トピックモデルを用いた訪日外国人周遊分析

辰巳　嘉大[1]・塚井　誠人[2]

[1]学生会員　広島大学大学院　工学部研究科（〒739-8527 広島県東広島市鏡山 1-4-1）
E-mail: m19@hiroshima-u.ac.jp

[2]正会員　広島大学大学院准教授　工学研究科（〒739-8527 広島県東広島市鏡山 1-4-1）
E-mail: mtukai@hiroshima-u.ac.jp

近年，新たな産業としてインバウンド観光が注目されている．国土交通省は，訪日外国人増加に向けた政策立案のために訪日外国人流動データ（以下 FF データ）[1)]を作成した．FF データは，周遊パターンや訪日目的などの属性の組み合わせが膨大である．そこで，FF データから効率よく代表的なパターンを抽出する手法が必要である．本研究では，データの潜在的なパターンを効率よく分析できるトピックモデルを FF データに適用することで，訪日外国人の旅行特性の実態を経年分析した．また，トピックモデルの既往研究の課題を考慮して，トピック数の決定手順，トピックの解釈手順を示した．関東や関西を訪問地とした旅行は，東京都や大阪府，京都府以外の訪問割合が増加しているなどの分析結果から，訪日外国人の旅行特性は多様化していることを明らかにした．

Key Words: *Latant Dirichlet Allocation, Topic Model, FF-Data, tour, foreign visitors to japan,*

１．はじめに

インバウンド観光は，経済，地域の活性化といった効果が期待される産業であり，都市部に留まらず地方都市にとっても重要な政策課題である．政府は訪日外国人増加に向けビジット・ジャパン事業 [2)]，訪日促進事業 [3)]などを実施しており，訪日外国人数は大きく増加している．国土交通省は，さらなる訪日外国人数増加に向けた政策を立案するために，訪日外国人流動データ（以下 FF データ）を作成した．

このデータでは，国籍，滞在日数などの各属性のクロス分析，および国内の周遊ルートに関する分析が可能である．ただし，2014～2016 年分が公開されている FF データのサンプル数は約 4.3～6.3 万人，総トリップ数は約 17.9～22.8 万にのぼる．このため，代表的な周遊パターン旅客の属性を抽出するには，それらの膨大な組み合わせから，効率よくパターンを抽出する手法が必要である．

本研究では，FF データから代表的な属性・周遊パターンを効率よく抽出するために，トピックモデルを適用する．トピックモデルへの入力データは，Bag of words（BOW）形式と呼ばれる，文書別語彙別のカウントデータである．後述する手順で加工することによって，FF データを．BOW 形式に変換する．さらに各年の FF データにトピックモデルを適用して，それぞれ代表的なパターンを抽出したうえで，訪日回数や国籍，訪日時期による訪日外国人旅行特性の違いや，経年的な訪日外国人旅行特性の変化を把握することを目的とする

本論文の構成は，以下のとおりである．2. では，本研究で用いる訪日外国人とトピックモデルに関する既往研究を整理する．3. では，本研究で用いるトピックモデルの概要を説明する．4. では，トピック抽出のための FF データの加工，トピック数の決定，トピックの解釈・名付け方法を説明する．5. では，抽出したトピックから訪日外国人旅行特性を把握する． 6. では，結論を述べる．

２．既往研究

本節では，訪日外国人とトピックモデルに関する研究を整理し，本研究の位置づけについて示す．

菱田ら [4)]は JNTO 訪日外国人消費行動調査を用いて，中国，台湾，韓国，香港，シンガポールからの訪日外国人の訪問地傾向を訪日経験別に経年分析した．このうち

中国については，居住地域別に分析をした．松井ら [5]は訪日外国人消費調査の個票データを用いて，個人属性別に訪問地と観光行動のパターンの違いをクロス集計によって分析した．その結果，近年増加する個人旅行者は，都市部を中心に多様な観光行動をする傾向にあることを明らかにした．古屋ら [6]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 24 の訪問パターンを導出した．クラス別構成比率と主要国籍・地域，旅行形態，旅行時期，訪日回数などの各要因との関連性について，一般化χ^2乗検定によって明らかにした．

以下，トピックモデルに関する研究を整理する．塚井ら [7]はトピックモデルによる討議分析の可能性について検討した．Web 上で公表されている各地の地域公共交通会議に関する討議録データベースに対して，トピックモデルを適用した．分析の結果，各地域の課題に即したトピックと地域間で共通のトピックが得られ，モデルの有効性を確認した．塚野 [8]は，各種用地面積や，人口・世帯，事業所など 23 属性の地理情報データにトピックモデルを適用して分析した．各メッシュが最も大きな確率で帰属するトピックを地図上に表示した．また，主成分分析・因子分析とトピックモデルとの比較を行い，固有値に負の値の点が出現しないトピックモデルの方が，トピック解釈の面で有用としている．一方で分析者が設定する抽出するトピック数の，より合理的な設定手順についてさらに検討が必要，としている．川野ら [9]はトピックモデルと離散連続モデルを結合した新たな自由記述データの分析手法を提案した．属性別の回答傾向の違いや，選択式設問の回答と自由回答中のトピックの対応が明確なことを確認して，同手法の有用性を確認した． 古屋ら [10]は GPS ログデータを用いて，訪日外国人旅行者の訪問パターン特性を分析した．トピックモデルを用いて訪問場所の組み合わせパターンを分類した．

既往研究の課題についてまとめる，訪日外国人旅行に関する研究では，訪日外国人の訪日時期，訪日目的，訪日経験回数，利用交通機関分担率，宿泊数，滞在日数，訪日手配方法といった旅行特性の全体を包括的に捉えた研究は見られない．トピックモデルに関する研究では，言語情報以外にも，地理情報や GPS ログなどへの適用が進んでいるが，FF データのようなアンケートデータへの適用例は見られない．とりわけ，本研究で加工を行うサンプル数が多く語彙数が少ないデータに適用した際のモデルの有効性は確認されていない．

本研究では，訪日外国人の都道府県間を跨ぐ周遊行動，訪日時期，訪日目的，訪日経験回数，利用交通機関分担率，宿泊数，滞在日数，訪日手配方法といった旅行特性について，分析者による実験的な旅行特性の仮定を行わずデータマイニングに基づいた分析を行い，その変化を経年的に明らかにする．分析手法としてトピックモデルを用いる．本研究では既往研究の課題を考慮して，トピック数の決定手順や，トピックの解釈手順を示す．また，サンプル数が多く語彙数が少ない FF データに対しても有効なことを示す．

