2J3-1

位置情報付きツイートによる観光行動のモデル化

Modeling of tourists' behavior using Geo-tagged tweets

劉 萌傑[*1] 吉田 孝志[*2] 和泉 潔[*1] 山田 健太[*1]
Mengjie LIU Takashi YOSHIDA Kiyoshi IZUMI Kenta YAMADA

[*1] 東京大学大学院工学系研究科 Graduate School of Engineering, The University of Tokyo
[*2] 日本電気株式会社 情報・ナレッジ研究所 Knowledge Discovery Research Laboratories, NEC Corporation

Recently, it becomes possible to analyze tourist behavior with the geo-tagged SNS data. In this study, we propose a method to figure out micro migrating tendency and the movement pattern of the tourist using sequential pattern mining from geo-tagged tweets. This method enables to grasp the tourist's behavior in a specific area by analyzing the movement patterns. An actual analysis was performed for Odaiba area and we compared the result with mobile spatial data for verification.

1. はじめに

2020 年東京オリンピックに向けて，訪日観光客の増加が予想される．このような背景の下で，観光ビジネスの機会損失を減らし、観光客の満足度を高めるために，どうように観光客の行動データを取得し分析するかが課題となっている．

観光客行動・動態の分析において，従来の研究では，データを取得する方法から，おおよそ二つのパターンがある．一つはアンケート票による紙面調査や，IC 乗車券の利用情報，もしくは宿泊施設の統計からマクロな視点で分析を行うパターンであり，もう一つは観光客に GPS が付いている機器を携帯してもらうことにより，観光客の位置情報を取得し，分析するパターンである．

しかし，その二つのアプローチについて，それぞれ課題が存在する．マクロな統計データを利用した研究では，観光客のおおよその行動を把握できるが、旅先でどういう路線で遊覧しているかはブラックボックスになっている（特に歩行の場合）．一方，GPS 機器を利用した行動分析は，ミクロなデータを入手できるが，機器のコスト・人的なコストが高いので，データの質と量が制限される．また，GPS データから抽出できる情報は軌跡だけなので，観光客の状態と関心の抽出は難しい．

近年，ユーザの位置情報が付いたソーシャルメディアが登場した．その中に，最も使われているツイッターは，日常的な行動や旅先での行動など，幅広い場面で，位置情報とともにテキストが投稿される．位置情報付きツイートにはジオタグとテキスト本文両方が付いていて，そのデータを解析することにより，観光客のミクロな行動を抽出することが行えるようだけでなく，活動内容および観覧状態も観測可能になった．

観光行動分析の領域において、位置情報付きツイートの利用はすでに始まっている．[桐村 13]では，位置情報付きツイートからユーザ行動の基本的な特徴，および京都市の観光客特徴を抽出した．位置情報付きツイートからは，観光行動を把握できることが示された．しかし，桐村の研究は拠点の利用傾向および二点間の移動傾向に止まって，回遊行動の抽出は行われなかった．

観光客の回遊行動は，順序のある一連の出来事であるので，シークエンスとして見なすのは可能である．そういうシークエンスを抽出するには，系列パターンマイニングという手法が有効であると考える．本研究では，位置情報付きのツイートデータから，系列パターンマイニング手法を用い，ミクロな移動傾向および観光客の移動パターンを抽出する手法を提案する．提案手法から抽出した移動パターンを分析することにより，特定エリアにおける観光客の観覧傾向と観覧経路を把握することが可能であると考える．

連絡先：東京大学大学院工学系研究科システム創成学専攻
〒135-0064 東京都文京区本郷 7-3-1
E-mail：moeketsulyu@gmail.com

2. 提案手法

2.1 本手法の概要

観光客がある場所で位置情報付きツイートを呟いたのは，単にそのエリアを通過するだけでなく，そのエリアに関心があることを意味しているので，それらのツイートから抽出した一連の位置情報は，ユーザの観覧順番を反映していると考えられる．

抽出した位置情報から回遊行動の抽出を行いたいと考え，そこで系列パターンマイニングを用いた．系列（シーケンシャル）パターンマイニングとは，全データセット中のアイテム系列に対し，最小支持度以上の割合で含まれるような系列を抽出する手法である．すなわち，アイテムの発生順序を保った頻出系列を抽出する（図 1）．この方法は，購買行動解析やアクセスログ解析など，アイテム間の順序が重要な場面でよく使われている．例えば[早川 06]では，この手法を用い，ユーザ毎のクリックストリームから Web ページ閲覧者の行動を分析した．

