# 外国人旅行者の行動分析および地域性の判定

新田 崇人† 難波 英嗣†† 石野 亜耶††† 竹澤 寿幸††

†広島市立大学 情報科学部 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
††広島市立大学大学院 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1
†††広島経済大学 ビジネス情報学科 〒731-0138 広島県広島市安佐南区祇園 5-37-1
E-mail: {nitta, nanba, ishino, takezawa}@ls.info.hiroshima-cu.ac.jp

**あらまし** 2008 年 10 月の観光庁の設置を起点として，日本は国際観光振興，とりわけ訪日外国人観光客誘致の重要性を再認識した．これにより，現在では特に観光を基幹産業と位置づけた多様な取り組みに力を入れている．2015 年には訪日外国人観光客数が 1800 万人を越え，2020 年の東京オリンピックに向け，さらに増加し続けることが期待できる．この増加に伴い，外国人旅行者のニーズにより適した観光支援が重要となってくる．そこで本研究では，旅行者がどこを訪れ，旅先で何を食べ，何を購入したのか，といった旅行者の行動情報を旅行ブログエントリから自動的に抽出し，その結果を地図上に表示するシステムを構築する．本システムを使用し，旅行者の行動を分析することで，外国人旅行者への地域ごとの観光推薦支援が可能となる．

**キーワード** 観光情報，行動分析，情報抽出

## 1. はじめに

近年，訪日外国人観光客の数は急激に増加している．2011 年の東日本大震災と福島第一原子力発電所事故の影響により一度は減少していた観光客数であるが，2012 年末から始まったアベノミクスにより円安が進んだことで大幅に増加を始めた．2013 年には 1036 万人，2014 年には 1341 万人と 1000 万人超えという記録を達成し，毎年記録を更新している．また，2014 年の訪日外国人観光客が使った金額も過去最高となる 2 兆 278 億円を記録しており，多大な経済効果が見込まれる．2015 年上半期（1-6 月期）の訪日外国人旅行者数は、前年同期比 46%増の約 914 万人となり上半期の過去最高を更新した.通年では 1900 万人超えを達成している．さらに，訪日外国人観光客の国籍の多様化が顕著になっている．そのため，外国人旅行者のニーズにより適した観光推薦支援が重要となってくる．

外国人旅行者のニーズに適した観光推薦を行うためには,過去の外国人旅行者がどこを訪れ,何を食べ，何を購入したのかという行動情報を収集し分析することが必要である．外国人旅行者の行動情報を収集するための情報源として,旅行ブログエントリに注目する．旅行ブログエントリには，旅行者の体験記や感想が記述されており，いつ，どこで，何を食べ，何を購入したかなどの情報が記載されている．そのため，旅行ブログエントリは，外国人旅行者の行動情報を抽出するための有用な情報源であると考えられる．しかし，旅行ブログエントリから人手で行動情報を抽出し整理するためには，多大なコストを要する．

そこで本研究では，旅行ブログエントリから，機械学習を利用することで，自動で旅行者の行動情報を抽出する手法を提案する．また，抽出した旅行ブログエントリを地図上に表示するシステムを構築する．本システムを利用することで，ある地域で旅行者が頻繁に訪れる場所や食べるもの，購入するものを把握し行動分析を行うことができる．さらに，旅行者の行動に基づいて地域性を判定することで，外国人旅行者への地域ごとの観光推薦支援が可能となる．

本論文の構成は以下の通りである．2 節では，システムの動作例を示す．3 節では，関連研究について述べる．4 節では，外国人旅行者の行動情報の自動抽出手法について説明する．5 節では評価実験について述べ，6 節で本論文をまとめる．

## 2. システムの動作例

本章では，構築するシステムの動作例について説明する．システムの動作例として，「oyster（牡蠣）」と入力した場合の出力結果例を図 1 に示す．本システムでは，「oyster（牡蠣）」を食べた場所にピンと，旅行ブログエントリへのリンクが提示される．旅行ブログエントリの本文を閲覧することで，どのような旅行者が牡蠣を食べたかや，感想など詳細な情報を閲覧することができる．