３． トピックモデル

トピックモデルの基本となる考え方は，確率的トピック生成モデル Latent Dirichlet Allocation(LDA)に示されている [11]．LDA では，トピックは潜在変数，かつ 1 文書は複数のトピックスによって構成されると仮定する．なお以下の説明は既往研究 [12]に基づいており，新規性はない．

D 個の文書集合を考える．各文書 d はN^d個の語から成り，文書dのn番目の語を$\{w^{n,d}\}_{n=1}^{N^d}$とする．それぞれの語は 1-of-V 表現$w_v^{n,d} \in \{e_v\}_{v=1}^{V}$で表す．*1-of-V*表現とは，$V$個の語彙に順に番号を割り振り，各文書の n 番目に出現した語の語彙番号が v のとき，ベクトル$w^{n,d}$の v 番目の要素を 1，それ以外の要素を 0 とする表現である [13]．$w^{n,d}$は，文書全体では$N \times V$の要素を持つ．ただし N は全単語数であり，文書別単語数N^dの和である．

LDA では，各語彙は潜在トピック$z^{n,d} \in \{e_k\}_{k=1}^{K}$に属すると仮定する．各文書は異なるトピック分布$\tilde{\theta}_{\mathrm{d}}$を，また各トピック kはそれぞれ異なる語彙分布β_k持つと考える．さらに各文書内での単語の並び替えが可能と仮定する．この仮定は，文書内で共起する語彙のまとまりが情報の基本単位であって，その出現順序には意味がないとする考えに基づく．潜在トピックあるいは共起語彙の出現確率は多項分布を用いて，式(1)，(2)で表わされる．

$$P(z^{n,d}|\tilde{\theta}_d) = Multi_{k,1}(z^{n,d};\tilde{\theta}_d) \tag{1}$$

$$p(w_v^{n,d}|z^{n,d},\beta_1,\dots,\beta_k) = \prod_{k=1}^{k}\{Multi_{V,1}(w^{n,d};\beta_k)\}^{z_k^{n,d}} \tag{2}$$

多項分布のパラメータ$\tilde{\theta}_{\mathrm{d}}$，$\beta_k$の推定のため，その共役事前分布であるディリクレ分布を仮定する．これらは，式(3)，(4)で表される．

$$p(\tilde{\theta}_d|\alpha) = Dir_K(\tilde{\theta}_d;\alpha) \tag{3}$$

$$p(\beta_k|\eta) = Dir_V(\beta_k;\eta) \tag{4}$$

各文書のトピック分布を D 行 K 列の文書パラメータ$\Theta = (\tilde{\theta}_1,\dots,\tilde{\theta}_d)^t$，各トピックの語彙分布を K 行 V 列のトピックパラメータ$B = (\beta_1,\dots,\beta_k)^t$と定義する．なお右肩の添え字 t は転置を表す．観測データ$W = [\{w^{n,d}\}_{n=1}^{N^d}]_{d=1}^{D}$と潜在変数$\mathrm{Z} = [\{z^{n,d}\}_{n=1}^{N^d}]_{d=1}^{D}$の同時分布は，式(5)のように表される．

$$
\begin{aligned}
&p(W,Z\,|\,\Theta.\Phi)\\
&=\prod_{d=1}^{D}\sum_{n=1}^{N^d}p(w^{n,d}\,|z^{n,d},\beta_k)p\left(z^{n,d}|\tilde{\theta}_m\right)\\
&=\prod_{d=1}^{D}\prod_{n=1}^{N^d}\prod_{k=1}^{K}(\Theta_{d,k}\prod_{v=1}^{V}B_{k,v}^{w_v^{n,d}})^{z_k^{n,d}}
\end{aligned}
\tag{5}
$$

モデルの特徴を明らかにするため，潜在トピック$z^{n,d}$を消去して，W（N 行 V 列）に関する周辺確率を求める．これは，式(6)で表される．さらに *1-of-V* 表現された W を，式(3.7)によって定義される文書単位の Bag-of-words(以下 BOW)表現のデータ M で書き改める．

$$
\begin{aligned}
&p(W|\Theta,B)\\
&=\sum_{Z}p(W,Z|\Phi,B)\\
&=\prod_{d=1}^{D}\prod_{n=1}^{N^d}(\sum_{z^{n,d}\in\{e_k\}_{k=1}^{K}}\sum_{k=1}^{K}(\Theta_{d,k}\sum_{v=1}^{v}B_{k,v}^{w_v^{n,d}})^{z_k^{n,d}}\\
&=\prod_{d=1}^{D}\prod_{v=1}^{V}((\Theta B)_{d,v})^{\sum_{n=1}^{N^d}w_v^{n,d}}
\end{aligned}
\tag{6}
$$

$$
M=(m_1,\dots,m_D)^t, M_{d,v}=\sum_{n=1}^{N^d}w_v^{nd} \tag{7}
$$

式(7)に示すように，この操作によって得られる M は D 行 V 列となる．ここで，u_d は行列 $U=(u_1,\dots,u_D)^t=\Theta\Phi$ の第 d 行ベクトルである．すると式(6)は，データ M に関する確率分布として式(8)に書き改められる．

$$
\begin{aligned}
&p(M|\Theta,B)\\
&=\prod_{d=1}^{D}N^d!\prod_{v=1}^{V}\frac{((\Theta B)_{d,v})^{M_{d,v}}}{M_{d,v}!}\\
&=\prod_{d=1}^{D}Mult\,i_{V,N^d}(m_d;u_d)
\end{aligned}
\tag{8}
$$

$$
M\approx\Theta B \tag{9}
$$

式(8)は，潜在パラメータ$\Theta\Phi$の積 U をハイパーパラメータとする，文書単位の BOW データ M の確率モデルである．この構造を単純化して表記すると，式(9)が得られる．同式より LDA は，トピック数 K をランクとする低ランク行列$\Theta\Phi$で観測データ M を近似する行列分解モデルとなっていることがわかる．トピックモデルのパラメータは，変分ベイズ法によって推定する．

