訪日外国人旅行者の鉄道利用データを用いた周遊パターン抽出に関する基礎的研究

稲場　亘[1]・中川　伸吾[2]・渡邉　拓也[3]・深澤　紀子[4]

[1]正会員　（公財）鉄道総合技術研究所　情報解析研究室（〒185-8540 東京都国分寺市光町 2-8-38）
E-mail: inaba.wataru.54@rtri.or.jp

[2]正会員　（公財）鉄道総合技術研究所　情報解析研究室（〒185-8540 東京都国分寺市光町 2-8-38）
E-mail: nakagawa.shingo.39@rtri.or.jp

[3]正会員　（公財）鉄道総合技術研究所　情報解析研究室（〒185-8540 東京都国分寺市光町 2-8-38）
E-mail: watanabe.takuya.42@rtri.or.jp

[4]正会員　（公財）鉄道総合技術研究所　情報解析研究室（〒185-8540 東京都国分寺市光町 2-8-38）
E-mail: fukasawa.noriko.11@rtri.or.jp

インバウンド需要の回復に向けて，鉄道事業者は訪日外国人旅行者の受入環境整備を目的とした施策に取り組んでいる．適切な施策を打つためには，最盛期の訪日外国人旅行者の鉄道利用行動を把握し，多くの知見を積み重ねていくことが重要である．本研究では，訪日外国人の代表的な鉄道利用行動を明らかにすることを目的に，駅の入出場を表すデータを活用した分析を行った．具体的には入出場データの中から降車した駅の組合せに着目し，トピックモデルを用いて代表的な周遊パターンを抽出した．その結果，大都市圏内を周遊する訪日外国人の周遊パターンの特徴が，駅事業者，周遊都道府県，出現時期の観点から把握可能となり，施策への活用可能性について知見が得られた．

Key Words: *foreign visitors, railway, latent dirichlet allocation, biterm topic model*

1. はじめに

観光産業は我が国の成長産業であり，鉄道事業者にとっても，訪日外国人旅行者の移動需要を満たすことは重要な課題である．鉄道事業者はインバウンド受入環境の整備を目的とした様々な施策を実施しており[1)]，旅行商品の開発や輸送計画の改善に取り組んでいる．その一方で，これらの施策の妥当性をデータから検証したり，定量的な根拠を基に新たな施策を提案したりする試みは十分に進んでいないのが現状である．訪日外国人観光客は日本国内を移動する際に，鉄道やバスを中心とした公共交通機関を利用する割合が大きく，特に関東地方では 57.5%，近畿地方では 64.4%と，高い鉄道分担率を示している[2)]．そのため，訪日外国人の鉄道利用行動を明らかにすることで，鉄道事業者が適切な施策を打つことができるだけでなく，訪日外国人の旅行全体の満足度向上に寄与すると考えられる．

以上を踏まえて，本研究では鉄道事業者が持つ駅の入出場を表すデータ（以下，入出場データと表記）から訪日外国人の鉄道利用行動を明らかにして，施策に活用可能な知見を得ることを目的とする．具体的には，訪日外国人のみが購入可能な商品（以下，訪日商品と表記）から得られる入出場データの中から降車した駅の組合せに着目し，トピックモデルを用いて訪日外国人の代表的な周遊パターンを抽出する．トピックモデルは文書の生成過程を確率的に記述するモデルであるが，応用範囲が広く本課題への適用も可能である．分析には入出場データの特徴を踏まえ，最も代表的なトピックモデルである LDA（Latent Dirichlet Allocation）[3)]に加えて，単語数が少ない文書への適用が有効な BTM（Biterm Topic Model）[4)]を活用する．大都市圏における一日単位の鉄道利用行動に対して LDA と BTM を適用し，説明力と解釈性について評価した結果を示す．さらに，得られた結果を施策に活用するために，抽出された周遊パターンの特徴を駅事業者，周遊都道府県，出現時期の観点から考察する．以上を通して，具体的な施策への展開方法について提案する．

第2章では，既往研究の整理と本研究の位置づけについて述べる．第3章では，本研究で周遊パターンの抽出に活用するトピックモデルの概要を述べる．第4章では，本研究で使用する入出場データの基礎分析を行った結果を示す．第5章では，入出場データにトピックモデルを適用した結果を示すとともに，抽出された周遊パターンを考察して施策への活用方法を検討する．第6章では，本研究のまとめと今後の課題を述べる．

2. 既往研究の整理と本研究の位置づけ

訪日外国人の鉄道利用行動に着目した研究事例は少ないが，訪日外国人の観光行動に着目した研究事例は多く存在する．そこで本章では，訪日外国人に関する観光行動特性の分析や周遊パターンの抽出を行った研究について整理した上で，本研究の位置づけを示す．

菱田ら [5]は，JNTO 訪日外客訪問地調査のデータにクラスター分析を適用し，居住地域や訪日回数による訪日中国人旅行者の観光行動の違いを明らかにした．矢部ら [6]は，訪日外国人向けの IC 乗車券データに配列解析を適用し，東京大都市圏を周遊する旅行者の行動を 10 パターンに類型化した．ただし，以上の研究で用いられた手法では周遊パターンを明示的にモデル化できず，旅行者の訪問地選択過程を把握できない．観測できない周遊パターンを潜在変数として導入し，観測可能な訪問地から旅行者間の関係性を説明できる手法が潜在クラス分析である．古屋ら [7]は，訪日外国人消費動向調査の個票データに潜在クラス分析を適用し，都道府県単位の旅行者の周遊を 24 パターンに類型化した．ただし，潜在クラス分析ではパラメータを定数として推定するため局所解や過学習の問題に陥りやすく [8]，少数の周遊パターンが過大推計される可能性がある．

これらの課題を解決するにはトピックモデルの活用が有効である．トピックモデルでは旅行者の訪問地選択過程をパラメータの過学習を押さえてモデル化できる．辰巳ら [9]は，訪日外国人流動データ（以下，FF データという）にトピックモデルを適用し，都道府県単位の周遊傾向の特徴を明らかにした．古屋ら [10]は，FF データよりも集計単位が細かい訪日外国人消費動向調査の個票データに階層的トピックモデルを適用し，訪問場所の組合せパターンを詳細に分類して国籍や訪日回数との関係を明らかにした．アンケート調査ではなく，旅行者から直接得られる行動履歴を分析した事例も存在する．古屋ら [11]は，GPS ログデータにトピックモデルを適用し，日本全国の代表的な 6 つの周遊パターンを抽出して，個人属性との関係や観光地の特徴を明らかにした．以上をはじめ，訪日外国人の周遊パターン把握に関する研究の多くは日本全国を対象としており，特定地域の周遊を対象とした事例は少ない．鉄道事業者が施策に活かせる知見を得るためには，より狭いエリアにおける駅単位，観光地単位の周遊パターンを詳細に把握できることが好ましく，入出場データの活用が有効だと考えられる．

一般に入出場データから移動目的や目的地を正確に把握することはできない．ただし，訪日商品の利用者の移動目的は基本的に観光であり，鉄道でアクセス可能な観光地が多い地域では目的地を推測しやすい．加えて入出場データには，長期間の定点観測が可能で変動傾向が分析しやすい，施策主体が独自に収集しているためデータ取得に追加のコストがかからない，といったメリットがある．

以上を踏まえて，本研究では入出場データにトピックモデルを適用する方法を提案し，大都市圏を周遊する訪日外国人旅行者の鉄道利用行動を対象にケーススタディを行う．地理的に近いエリアで解釈性の高い周遊パターンが抽出可能であるか検証し，具体的な施策への活用可能性についても言及する．

3. トピックモデルによる訪問地選択過程の定式化

トピックモデルとは人間が文書を書く過程を「著者が文書のトピックを決め，トピックと関連の深い単語を選ぶ」と仮定して，文書の生成過程を確率的に記述するモデルである．この考え方を旅行者が旅程を決める過程について適用すると，「旅行のトピックを決めてからトピックと関連の深い訪問地を選ぶ」と仮定でき，トピックを周遊パターン，著者を旅行者，文書を各旅行者の旅程，単語を訪問地と読み替えることができる．すなわち，旅行者の訪問地選択過程をトピックモデルで記述することが可能になる．この手法を用いて，本研究では代表的な周遊パターンの抽出や旅行者の分類を試みる．本章では旅行者が訪問地を決める過程について，LDA と BTM の 2 つのモデルを使った定式化を行う．BTM とは単語数の少ない文書データに対して，解釈性の高いトピックが抽出可能なトピックモデルである．なお，定式化の際の表記等に関して LDA は佐藤ら [12]の文献，岩田ら [13]の文献，BTM は Cheng ら [4]の報告を踏襲している．

(1) LDA による定式化

LDA では全体の選択過程を，旅行者が周遊パターンを選択する過程と周遊パターンに応じた訪問地を選択する過程の 2 つに分け，それぞれ多項分布を用いて式(1),(2)のように表す．

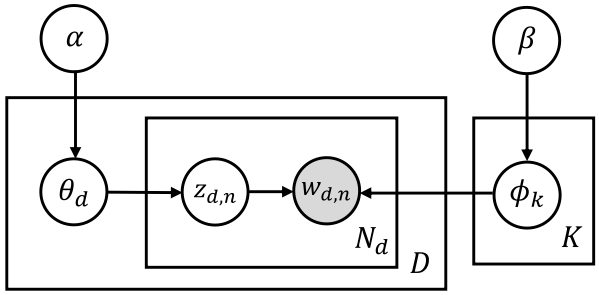

図-1　LDA のグラフィカルモデル

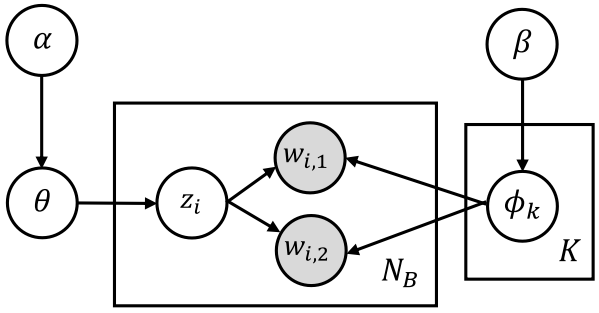

図-2　BTM のグラフィカルモデル

$$z_{d,n}\sim\mathrm{Multi}(\theta_d)=p(z_{d,n}|\theta_d) \quad (1)$$

$$w_{d,n}\sim\mathrm{Multi}(\phi_{z_{d,n}})=p(w_{d,n}|\phi_{z_{d,n}}) \quad (2)$$

ここに,

$z_{d,n}$：旅行者dのn番目の周遊パターン

$\theta_d=(\theta_{d,1},\dots,\theta_{d,K})$：旅行者$d$が周遊パターン 1,…,$K$をそれぞれ選択する確率のベクトル，トピック分布

$w_{d,n}$：旅行者dのn番目の訪問地

$\phi_k=(\phi_{k,1},\dots,\phi_{k,V})$：周遊パターン$k$で訪問地 1,…,$V$をそれぞれ選択する確率のベクトル，単語分布

D：旅行者の総数，N_d：旅行者dの訪問地の総数,

K：周遊パターンの総数，V：ユニークな訪問地の総数

である．**図-1** は LDA のグラフィカルモデルである．θ_dは旅行者dが持つ周遊パターンの嗜好と解釈することができ，LDA では各旅行者が異なるθを持つが，周遊パターンから各観光地を選択する確率ϕは旅行者によらないと仮定する．LDA ではこれらのパラメータを確率分布として推定するために，θ_dとϕ_kの生成過程を Dirichlet 分布を用いて式(3),(4)のように表す.

$$\theta_d\sim\mathrm{Dir}(\alpha)=p(\theta_d|\alpha) \quad (3)$$

$$\phi_k\sim\mathrm{Dir}(\beta)=p(\phi_k|\beta) \quad (4)$$

ここに,

α：θ_dを制御するK次元のパラメータベクトル

β：ϕ_kを制御するV次元のパラメータベクトル

であり，αとβはハイパーパラメータである．パラメータθ_d,ϕ_kと潜在変数$z_{d,n}$は実際には観測できないため，観測可能な各旅行者の訪問地から推定する．訪問地Wとハイパーパラメータα,βで条件付けられたパラメータΘ,Φと潜在変数Zの同時分布は，式(5)のように表せる．

$$p(Z,\Theta,\Phi|W,\alpha,\beta)=\frac{p(Z,\Theta,\Phi,W|\alpha,\beta)}{p(W|\alpha,\beta)}\propto p(Z,\Theta,\Phi,W|\alpha,\beta) \quad (5)$$

ここに,

$$Z=(z_{1,1},\dots,z_{D,N_D}),\ \Theta=(\theta_1,\dots,\theta_D),$$

$$\Phi=(\phi_1,\dots,\phi_K),\ W=(w_{1,1},\dots,w_{D,N_D})$$

である．式(5)は解析的に解くことが難しいため，周辺化ギブスサンプリングを用いてパラメータ推定を行う．ハイパーパラメータα,βの初期値はともに 1.0 とし，サンプリングの度に不動点反復法を用いて更新する．

(2)　BTM による定式化

BTM とは単語の共起を明示的にモデル化し，文書全体で共起関係を学習することで，単語数が少ない文書で解釈性の高いトピックの抽出を図る手法である．旅行の訪問地選択では，各個人の訪問地数が少ない場合への適用に適していると考えられる．BTM では各旅行者の訪問地の組合せ（以下，訪問地ペアと表記）を作成し，旅行者全体で周遊パターンと訪問地ペアを選択する過程をモデル化する．これら2つの過程を，LDA と同様に多項分布を用いて式(6),(7)のように表す.

$$z_i\sim\mathrm{Multi}(\theta)=p(z_i|\theta)\quad(i=1,\dots,N_B) \quad (6)$$

$$b_i\sim\mathrm{Multi}(\phi_{z_i})=p(b_i|\phi_{z_i})\quad(i=1,\dots,N_B) \quad (7)$$

ここに,

z_i：旅行者全体のi番目の周遊パターン

$\theta=(\theta_1,\dots,\theta_K)$：旅行者全体で周遊パターン 1,…,$K$をそれぞれ選択する確率のベクトル，トピック分布

$b_i=(w_{i,1},w_{i,2})\quad(i=1,\dots,N_B)$：旅行者全体の$i$番目の訪問地ペア

$w_{i,1}$：i番目の訪問地ペアの 1 つ目の訪問地

$w_{i,2}$：i番目の訪問地ペアの 2 つ目の訪問地

$\phi_k=(\phi_{k,1},\dots,\phi_{k,V})$：周遊パターン$k$で訪問地 1,…,$V$をそれぞれ選択する確率のベクトル，単語分布

N_B：訪問地ペアの総数

である．**図-2** は BTM のグラフィカルモデルである．BTM では旅行者全体で一つの周遊パターンの嗜好θを持ち，周遊パターンから各観光地を選択する確率ϕは旅行者によらないと仮定する．ただし，全ての旅行者が共通の嗜好を持つわけではなく，事後的に旅行者dの嗜好θ_dを算出できる．BTM でもこれらのパラメータの生成過程を Dirichlet 分布を用いて式(8),(9)のように表す.

$$\theta\sim\mathrm{Dir}(\alpha)=p(\theta|\alpha) \quad (8)$$

$$\phi_k\sim\mathrm{Dir}(\beta)=p(\phi_k|\beta) \quad (9)$$

ここに,
α：θを制御するK次元パラメータベクトル
β：ϕ_kを制御するV次元パラメータベクトル
であり，αとβはハイパーパラメータである．パラメータθ, ϕ_kと潜在変数z_iは実際には観測できないため，観測可能な訪問地ペアから推定する．訪問地ペアBとハイパーパラメータα, βで条件付けられたパラメータΘ, Φと潜在変数Zの同時分布は，式(10)のように表せる．

$$p(Z, \theta, \Phi | B, \alpha, \beta) = \frac{p(Z, \theta, \Phi, B | \alpha, \beta)}{p(B | \alpha, \beta)} \propto p(Z, \theta, \Phi, B | \alpha, \beta) \tag{10}$$

ここに,

$$Z = (z_1, \dots, z_{N_B}),\ \Phi = (\phi_1, \dots, \phi_V),\ B = (b_1, \dots, b_{N_B})$$

である．式(10)は解析的に解くことが難しいため，周辺化ギブスサンプリングを用いてパラメータ推定を行う．ハイパーパラメータα, βの値はともに 1.0 とした．

BTM では，各旅行者が持つ訪問地ペアに対して推定された周遊パターンを集計することで，旅行者dの周遊パターンkの嗜好$\theta_{d,k}$を式(11)から算出できる．

$$\begin{aligned} & p(z = k | d) \\ &= \sum_{i=1}^{N_d} p\left(z = k | b_i^{(d)}\right) p\left(b_i^{(d)} | d\right) \\ &= \sum_{i=1}^{N_d} \frac{\theta_k \phi_{k, w_{i,1}^{(d)}} \phi_{k, w_{i,2}^{(d)}}}{\sum_{k=1}^{K} \left(\theta_k \phi_{k, w_{i,1}^{(d)}} \phi_{k, w_{i,2}^{(d)}}\right)} \times \frac{N\left(b_i^{(d)}\right)}{\sum_{i=1}^{N_d} N\left(b_i^{(d)}\right)} \\ &= \theta_{d,k} \end{aligned} \tag{11}$$

ここに,
N_d：旅行者dの訪問地ペアの総数
$b_i^{(d)}$：旅行者dの訪問地ペア　$(i = 1, \dots, N_d)$
$N\left(b_i^{(d)}\right)$：旅行者dの$b_i^{(d)}$の個数　$(i = 1, \dots, N_d)$
である．

4. ケーススタディ

(1) データセットの概要と前処理

本研究では，2016 年 4 月 1 日から 2019 年 12 月 31 日の期間に，日本国内の大都市圏内を周遊する訪日外国人旅行者の入出場データを分析する．具体的には同じ地方区分に含まれ，特に利用実績が多い 4 都道府県のみを周遊する旅行者を対象とする．本研究の目的は訪問地の組合せから代表的な周遊パターンを抽出することであるため，乗換時に記録されたデータを削除し，トリップを表す入出場データ 5,674,230 件を集計してデータセットを作成する．なお，本研究では降車駅を訪問地，鉄道利用パターンを周遊パターンと仮定する．データセット作成時には旅行単位を決める必要がある．本研究では同一利用者のトリップを一日単位で集計して降車駅を抽出したデータセットを作成する．さらに，結果を解釈しやすくするために，以下の前処理を行う．
ステップ 1：各旅行者の最後の降車駅を削除
ステップ 2：出現割合が 0.001 未満の駅を削除
ステップ 3：降車駅数が 2 駅以上の旅行者を抽出
ステップ 1 は宿泊や出国のために降車した駅を削除するための処理である．以上の前処理を行った結果，旅行者数は 594,870，平均降車駅数は 2.46，ユニーク旅行者数は 260,261，ユニーク降車駅数は 157 である．

(2) 基礎分析

図-3に，降車駅数別の旅行者数を示す．**図-3**に示すように，降車駅数が 2 駅の旅行者が最も多く，降車駅数が 4 駅以下の旅行者が全体の約 98%を占めている．

図-4 に，周遊状況別の旅行者数を示す．4 都道府県を A，B，C，D と表現し，降車駅の所在地を都道府県に変換して周遊状況を集計した．**図-4**に示すように，A のみを周遊する旅行者が最も多く，次いで B のみ，A,B となっている．A と B の 2 都道府県に関する周遊で全体の約 83%を占めているが，C と D を周遊する旅行者も一定数存在する．

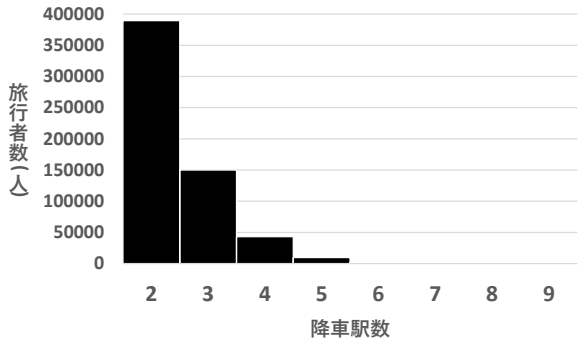

図-3　降車駅数別の旅行者数

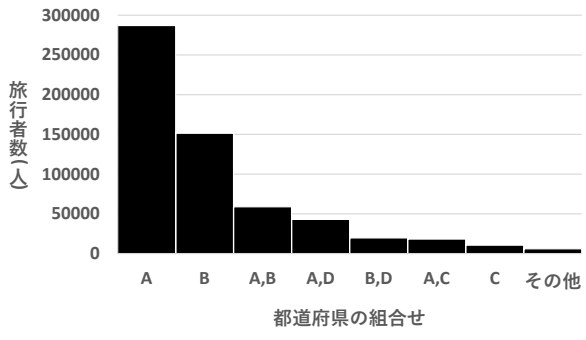

図-4　都道府県周遊状況別の旅行者数

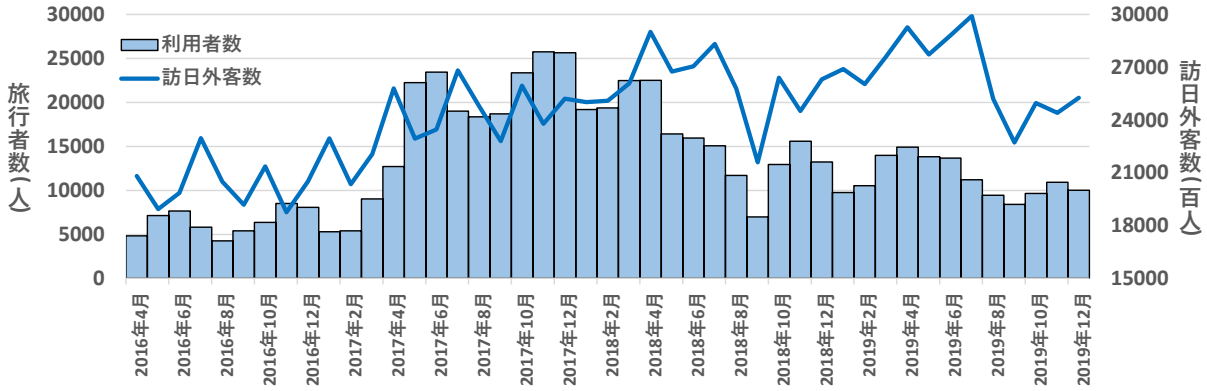

図-5　旅行者数と訪日外客数の推移

図-5 は日本全国の訪日外客数と入出場データから算出した旅行者数を 2016 年 4 月から 2019 年 12 月の期間，1 ヶ月単位で集計した結果である．訪日外客数のグラフは，JNTO（日本政府観光局）が公表しているデータ [14] から筆者が作成した．図-4 の結果から，訪日外客数は月ごとの変動が大きいものの，年単位で見ると増加傾向である．一方，旅行者数は2017年4月に急増した後，2018年前半にかけて横ばいで推移し，2018年5月以降はピーク時の 60%程度の水準で推移している．対象地域を訪れる訪日外国人は 2016 年から 2019 年にかけて増加しているため，分析データ特有の変動が存在していることが分かる．同一年度内で比較すると，旅行者数は春と秋に増加して，夏と冬に減少する傾向が確認できるが，訪日外客数は春と初夏に増加する傾向がある．なお，両データとも 2018 年 9 月と 2019 年 9 月は台風上陸の影響で大きく減少していると考えられる．

以上のように，本分析で用いる入出場データは訪日外国人のごく一部を対象としたものであることに留意する必要がある．しかしながら，最も多い月で25,000 人，一日当たり 800 人程度のデータが取れているため，対面調査等では得ることの難しい量のデータを継続的に取得できる貴重なサンプルだと考えられる．

5.　トピックモデルの分析結果

(1)　モデルの説明力の評価

トピックモデルでは分析者がトピックの数を設定する必要があるため，作成した複数のモデルで精度比較を行って最適なトピック数を決定する．精度検証の際には Perplexity を用いて予測精度を図ることが多い [12),13)] が，本研究では現況把握を目的としていること，各旅行者の降車駅数が少なく学習データと検証データの分割が困難であることから，尤度比を用いる．LDA と BTM の対数尤度は，パラメータが得られた上でのデータの生成確率として，それぞれ式(12),(13)で算出する．

$$
\begin{aligned}
L_{LDA} &= \log p(W|\theta_d, \phi_k) \\
&= \log \prod_{d=1}^{D} \prod_{n=1}^{N_d} \sum_{k=1}^{K} p(w_{d,n}|\phi_k) p(k|\theta_d) \\
&= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log \sum_{k=1}^{K} \theta_{d,k}\, \phi_{k,w_{d,n}}
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
L_{BTM} &= \log p(B|\theta, \phi_k) \\
&= \log \prod_{i=1}^{B} \sum_{k=1}^{K} p(b_i|\phi_k) p(k|\theta) \\
&= \sum_{i=1}^{B} \log \sum_{k=1}^{K} \theta_k\, \phi_{k,w_{i,1}} \phi_{k,w_{i,2}}
\end{aligned}
\tag{13}
$$

初期対数尤度は分類を行わない状態でのデータの生成確率として，式(14),(15)で算出する．

$$
L_{LDA_0} = \log \left(\frac{1}{V}\right)^{n} = -n \log V \tag{14}
$$

$$
L_{BTM_0} = \log \left(\frac{1}{V}\right)^{2N_B} = -2N_B \log V \tag{15}
$$

ここに，

n：W の総数

である．式(12),(14)の計算結果を式(16)に代入して LDA の尤度比を算出し，式(13),(15)の計算結果を式(17)に代入して BTM の尤度比を算出する．

$$
\rho_{LDA} = \frac{L_{LDA_0} - L_{LDA}}{L_{LDA_0}} \tag{16}
$$

$$
\rho_{BTM} = \frac{L_{BTM_0} - L_{BTM}}{L_{BTM_0}} \tag{17}
$$

周遊パターンの数を 10 から 100 まで 10 ずつ変化させて LDA と BTM を構築し，尤度比を算出する．**図-6**，**図-7** に，LDA，BTM における周遊パターンと尤度比の関係を示す．**図-6** の結果から，LDA では周遊パターン数が 50 の時に尤度比が最大となり，その後は減少に転じた．初めは周遊パターンの数を増やすほどモデルの説明力が上がるが，データを説明するのに十分な数の周遊パターンが生成された後は，有力な周遊パターンのパラメータの値が小さくなり尤度比が低下すると考えられる．**図-7 の**結果から，BTM では周遊パターン数が 30 まで尤度比が増加し，その後は頭打ちとなった．BTM では周遊パターンの数を一定以上に増やすと旅行者全体で極めて出現頻度が少ない共起関係を集約した周遊パターンが生成されるが，有力な周遊パターンのパラメータの値はほとんど低下しないため，尤度比は頭打ちになると考えられる．

(2) モデルの解釈性の評価

得られた結果を施策に活用するためには，モデルの説明力が高いだけでなく，周遊パターンが解釈しやすいことが重要である．解釈しやすい周遊パターンとは，単語分布が上位の駅に共通点が見られる周遊パターンである．加えて，抽出された周遊パターン同士の関係も重要である．類似度が高い周遊パターンの組合せが多いとそれらの差を解釈することは難しい場合が多く，解釈性が低下する．一方で，地理的に近いエリア内の周遊では共通する訪問地が存在し，周遊パターン同士に関連性があることも考えられる．そのため，類似度が低く独立性の高い周遊パターンが多くても実態を反映していない可能性がある．このことを確認するために，周遊パターンの類似度を単語分布間のコサイン類似度で評価する．周遊パターン数が 30 の LDA，BTM について，周遊パターンk_1と周遊パターンk_2のコサイン類似度$\cos(k_1, k_2)$を式(18)により算出する．

$$\cos(k_1, k_2) = \frac{\phi_{k_1,v} \cdot \phi_{k_2,v}}{|\phi_{k_1,v}||\phi_{k_2,v}|} \tag{18}$$

30 個の周遊パターンで 435 通りの計算を行い，集計した結果を**図-8** に示す．**図-8** に示す結果によると，コサイン類似度が 0 以上 0.1 未満，0.4 以上の階級では LDA の頻度が高く，0.1 以上 0.4 未満の階級では BTM の頻度が高い．LDA は周遊パターン同士の類似度が低い組合せと高い組合せが多く，各旅行者に過度に適合した周遊パターンが抽出されている可能性がある．**図-9** は周遊パターン数が 10，20，30 の各モデルでコサイン類似度が 0.7 以上の周遊パターンの組合せ数を集計した結果である．いずれの周遊パターン数でも BTM は LDA と比較して類似度が高い組合せ数が少なくなっている．

以上の結果から，今回の分析では BTM の適用によって，解釈性の高いモデル構築が行われたと考えられる．

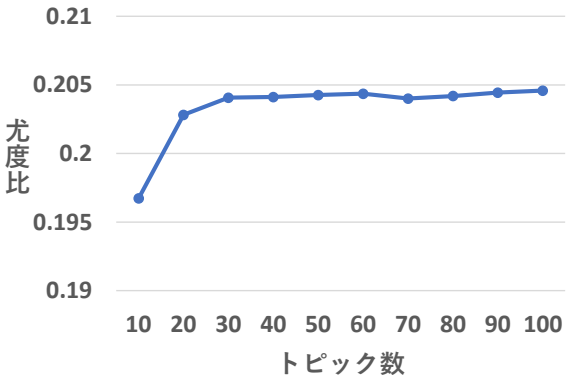

図-6　LDA の周遊パターン数と尤度比の関係

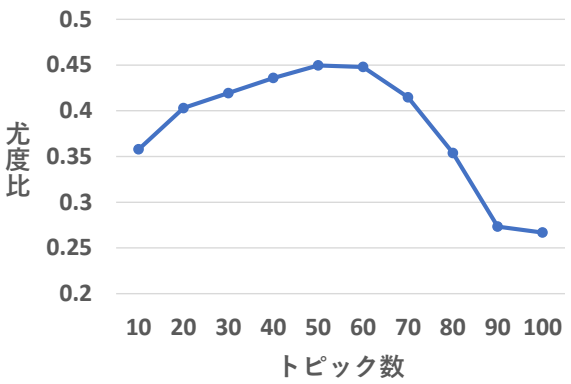

図-7　BTM の周遊パターン数と尤度比の関係

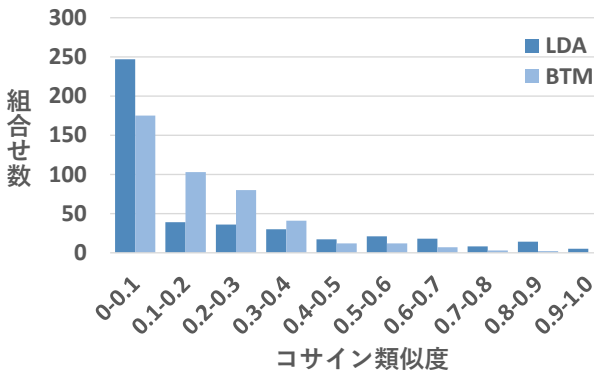

図-8　周遊パターン数 30 の時のコサイン類似度の分布

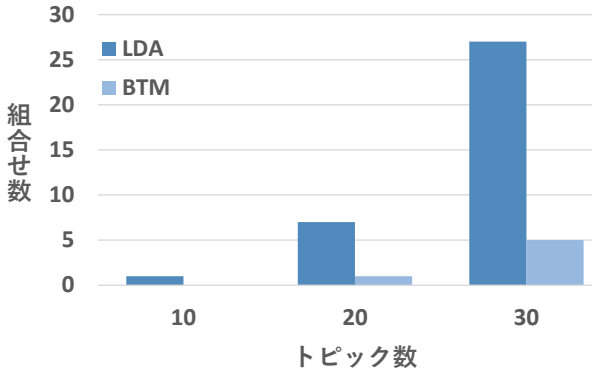

図-9　類似度が高い周遊パターンの組合せ数の比較

(3) 周遊パターンの解釈

(1),(2)の検討から，モデルの説明力とモデルの解釈性がともに高いと判断した周遊パターン数が 30 の BTM の結果を考察する．**図-10** はトピック分布とトピック分布の累積和をプロットした結果である．上位 3 パターンの値は，0.159，0.113，0.096 であり，周遊パターン 13 までの累積和で 0.8，周遊パターン 18 までの累積和で 0.9 を超える．トピック分布の値が大きい周遊パターンは既に人気のある周遊パターン，すなわち鉄道事業者が施策を判断する上で現状の強みであり，トピック分布の値が小さい周遊パターンは人気の低い周遊パターン，すなわち鉄道事業者が施策を判断する上で現状の弱みと言える．そのため，周遊パターンの解釈や施策活用の際には，目的に応じて着目する周遊パターンを決めることが重要である．

表-1 は各周遊パターンについて，単語分布上位 5 駅に着目して，特徴を集計した結果である．周遊パターン 1 から周遊パターン 30 について，都道府県の周遊状況，事業者数，駅から推定される観光地区分，周遊パターンの特徴について整理する．**図-11** に，**表-1** の都道府県周遊状況別の周遊パターン数を示す．**図-11** は**図-4** と概ね同じ形状をしており，都道府県単位の周遊状況が，駅単位の周遊パターンに集約されていることが確認できる．**表-1** の事業者数は上位 5 駅を管轄する鉄道事業者の数を集計した．乗換のしやすさなどから，同じ事業者の駅が同じ周遊パターンに集約されることが多いが，複数の事業者の駅が集約された周遊パターンが 20 個存在する．**表-1** の観光区分は，駅の周辺に存在する観光地を区分ごとに集計した．区分は，繁華街，寺社，景勝地，テーマパーク，城，水族館，公園，空港，買い物，温泉，動物園，球場，博物館の 13 種類である．一日単位の周遊を対象としているため，一つか二つ程度の観光区分が集計された周遊パターンが多くなっている．**表-1** の周遊パターンの特徴は，上位 5 駅の特徴を総合的にまとめたものである．地理的に近いエリアを周遊する周遊パターンを「周遊」，複数の都道府県を周遊する周遊パターンを「広域周遊」，地理的に近いエリアに複数の事業者の駅が集約されている周遊パターンを「一部競合」もしくは「競合」と整理した．トピックモデルでは，同じ周遊パターンに集約された駅

表-1　抽出された周遊パターンの特徴一覧

周遊パターン	周遊状況	事業者数	観光区分	特徴
1	A	1	繁華街	周遊
2	B	2	寺社	周遊、一部競合
3	A	1	繁華街	周遊
4	A,B	1	寺社、景勝地	広域周遊
5	A	2	テーマパーク、城	周遊
6	A	1	繁華街	周遊
7	B,D	2	寺社	広域周遊、一部競合
8	B	2	寺社	周遊
9	A,D	3	寺社、繁華街	広域周遊、一部競合
10	B	1	寺社	周遊
11	A,B	2	寺社	広域周遊
12	A,D	2	繁華街、寺社	広域周遊
13	A	3	テーマパーク、繁華街	周遊
14	A	1	水族館、繁華街、城	周遊
15	A	2	繁華街	周遊
16	A	1	繁華街	周遊
17	A,C	1	繁華街、城	広域周遊
18	A	1	繁華街	周遊
19	A	2	繁華街、公園	周遊
20	A	2	繁華街、買い物	周遊、一部競合
21	A	1	繁華街	周遊
22	B	2	景勝地	周遊
23	A	3	空港駅、買い物	競合
24	B,D	3	寺社	広域周遊、一部競合
25	B	3	寺社	周遊
26	A,C	3	繁華街、温泉	広域周遊
27	C	3	動物園、繁華街	周遊、一部競合
28	A,C	2	繁華街、球場	広域周遊、一部競合
29	A,B	4	景勝地、繁華街、博物館	広域周遊
30	A	3	城、繁華街	周遊

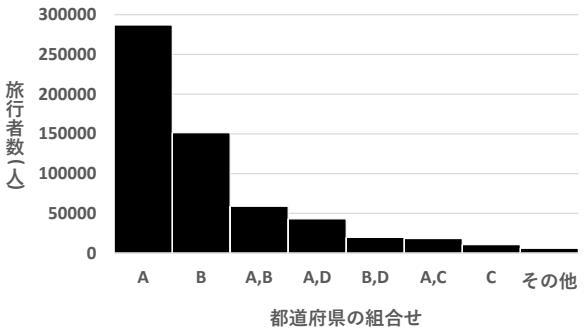

図-11　都道府県周遊状況別の周遊パターン数

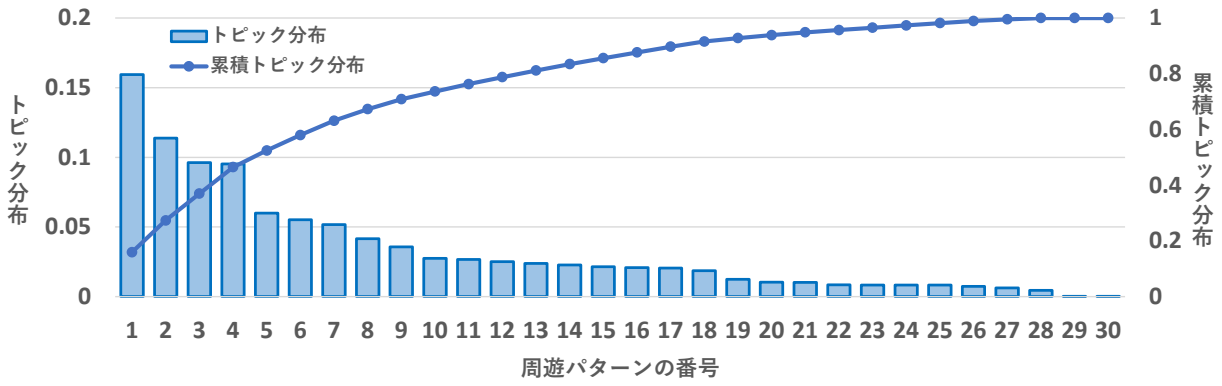

図-10　トピック分布と累積トピック分布

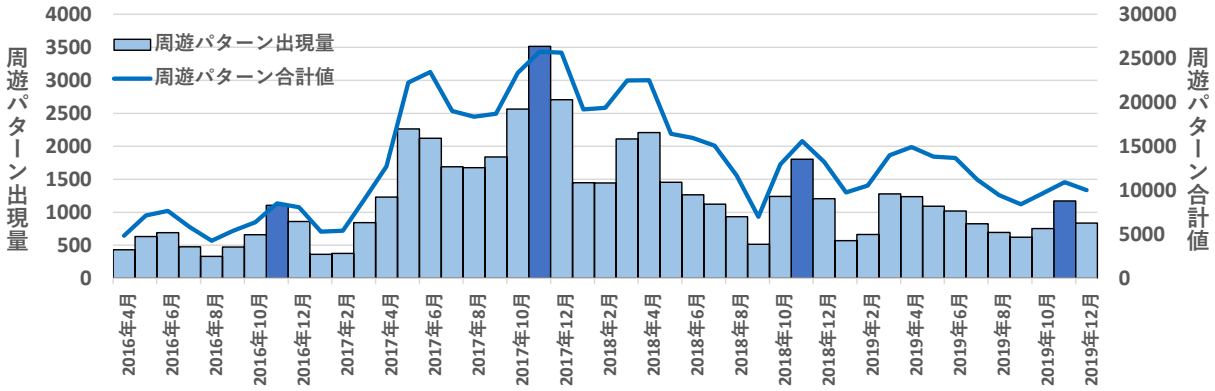

図-12　景勝地を周遊する周遊パターン出現量の推移

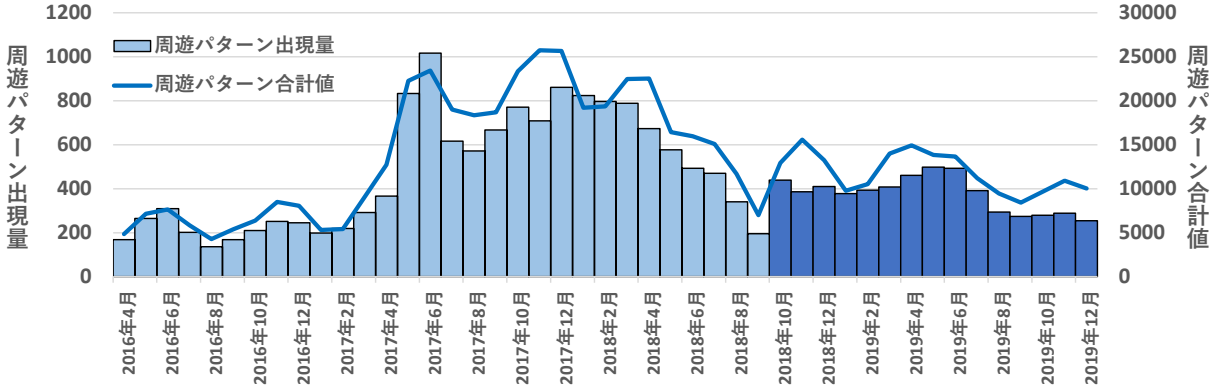

図-13　テーマパークを周遊する周遊パターン出現量の推移

が必ずしも併せて訪問されるとは限らず，潜在的な共起関係があると判断された駅が集約される場合がある．鉄道ネットワークが密で旅客の利用パターンが複雑になるエリアでは，異なる事業者の近接した駅が一つの周遊パターンに集約されると考えられる．

周遊パターンの特徴を明確にするために，式(11)から算出した各旅行者のトピック分布$\theta_{d,k}$を時期ごとに集計する．式(19)から，月mにおける周遊パターンkの出現量$T_{m,k}$を算出し，式(20)から，月mの周遊パターン合計値T_mを算出する．

$$T_{m,k} = \sum_{m \in d} \theta_{d,k} \tag{19}$$

$$T_m = \sum_{m \in d} \sum_{k=1}^{K} \theta_{d,k} \tag{20}$$

ここに，

$$m = (2016\ 年\ 4\ 月, \ldots, 2019\ 年\ 12\ 月)$$

$$k = (1, \ldots, 30)$$

である．観光区分ごとに特徴が良く表れた周遊パターンを抜粋してグラフ化した結果が**図-12**，**図-13**である．

図-12 は$T_{m,4}$を集計したグラフである．周遊パターン 4 は景勝地を周遊する周遊パターンであり，春と秋に卓越する傾向がある．特に青色で示した 11 月の周遊パターン出現量は前後の閑散期と比較して 2 倍程度になっており，主に秋の紅葉を目当てにした周遊であると考えられる．

図-13 は$T_{m,13}$を集計したグラフである．周遊パターン 13 はテーマパークを周遊する周遊パターンであり，全体的に月ごとの変動が少ない．特に青色で示した 2018 年 10 月以降の周遊パターン出現量は 400 前後で安定的に推移している．テーマパーク以外では繁華街を周遊する周遊パターンも月ごとの変動が少ない傾向があり，このような閑散期に需要が落ちにくい周遊パターンを把握することができる．

(4)　施策への活用可能性

周遊パターンの解釈により得られた知見を具体的な施策に活用する方法を提案する．

まず，周遊，広域周遊と判定された周遊パターンについては，同じ周遊パターンに分類された駅を周遊できる切符を発売する，同じ周遊パターンに分類された駅同士でおすすめの行先を案内するポスターを掲示する，といった施策が考えられる．周遊パターン出現量を時期ごとに集計した結果と併せて考えれば，上記の施策を重点的に実施する時期を定量的な根拠を基準に

判断することが可能となる．

一方で，競合と判定された周遊パターンについては，現状分析に活用できる．周遊パターンの単語分布を見ることで，競合路線，競合駅の利用状況から大まかなシェアを推測できる可能性がある．この結果は，各社がサービス改善に取り組むための基礎資料になるほか，事業者間連携を進めることで，旅客の利便性向上に資する施策を実行することが可能になると考えられる．

6. おわりに

本研究では，訪日商品の利用履歴から得られる入出場データを用いて，訪日外国人旅行者の代表的な周遊パターンの抽出を行った．大量のデータから事前の知識を用いることなく，効率的に特徴抽出を行うために，トピックモデルを活用した．モデル適用の前に，入出場データにトピックモデルを適用する際の前処理やデータセットの作成手順を述べた．モデル適用時には，LDA と BTM の 2 つの手法を活用し，両モデルの分析結果を比較した．分析の結果，降車駅数の少ない一日単位の鉄道利用行動の分類では，LDA よりも BTM の方が解釈性の高いトピックが抽出できる可能性があることが示された．抽出されたトピックの単語分布上位の駅に着目すると，地理的に近い駅が同一のトピックに集約されやすいが，異なる事業者の駅が集約されるトピックが少なくないことが明らかになった．さらに，各旅行者のトピック分布を月ごとに集計することで，季節の違いに現れる周遊パターンの特徴が定量的に把握可能となった．

以上の分析結果を踏まえ，施策への活用可能性についての検討も行った．具体的には，同一の周遊パターンに含まれる駅同士の訪問を促進する商品の販売や宣伝，事業者間の連携を進める根拠の提示などに活用可能だと考えられる．

最後に，今回の分析から明らかになった課題をまとめる．まず，複数日に渡る鉄道利用行動の分類が挙げられる．今回の分析では，行動範囲が限られる一日単位の鉄道利用行動を対象としたが，全旅程を対象とすることで，より詳細な周遊パターンが把握できる．旅程単位の周遊パターンを抽出できれば，周遊パターンと宿泊地，滞在日数との関係も分析可能になる．しかしながら，大都市圏の降車駅を複数日に渡って集計すると，集計単位が細かすぎて周遊パターンの分類や解釈が非常に困難となる．そのため，こうした分析を行うためには，入出場データの集計方法やデータセットの作成方法を工夫する必要がある．また，モデルの改良についても検討が必要である．今回使用したデータには個人属性が紐づいていなかったが，国籍や訪日回数などは旅客のセグメントにおいて重要な情報である．他のデータを用いて，入出場データにこれらの情報を付与することができれば，周遊パターンとの関係が分析可能となる．さらに，トピックモデルには文書に紐づいた補助情報をモデルに組み込む手法や，文書の時系列を考慮したモデルが提案されている．これらを適用することで，より詳細な周遊パターンを抽出できる可能性がある．

以上の方針で発展させることが今後の課題である．

謝辞：本稿で分析対象とした入出場データを提供いただいた鉄道事業者に，深く御礼申し上げる．

参考文献

1) 国土交通省鉄道局：鉄道分野におけるインバウンド受入環境整備について，https://www.mlit.go.jp/common/001240898.pdf（2023.2.24 閲覧）
2) 国土交通省：FF-Data（訪日外国人流動データ）分析例（運輸局ブロック別 ブロック内移動の交通機関分担率），https://www.mlit.go.jp/sogoseisaku/soukou/sogoseisaku_soukou_fr_000022.html（2023.2.24 閲覧）
3) Blei, D. M., Ng, A. Y. and Jordan, M. I.：Latent Dirichlet Allocation，Journal of Machine Learning Research, Vol. 3, pp.993-1022, 2003.
4) