分類することができる．

3.3 期間ごとのツイートの有無に基づいた分類手法（分類手法 2）

分類手法 2 では，期間ごとの日本国内でのツイートの有無に着目して，外国人ユーザを短期滞在外国人と長期滞在外国人に分類する．

分類手法 2 のイメージを図 3 に示す．分析する期間の開

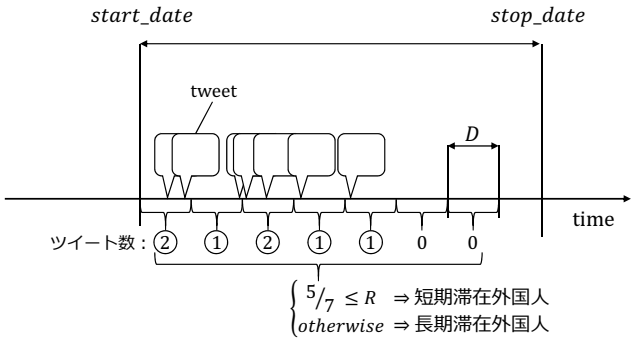

図 3 分類手法 2

始日 *start_date* と終了日 *stop_date* を設定し，その期間のツイートを取得する．次に，*start_date* から，*stop_date* までの期間を D 日ごとに分割することで，D 日間のツイートを 1 つのブロックとする．分割した結果，最後のブロックの長さが D 日よりも短くなった場合は，最後のブロックは用いない．分割されたそれぞれのブロックの期間内に，ユーザがツイートを投稿していたかどうかを確認する．

ユーザがツイートを投稿していたブロックの数から，ユーザが投稿していたブロックの分析期間全体に対する比率 r を求める．ここで，ユーザがツイートを投稿していたブロックの分析期間全体に対する比率 r が，閾値 R 以下であった場合は，ユーザは短期滞在外国人ユーザであると判定する．また，比率 r が，閾値 R よりも大きい場合は，長期滞在外国人ユーザであると判定する．これまでの研究で確認した結果，この手法は分類手法 1 と同様に，短期滞在外国人と長期滞在外国人の約 9 割をそれぞれ正しく分類することができる．

3.4 長期滞在外国人ユーザの滞在開始日の推定

ユーザの属性の分類手法を適用し，長期滞在外国人と分類されたユーザについて，日本国内での滞在開始日を推定する．

はじめに，ユーザごとに，ユーザの属性を分類する際に設定した，分析する期間の開始日 *start_date* から，分析する期間の終了日 *stop_date* までの期間に，日本国内で投稿されたツイートを取得する．次に，ユーザが分析する期間に日本国内で投稿した最も古いツイートの投稿日 D_{old} を取得する．また，分析する期間に日本国内で投稿されたツイートの総数 T_{all} を計算する．*start_date* から，*period* 日間の間に日本国内でツイートを投稿しておらず，日本国内で投稿されたツイートの総数 T_{all} が，閾値 *threshold* よりも高いユーザを *start_date* 以降に日本での滞在を開始したユーザであるとする．この時，日本国内で投稿された投稿時間の最も古いツイートの投稿日 D_{old} を，長期滞在外国人ユーザの滞在開始日であるとする．また，*period* は，分析期間より以前から日本国内に滞在していたかどうかを判定するための期間である．

表 1 ユーザの属性の分類手法のパラメータの値

分類手法 1		分類手法 2	
travel_period	14, 15, ..., 33	D	1, 2, ..., 30
j	3, 4, ..., 50	R	0.05, 0.1, 0.2, ..., 0.7
T_{min}	1, 2, ..., 10	—	

4. 評価実験

本章では，Twitter から取得したツイートを用いて，提案手法の評価を行う．はじめに，ユーザの属性の分類手法と，長期滞在外国人の滞在開始日の推定手法の適切なパラメータを探索する．次に，長期滞在外国人の滞在開始日の推定手法の評価を行う．