4． トピックの抽出

(1) FF データの加工

本研究では，Flows of Foreigners （以下，FF）データを用いる． FF データは，空海港から出国する外国人を対象とし調査票を用いたアンケート調査である．国土交通省が，出入国管理統計 [14]，国際航空旅客動態調査 [15]，訪日外国人消費動向調査 [16]で得たサンプル情報を基に拡大処理を行っている．拡大係数は，各サンプルから実流動量を推計するための係数であり，四半期，または年間の値が付されている．なお国内訪問地間の利用交通機関については，国際航空動態調査 [15]で取得した OD 別の交通機関分担率を全データに適用している．FF データの特徴は，入国海空港から国内訪問地，出国海空港までの一連のトリップチェーン情報が記録されている点である．データベースでは，一連のトリップチェーンを 2 地点間のトリップ単位に分割して記録するため，訪日外国人 1 サンプルについて，周遊目的地分の行数が使われる．つまり周遊が 3 トリップから成る場合は，3 行に及ぶ．

トピックモデルの入力データは，文書 D ごとに含まれる語彙 V の数をカウントした $D\times V$ の行列データである．すなわち，行方向に訪日外国人数 D，列方向に旅行特性 V をとった $D\times V$ 形式に加工する．これは，FF データを 1 サンプル 1 行に改めることを意味する．そこで，全サンプル中で最も多い周遊目的地分の列を新たに追加し，トリップチェーン内で 2 番目の訪問地を訪問地 2，3 番目の訪問地を訪問地 3 のように加工を行った．また，利用交通機関分担率は，各トリップの拡大係数を合計して，トリップチェーン全体の利用交通機関分担率を算出した．

各個人の FF データは，その旅行者の周遊全体で不変の個人属性と，周遊を構成するトリップごとに異なるトピック特性から成る．前者については，各変数のカテゴリカル値を表すダミー変数を作成してそれを BOW ベクトルの要素とする．後者については，まず複数行にわたる情報を行単位のデータ形式に改めた上で，連続データである宿泊数，滞在日数，利用交通機関分担率は，階級値として離散化して各カテゴリへの帰属を表すダミー変数を作成して，前者の BOW ベクトルに結合する．FF データの属性を表 1 に示す．以下の分析では，2014 年から 2016 年の 3 年分のデータを用いる．なお，2014 年の FF データには，旅行手配方法，訪日経験回数の属性はない．

(2) トピック数の決定

LDA の推定では，抽出するトピック数 K を，モデル推定に先立って設定しなくてはならない．K の設定を変更すると尤度が変わるので，候補とする K の範囲を定めて計算を繰り返し，その中で最適な K を選択する．ただし，トピックの最適性は統計的性質のみで定まるわけではなく，解釈の容易性も問題になる．

本研究は，まず尤度比を用いて K の検討範囲の上限となるトピック数 K_{max} を決定する．統計的な意味では尤度比が極大値をとるトピック数が最も適合性が高いが，そのときのトピックの解釈性も併せて検討しなくてはならない．そこで以下の分析では，K_{max} は検討範囲の上限として，推定したトピック群から類似したトピックを

合成する．本研究は類似性の指標として，トピックベクトル間のコサイン類似度を用いる．トピックkとトピックk'の類似度$M_{kk'}$は式(10)で定義される．

$M_{kk'}$の値が大きいほど，トピックkとトピックk'の類似性は高い．以下の分析では，類似を判定する閾値を設定して，K_{max}から順次類似トピック数を合成したときに，類似しないトピック数が最大となるトピック数を，最適なトピック数として採用する．なお，本研究はベクトル間の角度が cos45°となる 0.7 を閾値とする．

$$M_{kk'} = \frac{U_k \cdot U_{k'}}{|U_k||U_{k'}|} \tag{10}$$

今回のデータセットにおいて，尤度比は全年でトピック数 60 を超えてもなお増加傾向であり，極大値をとるトピック数を選択するのは困難だった．そこで，トピック数 5 から順次閾値 0.7 を超えるトピックを合成して，類似しないトピック数の最大をそれぞれ確認すると，2014～2016 年すべて 8 トピックであった．この結果に基づいて，トピック併合に関する探索漏れが起きないように，安全側を考慮してトピック数K_{max}は全年で 20 とする．トピック数5からK_{max}までの間での尤度比の極値，類似しないトピック数が最大となるトピック数を考えて採用するトピック数を検討する．

本研究はトピック数 10（2014 年），8（2015,2016 年）を採用した．なお，2014 年は類似しないトピック数に合成した結果を示す．

(3)　トピックの解釈・名付け

2014，2015，2016 年の各トピックの命名ルールは以下のとおりである．まず 2014 年に 14，2015 年に 15，2016 年に 16 のように，それぞれ年号の識別記号を付けた．次に構成比率が大きいトピックから順に 1，2，3 とした．最後に周遊地方が明確な場合にはその地方名，または都道府県名を付した．ただし，複数の地方，都道府県が現れる場合はそれらを名付けた．

また周遊している地方や，都道府県情報がほぼないトピックは，解釈困難と名付けた．たとえば，2014 年に構成比率が 4 番目に多く，関西を周遊しているトピックは，14.4.関西と名付けた．各トピックの名前を，表2～4に示す．

