

旅行記の LDA モデル分析による訪日中国人旅行者の旅行行動基礎的分析

宋紫龍^{†1} 古屋秀樹^{†2}

概要：2018 年に訪日中国人旅行者数は 838 万人、訪日旅行消費額はおよそ 1 兆 5,370 億円になり、日本のインバウンド観光の最大市場である。一方で、中国人旅行者の旅行行動は変化しつつあり、その実態を把握する必要がある。そこで本研究は、旅行者の意向が記されている旅行記を用いて分析し、訪日中国人旅行者の旅行意向・行動の把握を目的とする。SNS データから旅行地点・内容及び個人属性が記された旅行記 35,552 編を用いて、LDA モデルにより旅行地点及び旅行内容の類型化を行った。その結果、旅行記を 84 トピックに分類でき、旅行記作者の個人属性及び旅程情報と結びつけ、訪日中国人旅行者の旅行行動及び特徴を明らかにした。

キーワード：LDA モデル、トピックモデル、旅行記、旅行行動

1. はじめに

2018 年の訪日外国人旅行者数は 3,119 万人、訪日外国人旅行者消費額は 4 兆 5,064 億円を記録し、ともに過去の最高を更新した。その中で、図 1 に示すように、訪日中国人が訪日外国人の旅行者数の 27%ならびに消費額 34%を占め、いずれも第 1 位になっており、日本インバウンド観光市場におけるもっとも大きなマーケットと言える。一方、訪日中国人の 1 人当たり訪日消費額は 2015 年から年毎に減少しつつあり、2018 年に 1 人当たり消費額がおおよそ 22 万円になっており、すべての訪日外国人の中の第 4 位となった。訪日中国人旅行者は「爆買い」の沈静化をはじめ、訪日中国人旅行者のニーズ・行動が変化しつつあると考えられる。そのため、訪日中国人旅行者の旅行行動・特徴を把握することは喫緊の課題である。

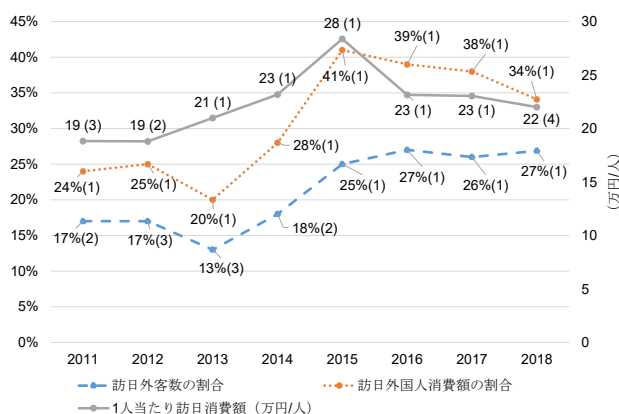


図 1 訪日中国人旅行者数及び消費額の割合

従来、アンケート調査によって得られたデータから多変量解析などを用いて旅行行動のパターン化や特徴把握を試みる研究が多くみられる。ところが、アンケート調査は被験者の確保が必要となり、大変なコストやデータハンドリングの時間がかかるなどの問題点がある。更に、2010 年代後半に、ビッグデータ、IoT、AI、ロボットといった技術をコアとする第 4 次産業革命の発展が始まった。データ分析手法の革新を観光研究・観光マーケティングに適用し、行動データを収集・活用しながら観光が抱える諸課題の解決に資することが期待されている。

そこで、本研究は、旅行者の旅行行動や意向が記されている旅行記を自動的に収集し、機械学習・自然言語処理の 1 つである LDA モデルを用い、記述内容の出現頻度やその類似度に基づく旅行記の類型化によって、訪日中国人旅行者の旅行行動・特徴の把握を目的とする。

2. 先行研究と本研究の位置づけ

文献 1 では、訪日外国人消費動向調査を用いながら、潜在クラスモデルによって外国人旅行者の訪問地を幾つかのパターンに分類した。さらに、文献 2、3 では、潜在クラス分析の強い仮定を緩和した一般モデルと位置付けできるトピックモデルが適用されている。これらで用いられた分析データは位置情報データであったが、旅行行動の研究は位置情報にとどまらず、旅行の内容、即ちコンテンツも含まれることが望ましい。現時点まで、コンテンツを含める旅行行動に関する研究は十分に行われていないと考えられることから、本研究では旅行者の旅行行動、嗜好、評価等のデータを広範に収集する方法を検討するとともに、LDA モデルを用いて旅行者の類型化を行う。なお、分析対象として中国人旅行者を設定する。