4.1 ユーザの属性の分類手法の適切なパラメータの探索

はじめに，分類手法の適切なパラメータを探索するために作成した正解データについて述べる．正解データを作成するため，Twitter から 2014 年 7 月 1 日から 2015 年 6 月 30 日の期間内に日本国内で投稿された 137,692,381 件のツイートを取得した．取得したツイートの中から，ユーザのツイート本文や，Twitter のプロフィール情報をもとに，主に英語を使用するユーザのなかから，短期滞在外国人ユーザと，長期滞在外国人ユーザを 100 人ずつ，合計 200 人を人手で抽出した．1 ユーザあたりの平均のツイート数は，短期滞在外国人ユーザが約 32 件，長期滞在外国人ユーザが約 186 件である．

次に，2 つの分類手法の適切なパラメータを探索する．外国人ユーザを短期滞在外国人と長期滞在外国人に分類する期間の開始日 *start_date* と終了日 *stop_date* は，ツイートを収集した期間と同様の 2014 年 7 月 1 日，2015 年 6 月 30 日とした．そのほかのパラメータについては，複数の値を選択し，表 1 に示すそれぞれの分類手法のすべてのパラメータの組み合わせについて，ユーザの属性の分類を行う．分類手法の性能が最も高くなった組み合わせを提案手法の適切なパラメータとする．

適切なパラメータを決定するための基準とした分類手法の性能の算出方法について述べる．提案手法の分類結果のうち，短期滞在外国人かどうかの分類結果について，F 値を算出する．ここで，F 値とは，適合率と再現率の調和平均である．また，長期滞在外国人ユーザかどうかの分類結果についても，同様に F 値を算出する．これら 2 つの F 値の平均値を，提案手法全体の性能とする．

表 1 のパラメータのそれぞれの組み合わせで，短期滞在外国人ユーザ 100 人，長期滞在外国人ユーザ 100 人の合計 200 人に対して分類手法 1 を適用した結果，最も性能の値が高くなるパラメータの組み合わせが複数存在した．最も性能の値の高いパラメータの組み合わせを表 2 に，またこの時の性能を表 3 に，実際の分類結果を表 4 に示す．表

表 2　分類手法 1 において最も性能が良くなったパラメータの一覧

$travel\_period$	j	T_{min}
26	29, 30, 31, 32	2
	24, 25, ..., 32	3
27	29, 30, 31, 32	2
	24, 25, ..., 32	3
28	30, 31, 32	2
	30, 31, 32	3
29	30, 31, 32	3
30	30, 31, 32	3
31	30, 31, 32	3

表 3　F 値：分類手法 1

短期滞在外国人	長期滞在外国人	全体
0.995	0.995	0.995

表 4　分類結果：分類手法 1

		分類結果		
		短期滞在	長期滞在	分類不能
正解データ	短期滞在	0.990	0.010	0.000
	長期滞在	0.000	1.000	0.000

表 5　F 値：分類手法 2

			F 値		
順位	D	R	短期滞在外国人	長期滞在外国人	全体
1	18	0.1	0.990	0.990	0.990
2	10	0.1	0.985	0.985	0.985
3	11	0.1	0.985	0.985	0.985
4	26	0.2	0.985	0.985	0.985
5	25	0.2	0.980	0.980	0.980
6	8	0.1	0.980	0.980	0.980
7	9	0.1	0.980	0.980	0.980
8	15	0.1	0.980	0.980	0.980
9	16	0.1	0.980	0.980	0.980
10	12	0.1	0.975	0.975	0.975

表 6　分類結果：分類手法 2

		分類結果	
		短期滞在	長期滞在
正解データ	短期滞在	0.980	0.020
	長期滞在	0.000	1.000

4 の短期滞在，長期滞在は，それぞれ短期滞在外国人と長期滞在外国人を指す．表 2 より，各パラメータについて，$travel\_period$ と j を 30 前後，T_{min} を 2 または 3 と設定した際に，分類手法 1 は性能が高くなる傾向にある．