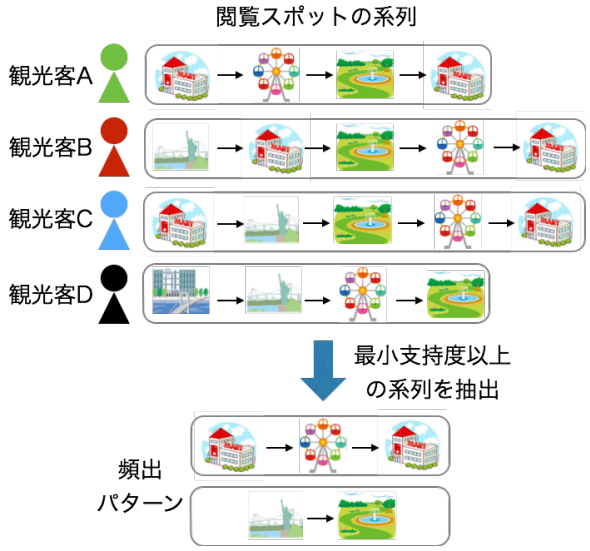

図 1: 本研究における系列パターンマイニングのイメージ

本研究では、対象地域をいくつかのエリアに分け，ツイートが付いているジオタグが対応しているエリア番号をアイテムとして扱う．もしある人が一日中同じエリアで連続ツイートした場合は，一個のみ残した．そうすると，一日中一人の観光客が行った一連の観光スポット（ショッピングセンター・商店街・遊ぶ施設・見学施設・パーク）が系列になり，系列パターンマイニングが適用できる．

2.2 対象エリア・データ

本研究では，お台場エリアの観光客を研究対象にした．お台場エリアには，住宅・会社が少なく，観光エリアとして認識されているので，観光活動以外の理由でお台場に来る場合は少ないと思われる．また，一般的な遊園地・商店街と違って，スポットの間にはある程度の距離があり，観光客の動きをはっきり観測することが可能である．

今回用いたジオタグツイートデータには，収集された 2013 年 1 月から 2015 年 12 月までお台場エリアで投稿されたものである．データに含まれるツイート数は 148,922 件，ユーザ数は 49,248 件であった．

2.3 エリア分け

アイテムとしてのエリアについて，分け方が異なると，結果が大きく変わるので，その境界線をどうやって決めるかは重要な課題となる．ここで二種類の手法を提案する．

一つ目は，先行研究[桐村 13]と同様に，標準地域メッシュ（250m）でエリアを区切る方法である．そうすると，人流の傾向を見ることができ，他のデータソースとの比較もできるようになった（図 2）．しかし，メッシュで区切る場合は，1 個のエリアには複数のスポットが入る，また 1 個のスポットが複数のエリアを跨いでしまうなど，観光客はどのスポットに行ったのは分からなくなって，頻出系列を得たとしても，意味のある結論が得られにくい．

図 2: 地域コードメッシュによる分け方

そこで，もう一つの手法としては，各スポットそれぞれを一つのアイテムのようにエリアを分けることである（図 3）．オフィシャルのパンフレットに載っている各スポットを，アイテムとして扱い，スポットの輪廓を境界線とする．この方法で抽出した場合，カバーされていないエリアがあるため一部のサンプルが除外されたが，観光客の行き先は明確になった．

本研究では，まずメッシュ地域コードメッシュによる分け方で提案手法の妥当性を検証した上で，スポットを独立するような分け方で意味のあるパターンを抽出する．

3. 妥当性検証

3.1 位置情報付きツイートの信頼性

まずは，抽出されたツイートは観光客実際の行動を反映できるかどうかを検証する．ここで，本研究の分析結果を比較するために，モバイル空間統計データを利用した．モバイル空間統計とは，携帯電話ネットワークの運用データを用いて作成された人口統計である[寺田 12]．その統計では，各時間帯・エリアごとに携帯電話数の期待値に基づいて在圏数という指標を算出し，算出した在圏数からさらにその時点の人口を推定する．

今回は一日（2014 年 8 月 16 日）の空間統計データおよびツイートデータから，各エリアにおける人数シェアを算出した．表 1 から見ると，ツイートの数が比較的に多いエリア（例えばエリア 22）は，スポットが含まれるエリアである一方，空間統計による人口の数が比較的に多いエリア（例えばエリア 11）は，駅・主要交通路が含まれる場合が多い．これは空間統計とツイートが持っている意味が異なると考えられる．前述のようにツイートは観光客の関心を意味しているが，空間統計の場合はそうでなく，ただの通過することを意味しているので，そのエリアにスポットがない場合，人数のシェアが低くなる．従って，ツイートだからこそ，観光客の興味・関心を抽出することができると言える．

