トピックモデルを用いた
訪日中国人旅行者による旅行記の基礎的分析

宋　紫龍[1]・古屋　秀樹[2]

[1] 学生会員 東洋大学大学院生 国際地域学研究科国際観光学専攻（〒112-8606 東京都文京区白山 5-28-20）
E-mail: s38201610023@toyo.jp

[2] 正会員 東洋大学教授 国際観光学部国際観光学科（〒112-8606 東京都文京区白山 5-28-20）
E-mail: furuya@toyo.jp

日本インバウンド観光の最大市場である中国人旅行者の需要が変化しつつあることから、彼らの旅行行動の把握が重要となっている。そこで、訪問地点に加え、訪問地に対する評価の把握によって、旅行行動の背景や嗜好を理解できるとともに、今後の動向も検討可能と考えられる。そこで、本研究は旅行内容ならびに訪問地点が記載されている訪日中国人旅行者が書いた旅行記に着目し、機械学習の1つであるトピックモデルを用いて、記述内容の類似度による旅行記の類型化を行い、訪日中国人旅行者の旅行意向・行動の把握を目的とする。収集した16,734編の旅行記を用いて分析した結果、152個のトピックに分類できるとともに、それに基づいて訪日中国人旅行者の旅行行動および特徴を明らかにした。

Key Words: *machine learning, natural language processing, Latent Dirichlet Allocation, travel literature*

1.　はじめに

観光産業は日本の成長戦略の柱、地方創生の切り札として位置づけられている。2016年3月に策定された「明日の日本を支える観光ビジョン」の諸目標を達成し、「観光先進国」に向けて取り組みが進む中、依然観光をめぐる課題は多岐にわたる。その中で、いかに日々変化しつつある旅行者らのニーズを把握し、それに応じて効率的かつ効果的に観光プロモーションを行うことは重要な論点の1つである。

2017年の訪日外客数は2,869万人を数えるとともに、訪日外国人消費額は4兆4,162億円となり、53年ぶりの国際旅行収支の黒字となった。その中で、訪日中国人の旅行者数（構成率：26%）、ならびに消費額（同：38%）はいずれも第1位となり（図-1）、インバウンド旅行市場で最も大きなマーケットと言える。一方、2015年から、消費額および1人当たり訪日消費金額は年々減少しており、「爆買い」の沈静化をはじめ、訪日中国人旅行者の需要・行動が変化しつつある。そのため、訪日中国人旅行者の需要・行動を把握することは喫緊の課題である。

さて、ICTを活用して取得されるビッグデータを分析・活用することにより、新たな経済価値が生まれている。そのため、これらを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。

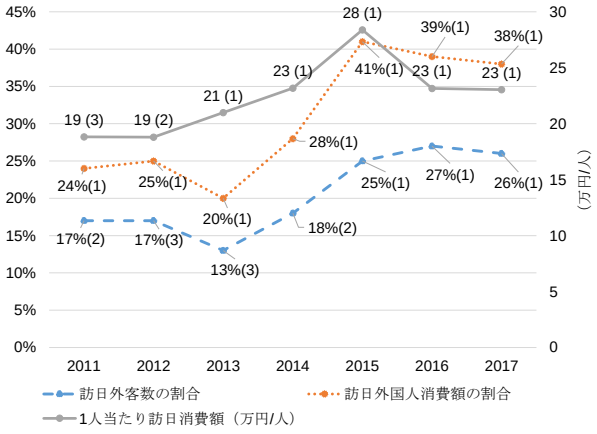

図-1　訪日中国人旅行者人数と消費額の割合

そこで、本研究は、訪日中国人旅行者による訪問地点ならびに旅行内容を記述したWeb上の旅行記を取り上げ、機械学習・自然言語処理の1つであるトピックモデルを用いて、旅行記の記述内容の出現頻度やその類似度により、旅行記の類型化を行い、訪日中国人旅行者の旅行行動および特徴の把握を目的とする。これらの分析を通じて、訪日中国人旅行者の旅行行動および特徴が明確になるとともに、効率的かつ効果的な観光プロモーションに資することが期待できる。

2. 先行研究と本研究の位置づけ

近年、訪日外国人旅行者の旅行行動データに機械学習等を用いて分析した研究事例が見られる。

文献1では、「訪日外国人消費動向調査（観光庁）」の訪問地点データを用いながら、潜在クラスモデルによって訪日外国人旅行者の訪問地のパターン化を通じて、旅行行動特性を明確にした。さらに、文献2、3では、訪問地点データを用いながら、潜在クラスモデル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を適用している。これらの「訪問地点のパターン化」に対して、「訪問地点に加えて、旅行内容やその評価のパターン化」が考えられる。それによって、満足や不満といった意向を反映することができ、効率的かつ効果的な観光プロモーションの実現が期待される。

そこで、本研究は、旅行者の旅行地点・旅行内容を含んでいるデータを広範に収集する方法を検討しながら、トピックモデルを用いて訪日中国人旅行者の旅行行動のパターン化に関する研究として位置づけできる。

3. 分析データ・収集手法

旅行行動把握のためにアンケート調査が用いられるが、その実施において、人手、コスト、時間の負担が大きく、大量なデータ取得は困難といえる。

それに対して、本研究は旅行地点に加えて旅行内容の情報が豊富に記載されているWeb上の旅行記に着目した。より多数の旅行記を集めることを意図して、本研究では、中国旅游研究院と連携し、「全球自由行報告」をはじめ毎年、観光に関する報告書を発行している中国における最大級の旅行ポータルサイト「馬蜂窩（Mafengwo・マァファンウォー。以下、Mafengwo）」にある旅行記に着目した。この旅行記を用いることによって、大量のデータを収集でき、より多くの評価，行動の分析が可能といえる。