^{†1} 東洋大学大学院国際観光学研究科国際観光学専攻
Toyo University, Graduate School of International Tourism Management

^{†2} 東洋大学国際観光学部国際観光学科
Toyo University, Department of Tourism

3. 分析データとデータ収集手法

本研究では、多様な訪日中国人旅行者の嗜好、旅行に対する評価、訪問地を把握するため、Web上の旅行記に着目した。一方で旅行記の執筆者に偏りを生じるおそれがあるため、より多数の旅行記を集めることによって、偏りを相対的に小さくすることを意図しながら、現在、中国における最大級の観光口コミサイトである「蚂蜂窝(Mafengwo)、以下 Mafengwo」を分析対象とした。Mafengwo では、それぞれの旅行者が記述した訪問先、旅行行動やその評価とともに、旅行記の題目、居住地、性別、出発時間、同行人物、滞在日数、旅行費用も収集できる。さらに、2013 年から、Mafengwo は中国旅游研究院と連携しながら、「全球自由行報告」を始め、毎年、観光に関する報告書を発行している。そのため、Mafengwo からのデータの一定程度の信頼性、説得力が期待できる。以上から、作者の個人属性を含めた情報が豊富的に記録されているため、本研究にとって適当な分析データと判断した。

そして、クローラを用い、Mafengwo に掲載されている訪日中国人が書いた旅行記に対する収集を行った。クローラとは、ウェブ上のデータを周期的に取得し、自動的にデータベース化するプログラムであり、これをプログラミング言語 (Python) で構築したことによって、訪日旅行記を 2017 年 4 月から 2018 年 6 月中旬にかけて合計 35,552 編を収集し、これらを用いてトピックモデル分析を行った。表 1 は収集したデータの概況を示している。

表 1 データ収集の概況

収集期間	2017 年 4 月上旬~2018 年 6 月上旬
収集した旅行記本文数	35,552 篇
収集した作者の個人属性	性別、居住地
収集した旅程情報	出発時期、同行者、滞在日数、コスト

4. LDA モデルについて^{[4], [5]}

LDA モデルは自然言語処理方法で、教師データなしの機械学習の 1 つである。文章内での形態素 (言語において、それ以上分解したら意味をなさなくなるところまで分割された最小単位) の出現の組み合わせから、尤度にもとづき類似した文章を非排他的かつ明示された導出過程に基づきセグメント (トピックの割り当て) するものである。1 つの文書は必ずしも単一のトピックに帰属するとは限らない点を考慮しながら、トピック別に形態素と出現頻度とのペアの集合をモデル化する方法である。

さて、LDA モデル分析を行う前に、各旅行記を形態素に分割する必要がある。本研究は 35,552 編の旅行記の中で位

置情報と旅行内容を表せる地名、観光スポット名及び品詞を対象として、形態素に分割した。その結果、位置情報「pos」 (観光スポット・都道府県・市町村名) を表す形態素 (単語) 1,910,458 個、2,862 種、ならびに旅行内容「n」 (名詞) を表す形態素 (単語) 10,675,514 個、129,538 種、合計 12,585,972 個、132,400 種の形態素を抽出でき、これを LDA モデルに導入した。LDA モデルの生成過程は以下である。

1. For トピック $k=1, \dots, K$
 - (a) 形態素分布を生成 $\phi_k \sim \text{Dirichlet}(\beta)$
2. For 文書 $d=1, \dots, D$
 - (a) トピック分布を生成 $\theta_d \sim \text{Dirichlet}(\alpha)$
 - (b) For 形態素 $n=1, \dots, N_d$
 - i. トピック生成 $z_{dn} \sim \text{Categorical}(\theta_d)$
 - ii. 形態素を生成 $w_{dn} \sim \text{Categorical}(\phi_{z_{dn}})$

トピック分布 θ_d と形態素分布集合 ϕ が与えられた際の文書 w_d の確率は下の公式で示される。

$$p(w_d | \theta_d, \phi) = \prod_{n=1}^{N_d} \sum_{k=1}^K p(z_{dn} = k | \theta_d) p(w_{dn} | \phi_k) \\ = \prod_{n=1}^{N_d} \sum_{k=1}^K \theta_{dk} \phi_{kw_{dn}} \quad \dots (1) \text{式}$$