表 1 のパラメータのそれぞれの組み合わせで分類手法 2 を適用した結果，性能の高い上位 10 件を表 5 に示す．分類手法 2 を用いて，ユーザの判定を行った結果，最も性能が高かったのは，分析期間を分割する際の日数 D を 18，ユーザがツイートを投稿していた期間の分析期間全体に対する比率 R を 0.1 と設定した場合であった．この時の実際の分類結果を表 6 に示す．表 6 の短期滞在，長期滞在は，それぞれ短期滞在外国人と長期滞在外国人を指す．表 5 より，分類手法 2 のパラメータの傾向として，D よりも R の値が性能に大きく影響し，R を小さめに設定すると，分類手法 2 の性能が高くなる傾向にある．

表 7　滞在開始日の推定手法のパラメータの値

$period$	5, 6, ..., 30
$threshold$	1, 2, ..., 15

2 つの分類手法を適用した結果を比較すると，どちらの手法においても，高い性能で分類ができていたが，分類手法 1 の方が，分類の性能が高かった．よって，本研究では，分類手法 1 を用いて長期滞在外国人を分類することとする．また，分類手法 1 に適用するパラメータは，$travel\_period$ を 29，j を 31，T_{min} を 3 と設定する．

4.2　長期滞在外国人の滞在開始日の推定手法の適切なパラメータの探索

分類手法 1 を用いて，長期滞在外国人に分類されたユーザが，2014 年 7 月 1 日から 2015 年 6 月 30 日の期間に投稿したツイートを用いて，長期滞在外国人ユーザの滞在開始日の推定手法の適切なパラメータを探索する．2014 年 10 月 1 日以降に日本での滞在を開始したユーザを推定するため，分析する期間の開始日 $start\_date$ を 2014 年 10 月 1 日，分析する期間の終了日 $stop\_date$ を 2015 年 6 月 30 日と設定する．分析期間より以前から日本国内に滞在していたかどうかを判定するための期間 $period$ と，ユーザが日本国内で投稿したツイートの総数の閾値 $threshold$ については複数の値を選択し，表 7 に示すパラメータの組み合わせについて，2014 年 10 月 1 日以降に日本で滞在を開始したユーザの推定を行う．推定結果の性能が最も高くなった組み合わせを推定手法の適切なパラメータであるとする．

適切なパラメータを決定するための基準とした，性能の算出方法について述べる．2014 年 7 月 1 日から，分析する期間の開始日 $start\_date$ の前日である 2014 年 9 月 30 日までに日本国内でツイートを投稿している，2014 年 10 月 1 日以前から日本に滞在していた 5,009 人のユーザは，不正解ユーザとする．また，2014 年 10 月 1 日以降にのみ，日本国内でツイートを投稿している 3,236 人のユーザは，2014 年 10 月 1 日以降に滞在を開始した，正解ユーザとする．提案手法の推定結果と比較することで，F 値を求め，これを提案手法の性能とする．

表 7 のパラメータのそれぞれの組み合わせで提案手法を適用した結果，F 値の高い上位 10 件を表 8 に示す．提案手法を用いて，ユーザの判定を行った結果，最も F 値が高かったのは，分析期間より以前から日本国内に滞在していたかどうかを判定するための期間 $period$ を 21，ユーザが日本国内で投稿したツイートの総数の閾値 $threshold$ を 3 と設定した場合であった．また，$threshold$ を 3 とし，$period$ を 20 前後に設定した際に，提案手法の性能が良く

表 8　F 値：滞在開始日の推定手法

順位	*period*	*threshold*	適合率	再現率	F 値
1	21	3	0.693	0.858	0.766
2	20	3	0.689	0.863	0.766
3	22	3	0.695	0.853	0.766
4	14	3	0.664	0.905	0.766
5	17	3	0.673	0.888	0.766
6	18	3	0.678	0.879	0.765
7	13	3	0.660	0.911	0.765
8	19	3	0.684	0.868	0.765
9	15	3	0.666	0.898	0.765
10	16	3	0.669	0.893	0.765