表 1 本研究で用いた属性

属性	内容/カテゴリ
出国港	新千歳空港，旭川空港，函館空港，青森空港，仙台空港，秋田空港
	茨城空港，羽田空港，成田空港，新潟空港，富山空港，小松空港，静岡空港，中部空港
	関西空港，米子空港，岡山空港，広島空港，高松空港，松山空港，福岡空港
	福岡空港，佐賀空港，長崎空港，熊本空港，大分空港，宮崎空港，鹿児島空港
	那覇空港，石垣空港，北九州海港，博多海港，下関海港，厳原海港，その他空海港
国籍	韓国，台湾，香港，中国
	タイ，シンガポール，マレーシア，インドネシア，フィリピン，ベトナム
	インド，その他アジア
	イギリス，ドイツ，フランス，イタリア，スペイン，ロシア，その他ヨーロッパ
	アメリカ，カナダ，その他北アメリカ，南アメリカ
	オーストラリア，アフリカ，，その他オセアニア，無国籍
旅行目的	観光・レジャー，家族知人の訪問，業務，研修・学会等，留学，乗り継ぎ，その他，不明
旅行手配方法	団体旅行，個人旅行，不明
訪日経験回数	1回目，2回目，3回目，4回目，5回目，6～9回目，10～19回目，20回以上，不明
出発地	47都道府県，不明
	新千歳空港，旭川空港，函館空港，青森空港，仙台空港，秋田空港
	茨城空港，羽田空港，成田空港，新潟空港，富山空港，小松空港，静岡空港，中部空港
	関西空港，米子空港，岡山空港，広島空港，高松空港，松山空港，福岡空港
	福岡空港，佐賀空港，長崎空港，熊本空港，大分空港，宮崎空港，鹿児島空港
	那覇空港，石垣空港，北九州海港，博多海港，下関海港，厳原海港，その他空海港
目的地	47都道府県，不明
	新千歳空港，旭川空港，函館空港，青森空港，仙台空港，秋田空港
	茨城空港，羽田空港，成田空港，新潟空港，富山空港，小松空港，静岡空港，中部空港
	関西空港，米子空港，岡山空港，広島空港，高松空港，松山空港，福岡空港
	福岡空港，佐賀空港，長崎空港，熊本空港，大分空港，宮崎空港，鹿児島空港
	那覇空港，石垣空港，北九州海港，博多海港，下関海港，厳原海港，その他空海港
トリップ数	サンプルIDごとのトリップ数
訪日時期	1～3月期，4～6月期，7～9月期，10～12月期
滞在日数	0日，1～2日，3日，4日，5日，6日，7日，8～10日
	11～14日，15～30日，31～90日，91～364日
宿泊数	0～2日，3～7日，8日以上
利用交通機関	バス，鉄道，国内線飛行機，自動車
利用交通機関分担率	0～9%，10～19%，20～29%，30～39%，40～49%，50～59%
	60～69%，70～79%，80～89%，90～99%
拡大係数	サンプルIDごとの拡大係数

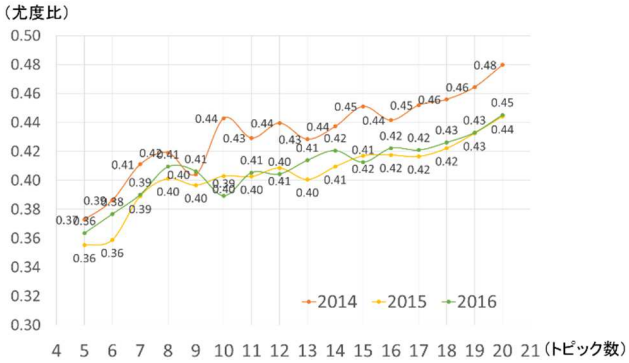

図 2　トピック数と尤度比

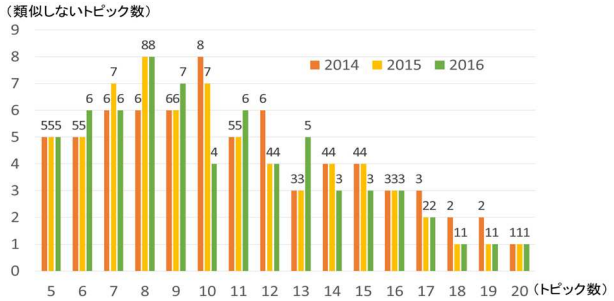

図 1　トピック数と類似しないトピック数

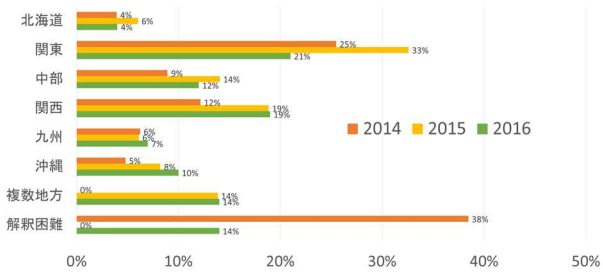

図 3　トピック構成比率

表 2　トピックと名前（2014 年）

順位	14.8. 北海道		14.2. 関東1		14.3. 関東2		14.5. 中部	
1位	北海道（0～2日宿泊）	0.18	東京都（0～2日宿泊）	0.74	国内線飛行機分担率0～9%	0.07	国内線飛行機分担率0～9%	0.06
2位	鉄道分担率10～19%	0.06	成田空港出国	0.14	東京都（0～2日宿泊）	0.06	愛知県（0～2日宿泊）	0.06
3位	バス分担率70～79%	0.04	成田空港入国	0.13	バス分担率20～29%	0.05	鉄道分担率40～49%	0.05
4位	新千歳空港入国	0.03	東京都（3～7日宿泊）	0.10	成田空港出国	0.05	観光・レジャー	0.04
5位	新千歳空港出国	0.03	自動車分担率10～19%	0.06	成田空港入国	0.05	中部空港出国	0.03
6位	バス分担率60～69%	0.02	神奈川県（0～2日宿泊）	0.05	自動車分担率10～19%	0.05	中国	0.03
7位	5日滞在	0.02	バス分担率20～29%	0.05	鉄道分担率50～59%	0.05	10-12月期	0.03
8位	台湾	0.02	千葉県（0～2日宿泊）	0.04	観光・レジャー	0.03	中部空港入国	0.03
9位	北海道（3～7日宿泊）	0.01	東京都（8日以上宿泊）	0.03	羽田空港入国	0.03	07-09月期	0.02
10位	5トリップ	0.01	鉄道分担率50～59%	0.02	羽田空港出国	0.03	2トリップ	0.02
11位	自動車分担率0～9%	0.01	鉄道分担率60～69%	0.02	東京都（3～7日宿泊）	0.03	自動車分担率10～19%	0.02
12位	旭川空港出国	0.01	7トリップ	0.02	04-06月期	0.02	6日滞在	0.02
13位	旭川空港入国	0.01	8トリップ	0.02	自動車分担率0～9%	0.02	自動車分担率20～29%	0.02
14位	函館空港出国	0.01	山梨県（0～2日宿泊）	0.02	神奈川県（0～2日宿泊）	0.02	岐阜県（0～2日宿泊）	0.01
15位	青森県（0～2日宿泊）	0.01	9トリップ	0.01	01-03月期	0.02	長野県（0～2日宿泊）	0.01
16位	観光・レジャー	0.01	15～30日滞在	0.01	業務	0.02	バス分担率30～39%	0.01
17位	函館空港入国	0.01	国内線飛行機分担率0～9%	0.01	中国	0.02	業務	0.01
18位	国内線飛行機分担率30～39%	0.00	04-06月期	0.01	07-09月期	0.02	鉄道分担率30～39%	0.01
19位	青森空港出国	0.00	カナダ	0.01	8～10日滞在	0.02	石川県（0～2日宿泊）	0.01
20位	秋田県（0～2日宿泊）	0.00	羽田空港入国	0.01	バス分担率10～19%	0.02	6トリップ	0.01
0.01以上の属性	なし		11～14日滞在，01-03月期		10-12月期，バス分担率30～39%，15～30日滞在，5トリップ，11～14日滞在，4トリップ，6トリップ，千葉県（0～2日宿泊），鉄道分担率60～69%		富山県（0～2日宿泊），3トリップ，台湾	