表 1:ツイートの数と推定人口の照合

エリア番号	ツイート	空間統計	エリア番号	ツイート	空間統計
1	0.36%	1.88%	14	3.74%	7.58%
2	0.38%	3.60%	15	7.01%	4.35%
3	2.57%	6.13%	16	0.80%	1.55%
4	1.78%	5.50%	17	7.08%	8.77%
5	7.24%	4.60%	18	13.99%	7.07%
6	1.10%	1.67%	19	5.75%	3.78%
7	1.89%	0.83%	20	4.33%	5.62%
8	5.02%	5.15%	21	8.33%	3.60%
9	1.09%	2.50%	22	7.27%	1.06%
10	1.06%	4.80%	23	0.56%	0.99%
11	0.30%	3.16%	24	5.00%	4.79%
12	8.86%	2.92%	25	2.11%	3.93%
13	2.36%	4.16%			

3.2 提案手法の妥当性

次に，手法の妥当性を検証する．地域コードメッシュによる分け方を用い，前章で述べた提案手法で実験を行った．各エリアの放出力・吸収力を掴めるために，長さが 2 のパターンに注目し、数式（1）と（2）により特定のエリア x から他のエリアに行くパターンの支持度の総和 S_x と，他のエリアからこの特定のエリアに行くパターンの支持度の総和 E_x をそれぞれ算出した．

$$S_x = \sum_{i=1}^{n} Support(< \{x\}, \{i\} >) \quad (1)$$

$$E_x = \sum_{i=1}^{n} Support(< \{i\}, \{x\} >) \quad (2)$$

お台場エリアに行った人は，半日以上遊ぶ場合が多いので，放出力が高いエリアは昼間に来る人が多い一方，吸収力が高いエリアなら夕方以降の人数が多いという仮説を立った．ここで，S_x と E_x の比を算出した．また，空間統計データから，昼間・夕方以降それぞれに対して，各エリアのシェア D_x と N_x を算出し，さらその比も算出した．($P10_x$: 10 時のエリア x での人数)

$$D_x = \frac{P10_x + P13_x}{\sum_1^n (P10_i + P13_i)} \quad (3)$$

$$N_x = \frac{P17_x + P20_x}{\sum_1^n (P17_i + P20_i)} \quad (4)$$

前述したように，一部のエリアにおいて，ツイートの分布と空間統計の分布が異なる場合があるので，それらのエリアで比較しても意味がない．ここで，スポットが含まれるエリアのみを選んで比較を行った．表 2 では，各エリアにおいて，D_x/N_x および S_x/E_x の結果をまとめた．

表 2: 空間統計を用いる検証

エリア番号	D_x/N_x	S_x/E_x
3	1.13	1.5
4	1.14	1.13
5	1.08	1
8	1.16	1.2
13	0.97	0.84
14	1.02	1.06
15	0.90	1.25
17	1.01	0.88
18	0.90	0.91
19	0.89	0.62
20	0.91	0.79
22	0.82	1.25
24	0.92	0.61

ほとんどのエリアにおいて，放出力と吸収力の比と，昼間人数と夕方以降人数比は近いので，仮説が検証された．二つの比が最も近いエリア 8 とエリア 18 を詳しく見てみると．エリア 8 の大部分は「日本科学未来館」である．その日は「～カガクの星をつなげよう！！～」という大規模科学イベントが開催された．このイベントの開催時間は 10 時から 17 時なので，昼間にそのエリアにいた人は，夕方以降は別のエリアに移動したと考えられる．一方，エリア 18 にあるショッピングセンター「Divercity」には、その日の夜「SUMMER BOMB」というライブがあった．その開催時が遅いなので，どこかに寄ってからライブに来る可能性が高いと思われる．

4. 観光経路の抽出

4.1 全体的な実験

前章では，ソースデータ及び提案手法は有効であることを明らかにしたので，ここで観光スポットを独立するようなエリア分けを行い，提案手法から観光客の行動パターンを抽出した．(表 3)

表 3: 支持度が上位のパターン

2-Len		3-Len	
Sequence	Support	Sequence	Support
<{8},{4}>	0.0083	<{9},{4},{9}>	0.0013
<{9},{4}>	0.0076	<{4},{9},{4}>	0.0010
<{4},{1}>	0.0076	<{4},{8},{4}>	0.0010
<{1},{4}>	0.0068	<{4},{7},{4}>	0.0007
<{4},{8}>	0.0062	<{8},{4},{8}>	0.0006