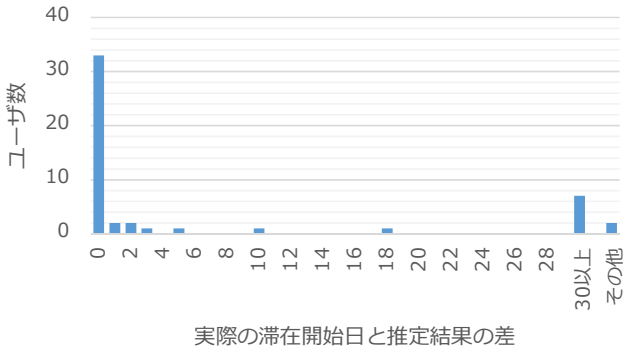

図 4　正解データと提案手法によって推定されたユーザの日本での滞在開始日の差

なることが分かった．本研究では，提案手法に適用するパラメータを *period* を 21，*threshold* を 3 と設定して，滞在開始日を推定する．

4.3　長期滞在外国人の滞在開始日の推定結果の評価

探索した適切なパラメータを用いて，日本での滞在開始日を推定し，人手で作成した正解データと比較を行うことで，提案手法の性能の評価を行う．

分類手法 1 によって分類された長期滞在外国人ユーザに対して提案手法を適用し，日本での滞在開始日が推定されたユーザから，50 人のユーザを無作為に選択する．選択されたユーザが投稿したすべてのツイートを確認することで，日本での滞在開始日を推定し，正解データとする．人手で作成された正解データと，提案手法によって推定されたユーザの日本での滞在開始日の差を求めた．正解データと推定結果の差を，図 4 に示す．

図 4 より，人手で作成した日本での滞在開始日と，提案手法を用いて推定した日本での滞在開始日が等しくなったユーザは，50 人中 33 人であった．ユーザは，日本に到着した際に，到着したことを知らせるため，位置情報を付与してツイートを行うことが多いため，多くのユーザについては，位置情報が付与されたツイートを用いて，滞在開始日を正しく推定することができた．また，図 4 より，人手で作成した日本での滞在開始日と，提案手法を用いて推定した日本での滞在開始日の差が 30 日以上であるユーザは 50 人中 7 人であった．これらのユーザを確認したところ，普段は位置情報を付与せずにツイートを投稿しているユーザであった．このようなユーザについては，位置情報を付与したツイートを投稿することが非常にまれであるため，位置情報が付与されたツイートのみを用いて滞在開始日の推定を行う提案手法では，正しく日本での滞在開始日が推定できなかったと考えられる．また，図 4 で「その他」と示されている 2 名のユーザは，日本でユーザ登録を行い，Twitter を始めた長期滞在外国人であった．提案手法では，日本でツイートを行う前のユーザの行動を考慮していないため，滞在開始日が推定できなかったことが原因である．これらのことより，位置情報が付与されたツイート数が多いユーザについては，滞在開始日の推定に提案手法は有効であると考えられる．

5.　滞在期間ごとの長期滞在外国人の訪問先の比較

本章では，長期滞在外国人ユーザの訪問先を日本に滞在している期間ごとに分析し，考察を行う．本研究では，東京都内の著名なエリアを訪れたユーザ数を分析する．

はじめに，2014 年 7 月 1 日から 2015 年 6 月 30 日の間に，日本での滞在を開始した長期滞在外国人ユーザの投稿したツイートをユーザの滞在開始日に従い滞在 1 日目から 180 日目まで 30 日ごとに割り当てる．日本での滞在を開始した長期滞在外国人ユーザの人数は 5,476 人で，ツイート数の合計は，361,796 件である．

次に，長期滞在外国人ユーザが投稿したツイートに付与された位置情報を用いて，日本での滞在を開始してからの時間の経過ごとに，長期滞在外国人ユーザの中で，東京都内の著名なエリアを訪れたユーザ数を求める．分析に用いる東京都内のエリアには，国別外国人旅行者行動特性調査 [*4] に記載されている，日本を訪れた外国人である訪日外国人の訪問先を調査する際に，東京都が設定した 32 エリアを用いる．分析に用いる東京都のエリア名を表 9 に示す．東京都のエリア名について，Geocoding.jp[*5] を用いてエリアの中心地の緯度・経度を取得し，その地点をエリアの中心とした．エリアの中心から 1 辺 1km の正方形のエリアの中でツイートを投稿したユーザをツイートに付与された位置情報を用いて抽出した．