さて、Web上の旅行記を効率的に取得するため、本研究はpythonを用いてWeb上の情報（文字または画像）を周期的・自動に収集できるWebクローラ[4]を適用した。データ収集は、2017年4月上旬～11月上旬にかけて行い、訪日旅行の出発日時は2015年1月から2017年11月までの16,734篇旅行記（総文字数：1.1億字、平均：6,619字/篇）を収集するとともに、各旅行記の「題目」、「作者性別」、「居住地」、「出発時間」、「同行人物」、「滞在日数」、「コスト（消費金額）」に関する情報も同時に収集した。

4. トピックモデルについて[5) 6)]

トピックモデルは自然言語処理方法の１つで、教師データなしの機械学習手法である。大量のデータを活用し、文書集合から頻出するトピック（パターン）や類似トピックを含む文書を抽出できる。

(1) BOWおよび形態素分析

トピックモデルでは、形態素分析によって各文書を1つ1つの単語の形で区切る。そして、区切られたすべての単語の語順を無視し、単語の多重集合であるBOW（bag-of-words）を構築する。次に、BOWの中の単語の出現の組み合わせから、尤度に基づき類似した旅行記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）し、結果であるトピックを多重集合（単語の集合）の形で表す。

さて、形態素分析を行う前に、分析精度を高めるため、日本の観光スポット名と「都道府県＋自治体名」のような位置情報を表す単語、さらに名詞のような旅行内容を表す単語を加える作業を行った。このような辞書を用い

図-2 データベースに保存される旅行記データ

id	articleName	authorPos	authorSex	startTime	people	duringDay	cost	articleInfo
	过滤	过滤	过滤	过滤	过滤	过滤	过滤	过滤
4719	['我们在旅行中寻找什么——…	甘肃省	Male	2016-02-03	家族出游	7 天	10000RMB	缘起最早对日本产生兴趣是2012读大学的时候，得益于《时尚旅游》杂志一期关于日本…
4722	['身未动，心已远。记录我们…	广东省	Male	2016-09-11	和朋友	8 天	10000RMB	时间:2016/9/11-2016/9/18 …
4726	['游购在日本']	None	None	2017-03-16	小两口	6 天	10000RMB	游购在日本2017年3月16日，我和老公打着给孩儿断奶的旗号，开始了为期6天的日本…
4729	['霓虹邂逅儿记——日本关东…	陕西省	FeMale	2016-10-15	小两口	12 天	10000RMB	阿孔与小番茄的梦想是环游世界，先来惯…
4742	['【青团亲子游】暑期OKINAW…	江苏省	Male	2016-07-08	带小孩	8 天	10000RMB	【GJG之卡帕莱KAPALAI之旅-2岁宝宝亲子游】:http://www.mafengwo.cn/i/1293639.…
4769	['日本の物语	樱花季的浪…	辽宁省	Male	2015-03-31	和朋友	7 天	10000RMB
4771	['日本6天5夜，静冈--京都-…	None	None	2016-03-30	和朋友	6 天	10000RMB	这是我的一场说走就走的旅行，也就是跟朋友聊天谈到出去。两人一拍即合就说好。…
4772	['【东京·镰仓·箱根】你猜我穿…	北京	Male	2017-04-28	和朋友	4 天	10000RMB	终章：走马灯2017年5月，东京-镰仓-横滨…
4779	['再游日本']	北京	FeMale	2017-04-28	小两口	6 天	10000RMB	第二次登上这个国家的土地，时隔两年不…
4787	['五月の北陸と関西，这是一…	辽宁省	FeMale	2017-04-29	一个人	9 天	10000RMB	说在前面首先，最最重要的，是感谢领导给了我这次日本出差的机会，使得我可以在…

て16,734篇の旅行記に対して形態素分析を行った。その結果、位置情報（観光スポット名と日本の都道府県＋自治体名）を表す単語2,495個、ならびに名詞96,386個、合計98,881個の形態素を抽出できた。それらをトピックモデルに導入し、分析を行った。

(2) トピックモデルの生成過程

1.For トピック k=1,・・・,K

(a) 形態素分布を生成$\emptyset_k$～Dirichlet（$\boldsymbol{\beta}$）

2.For 文書 d=1,・・・,D

(a) トピック分布を生成$\boldsymbol{\theta}_d$～Dirichlet（$\boldsymbol{\alpha}$）

(b) For 形態素 n=1,・・・,N_d

i.トピック生成z_{dn}～Categorical($\boldsymbol{\theta}_d$)

ii.形態素を生成w_{dn}～Categorical($\emptyset_{z_{dn}}$)

この中で、

$\emptyset_k$：トピックkの形態素分布

$\boldsymbol{\theta}_d$：文書dのトピック分布

N_d：文書dに含まれる形態素数

z_{dn}：文書dのn番目のトピック

w_{dn}：文書dのn番目の形態素

$\emptyset_{z_{dn}}$：文書dのn番目のトピックの分布

まず、生成するトピック数を設定し、BOW の中の単語がハイパーパラメータ$\boldsymbol{\beta}$を持つディリクレ分布に従ってトピックごとに単語分布$\emptyset_k$を生成する。つぎ、ハイパーパラメータ$\boldsymbol{\alpha}$を持つディリクレ分布に従って文書 d ごとのトピック分布$\boldsymbol{\theta}_d$を生成させる。そして、文書ごとのトピック分布$\boldsymbol{\theta}_d$を持つカテゴリ分布に従い、トピックZ_{dn}を生成させ、トピックz_{dn}を持つカテゴリ分布に従い、単語W_{dn}を生成する。

