トピックモデルによる訪日中国人観光客の旅行経験の分析
ー中国国内の観光ウェブサイト上の旅行記よりー

HUO　MINGMING[1]・榊原　弘之[2]

[1]非会員　山口大学大学院　創成科学研究科（〒755-8611 山口県宇部市常盤台 2-16-1）

E-mail: a502wdu@yamaguchi-u.ac.jp

[2]正会員　山口大学教授　創成科学研究科（〒755-8611 山口県宇部市常盤台 2-16-1）

E-mail: sakaki@yamaguchi-u.ac.jp

近年，訪日中国人観光客 1 人当たりの旅行消費額は減少傾向にある．「爆買い」の沈静化とともに，中国人観光客の旅行形態は買い物のみならず，美食，文化，景観などのテーマに多様化してきたと考えられる．本研究は中国語観光ウェブサイトに掲載されている，訪日中国人観光客が書き込んだ旅行記を分析し，訪問パターン，旅行テーマを特定し，するとともにその変化を分析する．具体的には，自然言語処理手法である LDA モデルと Word2Vec モデルを組み合わせ，訪日中国人観光客の旅行テーマや訪問パターンを明らかにする．さらに地方部を対象に，旅行者にとってより印象的，魅力的な地域資源を明らかにすることを試みる．

Key Words : *topic model, web travel notes, chinese tourists, inbound tourism, tourism behavior*

1.　はじめに

2019 年の訪日外国人数は 3,188 万人，訪日外国人の旅行消費額は 4 兆 8,135 億円であり，ともに過去最高となった．そのうち，訪日中国人観光客数は 857 万人，旅行消費額は 1 兆 7,704 億円であり，いずれも国別で最多であり，日本のインバウンド観光において重要な部分を占めている[1]．図-1[2]に示すように，2014 年から 2015 年にかけて訪日中国人の 1 人当たり旅行消費額は急増し，「爆買い」呼ばれる現象も生じたが，2016 年以降は訪日中国人の 1 人当たり旅行消費額は減少傾向にある．「爆買い」の沈静化とともに，訪日旅行の目的は買物のみならず，美食，文化，景観などに多様化してきたと考えられる[3]．一方，日本政府観光局（JNTO）の統計から，訪日中国人観光客の旅行形態は個人手配による旅行や訪日回数が複数回あるリピーターが増加し[4]，首都圏（東京都，神奈川県，埼玉県，千葉県），京阪神，愛知県以外の地方部への来訪が増加傾向にあり[5]，訪問地も多様化の傾向があると推測される．

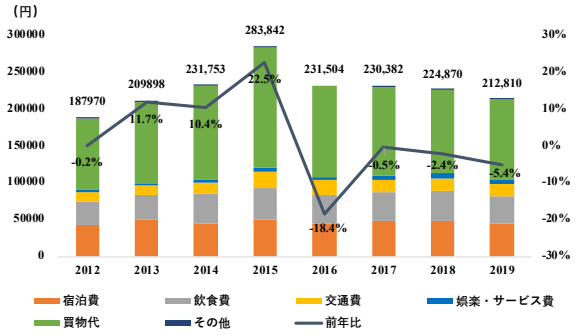

図-1　訪日中国人観光客 1 人あたり旅行消費額
（参考文献[2]より作成）

以上より，訪日中国人観光客の変化するニーズや観光行動を把握し，インバウンド観光についての施策を立案することが必要であると考えられる．このような情報を得るためには，一般的に訪日中国人観光客を対象にアンケート調査やヒアリング調査を実施する．しかしそれらの調査手法は，人手により行われており，時間や労力を要するという問題を抱えていた．この問題に対し，観光ウェブサイトに掲載されている観光客の旅行経過や感想が記されている旅行記，すなわち旅行ブログエントリを収集，分析するという方法が近年では広まりつつある．そこで，本研究は中国語観光ウェブサイトに掲載されている，訪日中国人観光客が書き込んだ旅行記のコンテンツを用

いて，自然言語処理手法である LDA モデルと Word2Vec モデルを組み合わせ，訪日中国人観光客の旅行テーマを特定，都道府県単位の訪問パターンと市町村単位の訪問パターンを明らかにする．

一方，2020 年 1 月末以降は，新型コロナウイルス感染症の拡大に伴い，観光業が大きな影響を受けるに至っている．2021 年現在においても，訪日中国人観光客は 2019 年以前と比較して大幅に減少している．感染の終息後も，旅行形態は多人数による集団行動が避けられ，個人・少人数行動を志向するようになり，団体旅行から個人旅行への変化が加速する可能性が高いと考えられる．

本研究は新型コロナウイルス感染の終息後，外国人訪日旅行の回復を目的とした政策立案のための基礎的知見を得るために，訪日中国人観光客の旅行テーマと，訪問パターンを特定し，それらの経年変化を明らかにすることを目的とする．さらに地方部を対象に，旅行者にとってより印象的，魅力的な地域資源を明らかにすることを試みる．

2. 既往研究と本研究の位置付け

本章では，訪日外国人観光客の行動に関する先行研究をまとめるとともに，本研究の位置付けを述べる．金 [6]は中国の旅行会社で販売されている日本旅行のツアー商品に着目して，訪日中国人観光客の観光行動に見られる訪問地の特徴を明らかにした．菱田ら [7]は観光庁の訪日外客訪問地調査データを用いて，訪日中国人観光客の居住地域別の観光行動の違いを分析して，訪問地選択の変化を明らかにした．一方，松井ら [8]は「訪日外国人消費動向調査」データを用いて，訪問地傾向と観光活動を組合せて分析するとともに，その変化を明らかにしている．古屋・劉 [9]も「訪日外国人消費動向調査」データを用いて，潜在クラスモデルによって類似性から外国人観光客の都道府県への訪問パターンを導出して，各訪問パターンの構成比率と主要国籍・地域，旅行形態，旅行時期，訪日回数などの要因との関連性を明らかにした．さらに古屋 [10]は，「訪日外国人消費動向調査」データを用いて，hPAM（Hierarical Pachinko Allocation Model）によって訪問場所の組合せパターンを分類するとともに，国籍・地域や訪日回数と訪問パターンとの関連性を明らかにした．特に都道府県区分より詳細な市町村，観光スポットを含む訪問パターンに着目している．

一方，本研究と同様に観光客の旅行記を分析した研究として，宋・古屋 [11]は旅行者の旅行経過や感想が記されている旅行記データを用いて，自然言語処理手法の LDA モデルにより訪問パターン及び旅行内容の類型化を行い，東京都を例として旅行記記入者の個人属性及び旅行情報と結びつけ，訪日中国人観光客の旅行行動及び特徴を明らかにした．