上記によって、教師なしデータである旅行記から、出現の仕方が類似した形態素の組み合わせ (形態素パターン) を、あらかじめ設定したトピック数のもとで抽出することができる。これより、あるトピックのもとで、旅行記の中で頻繁に使用される形態素パターンが明らかとなり、その形態素から旅行者の嗜好する目的地、観光資源を類推することが可能となる。

さて、このトピック数は複数設定した中で、モデルの説明力にもとづいて評価を行うのが一般的である。これを判断するために、分岐数または選択枝の数を表しており、確率の逆数で定義されているパープレキシティ (perplexity) を使う。パープレキシティが小さいほど高い説明力を有しているモデルであることを示す。

$$\text{Perplexity}(W|M) = \left(- \frac{\sum_{d=1}^D \log P(W_d|M)}{\sum_{d=1}^D N_d} \right) \dots (2) \text{式}$$

本研究では、トピック数を 1~200 まで設定して各々 20 回推定を行い、PPL の平均値を算出した。パープレキシティ値は 84 トピックの際に最小となり、説明力が最も高いトピック数であると判断した。したがって、本研究では、トピック数 84 個と設定して、各トピックの頻出形態素 15 個をもとに考察を行う。

5.3 旅行形態の抽出・特徴把握

各深度游地域及び広域ルート地域に関連するトピックに着目し、旅行形態の抽出を通じ、訪日中国人旅行者の旅行行動及び特徴の把握を行った。ここでは、東京都を例として紹介する。

東京都に関連するトピックに着目し（トピック 12、19、24、26、31、33、50、54、58、60、65 が該当）、位置情報を表す形態素に対する考察を通じ、それぞれの出現した立寄り場所を示したものが図4および表3である。



図4 東京都における旅行形態

表3 立寄り場所表

①	歌舞伎町	⑭	大江戸温泉物語
②	新宿御苑	⑮	銀座
③	明治神宮	⑯	皇居
④	原宿	⑰	千鳥ヶ淵
⑤	渋谷	⑱	築地
⑥	代官山	⑲	秋葉原
⑦	表参道	⑳	スカイツリー
⑧	六本木	㉑	浅草寺
⑨	目黒	㉒	上野公園
⑩	東京タワー	㉓	東京大学
⑪	お台場海浜公園	㉔	早稲田大学
⑫	フジテレビ	㉕	ディズニーランド
⑬	自由の女神	㉖	ジブリ美術館

各トピックの立寄り場所とそれらが対応している旅行内容の特徴を考察すると、図4に示すように、トピック12、19、26：歌舞伎町→新宿御苑→渋谷→東京タワー→スカイツリー→上野公園といったところを立寄り場所とする「定番巡り」、トピック24：上野公園→千鳥ヶ淵→目黒を立寄り場所とする「桜鑑賞」、トピック31：秋葉原→東京タワー（ワンピース）→ジブリ美術館を立寄り場所とする「アニメ」、トピック54、58、60：上野公園（博物館・動物園等）→東京大学→早稲田大学を立寄り場所とする「コンテンツツアー」、トピック33：ディズニーランドを立寄り場所とする「テーマパーク」、トピック50：お台場を中心と