図 1：「oyster（牡蠣）」と入力した際の出力結果例

## 3. 関連研究

本研究の関連研究として，3.1 節では，観光イメージに関する研究，3.2 節では，Web を情報源とした分析，3.3 節では，旅行ブログエントリを情報源とした分析についての研究をそれぞれ紹介する．

### 3.1. 観光イメージに関する研究

観光地のマーケティングを促進させていくための 1 つの手法として，観光イメージについて分析する研究が行われてきた．観光イメージについて分析した研究として，村上ら[1]の研究がある．村上らは，英語圏の訪日外国人の感想，とりわけブログでの訪日旅行の評判に着目し，テキストマイニングを用いて，イメージ分析を行っている．村上らの研究における，分析に Travel Blog[1]を用いている点や，行動パターンを抽出する点で本研究と類似している．しかし，主に人手による分析を行っているため，わずかなデータしか扱っていない。本研究では，自然言語処理を活用し，大量のデータを扱っている点で異なる．

大久保ら[2]は，海外版旅行ガイドブックである Lonely Planet と，旅行口コミサイトである TripAdvisor を対象に，その言語内容を解析し，有用な計画情報を抽出し，観光イメージを分析する手法を提案している．この研究では，データの電子化，収集においてテキストマイニングを用いている．また，収集したデータを元に Lonely Planet と，TripAdvisor の特徴分析，対応分析を行っている．さらに，特徴分析においては，観光地別，出身国別といったタイプ分類を行い，観光イメージを分析している．本研究では，Travel Blog を用いて，「訪れた場所」，「食べたもの」，「購入したもの」の 3 点に的を絞り，分析を行っている．また，抽出結果を地図上にマッピングし，可視化させての分析を行っている点で異なる．

倉田ら[3]は，観光客の観光地での行動を知るために観光客の行動履歴を分析する手法を提案している．観光客の行動履歴がわかることで，主要な滞在箇所と滞在時間，興味対象などが分かり，観光イメージを抽出できる．また，行動履歴をもとに，観光客の代表的行動パターンを抽出することも可能である．そのため倉田らは，訪日外国人観光客らの行動実態や関心を把握するための代表的手法として，既存観光統計の利用，アンケート形式による日誌調査，GPS ロガーを利用した調査，IC 乗車券の利用履歴を利用した調査，写真撮影箇所の位置情報を利用した調査の 5 つの観点から，それぞれの長所と短所について比較分析を行っている．倉田らは 5 つの行動分析調査から比較分析を行っているが，本研究では旅行ブログエントリを利用し行動分析を行っている．

[1] https://www.travelblog.org/

### 3.2. Web を情報源とした分析

Web 上の情報源として，ブログ，Wikipedia，YouTube，口コミサイト，旅行ガイドサイトなど様々なソーシャルメディアが挙げられる．近年では，SNS の代表格である Twitter，Facebook が Web 上の情報源として活躍している．このような時代背景もあり，ソーシャルメディアを活用した研究が盛んに行われている．

上記のようなソーシャルメディアを活用した研究では，藤原ら[4]の研究がある．藤原らは，旅行ブログエントリ，Twitter，YouTube のソーシャルメディアから，観光イベントに関するものを効率的に抽出する手法を提案し，分析を行っている．分析方法としては，意見文抽出器を用いて意見文を抽出する手法と，テキストマイニングツールを用いて，テキストの統計解析を実施する手法で行っている．これらを使用して分析を行い，ソーシャルメディアごとに分析結果を比較している．本研究では，Web 上の情報源として，Twitter，YouTube ではなく，旅行ブログエントリを用いて分析を行う．