トピック分布$\boldsymbol{\theta}_d$と形態素分布集合$\emptyset$が与えられた際の旅行記W_dの生起確率は、(1)式で示される。

$$P(w_d|\boldsymbol{\theta}_d, \Phi) = \prod_{n=1}^{N_d}\sum_{k=1}^{K} P(z_{dn}=k|\boldsymbol{\theta}_d)\, P(w_{dn}|\emptyset_k)$$

$$= \prod_{n=1}^{N_d}\sum_{k=1}^{K} \theta_{dk}\emptyset_{kw_{dn}} \quad \text{... (1)式}$$

なお、トピックモデルを生成する際に、ステップ1およびステップ2はハイパーパラメータ$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$を持つディリクレ分布に従って生成したが、ディリクレ分布は多項分布の「共役事前分布（conjugate prior）」であり、過学習を避けるために事前分布としてディリクレ分布を設定すること、データ数に対してパラメータが多い場合や比率が小さいセルが多い場合に偏った結果が導かれる危惧があることを避けるためである[2), 3)]。

ディリクレ分布は、$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_V)(\beta_i > 0)$（V：全文書の中で現れる形態素の種類数）をパラメータとして、

$$Dirichlet(\emptyset|\boldsymbol{\beta}) = \frac{\Gamma(\sum_{v=1}^{V}\boldsymbol{\beta}_v)}{\prod_{v=1}^{V}\Gamma(\boldsymbol{\beta}_v)}\prod_{v=1}^{V}\emptyset_v^{\boldsymbol{\beta}_v-1} \quad \text{... (2)式}$$

と定義される。

さて、トピック数は分析者が任意に設定した中で、モデルの説明力にもとづいて最終的に決定するのが一般的である。確率モデルの性能を評価する尺度とし、データに対するパープレキシティ（Perplexity, PPL）が使われる。パープレキシティは分岐数または選択肢の数を表しており、確率の逆数で定義されている。負の対数尤度から計算できる値で、低いパープレキシティはデータを高い精度で予測できるよい確率モデルであることを示す[2), 3)]。PPLは、下記のように示せる。

$$\text{Perplexity}(\boldsymbol{W}|M) = \left(-\frac{\sum_{d=1}^{D}\log\ P(W_d|M)}{\sum_{d=1}^{D}N_d}\right) \quad \text{... (3)式}$$

この中で、$\log\ P(W_d|M)$は確率モデルM（トピックモデル）の対数尤度である。

5. 分析結果

(1) トピック数の決定

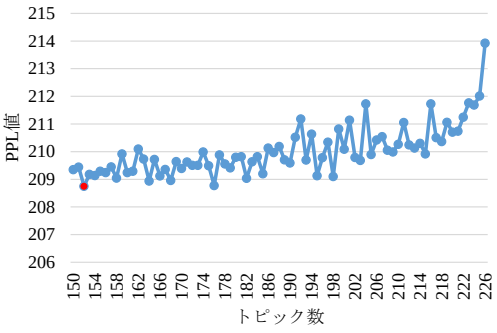

図-3 訪日中国人旅行者人数と消費額の割合

トピック数を決定するために、トピック数を1から276まで25ずつ変えて推定したところ、151~226までがPPLが低かった。そこで、この区間で１ずつトピック数を変化させて場合のPPLを算出した（図-3）。この時、152トピックの時に最小となった。

表-1　代表的なトピックの例

トピック	形態素（15個）				
トピック1 8.25%	地方 5%	感覚 4%	時間(time) 3%	旅程 2%	友達 2%
	写真 2%	観光スポット 2%	時間(hour) 1%	旅行記 1%	アドバイス 1%
	地図 1%	体験 1%	計画 1%	天気 1%	交通 1%
トピック2 6.88%	空港 4%	ホテル 3%	荷物 3%	時間 2%	行列 2%
	時間 2%	飛行機 2%	円 1%	地下鉄 1%	地方 1%
	従業員 1%	感覚 1%	ラーメン 1%	味 1%	スーツケース 1%
トピック22 1.40%	富士山 13%	忍野八海 5%	銀座 3%	ガイドさん 3%	空港 3%
	金閣寺 2%	ホテル 2%	大阪城公園 2%	時間(time) 2%	富士山五合目 2%
	公園 1%	地方 1%	買物 1%	五合目 1%	時間(hour) 1%
トピック49 0.37%	紅葉 14%	東福寺 9%	楓の葉 6%	禅林寺 4%	永観堂 3%
	琉璃 2%	常寂光寺 2%	時間(time) 2%	清水寺 2%	瑠璃光院 2%
	南禅寺 2%	宝厳院 1%	地方 1%	庭 1%	楓 1%
トピック75 0.16%	円 26%	高野山 12%	熊野古道 2%	奥之院 2%	金剛峯寺 1%
	大社 1%	おばさん 1%	人吉市 1%	ロープウェイ 1%	熊野那智大社 1%
	大師 1%	社寺 1%	生駒 1%	宿泊 1%	朝のお祈り 1%

(2)　トピックに対する集約

16,734篇の旅行記が152個のトピックに分類できたが、表-1は各トピックの全文書に占める構成比率や主要単語（下の数値：各トピックにおける形態素の構成比率（各トピックが表している主題を判断するための参考値））を示している。上位トピックはゴールデンルートでの伝統的な旅行に関連するトピックが抽出されたが、下位トピックほどトピックの構成比率が減少し、地方でのニッチな旅行が多くなる傾向を示した。