表4 東京都における旅行形態の特徴

旅行形態	居住地											
	北京市	上海市	東北	華北	華中	華南	西部	香港・マカオ	台湾	海外	日本	
定番巡り	1.30%	0.91%	0.82%	-0.41%	-1.29%	-0.70%	-1.85%	2.84%	-4.29%	-1.14%	1.97%	
桜鑑賞	-0.66%	-0.21%	-0.55%	-2.84%	-0.85%	-0.45%	0.29%	3.54%	-3.97%	0.18%	6.22%	
アニメ	0.36%	-0.40%	-0.05%	-1.58%	-0.36%	0.20%	0.17%	-1.02%	-2.31%	-0.34%	1.75%	
コンテンツツアー	0.16%	-0.09%	2.20%	-2.06%	-0.77%	-0.75%	-0.17%	-0.46%	-2.25%	-1.10%	3.99%	
テーマパーク	0.75%	-0.29%	2.34%	-2.21%	0.26%	-1.46%	-0.35%	-1.86%	-3.57%	-1.42%	1.03%	
台場巡り	0.16%	0.27%	-0.41%	0.12%	-0.12%	-0.08%	-0.75%	-0.16%	-0.81%	-1.49%	1.58%	
花火大会	-0.27%	1.00%	-0.92%	-1.72%	0.11%	-1.30%	-1.19%	-0.65%	-1.13%	-0.94%	1.95%	
旅行形態	性別		出発季節				同行者					
	男性	女性	春	夏	秋	冬	1人	友達	恋人	親子	家族	
定番巡り	0.92%	-0.31%	-0.09%	-1.66%	0.29%	0.98%	1.59%	-0.10%	0.10%	-1.49%	-0.91%	
桜鑑賞	0.77%	-0.26%	1.56%	-3.81%	-3.65%	-3.14%	1.42%	-0.48%	0.20%	-0.80%	-0.17%	
アニメ	0.19%	-0.08%	0.16%	-0.55%	0.43%	-0.12%	0.82%	0.04%	-0.15%	-0.40%	-0.50%	
コンテンツツアー	0.67%	-0.26%	-0.32%	-0.49%	-0.25%	0.68%	0.38%	-0.90%	-0.47%	0.13%	1.48%	
テーマパーク	-0.77%	0.23%	0.45%	0.34%	-0.69%	-0.19%	-2.46%	-0.80%	-1.32%	2.39%	-0.08%	
台場巡り	0.54%	-0.22%	-0.16%	-0.12%	-0.23%	0.47%	0.17%	-0.47%	-0.15%	0.97%	0.24%	
花火大会	0.09%	-0.03%	-1.12%	1.09%	-0.70%	-0.71%	0.12%	0.13%	-0.24%	-1.01%	0.52%	
旅行形態	滞在日数							コスト				
	3日以内	4~6日	7~13日	14~20日	21~27日	28~90日	91日以上	4万未満	4万~10万未満	10万~20万未満	20万~60万未満	60万以上
定番巡り	6.43%	4.44%	-1.85%	-5.32%	-3.61%	1.25%	-10.82%	7.08%	3.36%	0.97%	-1.75%	-4.88%
桜鑑賞	3.92%	-0.06%	-0.61%	0.91%	-0.35%	1.11%	3.42%	10.96%	0.27%	-0.22%	-0.29%	-1.98%
アニメ	3.91%	0.47%	-0.52%	-0.85%	-2.81%	-0.66%	-3.09%	3.69%	0.24%	-0.15%	0.05%	-2.24%
コンテンツツアー	4.25%	-0.11%	-0.62%	-0.92%	1.88%	0.53%	14.57%	4.06%	0.33%	-0.27%	0.02%	3.00%
テーマパーク	7.24%	0.14%	-0.66%	-1.99%	-4.63%	-2.31%	-5.00%	7.19%	-0.52%	0.35%	-0.51%	-3.59%
台場巡り	0.86%	0.56%	-0.28%	-0.58%	-0.45%	1.63%	-1.75%	-1.34%	0.50%	0.30%	-0.48%	-0.37%
花火大会	1.35%	-0.35%	-0.15%	0.42%	-0.71%	1.95%	0.00%	2.37%	0.45%	-0.26%	0.05%	-0.40%

する「台場巡り」、トピック 65：幕張花火大会といった花火大会に参加する「花火大会」の7つの旅行形態を分類でき、東京都における異なる旅行形態の差異を抽出できた。

次に、LDA モデルの結果である「各旅行記のトピック別構成比率」を使い、「特定属性に所属する各旅行記の特定旅行形態を主題とするトピック別構成比率」と「特定旅行形態を主題とする各旅行記のトピック別構成比率」の平均値の差を算出する。表4に示すように、カラスケールは濃ければ濃いほど、特定旅行形態に対する内容は平均値を上回って記述された傾向であると判断することから、各旅行形態の特徴を把握できる。各旅行形態の特徴として、下記のようにまとめることができる。

- ・ 定番巡り：男性、北京市・上海市・東北・香港マカオに居住、冬に1人旅または恋人を同行者とする。6日以下に滞在し、10万円以下を旅行コストとする。
- ・ 桜鑑賞：男性、西部・香港マカオ・海外に居住、春に1人旅または恋人を同行者とする。3日以下または91日以上に滞在し、4万円未満を旅行コストとする。
- ・ アニメ：男性、北京市・華南・西部に居住、春または秋に1人旅または友達を同行者とする。6日以下に滞在し、4万円未満を旅行コストとする。
- ・ コンテンツツアー：男性、北京市・東北に居住、冬に1人旅または親子・家族を同行者とする。3日以内・21~27日・91日以上に滞在し、4万円未満または60万以上を旅行コストとする。
- ・ テーマパーク：女性、北京市・東北・華中に居住、春または夏に親子を同行者とする。6日以下に滞在し、4万円未満または10万~20万円未満を旅行コストとする。
- ・ 台場巡り：男性、北京・上海市・華北に居住、冬に親子を同行者とする。6日以下または28~90日に滞在し、4万円~20万未満を旅行コストとする。
- ・ 花火大会：男性、上海市に居住、夏に家族を同行者とする。3日以内・14~20日・28~90日に滞在し、10万円以下を旅行コストとする。