Web 上の情報源を使用し，位置情報を用いた手法で旅行者の行動を分析している研究として笠原ら[5]や，佐伯ら[6]の研究がある．笠原らは，ネットワークを用いて，旅行者の GPS 移動軌跡で構成される遷移ネットワークを分析する手法を提案している。観光地ごとに同様の分析ができるよう，観光地におけるスポット間の関係を，スポット同士の遷移ネットワークに抽象化して扱っている．さらに，個々の旅行者の行動を遷移ネットワークの全体構造と比較し，分析することで，旅行者のタイプ分類を行っている．分析を通じて，距離要因が旅行者の遷移決定において重要であると判断し，距離依存型と，非依存型という 2 種類のタイプに分類できることを明らかにしている．佐伯らは，Twitter のツイート投稿時間，付与された位置情報など，使用言語に依存しない特徴量を用いることで，外国語を用いるユーザが訪日外国人なのか，在日外国人なのかを判別する手法を提案している．ツイートに対する言語判定として，Language-Detection を使用しており，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し，訪日外国人なのか，在日外国人なのかの判別を行っている．この判別後，日本人ユーザ，訪日外国人ユーザ，および在日外国人ユーザが訪問した観光スポットについて比較し，結果を地図上に可視化している．これらの研究では GPS などの位置情報を用いている．旅行ブログエントリは，どこの国に関する内容であるかをブロガーがあらかじめ登録して投稿している場合がある．また，旅行ブログエントリに，訪問地の情報が記載されている場合もある．本研究では，これらの情報を，旅行者の位置情報として使用する．

### 3.3. 旅行ブログエントリを情報源とした分析

本研究では，行動情報を抽出するための情報源として，旅行ブログエントリを使用している．旅行ブログエントリは，作成，読み手とのコミュニケーションが容易で匿名性も確保され，信頼性も高い．そのため，旅行や観光の記録や感想，意見を表現する手段として多くのユーザに扱われており，情報量が非常に多いという利点がある．

旅行ブログエントリが，観光情報の情報源として有益かどうかを分析した研究として，石野ら[7]の研究がある．石野らは，観光情報を収集するため，ブロガーが日記形式で綴った旅行記である旅行ブログに焦点を当てた．多くのブロガーが旅行記をこの形で記述するため，旅行ブログは観光情報を得るための有益な情報源であると考え，ブログデータベースから旅行ブログエントリを検出し，その中から観光情報を抽出している．結果として旅行ブログエントリは，観光情報の情報源として有益であることを示している.本研究では，石野らの研究と同様に，旅行ブログエントリから観光情報を抽出している．

藤井ら[8]は，旅行ブログエントリの属性に基づいた旅行者の行動分析を行い旅行者の特徴を明らかにするために,「性別」,「使用言語」,「観光タイプ」の 3 種類の属性を与え，自動的に属性を判定する方法を提案している．旅行ブログエントリを使用し，旅行者の行動分析を行う点では，本研究と非常に類似しているが，本研究では行動情報を抽出する点で異なる．

## 4. 旅行ブログエントリからの外国人旅行者の行動情報の自動抽出

本研究では，旅行ブログエントリから，訪れた場所，食べたもの，購入したものといった旅行者の行動情報を旅行ブログエントリから自動で抽出する手法を提案する．

外国人旅行者の行動情報を抽出する際，本文中の「訪れた場所」,「食べたもの」,「購入したもの」に関する語句の有無を判定に用いる．しかしながら旅行ブログエントリは膨大にあるため，まず，表 1 に示す手掛かり語を含む文を，訪れた場所，食べたもの，購入したものの情報が記載されている文の候補として収集する．