次に、152個のトピックは15の旅行地点あるいは旅行内容の情報を持つ単語から組み合わせるため、理解が用意になるように旅行地点を都道府県ごとに集約した。また、旅行内容は、訪日外国人消費動向調査（観光庁）や国民の観光に関する動向調査（観光の実態と志向。日本観光振興協会）等を参考にしながら、主要な旅行行動を1)総合的な観光、2)歴史文化、3)グルメ、4)テーマパーク、5)大都市遊覧、6)季節・自然風景、7)購買行動、8)アニメ・ポップカルチャー、9)クルーズ、10）温泉旅館、11)大学文化体験、12)リゾート、13)アートツアー、14)親子旅行、15)イベント参加、16)スポーツツアー，17)その他、以上の17区分に集約した。具体的な集約の過程の例は表-2に示し、一部のトピックの集約した結果は表-3のように示される。

表-3　一部のトピックの集約した結果

トピック	訪問地	コンテンツ	構成比率
1	-	通常の旅行コンテンツ	8.25%
2	-	通常の旅行コンテンツ	6.88%
11	大阪府	テーマパーク	2.55%
19	京都府	歴史文化	1.78%
40	-	通常の旅行コンテンツ	0.47%
41	東京都	大都市遊覧	0.47%
49	京都府	季節・自然風景	0.37%
60	神奈川県	アニメ	0.26%
67	-	通常の旅行コンテンツ	0.22%
70	静岡県	総合的な観光	0.21%
75	和歌山県	歴史文化	0.16%
89	島根県＋山口県	総合的な観光	0.12%
114	栃木県	季節・自然風景	0.07%
116	兵庫県	温泉旅館	0.07%
123	北海道	季節・自然風景	0.06%

表-2　集約の過程の例

トピック	形態素（15個）							
トピック1 8.25%	地方 5%	感覚 4%	時間(time) 3%	旅程 2%	友達 2%	写真 2%	観光スポット 2%	時間(hour) 1%
	旅行記 1%	アドバイス 1%	地図 1%	体験 1%	計画 1%	天気 1%	交通 1%	
トピック2 6.88%	空港 4%	ホテル 3%	荷物 3%	時間 2%	行列 2%	時間 2%	飛行機 2%	円 1%
	地下鉄 1%	地方 1%	従業員 1%	感覚 1%	ラーメン 1%	味 1%	スーツケース 1%	
トピック22 1.40%	富士山 13%	忍野八海 5%	銀座 3%	ガイドさん 3%	空港 3%	金閣寺 2%	ホテル 2%	大阪城公園 2%
	時間(time) 2%	富士山五合目 2%	公園 1%	地方 1%	買物 1%	五合目 1%	時間(hour) 1%	
トピック49 0.37%	紅葉 14%	東福寺 9%	楓の葉 6%	禅林寺 4%	永観堂 3%	琉璃 2%	常寂光寺 2%	時間(time) 2%
	清水寺 2%	瑠璃光院 2%	南禅寺 2%	宝厳院 1%	地方 1%	庭 1%	楓 1%	
トピック75 0.16%	円 26%	高野山 12%	熊野古道 2%	奥之院 2%	金剛峯寺 1%	大社 1%	おばさん 1%	人吉市 1%
	ロープウェイ 1%	熊野那智大社 1%	大師 1%	社寺 1%	生駒 1%	宿泊 1%	朝のお祈り 1%	

訪問地	-
旅行コンテンツ	通常の旅行コンテンツ
訪問地	-
旅行コンテンツ	通常の旅行コンテンツ
訪問地	東京都＋山梨県＋大阪府＋京都府
旅行コンテンツ	総合的な観光
訪問地	京都府
旅行コンテンツ	季節・自然風景
訪問地	和歌山県
旅行コンテンツ	歴史文化