以上より，金 [6]菱田ら [7]の研究は，主に都道府県単位の中国人観光客の行動実態を分析している．また松井ら [8] 古屋・劉 [9] 古屋 [10] の研究は，観光庁統計の「訪日外国人消費動向調査」データを用いて，訪日外国人の訪問パターンや観光活動と個人属性の関係を明らかにしている．それに対して本研究と宋・古屋 [10]は，旅行記データに自然言語処理手法を適用している．旅行記は，記載内容の選択が記入者本人に委ねられている．従ってその内容は記入者本人に強い印象を与えた訪問地や体験に関する記述をより多く含んでいる可能性がある．そのような旅行記データを分析することで，訪日中国人観光客にとってより魅力的な訪問地，観光スポットや体験を理解することができると考えられる．宋・古屋 [10]は LDA モデルにより抽出されたトピックの主要な形態により，個別の訪問地での行動実態を明らかにしているが，旅行形態の詳細な分析は首都圏，関西圏地方などが中心である．一方本研究では，特に地方部への訪問パターンに着目する．旅行記で地方部の訪問パターンを特定し，それらの訪問地言及された地域資源を特定することができれば，地方部でも訪日中国人観光客にとって魅力的なコース設定が可能となると期待される．

本研究では，二つの中国語観光ウェブサイト MaFengWo（以下 MFW）と QiongYouWang（以下 QYW）に掲載されている訪日中国人観光客の旅行経過や感想が記されている旅行記の筆者の属性と旅行情報を収集，統計した．その結果，より多数の旅行記が得られる MFW のコンテンツを用いて，自然言語処理手法である LDA モデルと Word2Vec モデルを組み合わせ，訪日中国人観光客の旅行テーマと訪問パターンの分析を試みる．

3. 旅行記データ

(1) 旅行記データの概要

本研究では，プログラミング言語（Python）で構築した Web クローラーテクノロジーを利用して，二つの中国語観光ウェブサイト MFW と QYW に掲載されている訪日中国人観光客が記されている旅行記と記入者の属性や旅行情報を収集した．MFW は若い中国人の間で人気のある旅行ウェブサイトであり，2013 年には中国旅行研究院と連携し，「全球自由行報

告」を初めて発表した．また 2018 年に中国旅行研究院とともに「自由行大数据聯合実験室」を設立し，ユーザーの旅行情報，旅行記，コメント，トランザクションデータに基づいて，中国人観光客の観光行動を深く掘り下げ，定期的にレポートを作成，発表している．以上より，MFW からの旅行記データは信頼性や説得力が期待できると考えられる．一方，QYW は中国の海外旅行客に人気がある．さまざまな分野との連携があり，2016 年にはそれぞれ LinkedIn および日本政府観光局の北京事務所と連携し，「中国人観光客出境遊趨勢報告―日本篇」と「中国職場人出境遊行為報告」を発表した．2017 年には中国の招商銀行と連携し，「出境遊 85 後新青年旅行報告」を発表した．2019 年には LianTong ビッグデータ株式会社や銀聯智慧株式会社と連携し，「会玩的中国人―2019 五一旅遊大数据報告」を発表した．以上より，QYW からの旅行記データも一定程度の信頼性が期待できる．収集した旅行記データの概要を表-1 に示す．

表-1 収集した旅行記データの概要

観光ウェブサイト名		MaFengWo	QiongYouWang
収集期間		2021/4/14~2021/05/30	2021/3/22~2021/04/08
個人属性	性別	○	○
	年齢	×	○
	居住地	○	○
旅行情報	出発時間	○	×
	滞在期間	○	○
	同行者	○	○
	コスト	○	×
篇数		37259篇	2998篇

(2) 旅行記データの集計

本節では，(1) に示した二つの観光ウェブサイトに掲載されている旅行記の記入者の属性と旅行情報を集計し，それぞれ観光庁の統計データと比較し，差異の有無を確認する．

a) 個人属性

図-2 に訪日中国人観光客と旅行記記入者の性別分布を示す．MFW や QYW から集計した旅行記記入者と，観光庁統計の訪日中国人観光客は，どちらも女性の比率が高い．また図-3 に示すように，QYW から集計した記入者の年齢分布は，観光庁統計の訪日中国人観光客の年齢分布と近似している．すなわち，20，30 歳代の中国人観光客が訪日旅行の主力と考えられる．以上より，旅行記記入者の個人属性は，訪日中国人全体と比較的類似していると考えられる．

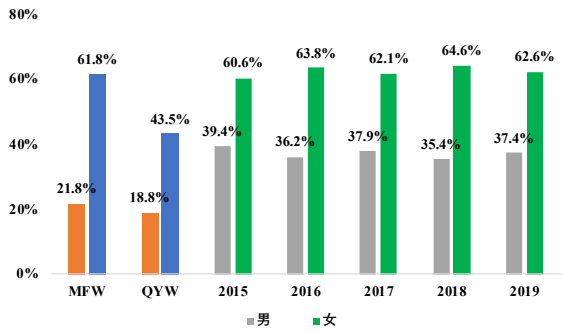

図-2 旅行記データと観光庁統計の性別

(参考文献 2)より作成)

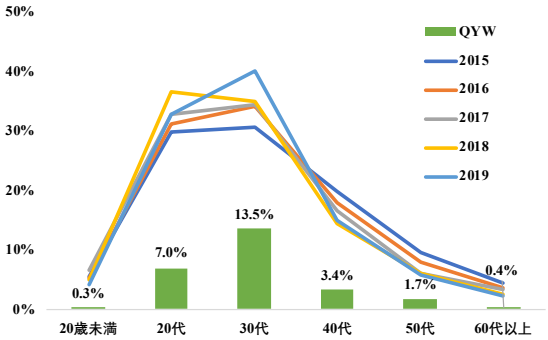

図-3 旅行記データと観光庁統計の年齢層

(参考文献 2)より作成)

b) 旅行情報

図-4 に月別で訪日中国人観光客と旅行記記入者の出発時期を示す．観光局統計では夏期の 7，8 月の訪日が最も多くなっているのに対し，MFW のデータで確認可能な旅行記で最多は 4 月であり，冬期（1，2 月），秋期（9，10 月）も比較的多い．この差異の要因として，冬期の雪，春期の桜，秋期の紅葉など，印象的な写真撮映が可能な季節には比較的多くの旅行記が投稿されるのでないかと推測される．また図-5 に示すように， MFW や QYW から集計した訪日中国人観光客の滞在期間は，7~13 日の比率が高いのに対し，観光庁統計の滞在期間は4~6 日の比率が高い．両者の差異の要因としては，MFW や QYW から集計した滞在期間は出発日から中国に戻るまでの期間を記録しているのに対し，観光庁統計のデータは日本国内の滞在のみを記録しているためと考えられる．さらに図-6 に示すように，MFW や QYW から集計した訪日中国人観光客の同行者は，恋人・配偶，家族と友人がほぼ平均的に分布しているが，観光庁統計の訪日中国人観光客の同行者は，家族の比率がもっ