表 1：行動情報を含む文を収集するための手掛かり語

| | |
|---|---|
| 訪れた場所 | Fushimi-Inari（伏見稲荷神社），Ginkakuji（銀閣寺），Kenrokuen（兼六園），Mt.Fuji（富士山），Kiyomizu-Dera（清水寺），Himeji-Castle（姫路城），Kinkakuji（金閣寺），Sensoji（浅草寺），Todaiji（東大寺），Miyajima（宮島），Tokyo-Tower（東京タワー），Yakushima（屋久島） |
| 食べたもの | ramen（ラーメン），sashimi（刺身），sushi（寿司），takoyaki（たこ焼き），yakisoba（焼きそば），yakitori（焼き鳥），curry（カレー），tofu（豆腐），wasabi（わさび），tempura（天ぷら），oyster（牡蠣） |
| 購入したもの | matcha（抹茶），chopsticks（箸），laver（海苔），mocha（餅），stamp（判子），wine（ワイン），keychains（キーホルダー），magnet（マグネット），momiji（もみじ饅頭），sake（酒），yukata（浴衣） |

収集した文から行動情報を抽出するため，3 種類のタグを定義する．また，これらのタグを旅行ブログエントリに付与した例を図 2 に示す．

- visit：訪れた場所
- eat：食べたもの
- buy：購入したもの

- Today we took a ferry over to the island of <visit>Miyajima</visit>.
- We had lunch at an <eat>Okonomiyaki</eat> place, which was great.
- I also bought a set of <buy>silk placemats</buy> and <buy>chopsticks</buy> for the house.

図 2：旅行ブログエントリにタグを付与した例

本研究では機械学習として CRF を使用した．CRF 基本手法は与えられた文に含まれる語を分類するのに使用した．素性とタグは以下のように CRF に与える．

(1) ターゲットとなる単語より前の k 個の単語に付与されたタグ
(2) ターゲットとなる単語の前に存在する，ターゲットからの距離が k 以内に現れる単語
(3) ターゲットとなる単語の後に存在する，ターゲットからの距離が k 以内に現れる単語

我々は予備実験の結果から，k=2 と定めた．また，表 1 の手掛かり語を用い，以下の 17 つの素性を決定し，

機械学習に使用した.

- 単語
- 品詞
- 括弧（( )，< >など）
- 主語：“you”，“his”，“he”など，文章の主語となる単語 （67 語）
- 述語：“is”，“was”，“were”など，文章の述語となる単語 （137 語）
- 接続詞：“and”，“but”，“or”など，文章の接続詞となる単語（36 語）
- 訪れた場所：“japan”，“tokyo”，“asakusa”など，訪れた場所を表す単語（53 語）
- 訪れた際に使用する語：“from”，“went”，“see”など，観光地を訪れ，観光した際に使用する語（68 語）
- 観光地の名称に関する語：“temple”，“castle”，“pavilion”など，観光地や観光スポットの名称に頻出する語（28 件）
- 食べた際に使用する語：“eat”，“ate”，“had”など，食べた際に使用する語（46 語）
- 食材に関する語：“chicken”，“fish”，“meat”など，食べものに使われる食材に関する語（83 語）
- 調理表現：“grilled”，“steamed”，“boiled”など，食べものの調理表現（11 語）
- 食べ物の感想に関する語：“fresh”，“great”，“yummy”など，食べた感想や食べものの状態を含んだ，食べものの感想に使用する語（38 語）
- 食べた場所に関する語：“bar”，“hotel”，“restaurant”など，食べものを提供するお店に関する語（23 語）
- 購入した際に使用する語：“buy”，“shopping”，“get”など，土産物などを購入した際に使用する語（43 語）
- 購入したものに関する語：“tea”，“kimono”，“glass”など，土産物など購入したものに関する語（38 語）
- 購入した場所に関する語：“shops”，“store”，“market”など，購入するお店や場所の名称に頻出する語（23 語）

## 5. 実験

本研究では，旅行ブログエントリからの行動情報の自動抽出に関する実験を行った．5.1 節では，実験方手法，5.2 節では実験結果と考察について述べる.

### 5.1. 実験方法

**【実験データ】**

実験には，Travel Blog に登録されている旅行ブログエントリを使用した．Travel Blog とは，旅行ブログエントリを主体とした海外の Web サイトであり，世界各国に関する旅行ブログエントリが写真とともに掲載されている．また，ブロガーが旅行ブログエントリを投稿する際，どこに関する旅行ブログエントリであるかを「大陸」,「国」,「都市」の階層であらかじめ決めて投稿する仕様となっている．例えば,「広島市」に関する内容を投稿しようとした場合，「Asia / Japan / Hiroshima / Hiroshima」と登録して投稿される.