長さが 2 のパターンから，観光客がお台場の中心位置にある「Divercity」を中心に活動をしていることが分かった．また，長さが 3 のパターンを見てみると，お台場エリアに長い時間を過ごす観光客は，往復の場合が多く，どこかに行ってから先のところに戻る傾向があることが判明した．

4.2 平日と休日の違い

次に，曜日によって観光客の観光経路がどう変わったかを調べた．全部のツイートデータを平日・休日二つのグループに分け，提案手法を行い，それぞれの支持度を算出した．また，各ルートの平日の支持度と休日の支持度の比を算出した．

表 4: 平日・休日の人気ルート

休日の人気ルート		平日の人気ルート	
Sequence	平日/休日	Sequence	平日/休日
<{9},{4},{9}>	0.08	<{6},{1}>	4.74
<{9},{7}>	0.09	<{7},{15}>	3.08
<{9},{9}>	0.10	<{12},{5}>	2.61
<{7},{9}>	0.10	<{12},{8},{4}>	2.61
<{4},{9},{4}>	0.14	<{14},{12},{14}>	2.53

表 4 から，休日の台場エリアは，観光客が「イベントエリア」，「日本科学未来館」，「Divercity」というイベントが多いエリアに集中し，各イベントを中心に遊覧する傾向がある一方，平日には各スポットに分散することが明らかになった．

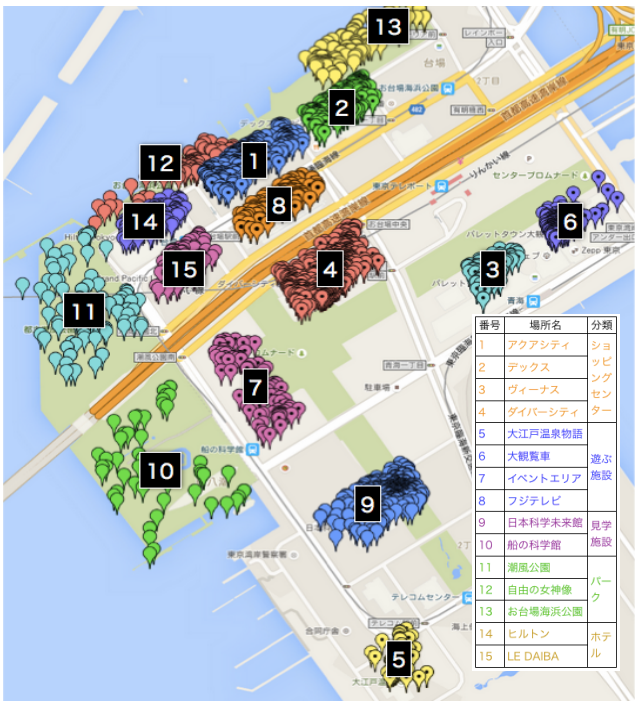

図 3: スポットを独立するような分け方

5. まとめと今後の課題

本研究では，位置情報付きツイートからミクロな観光行動を抽出する可能性を示した．お台場エリアで実験を行い，空間統計データを用いて検証を行った．また，お台場エリアでの観覧経路，および平日と休日の違いを把握した．観光客の関心を把握するには，空間統計データより，位置情報付きツイートのほうが優れていることを明らかにした．

今後の課題としては，以下三点を考えている．一つ目はユーザに関する属性（例えば性別や年齢）を推定することにより，属性によって行動の違いを究明する．二つ目，これからツイートのテキスト本文を用い，機械学習の手法でユーザの行動をラベリングする．最後に，今回の提案手法を回遊行動シミュレーションに応用し，パターンマイニングの結果をデータフィッティングの段階で使う．

参考文献

[桐村 13] 桐村 喬: 位置情報付きツイッター投稿データにみるユーザ行動の基本的特徴ー観光行動分析への可能性ー，地理システム学会講演論文集 22，2013．

[早川 06] 早川 潤一：頻出系列パターンマイニング手法を用いたWeb利用パターン発見に関する研究，名古屋工業大学修士論文，2006．

[寺田 12] 寺田 雅之，永田 智大，小林 基成：モバイル空間統計における人口推計技術 NTT DOCOMO テクニカル・ジャーナル，Vol. 20，No. 3，2012．