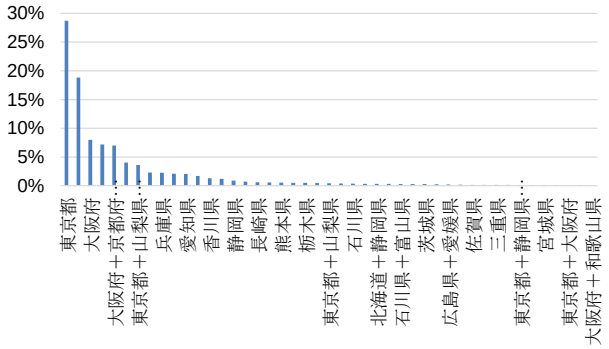

図-4　集約した旅行地点の構成比率図

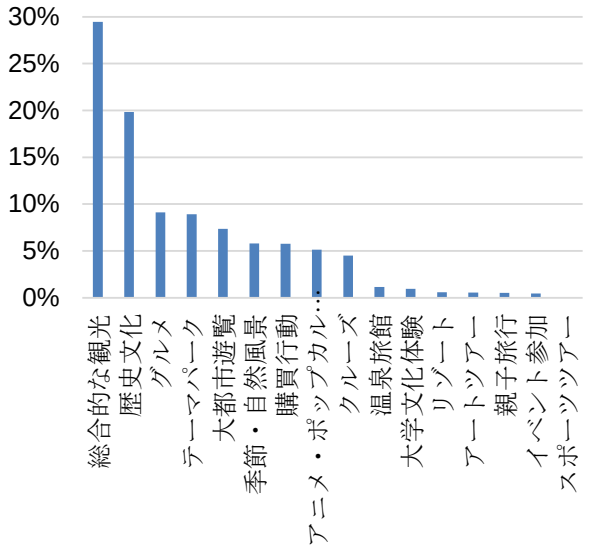

図-5　集約した旅行内容の構成比率図

集約した旅行地点の構成比率（図-4）をみると、訪問地点の上位10位の中、ゴールデンルート（多数の都道府県を訪問）は全体の12.7%を占めるのに対し、1つの都道府県のみ訪問する形は全体の71.3%を占めている。そのため、現在の訪日中国人の全体的な訪問地点の主な特徴は1都道府県を重点的に訪問することであると分かった。本研究は、このような訪問パターンを「深度游地域」と呼ぶ。そして、訪問地点の上位3位（55.5%）は東京都、京都府、大阪府のため、中国人旅行者は都市部への訪問が多く、地方への誘致は不十分と言える。また、旅行内容を集約した図-5から、「総合的な観光、歴史文化、グルメ、テーマパーク、大都市遊覧、季節・自然風景、購買行動、アニメ・ポップカルチャー、クルーズ」の旅行内容は全体の9割以上を占め、訪日中国人旅行者にとって大衆的な旅行内容であると考えられる。一方「温泉旅館、大学文化体験、リゾート、アートツアー、親子旅行、イベント参加、スポーツツアー」の旅行内容の割合は低く、訪日中国人旅行者にとってニッチな旅行内容であると考えられる。

(3)　「深度游地域」での旅行形態の抽出・特徴の把握

訪日中国人旅行者は「深度游地域」でいかに旅行を行っているか、さらに、収集した属性（性別・居住地・訪問季節・同行者）による特徴を明確するため、「深度游地域」での旅行形態の抽出および特徴の把握を行った。

a)　東京都における旅行形態の抽出および特徴把握

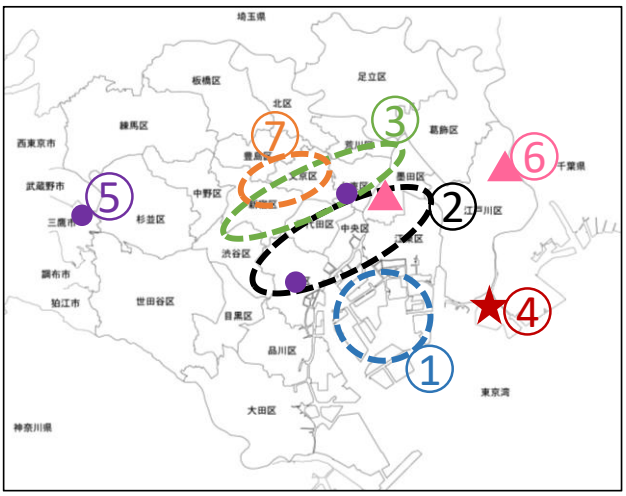

図-6　東京都における7つ旅行形態

表-4　東京都における各旅行形態の特徴

主題と代表トピック	性別		居住地						
	男性	女性	北京	上海	東北	華北	華中	華南	西部
①お台場巡り 36・139	0.07%	−0.02%	0.08%	0.01%	0.01%	0.00%	0.02%	0.03%	−0.02%
②定番巡り 13・14・29・34・37・54	−0.05%	0.03%	0.22%	−0.11%	0.03%	−0.06%	0.00%	0.07%	0.09%
③桜鑑賞 51・101	0.04%	0.00%	0.06%	−0.03%	0.03%	0.02%	−0.04%	0.03%	0.00%
④テーマパーク 25	−0.28%	0.07%	0.41%	0.03%	0.73%	0.25%	−0.06%	−0.46%	−0.15%
⑤アニメ 56・60・98	0.06%	−0.02%	0.07%	−0.07%	−0.04%	0.02%	−0.03%	0.04%	0.06%
⑥花火大会体験 68	−0.01%	0.01%	−0.02%	0.14%	−0.09%	−0.09%	−0.03%	−0.11%	−0.03%
⑦大学文化体験 44	0.23%	−0.07%	−0.06%	−0.11%	0.06%	−0.08%	−0.07%	0.09%	0.06%
主題と代表トピック	**季節**				**同行者**				
	春	夏	秋	冬	1人	夫婦	友達	親子	家族
①お台場巡り 36・139	0.00%	0.05%	0.00%	0.12%	0.07%	0.03%	0.00%	0.19%	−0.01%
②定番巡り 13・14・29・34・37・54	0.08%	−0.04%	−0.04%	0.09%	−0.11%	0.16%	0.07%	−0.30%	−0.15%
③桜鑑賞 51・101	0.18%	−0.04%	−0.08%	−0.04%	0.03%	−0.01%	0.01%	0.02%	0.03%
④テーマパーク 25	0.06%	0.35%	−0.28%	−0.12%	−0.80%	−0.23%	−0.31%	3.22%	−0.29%
⑤アニメ 56・60・98	0.01%	0.03%	0.01%	−0.01%	0.10%	0.00%	−0.01%	0.12%	−0.08%
⑥花火大会体験 68	−0.18%	0.47%	−0.15%	−0.13%	0.01%	−0.08%	0.12%	−0.11%	−0.02%
⑦大学文化体験 44	−0.06%	0.08%	−0.11%	0.02%	0.32%	−0.11%	−0.02%	−0.23%	−0.04%

東京都に関連するトピックに着目し（トピック13、14、25、29、34、36、37、44、51、54、56、60、68、98、101、139が該当）、それぞれの頻出地点・区域を示したものが図-6である。7つの地点・区域と同時に出現する旅行内容（名詞）の特徴を考えると、図-6と表-4に示すように、トピック36、139：お台場を中心とする①お台場巡り、トピック13、14、29、34、37、54：渋谷→東京タワー→浅草寺→スカイツリーを訪問地とする②定番巡り、トピック51、101：代々木＋新宿御苑＋上野を訪問地とする③桜鑑賞、トピック25：ディズニーランドを訪問する④テーマパーク、トピック56、60、98：三鷹の森ジブ