以上のことから、訪日中国人旅行者が東京における旅行形態の抽出が実現された。更に、各旅行形態における旅行者の個人属性・旅行情報の異なりによる特徴も明らかにされた。

東京だけではなく、他の深度游地域（東京を含む5つ）及び広域ルート（4つ）の旅行形態抽出及び特徴把握を行った。表5は東京都を除いた深度游地域・広域ルールにおける旅行形態である。

表 5 深度游地域・広域ルートにおける旅行形態

地域	旅行形態（立寄り場所）
北海道（深）	①広域巡り（小樽・札幌・旭川市・美瑛・阿寒湖など） ②道南巡り（函館市・地球岬・洞爺湖など） ③登別巡り（登別温泉・のぼりべつクマ牧場・登別地獄谷など）
大阪府（深）	①定番巡り（梅田スカイビル・大阪城・四天王寺・通天閣・道頓堀など） ②グルメ（法善寺横丁を中心とする店） ③テーマパーク（ユニバーサル・スタジオ・ジャパン）
神奈川県（深）	①湘南巡り（明月院・鶴岡八幡宮・小町通り・鎌倉大仏・鎌倉高校前・江ノ島など） ②箱根巡り（箱根を中心とする） ③横浜巡り（赤レンガ倉庫・中華街・山下公園）
京都府（深）	①紅葉鑑賞（嵐山・嵯峨野・高山寺・金閣寺・南禅寺・東福寺・西芳寺など） ②古都巡り（北野天満宮・二条城・京都御所・清水寺・醍醐寺・三十三院など） ③宇治巡り（源氏物語ミュージアム・宇治上神社・伊平等院・伊藤久右衛門など）
九州地方（広）	①北九州巡り（ハウステンボス・熊本市・黒川温泉・別府市・門司市・太宰府など） ②南九州巡り（青島・桜島・鹿児島市・指宿市・屋久島など）
中部地方（広）	①定番巡り（犬山城・トヨタ産業記念館・名古屋城・熱田神宮など） ②広域巡り（金沢城・白川郷・下呂温泉・高山陣屋・飛騨市・黒部ダムなど） ③スポーツ（長野県長野市を中心とする）
瀬戸内（広）	①香川巡り（小豆島・豊島・直島・栗林公園・金刀比羅宮など） ②瀬戸内広域巡り（岡山城・岡山後楽園・大久野島・松山城・厳島など）
近畿地方（広）	①歴史巡り（琵琶湖・船山神社・伊勢神宮・熊野・高尾山・和歌山城など） ②兵庫巡り（有馬温泉・六甲山・北野異人館・神戸港・姫路城など）

6. まとめと今後の課題

本研究はクローラによって自動的にWeb上にある35,552編の旅行記を収集し、13万種類・1200万個の形態素を抽出した。これを用いて、LDAモデル分析によって84個トピックを得た。分析結果に対する考察によって、コンテンツを含める訪日中国人旅行者の旅行行動の実態と特徴を把握することができた。

今後の課題として、さらにデータ量を増やした分析を行うこと、トピック数の縮約方法の検討、他言語を用いた分析を行うとともに、比較検証を行い旅行者の国籍・地域による差異の把握が考えられる。

謝辞 本研究の分析にあたり、王旭氏（東京工科大学）に協力を頂いた。ここに深謝の意を表します。

参考文献

- [1] 古屋秀樹・劉瑜娟（2016）：潜在クラス分析を用いた訪日外国人旅行者の訪問パターン分析、土木学会論文集 D3、Vol. 72、No. 5、p. I 571-I 583
 - [2] 古屋秀樹（2016B）：トピックモデルによる訪日外国人旅行者の訪問パターンの基礎分析、第53回土木計画学研究発表会講演集、No.53
 - [3] 古屋秀樹・岡本直久・野津直樹（2017）：GPSログデータを用いた訪日外国人旅行者の訪問パターンの分析手法の開発、運輸政策研究、Vol.76
 - [4] 佐藤一誠（2015）：トピックモデルによる統計的潜在意味解析、コロナ社
- 岩田具治（2015）：トピックモデル（機械学習プロフェッショナルシリーズ）、講談社