図 5 に，2014 年 7 月 1 日から 2015 年 6 月 30 日の間に新たに日本での滞在を開始した長期滞在外国人が東京都内で訪れた上位の 10 エリアを滞在開始日からの時間の経過ごとに示す．各エリアを訪れた割合は，滞在開始日からの時

[*4] 東京都 平成 25 年度国別外国人旅行者行動特性調査：http://www.metro.tokyo.jp/INET/CHOUSA/2014/09/DATA/60o99102.pdf
[*5] http://www.geocoding.jp/

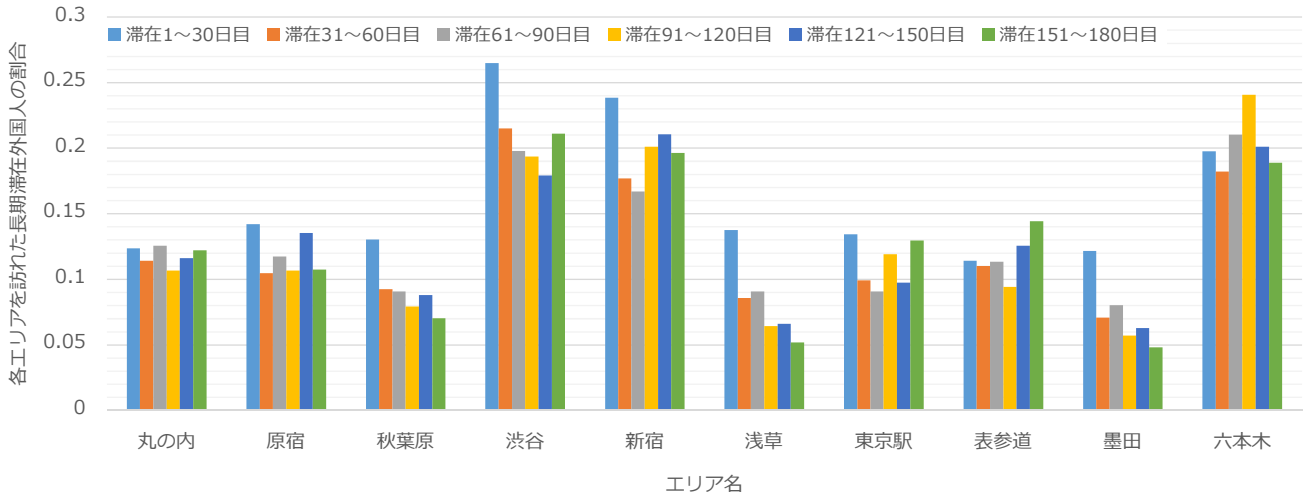

図 5 滞在開始日からの時間の経過ごとの各エリアを訪れた長期滞在外国人の割合

表 9 分析に用いるエリア名の一覧

新宿	秋葉原	六本木	墨田
大久保	上野	赤坂	両国
銀座	原宿	池袋	吉祥寺
浅草	表参道	品川	三鷹
渋谷	青山	築地	八王子
東京駅周辺	新橋	汐留	高尾山
丸の内	お台場	恵比寿	伊豆諸島
日本橋	東京湾	代官山	小笠原諸島

表 10 東京都内の著名なエリアを 1 度以上訪れたユーザ数

滞在開始日からの経過日数	ユーザ数
1～30	1,513
31～60	735
61～90	485
91～120	403
121～150	318
151～180	270

間の経過ごとに，東京都内の著名なエリアを 1 度でも訪れたユーザを全体のユーザ数とし，対象エリアを訪れたユーザの人数を全体のユーザ数で割ることで算出している．滞在開始日からの時間の経過ごとに，東京都内の著名なエリアを 1 度でも訪れたユーザ数を表 10 に示す．