とも高く，友人の比率も比較的高い．

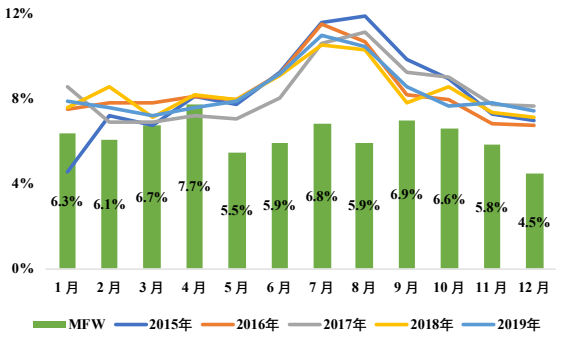

図-4 旅行記データと観光局統計の月別訪日中国人観光客
（参考文献 12)より作成）

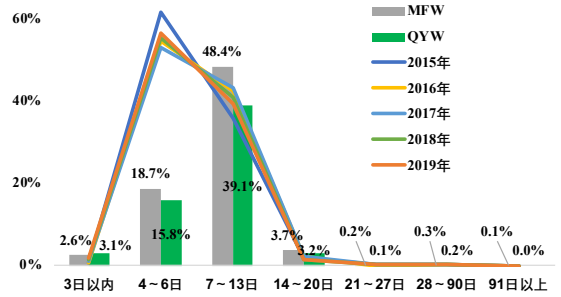

図-5 旅行記データと観光庁統計の滞在期間
（参考文献 2)より作成）

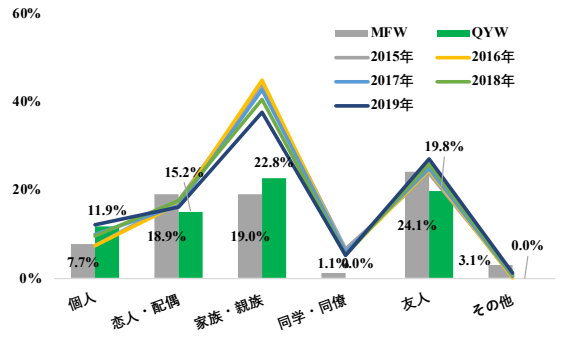

図-6 旅行記データと観光庁統計の同行者
（参考文献 2)より作成）

4. 分析手法

本研究では，中国国内の観光ウェブサイトに掲載されている，訪日中国人観光客が投稿した旅行記のコンテンツを分析する．まず，図-7 に示すように，自然言語処理手法である LDA モデルを適用して，各トピック特徴語の確率分布によって表される視点を取得し，訪日中国人観光客の旅行テーマと，都道府県単位の訪問パターンを抽出する．次に，Word2Vec モデルより，各訪問パターン（都道府県単位）中の地名に類似した単語および単語の確率分布を取得し，より詳細な市町村単位の訪問パターンを明らかにする．本章では各モデルについて説明する．

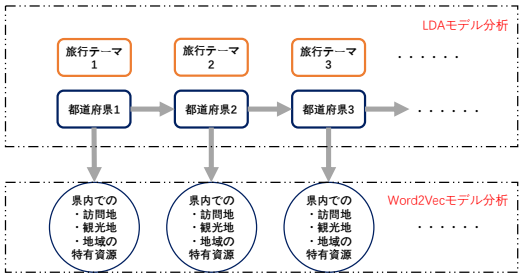

図-7 LDA モデルと Word2Vec モデルの組み合わせ

(1) LDA モデルについて

LDA （Latent Dirichlet Allocation；潜在的ディレクリ配分法）は，トピックモデルと呼ばれる手法の一種である．トピックモデルとは，文書のトピックを推定するモデルで，文書の分類や検索などに利用される手法である．LDA とは，自然言語処理方法で，教師データなしの機械学習の 1 つであり，各文書の複数のトピック確率分布と，各トピックの単語確率分布を予測できる 13) 14)．

LDA モデルを適用する前に，収集した旅行記データを前処理する必要がある．まず，各旅行記中のシンボル，数字を取り除く．次に Python の jieba（中国語の形態素解析器）パッケージを用い，以下の分析に使用しないストップワード（61991 個）を決定する．一方本研究のテーマの文脈特有のカスタムワード（5841 個）を jieba の辞書に追加し，各旅行記のテキストを形態素に分類する．最後に，頻度が 10 未満の形態素を除外し，合計 80,886 種類，27,814,431 個の形態素を得た．これらの形態素に LDA モデルを適用，分析を行う．

本研究では，Python の gensim パッケージを用いて，LDA モデル分析を行った．その際トピック数を決定する必要がある．今回は，LDA モデルの評価指標 Coherence[15] を用いて，トピック数を検討した．Coherence 値が高ければ高い程，良いモデルであると考えられる．本研究で使用したデータの場合，図-8 で示すように，トピック数 40 で，Coherence 値が最大となったため，計 40 個のトピックを設定した．また機械学習の繰り返し回数（passes）を 10 に設定し，

各旅行記に含まれるトピックの確率値（minimum_probability）を 0.01 以上と設定した．一方，複数回で LDA モデルを実行するときの結果を再現するために，乱数値（random_state）を 80 に設定して，LDA モデル分析を行った．表-2 に抽出されたトピックの一部を示す．各トピックで 15 個の主要な形態素を示している．トピックには，「桜」「花火大会」のような旅行テーマを示すものと，「広島―岡山―鳥取」のような訪問パターンを示すものが含まれる．