Travel Blog に登録されている旅行ブログエントリ 176,843,000 件の中から，表 1 の手掛かり語を含んだ文を約 3,000 件抽出した．その文に visit タグ，eat タグ，buy タグを人手で付与したデータを実験に使用した．人手で付与したタグの件数を表 2 に示す.

表 2：人手で付与したタグの件数

| タグ | 件数（文） |
|---|---|
| eat | 703 |
| visit | 694 |
| buy | 342 |
| 総数 | 1,739 |

**【機械学習と評価尺度】**

機械学習には CRF を用い，5 分割交差検定を行った．また，以下の式に示す精度，再現率を使用し，評価を行った.

$$\text{精度} = \frac{\text{システムと人手により正解と判定された旅行ブログエントリ数}}{\text{システムにより正解と判定された旅行ブログエントリ数}}$$

$$\text{再現率} = \frac{\text{システムと人手により正解と判定された旅行ブログエントリ数}}{\text{人手により正解と判定された旅行ブログエントリ数}}$$

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

**【比較手法】**

旅行ブログエントリに含まれる単語，品詞のみを素性として機械学習を行った場合をベースラインとした．また，ベースラインで用いた素性に，4 節で示した素性を加えて機械学習を行った場合を提案手法とした．これらを用いて実験を行い，比較を行う.

### 5.2. 実験結果と考察

ベースラインによる実験結果を表 3 に，提案手法による実験結果を表 4 に示す．実験結果を見ると，提案

手法により，若干の精度の低下はあったものの，大幅に再現率を向上させることができた.「visit」においては，精度は 0.013 ポイント下がったが，再現率は 0.085 ポイント上がった.「eat」においては，精度は 0.011 ポイント下がったが，再現率は 0.142 ポイント上がった.「buy」においては，精度は 0.013 ポイント下がったが, 再現率は 0.058 ポイント上がった. これにより，提案手法の有効性が確認できたといえる.

表 3：ベースラインによる行動情報抽出の実験結果

| タグ | 精度 | 再現率 | F 値 |
|---|---|---|---|
| eat | 0.829 | 0.550 | 0.661 |
| visit | 0.925 | 0.793 | 0.854 |
| buy | 0.963 | 0.512 | 0.669 |
| 平均 | 0.906 | 0.618 | 0.728 |

表 4：提案手法による行動情報抽出の実験結果

| タグ | 精度 | 再現率 | F 値 |
|---|---|---|---|
| eat | 0.912 | 0.878 | 0.895 |
| visit | 0.818 | 0.692 | 0.750 |
| buy | 0.950 | 0.570 | 0.712 |
| 平均 | 0.893 | 0.713 | 0.786 |

実験結果を分析したところ，人手ではタグを付与したが，システムではタグが付与されていないもの，また，人手ではタグを付与していないが，システムではタグが付与されているものが多数存在した．以下に検出誤りの主要な原因を示す.

（1） 2 単語以上の固有名詞化した単語
（2） 英語以外の言語で記載されたテキストデータ

以下に，それぞれの検出誤りについて説明する.

（1） 2 単語以上の固有名詞化した単語

「chichen」や「ramen」など 1 単語にはタグを付与しているが,「chicken patties」や「ramen noodles」など 2 単語以上で固有名詞化している単語にタグがついていない例が多数検出された．抽出の失敗例を以下の図 3 に示す.

I am need of a really good meal because the last four days I was surviving off <eat>chicken </eat>patties, <eat>ramen</eat> noodles, breakfast <eat>sausages</eat> and granola bars.