リ美術館＋東京タワー（ワンピース）＋秋葉原などを訪問地とする⑤アニメ、トピック68：隅田川または江戸川を訪問地とする⑥花火大会体験、トピック44：早稲田大学、東京大学などの大学を訪問する⑦大学文化体験に分類でき、東京都における旅行形態の差異を抽出できた。

そして、トピックモデルの結果とする「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」の平均値と「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、カラースケールは濃ければ濃いほど、所属している旅行形態のニーズが強いと判断できる。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。

①お台場巡り：男性、北京居住、冬に訪問、1人または親子が同行者とする。

②定番巡り：女性、北京または西部居住、春または冬に訪問、夫婦または友達を同行者とする。

③桜鑑賞：性別の特徴がなく、北京・東北または華南居住、春に訪問、1人旅または家族を同行者とする。

④テーマパーク：女性、北京・華北・東北居住、夏に訪問、親子を同行者とする。

⑤アニメ：男性、北京または西部居住、夏に訪問、1人旅または親子を同行者とする。

⑥花火大会参加：女性、上海居住、夏に訪問、1人旅または友達を同行者とする。

⑦大学文化体験：男性、東北・西部または華南居住、夏または冬に訪問、1人旅の特徴。

b)　京都府における旅行形態の抽出および特徴把握

次に、京都府で頻出するトピックをみると、図-7および表-5に示すように、トピック12、19、39、76、82：京都市中心部の社寺を主な訪問地とする①古都巡り、トピック30、49、52、99：京都市の西にある亀岡市の保津峡から京都市の南禅寺までの区域を訪問地とする②紅葉鑑賞、トピック55：宇治市を主な訪問地とする③抹茶体験の3つ旅行形態を抽出できた。

東京都と同様に各旅行形態の特徴を抽出すると、下記のようにまとめることができる。

①古都巡り：性別の特徴がなく、華中・華南または上海居住、春に訪問、夫婦または友達を同行者とする。

②紅葉鑑賞：男性、上海または華南居住、秋に訪問、1人旅の特徴。

③抹茶体験：男性、北京または華南居住、夏に訪問、家族を同行者とする。

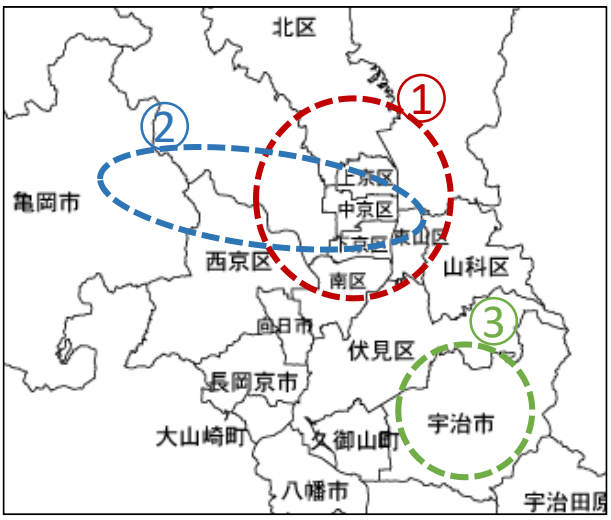

図-7　京都府における3つ旅行形態

表-5　京都府における各旅行形態の特徴

旅行形態と代表トピック	性別		居住地						
	男性	女性	北京	上海	東北	華北	華中	華南	西部
①古都巡り 12・19・39・76・82	−0.14%	−0.05%	−0.05%	0.12%	−0.33%	−0.10%	0.10%	0.17%	0.04%
②紅葉鑑賞 30・49・52・99	0.05%	0.00%	0.00%	0.06%	−0.14%	−0.05%	−0.08%	0.05%	−0.02%
③抹茶体験 55	0.06%	0.02%	0.02%	−0.01%	−0.20%	−0.18%	−0.05%	0.04%	−0.09%
旅行形態と代表トピック	季節				同行者				
	春	夏	秋	冬	1人	夫婦	友達	親子	家族
①古都巡り 12・19・39・76・82	0.25%	0.10%	0.15%	−0.06%	−0.15%	0.25%	0.24%	−0.32%	0.18%
②紅葉鑑賞 30・49・52・99	−0.03%	−0.08%	0.29%	−0.14%	0.11%	0.02%	0.01%	−0.10%	0.02%
③抹茶体験 55	0.01%	0.03%	0.00%	−0.02%	0.05%	−0.07%	0.04%	−0.10%	0.16%