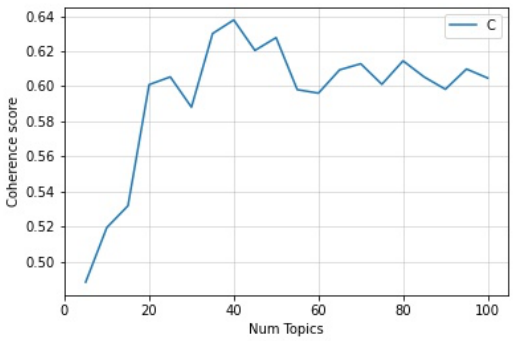

図-8 LDA モデルの評価指標（Coherence 値）

表-2 抽出されたトピックの例

トピック id	形態素（15個）								旅行テーマ・訪問パターン
5	0.172*"樱花" 0.009*"神社"	0.047*"公园" 0.009*"目黑川"	0.024*"樱花树" 0.008*"赏樱"	0.021*"哲学之道" 0.008*"上野公园"	"0.016*"樱花季" 0.008*"花瓣"	0.013*"盛开" 0.007*"造币"	0.012*"新宿御苑" 0.007*"平野神社"	0.011*"醍醐寺"	桜
6	0.145*"环球影城" 0.011*"门票"	0.038*"哈利波特" 0.010*"主题"	0.028*"排队" 0.010*"电影"	0.026*"大阪" 0.009*"城堡"	0.017*"usj" 0.008*"体验"	0.016*"过山车" 0.008*"魔法"	0.012*"小黄人" 0.007*"大白鲨"	0.012*"蜘蛛侠"	USJ
7	0.056*"花火大会" 0.017*"夏季"	0.040*"浴衣" 0.017*"球场"	0.034*"夏天" 0.016*"花火"	0.034*"烟花" 0.016*"田川"	0.026*"夏日" 0.014*"祭典"	0.023*"烟火" 0.014*"会场"	0.022*"大会" 0.011*"舞台"	0.019*"演唱会"	花火大会
14	0.066*"广岛" 0.018*"北荣町"	0.040*"柯南" 0.015*"jr"	0.037*"宫岛" 0.015*"严岛神社"	0.028*"冈山" 0.012*"美观"	0.025*"小镇" 0.012*"神社"	0.023*"仓敷" 0.011*"松山"	0.021*"鸟取" 0.010*"野岛"	0.019*"后乐园"	広島―岡山―鳥取
23	0.048*"仙台" 0.012*"东北"	0.030*"青森" 0.012*"角馆"	0.027*"滑雪" 0.012*"青森县"	0.021*"滑雪场" 0.011*"新干线"	0.015*"松岛" 0.011*"弘前公园"	0.014*"东京" 0.009*"雪场"	0.013*"jr" 0.009*"牛舌"	0.013*"秋田"	東北地方（スキー）
29	0.048*"福冈" 0.016*"熊本城"	0.038*"熊本" 0.016*"温泉"	0.033*"博多" 0.015*"别府"	0.026*"九州" 0.014*"jr"	0.024*"由布院" 0.012*"太宰府"	0.023*"鹿儿岛" 0.010*"佐贺"	0.020*"地狱" 0.010*"阿苏"	0.017*"长崎"	九州地方
36	0.140*"冲绳" 0.012*"公园"	0.057*"水族馆" 0.012*"酒店"	0.033*"那霸" 0.012*"美之海"	0.024*"美国村" 0.011*"首里城公园"	0.022*"国际通" 0.011*"潜水"	0.016*"海豚" 0.010*"首里城"	0.015*"海滩" 0.009*"美丽海水族馆"	0.013*"万座毛"	沖縄
37	0.074*"美术馆" 0.013*"高松市"	0.040*"直岛" 0.012*"岛上"	0.033*"高松" 0.012*"栗林"	0.027*"公园" 0.011*"橄榄"	0.019*"丰岛" 0.011*"南瓜"	0.017*"作品" 0.008*"天使"	0.014*"濑户内海" 0.008*"建筑"	0.014*"艺术"	瀬戸内海・高松
・・・	・・・・・・								・・・
40	0.060*"孩子" 0.014*"老公"	0.035*"小朋友" 0.014*"女儿"	0.026*"妈妈" 0.013*"爸爸"	0.022*"宝宝" 0.012*"大人"	0.020*"儿子" 0.011*"小孩"	0.017*"动物园" 0.011*"海游馆"	0.015*"儿童" 0.010*"老妈"	0.015*"博物馆"	家族旅行

(2)　Word2Vec モデルについて

Word2Vec（Word to Vector）は，文書中の単語を数値ベクトルに変換して，空間内における単語の距離の近さが単語の意味を把握する自然言語処理の手法であり，文章の単語の列から，間にある単語を予測する CBOW（Continuous Bag-of-Words）モデルと，ある単語からその周辺の単語を予測する Skipgram モデルを内包している [13]．Skipgram モデルは低頻出の単語や類推問題の性能の点において，良い結果が得られる傾向にあるが，今回は頻出語や類似度が高い単語を見つけることができる CBOW モデルを用いて，Word2Vec モデル分析を行う．

本研究では，LDA モデル分析で利用したものと同一の形態素データを用いて，Word2Vec モデル分析を行なった．単語ベクトルの次元数（size）を 300，ウィンドウサイズ（window）を 5，機械学習の繰り返し回数（epochs）を 10，乱数値（seed）を 20 と設定した．一方，複数回で Word2Vec モデルを実行するときの結果が再現するために，モデルを単一のワーカースレッド（workers = 1）に制限する必要がある．

図-9 に Word2Vec モデルの生成過程を示す．

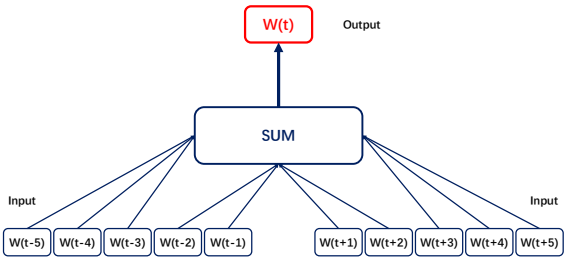

図-9 Word2Vec モデルの生成過程

5.　分析結果

本章では，MFW の旅行記データを利用し，LDA モデルと Word2Vec モデル分析の結果を示す．以下における「都市部」は，東京都，神奈川県，埼玉県，千葉県，愛知県，大阪府，京都府，兵庫県の 8 都府県，

「地方部」は，「都市部」以外の道県を指すものとする．また，各旅行記においてトピックの確率値が 0.01 以上の場合，旅行記はそのトピックに「該当」すると呼び，該当する旅行記の比率を「該当率」と呼ぶ．

(1)　LDA 分析結果

a)　訪問パターン（都道府県単位）

図-10 はトピックのうち主に訪問パターンを示すものの該当率を示している．1 つの旅行記は複数の訪問パターンに該当する可能性があるため，訪問パターンの該当率合計は 100%を上回る．LDA モデル分析の結果より，関西圏（大阪府，京都府，奈良県），東京都およびその周辺（神奈川県，静岡県，千葉県）の訪問パターンの該当率が高く，特に関西圏への訪問パターンに該当する旅行記が多いことがわかる．一方，地方部においては，伊豆半島，中国の上海から長崎へのクルーズの該当率が比較的高い．また該当率はやや低いものの，東北地方（青森県，秋田県，宮城県），北関東（長野県，群馬県，栃木県），広島―岡山―鳥取，瀬戸内海・高松など地方部の訪問パターンも見出された．

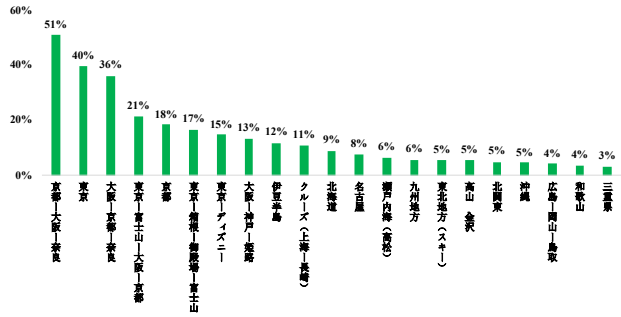

図-10 訪問パターン該当率

b)　旅行テーマ

図-11 はトピックのうち主に旅行テーマを示すものの該当率を示している．訪問中国人観光客の旅行テーマは多岐にわかっていることが示された.また，交通手段の割合は 60%と最も高く，「JR」や「地下鉄」の確率値が高いため，中国人観光客は一定程度鉄道などの公共交通を利用していることが推測される．