順位	14.4. 関西		14.6. 九州		14.7. 沖縄		14.1. 解釈困難	
1位	大阪府（0～2日宿泊）	0.40	鉄道分担率20～29%	0.06	沖縄県（0～2日宿泊）	0.07	国内線飛行機分担率0～9%	0.19
2位	京都府（0～2日宿泊）	0.32	福岡県（0～2日宿泊）	0.06	鉄道分担率0～9%	0.05	観光・レジャー	0.14
3位	関西空港出国	0.26	韓国	0.04	バス分担率0～9%	0.03	自動車分担率0～9%	0.11
4位	関西空港入国	0.24	バス分担率40～49%	0.04	不明空海港入国	0.03	07-09月期	0.07
5位	鉄道分担率60～69%	0.16	福岡空港出国	0.04	訪問地不明（0～2日宿泊）	0.03	10-12月期	0.07
6位	バス分担率20～29%	0.14	福岡空港入国	0.03	0日滞在	0.03	2トリップ	0.06
7位	自動車分担率0～9%	0.13	自動車分担率20～29%	0.03	鉄道分担率10～19%	0.02	バス分担率30～39%	0.06
8位	兵庫県（0～2日宿泊）	0.08	バス分担率50～59%	0.03	那覇空港出国	0.02	自動車分担率10～19%	0.05
9位	大阪府（3～7日宿泊）	0.07	大分県（0～2日宿泊）	0.03	那覇空港入国	0.02	鉄道分担率50～59%	0.05
10位	奈良県（0～2日宿泊）	0.07	熊本県（0～2日宿泊）	0.02	2トリップ	0.02	5日滞在	0.05
11位	観光・レジャー	0.05	3日滞在	0.02	自動車分担率40～49%	0.01	台湾	0.04
12位	京都府（3～7日宿泊）	0.05	長崎県（0～2日宿泊）	0.02	自動車分担率50～59%	0.01	韓国	0.04
13位	国内線飛行機分担率0～9%	0.04	鹿児島県（0～2日宿泊）	0.01	沖縄県（3～7日宿泊）	0.01	中国	0.04
14位	広島県（0～2日宿泊）	0.02	香川県（0～2日宿泊）	0.01	家族・知人の訪問	0.01	3トリップ	0.04
15位	8～10日滞在	0.02	観光・レジャー	0.01	バス分担率30～39%	0.00	4トリップ	0.04
16位	和歌山県（0～2日宿泊）	0.02	博多海港入国	0.01	茨城空港出国	0.00	4日滞在	0.04
17位	15～30日滞在	0.02	博多海港出国	0.01	新潟空港出国	0.00	業務	0.03
18位	11～14日滞在	0.01	福岡県（3～7日宿泊）	0.01	07-09月期	0.00	バス分担率20～29%	0.03
19位	04-06月期	0.01	宮崎県（0～2日宿泊）	0.01	新潟空港入国	0.00	01-03月期	0.03
20位	4トリップ	0.01	高松空港出国	0.01	香港	0.00	5トリップ	0.02
0.01以上の属性	鉄道分担率70～79%，大阪府（8日以上宿泊），中国，鉄道分担率50～59%，バス分担率10～19%，5トリップ，10-12月期		なし		なし		04-06月期，成田空港入国，成田空港出国，6日滞在，3日滞在，家族・知人の訪問，8～10日滞在，7日滞在，大阪府（0～2日宿泊），関西空港出国，香港，関西空港入国	

表 3　トピックと名前（2015 年）

順位	15.8. 北海道		15.2. 東京		15.3. 関東		15.4. 中部・関西	
1位	北海道（0～2日宿泊）	0.19	個人旅行	0.11	東京都（0～2日宿泊）	0.52	観光・レジャー	0.10
2位	団体旅行	0.08	国内線飛行機分担率0～9%	0.09	成田空港出国	0.23	自動車分担率0～9%	0.10
3位	鉄道分担率0～9%	0.06	鉄道分担率50～59%	0.08	成田空港入国	0.23	国内線飛行機分担率0～9%	0.08
4位	5日滞在	0.05	自動車分担率10～19%	0.06	東京都（3～7日宿泊）	0.11	中国	0.06
5位	観光・レジャー	0.05	業務	0.05	神奈川県（0～2日宿泊）	0.07	団体旅行	0.06
6位	台湾	0.04	2トリップ	0.05	千葉県（0～2日宿泊）	0.07	訪日回数不明	0.05
7位	訪日回数不明	0.04	羽田空港入国	0.04	訪日1回目	0.06	10-12月期	0.05
8位	新千歳空港入国	0.04	羽田空港出国	0.04	鉄道分担率40～49%	0.05	6日滞在	0.04
9位	新千歳空港出国	0.03	01-03月期	0.04	15～30日滞在	0.04	07-09月期	0.04
10位	自動車分担率0～9%	0.03	バス分担率20～29%	0.03	バス分担率20～29%	0.04	鉄道分担率0～9%	0.03
11位	バス分担率70～79%	0.03	04-06月期	0.03	11～14日滞在	0.04	大阪府（0～2日宿泊）	0.03
12位	5トリップ	0.03	3トリップ	0.02	鉄道分担率50～59%	0.03	京都府（0～2日宿泊）	0.03
13位	鉄道分担率10～19%	0.02	バス分担率30～39%	0.02	個人旅行	0.03	静岡県（0～2日宿泊）	0.03
14位	鹿児島県（0～2日宿泊）	0.02	韓国	0.02	米国	0.03	5トリップ	0.02
15位	07-09月期	0.02	訪日20回以上	0.02	自動車分担率10～19%	0.02	6トリップ	0.02
16位	香川県（0～2日宿泊）	0.02	東京都（0～2日宿泊）	0.02	山梨県（0～2日宿泊）	0.02	関西空港入国	0.02
17位	10-12月期	0.01	自動車分担率0～9%	0.02	東京都（8日以上宿泊）	0.02	4トリップ	0.02
18位	北海道（3～7日宿泊）	0.01	訪日10～19回目	0.02	04-06月期	0.02	7日滞在	0.01
19位	宮崎県（0～2日宿泊）	0.01	10-12月期	0.01	国内線飛行機分担率0～9%	0.02	5日滞在	0.01
20位	高松空港出国	0.01	3日滞在	0.01	バス分担率10～19%	0.02	愛知県（0～2日宿泊）	0.01
0.01以上の属性	なし		訪日2回目，07-09月期，訪日6～9回目，訪日1回目，観光・レジャー		8～10日滞在，01-03月期，広島県（0～2日宿泊），7トリップ，国内線飛行機分担率10～19%，オーストラリア，イタリア，栃木県（0～2日宿泊），8トリップ		台湾，岐阜県（0～2日宿泊），石川県（0～2日宿泊），関西空港出国	