c)　大阪府における旅行形態の抽出および特徴把握

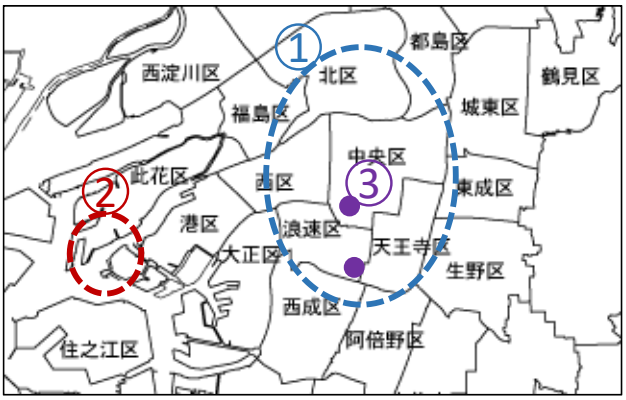

図-8　大阪府における3つ旅行形態

表-6　大阪府における各旅行形態の特徴

旅行形態と代表トピック	性別		居住地						
	男性	女性	北京	上海	東北	華北	華中	華南	西部
①定番巡り 18	0.11%	−0.07%	−0.24%	0.70%	−0.99%	0.69%	0.23%	−0.25%	−0.13%
②テーマパーク 11、45	−0.68%	0.87%	0.19%	−0.33%	−0.61%	1.41%	−0.70%	−0.86%	1.03%
③グルメ体験 91	0.00%	0.00%	0.07%	−0.09%	0.16%	−0.19%	0.05%	−0.02%	−0.01%
旅行形態と代表トピック	季節				同行者				
	春	夏	秋	冬	1人	夫婦	友達	親子	家族
①定番巡り 18	0.53%	−0.42%	0.17%	−0.17%	−0.38%	1.03%	−0.45%	−0.90%	1.10%
②テーマパーク 11、45	−0.68%	0.59%	0.25%	−0.86%	−1.25%	2.38%	−0.41%	−0.07%	−1.14%
③グルメ体験 91	0.01%	−0.01%	0.09%	−0.07%	−0.06%	0.14%	−0.06%	−0.10%	0.05%

大阪府では、図-8および表-5に示すように、トピック18：梅田スカイビル→大阪城公園→通天閣→四天王寺を主な訪問地とする①定番巡り、トピック11、45：USJを訪問地とする②テーマパーク、トピック91：法善寺横丁

または新世界を訪問地とする③グルメ体験の3つ旅行形態を抽出できた。

また、各旅行形態の特徴は表-6に示す通りである。

①定番巡り：男性、華北・華中または上海居住、春または秋に訪問、夫婦または家族を同行者とする。

②テーマパーク：女性、北京・華北または西部居住、夏または秋に訪問、夫婦を同行者とする。

③グルメ体験：性別の特徴がなく、北京・華中または東北居住、秋に訪問、夫婦または家族を同行者とする

d) 神奈川県における旅行形態の抽出および特徴把握

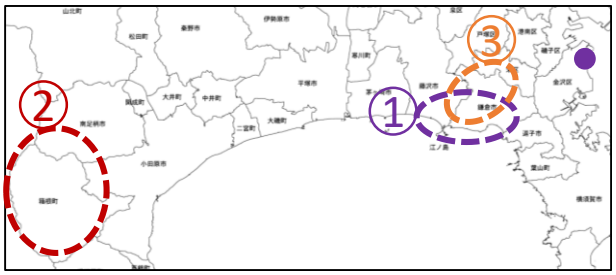

図-9 神奈川県における3つ旅行形態

表-7 神奈川県における各旅行形態の特徴

旅行形態と代表トピック	性別		居住地						
	男性	女性	北京	上海	東北	華北	華中	華南	西部
①アニメ 21、60、112	0.06%	0.00%	0.12%	−0.15%	0.05%	−0.17%	0.05%	0.18%	−0.12%
②箱根巡り 32、73、106、108	0.00%	0.00%	0.08%	−0.02%	0.03%	0.07%	−0.03%	0.00%	−0.08%
③紫陽花鑑賞 115	0.02%	−0.03%	0.03%	0.00%	−0.08%	0.06%	−0.05%	0.01%	0.02%
旅行形態と代表トピック	季節				同行者				
	春	夏	秋	冬	1人	夫婦	友達	親子	家族
①アニメ 21、60、112	0.18%	0.32%	−0.46%	−0.09%	0.76%	−0.66%	0.33%	−0.57%	−0.22%
②箱根巡り 32、73、106、108	0.01%	−0.03%	0.00%	0.04%	−0.10%	−0.03%	−0.01%	0.17%	0.04%
③紫陽花鑑賞 115	−0.01%	0.11%	−0.11%	0.00%	0.11%	−0.08%	0.01%	−0.07%	0.00%

神奈川県では、図-9および表-7に示すように、トピック21、60、112：鎌倉高校前＋湘南海岸＋江ノ島を主な訪問地とする①アニメ、トピック32、73、106、108：箱根町を主な訪問地とする②箱根巡り、トピック115：明月院をはじめ、北鎌倉を訪問地とする③紫陽花鑑賞の3つ旅行形態を抽出できた。