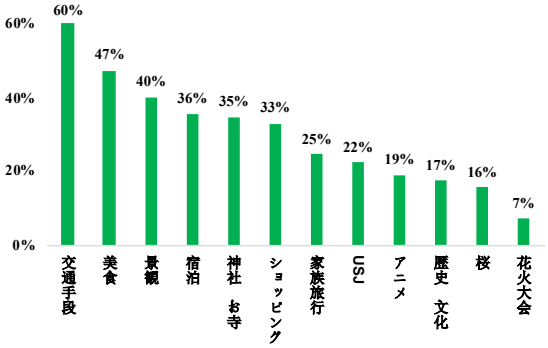

図-11 旅行テーマ該当率

c)　旅行テーマの経年変化

図-12 は旅行記における各旅行テーマの該当率の経年変化を示している．美食，神社・お寺，ショッピングなど該当率が高い，旅行テーマは経年減少しているが，家族旅行，USJ，花火，アニメおよび桜の旅行テーマは増加傾向にある．そのため，訪日中国人観光客は従来の美食，神社・お寺およびショッピング中心から，桜，花火，アニメ，USJ などの体験型旅行に多様化しつつあると考えられる．一方，図-1[2)]に示すように，2016 年以降中国人観光客の 1 人当たり旅行消費額が急速に縮小したものの，ショッピングの該当率は依然として高い．また 2019 年には訪日中国人観光客 1 人当たりの旅行消費額のうち，買い物代が５割以上を占めた．そのため，訪日中国人の「爆買い」現象は落ち着きつつあるが，買物は依然として中国人観光客の旅行支出の最も重要な部分であると考えられる．

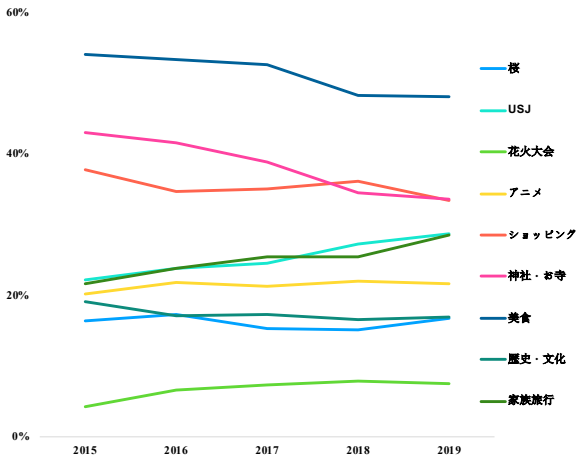

図-12 旅行テーマ該当率の経年変化

d)　訪問パターンの経年変化

図-13 は都市部における訪問パターン該当率の経年変化を示す．名古屋以外の関西圏と首都圏の訪問パターンの該当率は減少傾向にある．図-14 に示すように，北海道以外の地方部の訪問パターンの該当率はおおむね増加傾向にあり，特に瀬戸内海・高松，上海から長崎へのクルーズ，東北地方に関する旅行パターンの該当率が増加している．これは，個人手配による旅行や訪日回数が複数回あるリピーターの訪日中国人観光客の増加に伴い，旅行行動も都市部中心から地方部を含んだものに多様化してきているためと考えられる．

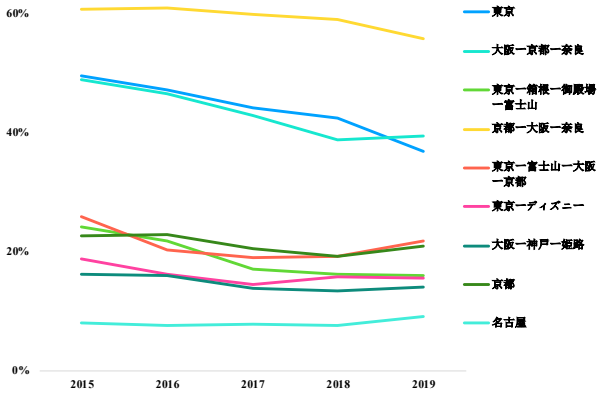

図-13 訪問パターン該当率の経年変化（都市部）

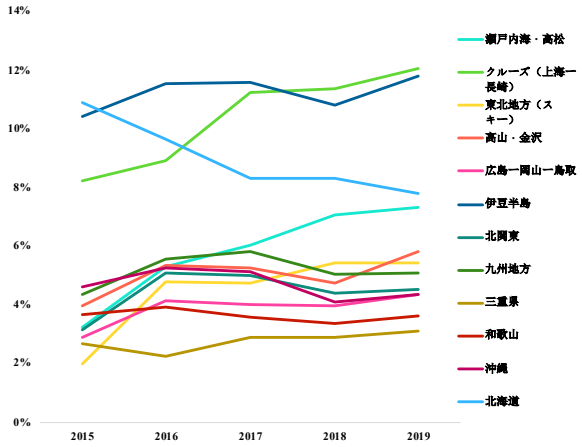

図-14 訪問パターンの経年変化（地方部）

e)　各訪問パターン間の関係

表-3 は，行に示す訪問パターンに該当する旅行記のうち，列の訪問パターンにも該当する旅行記の比率を示している．これにより，訪問パターン間の重複関係を示す．全体として，関西圏（トピック：19, 32）および東京都（トピック：34）の該当率が高く，関西および東京からのインバウンドの比率が高いと考えられる．また高山・金沢の訪問パターンは関西圏および名古屋の関連性が比較的高く，広島—岡山—鳥取と和歌山の訪問パターンは関西圏との関連性が高いため，関西圏からのインバウンドの可能性が高いと考えられる．九州地方は他の地方の訪問パターンとの重複が少なく，クルーズなどで直接来訪する傾向にあると考えられる．