順位	15.1. 関西		15.5. 愛知・広島		15.7. 九州		15.6. 沖縄	
1位	大阪府（0～2日宿泊）	0.11	国内線飛行機分担率0～9%	0.19	福岡県（0～2日宿泊）	0.06	沖縄県（0～2日宿泊）	0.05
2位	自動車分担率0～9%	0.10	個人旅行	0.17	韓国	0.05	4日滞在	0.05
3位	関西空港出国	0.08	2トリップ	0.14	鉄道分担率20～29%	0.03	観光・レジャー	0.05
4位	バス分担率30～39%	0.08	自動車分担率0～9%	0.11	大分県（0～2日宿泊）	0.03	韓国	0.05
5位	関西空港入国	0.08	07-09月期	0.10	福岡空港出国	0.03	国内線飛行機分担率0～9%	0.03
6位	国内線飛行機分担率0～9%	0.08	訪日回数不明	0.10	福岡空港入国	0.03	バス分担率30～39%	0.03
7位	観光・レジャー	0.07	10-12月期	0.09	3日滞在	0.02	鉄道分担率10～19%	0.03
8位	京都府（0～2日宿泊）	0.06	バス分担率0～9%	0.09	バス分担率40～49%	0.02	個人旅行	0.02
9位	個人旅行	0.06	1～2日滞在	0.06	自動車分担率20～29%	0.02	那覇空港入国	0.02
10位	訪日1回目	0.06	愛知県（0～2日宿泊）	0.05	熊本県（0～2日宿泊）	0.02	那覇空港出国	0.02
11位	鉄道分担率60～69%	0.04	家族・知人の訪問	0.05	観光・レジャー	0.01	自動車分担率50～59%	0.02
12位	3トリップ	0.04	鉄道分担率30～39%	0.05	バス分担率50～59%	0.01	5日滞在	0.01
13位	4トリップ	0.03	バス分担率50～59%	0.04	長崎県（0～2日宿泊）	0.01	台湾	0.01
14位	大阪府（3～7日宿泊）	0.03	中部空港出国	0.04	自動車分担率10～19%	0.01	訪日2回目	0.01
15位	兵庫県（0～2日宿泊）	0.02	鉄道分担率0～9%	0.04	4日滞在	0.01	10-12月期	0.01
16位	04-06月期	0.02	バス分担率40～49%	0.04	国内線飛行機分担率0～9%	0.01	07-09月期	0.01
17位	8～10日滞在	0.02	中部空港入国	0.04	博多海港入国	0.01	自動車分担率40～49%	0.01
18位	鉄道分担率50～59%	0.02	観光・レジャー	0.03	博多海港出国	0.01	香港	0.01
19位	訪日2回目	0.02	中国	0.03	4トリップ	0.00	沖縄県（3～7日宿泊）	0.01
20位	01-03月期	0.02	不明空海港入国	0.03	福岡県（3～7日宿泊）	0.00	訪日3回目	0.01
0.01以上の属性	韓国，07-09月期，10-12月期，5日滞在，奈良県（0～2日宿泊），京都府（3～7日宿泊）		研修・学会等，3トリップ，米国，その他，広島県（0～2日宿泊），8～10日滞在，7日滞在，訪日1回目，業務，フィリピン，鉄道分担率40～49%，31～90日滞在，留学		なし		なし	

表 4　トピックと名前（2016 年）

順位	16.8.北海道		16.1.関東		16.4.関東・中部・関西・広島		16.5.中部	
1位	北海道（0～2日宿泊）	0.12	東京都（0～2日宿泊）	0.17	訪日1回目	0.23	国内線飛行機分担率0～9%	0.11
2位	鉄道分担率10～19%	0.02	成田空港出国	0.09	自動車分担率0～9%	0.18	団体旅行	0.11
3位	新千歳空港入国	0.02	成田空港入国	0.09	観光・レジャー	0.17	中国	0.10
4位	バス分担率60～69%	0.02	鉄道分担率50～59%	0.08	京都府（0～2日宿泊）	0.16	観光・レジャー	0.10
5位	新千歳空港出国	0.02	個人旅行	0.05	東京都（0～2日宿泊）	0.15	自動車分担率0～9%	0.08
6位	自動車分担率10～19%	0.02	業務	0.04	個人旅行	0.15	訪日回数不明	0.07
7位	団体旅行	0.02	自動車分担率10～19%	0.04	国内線飛行機分担率0～9%	0.14	10-12月期	0.06
8位	観光・レジャー	0.01	バス分担率30～39%	0.04	バス分担率20～29%	0.12	鉄道分担率0～9%	0.06
9位	台湾	0.01	羽田空港入国	0.04	8～10日滞在	0.10	愛知県（0～2日宿泊）	0.06
10位	5日滞在	0.01	東京都（3～7日宿泊）	0.04	11～14日滞在	0.07	6日滞在	0.06
11位	香川県（0～2日宿泊）	0.01	羽田空港出国	0.04	15～30日滞在	0.07	07-09月期	0.06
12位	5トリップ	0.01	国内線飛行機分担率0～9%	0.03	東京都（3～7日宿泊）	0.06	5日滞在	0.04
13位	バス分担率70～79%	0.01	千葉県（0～2日宿泊）	0.02	04-06月期	0.06	静岡県（0～2日宿泊）	0.04
14位	北海道（3～7日宿泊）	0.01	神奈川県（0～2日宿泊）	0.02	5トリップ	0.06	台湾	0.04
15位	函館空港出国	0.01	01-03月期	0.02	広島県（0～2日宿泊）	0.05	中部空港出国	0.03
16位	高松空港出国	0.01	米国	0.02	バス分担率10～19%	0.05	6トリップ	0.03
17位	高松空港入国	0.00	04-06月期	0.01	京都府（3～7日宿泊）	0.05	中部空港入国	0.03
18位	愛媛県（0～2日宿泊）	0.00	訪日20回以上	0.01	成田空港入国	0.04	岐阜県（0～2日宿泊）	0.02
19位	旭川空港出国	0.00	鉄道分担率40～49%	0.01	4トリップ	0.04	7日滞在	0.02
20位	函館空港入国	0.00	東京都（8日以上宿泊）	0.01	7トリップ	0.04	石川県（0～2日宿泊）	0.02
0.01以上の属性	なし		なし		07-09月期，山梨県（0～2日宿泊），鉄道分担率60～69%，成田空港出国，鉄道分担率70～79%，奈良県（0～2日宿泊），鉄道分担率40～49%など		鹿児島県（0～2日宿泊），鉄道分担率30～39%，5トリップ，富山県（0～2日宿泊），バス分担率40～49%，香港，個人旅行，4トリップ	