図 5 より，来日から時間が経過するにつれて，訪れる長期滞在外国人の割合が変化するエリアが存在した．例えば，浅草や秋葉原は，長期滞在外国人が来日した直後に訪れる割合が高く，来日から時間が経過するにつれて，訪れる長期滞在外国人の割合が少なくなった．浅草や秋葉原は，日本の歴史や文化を感じることのできる場所として有名であり，多くの訪日外国人が訪れる傾向にある [*4]．来日から時間が浅い長期滞在外国人は，短期滞在外国人と同様に，浅草や秋葉原を訪れるが，1 度目の訪問以降に再び訪れることがないためであると考えられる．また，渋谷については，浅草や秋葉原などに比べると，時間の経過による割合の減少が少ない．これは，長期滞在外国人が来日した直後に訪れる割合が高いことに加え，渋谷にはスクランブル交差点などの，短期滞在外国人に人気の観光スポットも存在するが，渋谷駅が交通の要所であるため，移動の際に渋谷を訪れる長期滞在外国人が多いためであると考えられる．

また，図 5 より，原宿や丸の内など，来日から時間が経過した場合にも，訪れる長期滞在外国人の割合がほとんど変化しないエリアが存在した．原宿は，ショッピングセンターなどの商業施設の多いエリアである．そのため，買い物を目的として原宿を訪れる長期滞在外国人が多く存在するために，来日から時間が経過しても，訪れる長期滞在外国人の割合がほとんど変化しなかったと考えられる．同様に，丸の内についても，割合の変化はほとんど見られなかった．これについて，丸の内は，オフィス街として発展しているため，エリア内にオフィスがある企業に勤めている人など，ツイートを投稿する長期滞在外国人が固定化している可能性が考えられる．

短期滞在外国人の東京での訪問先と長期滞在外国人の東京での訪問先を日本での滞在を開始してからの時間の経過ごとに比較することで，時間の経過によって訪問先が変化することを確認するため，それらの相関係数を求める．相関の指標は，スピアマンの順位相関係数を用いる．この指標は，2 つのランキングの相関係数が 1 に近いほど，2 つのランキングに正の相関がある．短期滞在外国人と日本での滞在を開始してからの時間の経過ごとの長期滞在外国人の東京での訪問先の順位相関係数を表 11 に示す．表 11 より，短期滞在外国人と長期滞在外国人の東京での訪問先には正の相関があった．そのなかでも，滞在 1 日目から 30 日目の長期滞在外国人の訪問先が，短期滞在外国人の訪問先との相関が最も高く，来日からの時間が経過するにつれて，相関が低くなる．これは，前述したように，来日してから時間が経過していない期間は，短期滞在外国人と同様

表 11　短期滞在外国人と長期滞在外国人の東京での訪問先の順位相関係数

滞在開始日からの経過日数	順位相関係数
1〜30	0.946
31〜60	0.868
61〜90	0.822
91〜120	0.855
121〜150	0.827
151〜180	0.744

に，浅草などの日本の有名な観光地を訪れるが，時間の経過につれて，そのようなエリアに訪問しなくなることが原因であると考えられる．

6. おわりに

本研究では，外国人 Twitter ユーザを短期滞在外国人と長期滞在外国人の属性に分類し，長期滞在外国人ユーザについて，日本での滞在開始日を推定する手法を提案した．また，来日からの時間が経過すると，長期滞在外国人の訪問先に差異が生まれることを確認するため，長期滞在外国人の東京での訪問先を来日からの時間の経過ごとに分析し，考察を行った．その結果，来日から時間が経過するにつれて，訪れる長期滞在外国人の割合が変化する浅草や秋葉原などのエリアと，来日から時間が経過した場合にも，訪れる長期滞在外国人の割合が変化しない原宿や丸の内などのエリアが存在することを確認した．また，短期滞在外国人の東京での訪問先と比較することで，来日直後については，長期滞在外国人と短期滞在外国人の訪問先の傾向は類似していることを確認した．さらに，来日から時間が経過するにつれて，長期滞在外国人と短期滞在外国人の訪問先の傾向には差異が生じることを確認した．これらことから，滞在期間を考慮した，観光情報の抽出は，有用であると考えられる．