表-3 各訪問パターン間の関係（ピンクは全体の上位 1 割を示す）

トピックid	ピックid ーン / 訪問パタ	1	2	3	10	11	12	13	14	17	19	22	23	24	26	28	29	30	32	34	36	37
1	大阪一神戸一姫路		17.3%	8.6%	10.9%	7.2%	33.9%	10.3%	12.8%	9.9%	71.5%	10.6%	6.8%	6.5%	12.7%	17.1%	10.6%	9.3%	56.2%	27.0%	6.5%	11.7%
2	東京一富士山一大阪一京都	10.8%		7.0%	12.2%	5.8%	14.3%	10.1%	4.3%	3.7%	51.8%	20.9%	6.0%	4.8%	28.0%	15.9%	5.3%	9.2%	30.2%	52.4%	4.4%	4.8%
3	高山・金沢	21.3%	27.7%		38.8%	19.5%	18.1%	12.1%	14.0%	9.9%	41.9%	13.9%	20.9%	12.3%	17.8%	29.2%	14.9%	15.0%	21.6%	29.1%	10.3%	14.5%
10	名古屋	19.2%	34.4%	27.7%		9.3%	19.4%	11.6%	8.3%	7.1%	60.9%	18.1%	9.3%	11.3%	18.6%	20.1%	9.7%	11.5%	35.7%	36.4%	9.3%	8.7%
11	北関東	19.9%	25.4%	21.7%	14.5%		20.0%	14.2%	15.3%	10.0%	34.7%	19.9%	30.4%	14.5%	26.9%	35.2%	17.2%	19.7%	21.5%	48.6%	12.9%	17.2%
12	京都	24.7%	16.7%	5.3%	8.0%	5.3%		6.3%	4.7%	6.1%	82.8%	11.1%	4.3%	3.6%	17.0%	10.0%	4.1%	6.5%	64.9%	34.1%	4.4%	6.6%
13	クルーズ（上海一長崎）	12.6%	19.8%	6.0%	8.0%	6.3%	10.6%		7.4%	4.8%	26.3%	17.6%	7.8%	6.2%	11.9%	17.8%	17.4%	10.5%	21.6%	26.0%	14.3%	9.5%
14	広島一岡山一鳥取	40.3%	21.8%	17.9%	14.8%	17.5%	20.2%	19.0%		14.5%	47.9%	15.7%	16.4%	12.8%	16.9%	38.2%	28.3%	15.9%	29.8%	29.7%	12.3%	34.3%
17	和歌山	37.3%	22.2%	15.2%	15.2%	13.7%	31.7%	14.6%	17.4%		62.2%	14.6%	13.1%	18.7%	15.1%	28.2%	15.7%	14.0%	47.6%	23.2%	14.5%	17.0%
19	京都一大阪一奈良	18.8%	21.8%	4.5%	9.1%	3.3%	29.9%	5.7%	4.0%	4.3%		14.4%	3.1%	2.6%	18.4%	8.5%	3.4%	7.0%	55.5%	45.5%	3.7%	5.4%
22	東京一ディズニー	9.6%	30.1%	5.1%	9.3%	6.5%	13.7%	12.9%	4.5%	3.5%	49.5%		6.0%	4.3%	30.4%	12.0%	5.6%	10.8%	33.4%	64.3%	6.8%	5.9%
23	東北地方（スキー）	16.7%	23.7%	20.8%	13.0%	27.1%	14.4%	15.7%	12.8%	8.5%	28.5%	16.2%		10.6%	16.9%	27.7%	16.2%	32.2%	15.0%	37.9%	10.8%	14.4%
24	三重県	26.3%	31.4%	20.2%	26.1%	21.3%	20.1%	20.8%	16.6%	20.2%	40.6%	19.2%	17.6%		18.8%	32.9%	19.3%	16.7%	26.7%	37.3%	17.2%	16.8%
26	東京一箱根一御殿場一富士山	10.3%	36.2%	5.8%	8.5%	7.9%	18.8%	7.9%	4.3%	3.2%	56.5%	27.3%	5.6%	3.7%		16.8%	4.6%	9.8%	35.8%	75.7%	4.8%	6.0%
28	伊豆半島	19.6%	29.1%	13.5%	13.0%	14.6%	15.6%	16.7%	13.9%	8.6%	37.1%	15.3%	12.9%	9.2%	23.8%		14.8%	13.6%	21.1%	37.7%	9.5%	19.1%
29	九州地方	25.7%	20.5%	14.6%	13.3%	15.1%	13.6%	34.4%	21.7%	10.1%	30.8%	15.1%	15.9%	11.5%	13.8%	31.3%		17.3%	17.4%	25.3%	15.5%	16.2%
30	北海道	14.0%	22.1%	9.1%	9.9%	10.8%	13.3%	12.9%	7.6%	5.6%	40.2%	18.0%	19.7%	6.2%	18.2%	17.9%	10.8%		24.2%	46.8%	9.6%	8.0%
32	大阪一京都一奈良	20.9%	17.9%	3.3%	7.5%	2.9%	33.1%	6.6%	3.5%	4.7%	78.5%	13.8%	2.3%	2.4%	16.5%	6.9%	2.7%	6.0%		43.1%	4.1%	5.2%
34	東京	9.1%	28.4%	4.0%	7.0%	6.0%	15.8%	7.2%	3.2%	2.1%	58.6%	24.2%	5.2%	3.1%	31.7%	11.2%	3.5%	10.5%	39.3%		3.9%	5.4%
36	沖縄	18.5%	20.1%	11.8%	15.0%	13.3%	17.2%	33.4%	11.1%	10.9%	40.1%	21.6%	12.5%	12.0%	16.9%	23.7%	18.2%	18.1%	31.0%	32.4%		13.1%
37	瀬戸内海（高松）	24.7%	16.1%	12.3%	10.4%	13.1%	18.9%	16.3%	22.9%	9.5%	42.8%	13.7%	12.3%	8.7%	15.6%	35.1%	14.1%	11.1%	29.3%	33.4%	9.7%	

f)　各訪問パターンの出発時間（月別）

表-4 は各訪問パターンに該当する旅行記投稿者の月別の出発時期を示す．高山・金沢，名古屋，東北地方および北海道は 1，2 月の出発が比較的多い．これはこれらの訪問パターンの主目的が雪見，スキーであるためと考えられる．4 月は桜の季節のため，訪日中国人観光客がアクセスしやすい関西圏，東京都およびその周辺（神奈川県，静岡県，千葉県）への来訪が比較的多い．7 月は中国の夏休み，上海から長崎へのクルーズと東京・ディズニーランドへの来訪が多い．また 10, 11 月は紅葉の季節であり，京都，和歌山と瀬戸内海への来訪が多い．

表-4 各訪問パターンの出発時期（ピンクは全体の上位 1 割を示す）

トピックid	訪問パターン	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1	大阪ー神戸ー姫路	6.1%	6.0%	7.8%	8.2%	5.5%	6.3%	6.8%	5.9%	7.4%	6.8%	8.1%	4.8%
2	東京ー富士山ー大阪ー京都	6.5%	6.0%	6.9%	8.3%	5.5%	5.6%	6.3%	5.9%	6.5%	6.0%	5.8%	4.7%
3	高山・金沢	11.4%	8.6%	5.4%	7.9%	4.4%	3.3%	4.7%	3.0%	5.8%	6.5%	6.6%	5.9%
10	名古屋	9.2%	6.9%	6.9%	8.4%	5.0%	5.5%	6.7%	6.0%	7.3%	6.2%	5.8%	5.1%
11	北関東	6.7%	6.5%	6.0%	8.8%	4.7%	4.6%	5.0%	3.8%	5.7%	7.7%	6.0%	4.4%
12	京都	2.9%	2.9%	8.2%	9.1%	5.7%	5.8%	6.2%	5.4%	7.4%	9.1%	15.8%	4.8%
13	クルーズ（上海ー長崎）	5.1%	5.4%	5.4%	6.0%	5.7%	7.0%	8.8%	8.2%	7.3%	6.7%	4.6%	4.1%
14	広島ー岡山ー鳥取	6.5%	4.3%	8.1%	7.6%	6.2%	5.4%	5.9%	6.4%	6.2%	6.7%	4.5%	3.9%
17	和歌山	4.8%	4.9%	6.2%	7.4%	5.9%	5.8%	5.8%	5.8%	6.9%	7.3%	8.7%	5.2%
19	京都ー大阪ー奈良	6.6%	6.1%	7.9%	9.1%	6.3%	6.5%	7.7%	6.4%	7.9%	7.3%	6.9%	4.4%
22	東京ーディズニー	6.8%	7.3%	6.2%	7.5%	5.4%	6.5%	9.7%	7.5%	7.7%	6.4%	3.9%	4.1%
23	東北地方（スキー）	12.8%	10.1%	4.9%	8.6%	3.4%	2.8%	3.4%	3.7%	3.9%	5.6%	4.1%	5.5%
24	三重県	5.3%	5.3%	5.5%	6.5%	4.6%	4.3%	5.8%	5.5%	5.0%	7.2%	6.6%	4.6%
26	東京ー箱根ー御殿場ー富士山	7.0%	7.1%	8.2%	9.7%	5.4%	5.6%	7.0%	5.2%	7.4%	6.8%	6.2%	5.2%
28	伊豆半島	5.7%	6.1%	6.5%	7.7%	5.7%	5.5%	6.9%	6.1%	6.5%	7.2%	5.3%	4.0%
29	九州地方	5.9%	5.1%	6.6%	6.8%	5.5%	5.9%	6.6%	5.6%	7.2%	6.7%	4.7%	4.4%
30	北海道	15.2%	13.5%	4.0%	4.6%	3.0%	3.8%	7.1%	4.4%	4.2%	3.7%	3.7%	7.6%
32	大阪ー京都ー奈良	6.3%	6.1%	8.6%	9.5%	6.5%	7.0%	7.6%	6.4%	8.5%	7.3%	6.8%	4.4%
34	東京	6.9%	7.1%	7.8%	8.2%	5.6%	6.0%	8.2%	6.2%	7.5%	6.8%	5.1%	4.7%
36	沖縄	6.1%	5.5%	4.9%	6.7%	5.1%	6.0%	8.2%	7.1%	8.0%	6.1%	3.6%	4.8%
37	瀬戸内海（高松）	6.1%	4.9%	5.9%	7.2%	6.1%	6.0%	7.7%	6.2%	7.4%	8.5%	5.2%	3.4%