順位	16.2.関西		16.7.九州		16.6.沖縄		16.3.解釈困難	
1位	大阪府（0～2日宿泊）	0.14	福岡県（0～2日宿泊）	0.08	沖縄県（0～2日宿泊）	0.12	2トリップ	0.11
2位	関西空港出国	0.09	韓国	0.06	観光・レジャー	0.11	国内線飛行機分担率0～9%	0.09
3位	関西空港入国	0.09	福岡空港出国	0.04	4日滞在	0.10	個人旅行	0.09
4位	自動車分担率0～9%	0.09	福岡空港入国	0.04	国内線飛行機分担率0～9%	0.09	訪日回数不明	0.07
5位	鉄道分担率60～69%	0.07	鉄道分担率20～29%	0.04	韓国	0.09	10-12月期	0.06
6位	国内線飛行機分担率0～9%	0.07	大分県（0～2日宿泊）	0.03	個人旅行	0.08	07-09月期	0.06
7位	個人旅行	0.06	3日滞在	0.03	バス分担率20～29%	0.07	自動車分担率0～9%	0.05
8位	観光・レジャー	0.06	バス分担率40～49%	0.03	鉄道分担率0～9%	0.05	バス分担率0～9%	0.04
9位	3トリップ	0.05	バス分担率50～59%	0.03	那覇空港入国	0.04	1～2日滞在	0.03
10位	京都府（0～2日宿泊）	0.04	自動車分担率20～29%	0.02	那覇空港出国	0.04	鉄道分担率0～9%	0.02
11位	バス分担率30～39%	0.03	自動車分担率10～19%	0.02	訪日2回目	0.03	家族・知人の訪問	0.02
12位	大阪府（3～7日宿泊）	0.03	観光・レジャー	0.02	5日滞在	0.03	観光・レジャー	0.02
13位	韓国	0.03	長崎県（0～2日宿泊）	0.01	4トリップ	0.03	バス分担率50～59%	0.01
14位	バス分担率20～29%	0.03	国内線飛行機分担率0～9%	0.01	01-03月期	0.02	不明空海港入国	0.01
15位	4トリップ	0.03	個人旅行	0.01	自動車分担率40～49%	0.02	中国	0.01
16位	訪日1回目	0.02	熊本県（0～2日宿泊）	0.01	04-06月期	0.02	3日滞在	0.01
17位	兵庫県（0～2日宿泊）	0.02	4日滞在	0.01	台湾	0.02	韓国	0.01
18位	04-06月期	0.02	01-03月期	0.01	5トリップ	0.02	自動車分担率50～59%	0.01
19位	訪日2回目	0.02	佐賀県（0～2日宿泊）	0.01	訪日3回目	0.02	その他	0.01
20位	01-03月期	0.02	博多海港入国	0.01	07-09月期	0.01	5日滞在	0.01
0.01以上の属性	5日滞在，奈良県（0～2日宿泊）		なし		訪日4回目，沖縄県（3～7日宿泊），3トリップ，訪日1回目，自動車分担率30～39%		なし	

5．　訪日外国人旅行特性と経年変化

(1)　地方別訪日外国人旅行特性

北海道を周遊するトピックは，台湾，観光レジャー目的，団体旅行，5 日滞在，バス分担率 60～69，70～79%などが上位である．関東を周遊するトピックは，中国，アメリカ，業務，4～6月期，鉄道分担率50～59%などが上位である．中部を周遊するトピックは，中国，台湾，観光レジャー，6 日滞在，7～9 月期，10～12 月期などが上位である．関西を周遊するトピックは，韓国，観光レジャー，個人旅行，訪日 2回目，4～6 月期，バス分担率 20～29%，鉄道分担率 60～69%などが上位である．九州を周遊するトピックは，韓国，観光レジャー，3 日滞在，バス分担率 40～49，50～59%，自動車分担率 20～29%などが上位である．沖縄を周遊するトピックは，台湾，観光レジャー，個人旅行，訪日 2回目，自動車分担率 40～49，50～59%などが上位である．複数地方を周遊するトピックは，観光レジャー，訪日 1 回目，8~10，11～14，15～30 日滞在などが上位である．解釈困難トピックは，訪問地属性が上位に現れないトピックであり，中国，韓国，台湾，観光レジャー，業務，家族・知人の訪問などが上位である．

(2)　訪日外国人旅行特性の経年変化

北海道，中部，九州を周遊するトピックは，3 年ともほぼ同じ属性で構成される．関東，関西を周遊するトピックは 2014 年では東京都（0～2 日宿泊），大阪府（0～2 日宿泊），京都府（0～2 日宿泊）の属性の寄与が大きいが，2015，2016 年ではそれらの寄与は減少している．沖縄を周遊するトピックは，2014 年では 0 日滞在が上位である．同トピックは 2015 年より 4 日滞在，5 日滞在が上位である．