今後の課題として，位置情報を付与してツイートを投稿することが少ないユーザの滞在開始日の推定結果の改善があげられる．例えば，提案手法では，日本国内で投稿された位置情報が付与されたツイートのみに基づいて長期滞在外国人，短期滞在外国人の分類や，滞在開始日の推定を行っているが，日本国外でのツイートや，位置情報が付与されていないツイートを考慮することによって，分類や推定の結果がより正確になると考えられる．また，提案手法による分類結果の他の手法への応用があげられる．例えば，観光ルートの推薦について，短期滞在外国人や，来日してから日が浅い外国人には，移動時間よりも分かりやすさを優先した観光ルートを推薦する手法や，観光スポットの推薦について，長期滞在外国人の滞在期間ごとに，ユーザの評価の高い観光スポットを推薦する手法などを検討している．

謝辞　本研究（の一部）は傾斜的研究費（全学分）学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」による．

参考文献

[1] S. Choi, X. Y. Lehto, A. M. Morrison, “Destination image representation on the web: Content analysis of Macau travel related websites”, Tourism Management, Vol. 28, pp. 118-129 (2007)
[2] S. Stepchenkova, A. M. Morrison, “The destination image of Russia: From the online induced perspective”, Tourism Management, Vol. 27, pp. 943-956 (2006)
[3] 村上 嘉代子，川村 秀憲：外国人から見た日本旅行 –英語ブログからの観光イメージ抽出–，人工知能学会誌，Vol. 26，No. 3，pp. 286-293 (2011)
[4] A. Wenger, “Analysis of travel bloggers’ characteristics and their communication about Austria as a tourism destination”, Journal of Vacation Marketing, Vol. 14, No. 2, pp. 169-176 (2008)
[5] 青山 賢，廣田 雅春，石川 博，横山 昌平：写真に付与されたジオタグに基づいた道草発見，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，E4-2 (2014)
[6] 中嶋 勇人，新妻 弘崇，太田 学：位置情報付きツイートを利用した観光ルート推薦，情報処理学会研究報告，Vol. 2013-DBS-158，No. 28，pp. 1-6 (2013)
[7] 佐伯 圭介，遠藤 雅樹，廣田 雅春，倉田 陽平，石川 博：Twitter データを利用した訪日外国人の訪問先の言語別分析，観光情報学会誌「観光と情報」，Vol.11，No.1，pp. 45-56 (2015)
[8] 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，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，C4-3 (2015)
[9] Z. Cheng, J. Caverlee, K. Lee, “You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users”, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759-768 (2010)
[10] J. D. Burger, J. C. Henderson, “An Exploration of Observable Features Related to Blogger Age”, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 15-20 (2006)
[11] 伊藤 淳，西田 京介，星出 高秀，戸田 浩之，内山 匡：Twitter と Blog の共通ユーザおよび会話ユーザの同類性に着目した Twitter ユーザ属性推定，日本データベース学会論文誌，Vol.12，No.1，pp. 31-36 (2013)
[12] M. Pennacchiotti, A. Popescu, “Democrats, republicans and starbucks afficionados: user classification in twitter”, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 430-438 (2011)
[13] 奥谷 貴志，山名 早人：メンション情報を利用した Twitter ユーザプロフィール推定，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，B2-3 (2014)
[14] 岩井 宏道，道満 恵介，井手 一郎，出口 大輔，村瀬 洋：マイクロブログへの投稿に基づく政治家の立場推定，人工知能学会全国大会論文集，Vol.28，pp. 1-4 (2014)
[15] 野呂 勇太，廣田 雅春，野澤 浩樹，横山 昌平：記事単位でブロガーの立場を推定する手法の提案，第 7 回 Web とデータベースに関するフォーラム (2014)