図 3：(1) の検出誤り例

図 3 では，食べたと分かる 1 単語の固有名詞はタグが付与されているが，2 単語以上で意味を成している固有名詞には付与されていない．また「granola bars」など，食べたと分かるもので全くタグが付与されていない例もあった．これらは人手で付与した正解データの件数を増やし，2 単語以上の固有名詞化した単語に可能な限りタグ付けを行うことで解消できると思われる.

（2） 英語以外の言語で記載されたテキストデータ

旅行ブログエントリには，英語以外の言語で記載されたテキストデータがいくつか存在する．そのような文に対してタグが付与されている例がいくつか存在した．以下の図 4 に抽出の失敗例を示す.

Vi allierede os med Josh og Julie, der har arbejdet på Rudehøj, og som <buy>laver</buy> ture for grupper på bjerget og på floderne omkring.

図 4：(2) の検出誤り例

本研究では，主に英語のテキストデータを使用して，行動経路抽出のためのモデルを作成した．そのため，英語以外の言語で記載されたテキストデータに対して，正確にタグ付けを行うことができなかった．本研究で使用した手掛かり語を翻訳することで，英語以外の言語で記載されたテキストデータに対してもタグ付けを行い，自動的に旅行者の行動経路を抽出することが可能だと考えられる.

## 6. おわりに

本研究では, 旅行者が「訪れた場所」,「食べたもの」,「購入したもの」の情報を，旅行者の行動情報として旅行ブログエントリから自動的に抽出し，その結果を地図上に表示するシステムを構築した．旅行者の行動情報の自動抽出実験を行った結果，提案手法では，精度平均 0.893, 再現率平均 0.713 を得た．ベースラインの再現率平均 0.618 と比べて再現率を 0.095 ポイント向上させることができ，提案手法の有効性を確認することができた．今後の展望としては今回提案した手法を用いてタグ付けを行い，その結果を地図上に表示し分析を行うことでより正確な旅行者の行動分析および地域性の判定が行えると考えられる．また，これらの手法と，ブロガーの属性（性別，居住地域など）を組み合わせて利用することで，利用者ごとの最適な観光情報を提供することができるようになる．本研究では「訪れた場所」,「食べたもの」,「購入したもの」に着目をして研究を行ったが,「見た」,「体験した」など他の行動情報にも着目し抽出を行うことで，よりニーズに適した観光推薦を行うことができる．さらに各地域

をモデル化することで，ある地域と類似した他の地域を推薦することも可能となる．

## 謝辞

本研究の一部は，総務省による戦略的情報通信研究開発推進制度（SCOPE）の支援を受けて行われた．

## 参 考 文 献

[1] 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分析-”，人工知能学会誌，vol.26（3），pp.286-293，2011.

[2] 大久保立樹，室町泰徳，“旅行ガイドブックと口コミの言語解析による訪日外国人の観光地イメージに関する研究”，都市計画論文集，vol.49（3），pp.573-578，2014.

[3] 倉田陽平，矢部直人，駒木伸比古，有馬貴之，杉本興運，室町泰徳，“何を，いつ，どれくらい見て，どこに興味を示すのか？－訪日外国人観光客のより詳細な行動調査に向けて－”，観光情報学会第 2 回研究発表会，2010.

[4] 藤原泰士，難波英嗣，竹澤寿幸，“ソーシャルメディアの分析によるイベント開催支援”，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（DEIM2014），2014.

[5] 笠原秀一，森幹彦，椋木雅之，美濃導彦，“遷移ネットワークを用いた大規模観光地の旅行者行動分析”，人工知能学会「社会における AI」研究会 第 17 回研究会，2013.

[6] 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”，観光情報学会誌「観光と情報」，vol.11（1），pp.45-56，2015.

[7] 石野亜耶，難波英嗣，竹澤寿幸，“旅行ブログエントリからの観光情報の自動抽出”，日本知能情報ファジィ学会誌，vol22（6），pp.667-679，2010.

[8] 藤井一輝，難波英嗣，竹澤寿幸，石野亜耶，“旅行ブログエントリの属性に基づいた旅行者の行動分析”，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（DEIM2015），2015.