(3)　トピック構成比率の経年変化

トピック内の属性の構成比率はサンプル数を基に算出しているが，トピックごとの構成比率は FF データの拡大係数を基に算出した．以下に，2014～2016 年のトピックの構成比率を示す．関東，関西を周遊するトピックの構成比率が大きく，観光地である北海道，沖縄を周遊しているトピックも確認できた．これらの結果から，トピック解釈の妥当性が確認できる．

関西，九州，沖縄，複数地方トピックの構成比率は，経年的に増加している．北海道，関東，中部トピックの構成比率は増減している．解釈困難トピックの構成比率は，年によるばらつきが大きい．

６． おわりに

本研究では，訪日回数や国籍，季節，年別による訪日外国人の旅行特性の違いを把握するため，FF データを用いて彼らの周遊に関する特性を経年分析した．分析手法として，元来は文章解析に用いられるトピックモデルを活用した，このため，FF データに含まれる連続変数について離散化処理を行うことによって，BOW 表現を得る手法を提案した．その結果，膨大な訪日外国人の訪問パターンや訪問時期，利用交通機関等の組み合わせから，特徴的なパターンを抽出することがでえきた．また，トピックモデルを用いた既往研究で課題としてあげられていたトピック数の決定手順や，トピックの解釈手順については，以下の工夫を行うことにした．すなわち，尤度比とトピック間の類似度を指標としてトピック数の候補を選定し，各トピック数から得られる分析結果に基づいて，その解釈性に配慮したトピック数を採用した．また，本データへのトピックモデルの適用を通して，サンプル数が少なく語彙数が多い文章データではなく，サンプル数が多く語彙数が少ない FF データに対してもトピックモデルが有効なことを示した．

分析結果から，訪日外国人旅行特性は，周遊する地方で異なることがわかった．北海道，中部，九州を周遊するるお香では，経年的な特性の変化は少なかった．関東，関西を目的地とする旅行では東京都，大阪府，京都府以外の訪問割合が高い，沖縄を目的地とする旅行では滞在日数が増加している．さらに，2015 年に複数地方を周遊するトピックが現れ，2015 年，2016 年では，関東以外を目的地とする旅行が大きく増加しており，全体的な旅行特性は，多様化していた．以上より，トピックモデルを用いれば訪日外国人の旅行特性の経年変化を効果的に把握できることが示された．

今後の課題を以下にまとめる．本研究では，2014 年，2015 年，2016 年ごとにデータを分割してトピックモデルを適用した．しかし，たとえば季節ごとに分ける方法，出国海空港ごとに分ける方法など，他のデータの分割方法では，異なるトピックが得られる可能性がある．これらの方法から得られるトピックとの比較により，類似しないトピック数が最大となる入力データの設定手法を検討する必要がある．

本研究では，類似しないトピック数が最大となるトピック数をもとに最大トピック数を決め，その範囲での尤度比の極大値と，類似しない最大のトピック数を参照して，トピック数を選択した．ただし，その範囲では尤度比は上昇傾向にあり，計算時間の制約からより大きなトピック数で現れる可能性のある尤度の極大値については検討できていない．また，類似判定の基準としてコサイン類似度に閾値を設定したが，その値の妥当性の検証は十分ではない．今後は，尤度比の極大値の選択，閾値の選択について，検討を行う必要がある．

参考文献

1) 国土交通省：FF-Data（訪日外国人流動データ），http://www.mlit.go.jp/sogoseisaku/soukou/sogoseisaku_soukou_fr_000022.html（アクセス：2019 年 2 月 28 日）
2) 日本政府観光局（JNTO）：ビジット・ジャパン事業について，https://www.jnto.go.jp/jpn/projects/promotion/vj/index.html（アクセス：2019 年 2 月 28 日）
3) 観光庁：訪日旅行促進事業について，http://www.mlit.go.jp/kankocho/shisaku/kokusai/vjc.html（アクセス：2019 年 2 月 28 日）
4) 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，Vol.68，No.5，pp667-677，2012.
5) 松井裕樹・日比野直彦・森知茂・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行動に着目した観光行動分析，土木学会論文集 D3，Vol.72，No.5，pp533-546，2016.
6) 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の分析，土木学会論文集 D3，Vol.72，No.5，pp571-583，2016.
7) 塚井誠人・椎野 創介：討議録に対するトピックモデルの適用，土木学会論文集 D3，Vol.72，No.5，pp341-352，2016.
8) 塚野裕太：トピックモデルを用いた土地利用メッシュ特性の抽出，平成 28 年度卒業論文
9) 川野倫輝・佐藤嘉洋・円山琢也：トピックモデルと離散連続モデルを用いた自由記述の量的分析，土木学会論文集 D3，Vol.74，No.5，pp277-284，2018.
10) 古屋秀樹・岡本直久・野津直樹：GPS ログモデル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.20，pp20-29，2018.
11) 岩田具治：トピックモデル，講談社，2015.
12) 中島伸一：変分ベイズ学習，講談社，2016.
13) 坪井祐太，海野裕也，鈴木潤：深層学習による自然言語処理，講談社，2017.
14) 法務省：出入国管理統計，http://www.moj.go.jp/housei/toukei/housei05_00016.html（アクセス：2019 年 2 月 28 日）
15) 国土交通省：航空局実施の統計調査，http://www.mlit.go.jp/koku/koku_tk6_000001.html（アクセス：2019 年 2 月 27 日）
16) 観光庁：訪日外国人消費動向調査，http://www.mlit.go.jp/kankocho/siryou/toukei/syouhityousa.html（アクセス：2019 年 2 月 27 日）

（2019. 03. 10 受付）

A TOUR OF ANALYSIS OF FOREIGN VISITORS TO JAPAN BY APPLYING TOPIC MODEL

Yoshihiro TATSUMI and Makoto TSUKAI

In recent years, Japanese government implements Visit JAPAN Campaign and Promotion of travel to Japan. Further policy making requires to identify the tour trip characteristics in terms of visiting spots, tour schedule, accommodation or trip modes. In order to find significant and representative tour or trip patterns in foreign visitors to Japan, an efficient analytical tool to the above problem should be developed. This study applies a topic model, which can efficiently analyze the latent patterns in the Bag-of-words data set, to the data in Flow of Foreigners visiting to Japan, and clarifies the change of tour trip patterns. In conclusion, the topic model can clearly show the tour trip patterns and their characteristics of the foreign visitors to Japan. The tour trip patterns visiting to Kanto and Kansai has been diversified, and the other tour trip patterns visiting several different regions called “Golden Route” has increased.