また，各旅行形態の特徴は表-7に示す通りである。

①アニメ：男性、北京または華南居住、春または秋に訪問、1人旅または友達を同行者とする。

②箱根巡り：女性、北京・華北または東北居住、春または冬に訪問、親子または家族を同行者とする。

③紫陽花鑑賞：性別の特徴がなく、北京、華北または西部居住、夏に訪問、1人旅の特徴。

e) 北海道における旅行形態の抽出および特徴把握

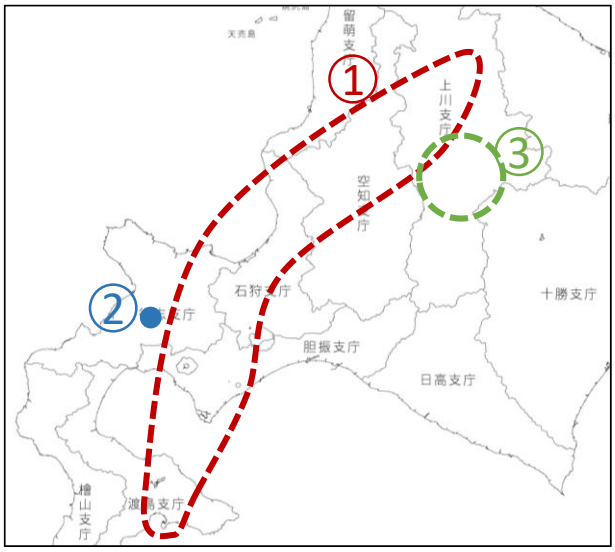

図-10 北海道における3つ旅行形態

表-8 北海道における各旅行形態の特徴

旅行形態と代表トピック	性別		居住地						
	男性	女性	北京	上海	東北	華北	華中	華南	西部
①定番巡り 23	0.33%	−0.04%	0.00%	0.00%	−0.22%	0.30%	−0.09%	0.35%	0.15%
②スポーツ 152	0.00%	0.00%	−0.01%	0.00%	−0.01%	0.00%	0.00%	0.01%	0.00%
③道中花巡り 71	0.08%	−0.02%	−0.04%	−0.01%	−0.15%	−0.05%	0.03%	−0.01%	−0.01%
旅行形態と代表トピック	季節				同行者				
	春	夏	秋	冬	1人	夫婦	友達	親子	家族
①定番巡り 23	−0.60%	−0.14%	−0.49%	2.13%	0.04%	0.35%	−0.02%	−0.01%	0.07%
②スポーツ 152	−0.01%	−0.01%	0.00%	0.01%	0.00%	0.00%	0.00%	−0.01%	0.00%
③道中花巡り 71	−0.13%	0.29%	−0.04%	−0.11%	0.04%	0.03%	−0.02%	0.00%	0.00%

北海道では、図-10に示すように、トピック23：北海道の南から函館→登別→小樽→旭川市までの区域を訪問地とする①定番巡り、トピック152：ニセコを訪問する②スポーツ、トピック71：富良野または美瑛町を主な訪問地とする③道央花巡りの3つ旅行形態を抽出できた。

各旅行形態の特徴は、表-8に示す通りである。

①定番巡り：男性、華北または華南居住、冬に訪問、夫婦を同行者とする。

②スポーツ：男性、華南居住、秋または冬に訪問、1人旅または友達を同行者とする。

③道央花巡り：男性、華中居住、夏に訪問、1人旅または夫婦を同行者とする。

以上から、5つの「深度游地域」における旅行形態を抽出し、特徴の把握ができた。

6. まとめ

本研究は、Web クローラを用いて収集された 16,734 篇の旅行記を収集し、形態素分析によっておよそ 10 万個の形態素を抽出した。それを用いて、トピックモデルによって旅行記を構成するトピックの抽出を行ったところ、152 個のトピックを抽出できるとともに、その集約を行いながら個人属性とのクロス集計を行った。その結果、訪日中国人旅行者が書いた旅行記の特徴を明確し、それによって訪日中国人旅行者の旅行行動を明確できた。具体的には、ゴールデンルートの構成比率が 12.7%に留まるのに対して、1 つの都道府県のみ訪問する深度游パターンが 71.3%を占めることがわかった。

これまで訪問地点の情報から訪問パターンの分析事例が多かった中で、実際の行動内容や評価が記述された旅行記を用いることによって、それらを反映したセグメント構築が出来る点が特徴と言える。一方で、分析データであるネット上の旅行記は、必ずしも旅行者全体と比較して偏りない代表性を有するとは言えない問題点も考えられる。他のデータと比較しながら、分析結果の特徴を比較し、データごとの特性を明らかにすることが今後、必要不可欠といえる。

さらに今後の課題として、データ数を増やしながら、トピック数の集約方法の検討を上げることができる。また、トピックモデルはトピックの間の関係性を考慮しないため、分析データが増えるにつれ、トピックの数が多くなり、考察することがさらに難しくなる恐れがある。以上の問題を改善することも今後の課題である。

参考文献

1) 古屋秀樹・劉瑜娟：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, pp.I_571-I_583, 2016.
2) 古屋秀樹：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, 2016.
3) 古屋秀樹・岡本直久・野津直樹：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, 2017.
4) 加藤耕太：Pythonクローリング&スクレイピング:データ収集・解析のための実践開発ガイド, 技術評論社, 2017.
5) 佐藤一誠：トピックモデルによる統計的潜在意味解析, コロナ社, 2015.
6) 岩田具治：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, 2015.
7) 宋紫龍,古屋秀樹：訪日中国人旅行者の旅行記を用いた旅行情報抽出方法の基礎的分析, 日本観光研究学会全国大会学術論文集, 32, pp.109-112, 2017.

(2018. 4. 25 受付)

ANALYSIS OF CHINESE TOURISTS' TRAVEL LITERATURES BY LATENT DIRICHLET ALLOCATION MODEL

Zilong SONG and Hideki FURUYA

It is necessary to grasp their travel behaviors because of the situation of demand changing of Chinese tourists who is the one of the largest markets in Japanese in-bound tourism. Furthermore, according to the sites where tourists have traveled and evaluation of tourists' intention for visiting area, it is possible to understand the reason and preference of the tourism behavior. It is also possible to consider the future tourism behavior trends through tourists' decision making on the Japanese travel. So, we focus on the contents of the trip and travel literatures which were written by Chinese travelers of the experience in Japanese visiting travel by using a topic model. The Latent Dirichlet Allocation model which is one of the topic typed model, are applied for analyzing the pattern of the visiting places combination. This model is one of the functions for machine learning to estimate the similarities between the travel literatures contents. The main purpose of this paper is to identify the characteristics of the literatures by Chinese tourists. Using the travel literatures on the Web site, 152 topics are identified which are different for each other topics. Based on the classified result, we clarified the intention and behavior of Chinese tourists who visited Japan.