(2) Word2Vec 分析結果

本節では，地方部でのより詳細な旅行行動に着目するために，LDA モデル分析で抽出した地方部への訪問パターンの一部（東北地方，広島—岡山—鳥取県，瀬戸内海・高松，九州地方と沖縄）を例とし，Word2Vec モデル分析を実施し，入力単語に類似した上位 30 単語を取得し，都道府県単位，市町村単位及び観光スポットに関する地名と特有の地域資源を明らかにした．表-5 に示すように，県内での周遊の訪問地や観光スポットおよび一部訪問地での美食と二次交通の利用状況を把握することができる．またその訪問地に特有の地域資源も把握することができる．例えば，仙台市の食べ物（牛タン），青森県の星野，秋田県の柴犬，岡山の桃太郎，鳥取県のコナン，瀬戸内海・高松の芸術祭，香川県のうどんと熊本県の馬刺しなどの特有の地域資源を示す語が抽出された．これらの結果に基づいて，地方部での周遊ルートの検討が可能となると考えられる．

表-5 訪問パターン（市町村単位）

トピック	Input word	地名			地域資源	その他
		都道府県・（県外）	市町村等・（県外）	観光スポット・（県外）		
東北地方（スキー）	宮城	宮城县 宮城 （岩手县 岩手 山形县 福島県 福岛）	宮城郡 柴田町 大河原町 大崎市 仙台市（花巻市 庄内 福岛市 新庄 鶴冈市 松城）	秋保 秋保温泉		良田 六县 刚宪 出羽国 三县 东北地区 奥国 六省 贵为 县境
	仙台	宮城县（山形 青森 新瀉）	仙台市 石巻 石巻市 青叶 青叶区 大河原（盛冈 伊达 郡山 宇都宫）	松岛 仙台城 秋保 秋保温泉 东北大学 瑞凤殿 鸣子温泉	牛舌（牛タン）萃萃 利久 喜助 伊之助（牛タン店）	仙台站 looplebus 东北地区 羽前
	青森	青森县 青森県（岩手）	弘前 弘前市 青森市 八戸市 五所川原 五所川原市 三泽 十和田 十和田市（盛冈）	界津轻 睡魔之家 白神 十和田湖 八甲田山 星野度假村 苹果园 休屋 佞武多 浅虫 八甲 浮汤	星野 菩哒（ねぶた）	五能线 东北地区 溪流
	秋田	秋田县 秋田県（山形 青森县）	大馆 大馆市 秋田市 羽后 能代 男鹿 角馆 角馆町 仙北市 横手 横手市 大曲 鹰巢 （弘前 新庄 五所川原）	千秋公园 田泽湖 乳头	柴犬 狗狗 土狗 稻庭（うどん店）	想养 五能线 美之国
広島―岡山―鳥取	広島	広島 （冈山）	广岛市 广岛机场 东广岛 西条 福山 福山市 三原 三原市 竹原市 吴市 吴港 纸屋	宫岛 广岛城 严岛神社 元安川 原子弹爆炸圆顶屋 和平纪念公园 缩景园 宫岛口	原爆 核爆 原子弹 三箭（サンフレッチェ）	广电 户岛 吴线 爆点
	岡山	岡山 岗山 冈山县	冈山市 仓敷 宇野 倉敷 玉野 早岛町 仓敷市 伊部（福山 高松 米子 姬路）	冈山城 儿岛 児岛 后乐园 宇野港	桃太郎	daiwaroynethotelokayamaekimae 岡山站 山阳 从米子 高松站 备中 阳线 多津(宇多津) 艾克玛
	鳥取	鳥取 （岛根 岛根县）	东伯郡 鸟取市 由良 仓吉 仓吉市 岩美 米子 米子市 境港 境港市 智头 倉吉 三朝町 由良町 丰冈 （安来）	沙丘 砂丘 鸟取城 皆生温泉 白兔 中田岛 山阴	柯南（コナン）	良站 从米子 superhakuto
瀬戸内海・高松	瀬戸内海	濑户 四国 四国岛 香川	高松 下滩 今治 竹原市 尾道 玉野市	小豆岛 濑户大桥 犬岛 直岛 因岛 生口岛	艺术节（芸術祭）	跳岛游 岛游 跳岛 岛屿 内岛 各岛 波海 离岛 老山 羽山 豆岛 多津(宇多津) 个岛
	香川	香川県 香川郡（四国 四國 四国岛 愛媛 爱媛 德岛 冈山县）	仲多度郡 片原 小豆郡 坂出市 丸亀市 琴平 土庄町（三好市 盛冈市）	赞岐国 金刀（备中）	乌冬之乡 乌冬县（うどん県） 川福 一鹤 一鹤骨付	知县 鸟味 赞歧 多津(宇多津)
	高松	香川（冈山 德岛 濑户内海）	高松市 高松站 丸龟 琴电 琴平 土庄瓦 町站 片原 （宇野 玉野 玉野市 宇野港）	小豆岛 栗林 玉藻公园（岛山）	一鹤 一鹤骨付	jrhotelclementtakamatsu 跳岛游 多津(宇多津) 跳岛 鸟味 赶船 各岛 paicei
九州地方	福岡	福岡 福岗 （北九州 佐贺 长崎 鹿儿岛）	博多 博多区 福冈市 中洲 中州 太宰府市 久留米市（北九州市 小仓 佐贺市 鸟栖 佐世保 别府 门司港）	太宰府 大濠 大濠公园 博多湾	牛肠锅	九州 反射镜 不登塔 宰府 舞鹤
	佐賀	佐贺县 佐賀（福冈 北九州 鹿儿岛 宫崎 大分）	鸟栖 武雄市 佐贺市 唐津 黑川 肥前（别府 由布院 久留米 久留米市 小仓 谏早 佐世保 人吉 日田 阿苏 今宿）	佐贺城 善寺	牛肠锅 季乐（佐賀牛レストラン）	武雄 九州
	長崎	长崎县（福冈）	佐世保 佐世保市 长崎市 谏早 西滨 浦上 云仙 大村市	稻佐山 哥拉巴公园 大浦天主堂 荷兰坂 豪斯登堡 思案桥 和平公园 原爆资料馆 九十九岛	福砂（カステラ） 角煮	哥拉巴园 大埔 天主堂 天主教堂 出岛 回船 靠岸 荷兰 登岸
	熊本	熊本县 熊本県（鹿儿岛 北九州 长崎）	熊本市 新市街 下通 阿苏 高森（鸟栖 久留米）	熊店 熊馆 熊本城 城彩苑 辛岛 竹笛 本妙寺 黑川 阿苏火山 水前寺成趣园 趣园	马肉（馬刺）菅乃屋（馬刺店） 萌熊 黑亭（ラーメン）	部长 九州 鹤屋（百貨店）
	大分	大分県（佐賀 佐贺县 宫崎县 熊本県 熊本县 长崎县 福岡県 福冈）	由布 由布市 汤布院 别府 日田市 丰后 汤布院町（人吉 久留米 筑前 北九州市 云仙 霧島）	别府温泉 耶马溪 深耶马溪	八汤（別府温泉）	分站 肥萨线 久大本線 久大本线
	鹿児島	鹿児島（熊本 佐贺 宫崎 福冈 长崎）	鹿儿岛市 金生 屋台村 指宿 指宿市 雾岛市（佐世保 高森 吉松 熊本市 天文馆）	屋久岛 仙岩园 雾岛 樱岛 城山 锦江湾 严园	黑豚 黑猪	返船 毛猪 amuplaza 九州
沖縄	沖縄	沖縄 冲绳岛 冲绳县 琉球	那霸 那霸市 那霸市 恩纳 恩纳村 渡嘉敷 名护市 本部町 北谷町 照屋	国际通 海洋博公园 宫古岛 离岛 石垣岛 首里城 古宇利岛 万座毛 知念	观鲸 雪盐 恋战（映画）	本岛 琉球人 那霸机场 东北三省

6. おわりに

本研究では，中国語の観光ウェブサイト（MFW）に掲載されている訪日中国人観光客が書き込んだ旅行記のコンテンツを用いて，自然言語処理手法である LDA モデルと Word2Vec モデルを組み合わせ，訪日中国人観光客の旅行テーマと，訪問パターンを明らかにした．Word2Vec モデルの分析では，都道府県単位，市町村単位および観光スポットに関する地名と特有の地域資源を明らかにした．

属性・旅行情報の集計結果から，旅行記記入者は，20，30 代，女性の比率が高く，恋人・配偶者，家族および友人と一緒の旅行の比率が高い．また雪，桜および紅葉の季節に対応する時期来訪が比較的多く，出発日から中国に戻るまでの滞在期間は，7~13 日の比率が高い．

LDA モデル分析の結果から，訪日中国人観光客は都市部への訪問パターンの該当率が高い一方で，関西圏と首都圏の訪問パターンの該当率は減少傾向にあることが示された．一方，図-14 に示すように，北

海道以外の地方部への訪問パターンの該当率はおおむね増加傾向にあり，特に瀬戸内海，上海から長崎へのクルーズ，東北地方の訪問パターンの該当率が増加している．これは，個人手配による旅行や，訪日回数が複数回のリピーターの増加に伴い，旅行行動も都市部中心から地方部を含んだものに変化しているためと考えられる．また旅行テーマについては，美食，神社・お寺，ショッピングなどの該当率は減少傾向にある一方，家族旅行，USJ，花火，アニメおよび桜の旅行テーマは増加傾向にある．そのため，訪日中国人観光客は従来の美食，神社・お寺およびショッピング中心から，桜，花火，アニメ，USJ などの体験型旅行に多様化しつつあると考えられる．

さらに Word2Vec モデル分析により，県内での周遊の訪問地や観光スポットおよび一部訪問地での美食と二次交通の利用状況を明らかにした．訪問地に特有の地域資源を示す語が抽出でき，それより訪日中国人観光客の県内の周遊ルートを検討することが可能となると考えられる．

参考文献

1) 国土交通省：「観光白書」，令和２年版 https://www.mlit.go.jp/statistics/file000008.html（最終閲覧日 2021.9.30）
2) 観光庁：「訪日外国人消費動向調査」 https://www.mlit.go.jp/kankocho/siryou/toukei/syouhityousa.html（最終閲覧日 2021.9.30）
3) 李 温慧：日本インバウンドにおける中国人観光行動の多様化と深化，滋賀大学大学院教育学研究科論文集 49 第 20 号, p.49～p.59, 2017.
4) 日本政府観光局（JNTO）：「日本の観光統計データ」 https://statistics.jnto.go.jp/（最終閲覧日 2021.9.30）
5) 観光庁：「宿泊旅行統計調査」 https://www.mlit.go.jp/kankocho/siryou/toukei/shukuhakutoukei.html（最終閲覧日 2021.9.30）
6) 金 玉実: 日本における中国人旅行者行動の空間的特徴，地理学評論 82-4, p.332～p.345, 2009.
7) 菱田 のぞみ，日比野 直彦，森地 茂：訪問地選択の多様性に着目した訪日中国人旅行者の居住地別観光行動の時系列分析，土木学会論文集 D3(土木計画学), VoI.68, No.5(土木計画学研究・論文集第 29 巻), I_667-I_667, 2012.
8) 松井 祐樹，日比野 直彦，森地 茂，家田 仁：訪日外国人旅行者の個人行動データを用いた訪問地および観光活動に着目した観光行動分析，土木学会論文集 D3 (土木計画学), Vol.72, No.5 (土木計画学研究・論文集第 33 巻), I_533-I_546, 2016.
9) 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 (土木計画学), Vol.72, No.5 (土木計画学研究・論文集第 33 巻), I_571-I_583, 2016.
10) 古屋 秀樹：hPAM による類似性を考慮した訪日外国人 旅行者の訪問パターン抽出に関する基礎的研究，土木学会論文集 D3 (土木計画学), Vol.75, No.5 (土木計画学研究・論文集第 36 巻), I_507-I_517, 2019.
11) 宋紫龍，古屋秀樹：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，研究報告ドキュメントコミュニケーション(DC), 2019-DC-112 巻 6, p.1～p.5, 2019.
12) 日本政府観光局（JNTO）：「訪日外客数」 https://www.jnto.go.jp/jpn/statistics/since2003_visitor_arrivals.pdf
13) 加藤公一：「Gensim」による機械学習を使った自然言語分析の基本――「NLTK」「潜在的ディリクレ配分法（LDA）」「Word2vec」とは, Python で始める機械学習入門（9）, 2019. https://atmarkit.itmedia.co.jp/ait/articles/1905/22/news007.html（最終閲覧日 2021.9.30）
14) 岩田具治：トピックモデル，講談社，2015.
15) Selva Prabhakaran : Topic Modeling with Gensim (Python), 2018. https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#17howtofindtheoptimalnumberoftopicsforlda（最終閲覧日 2021.9.30）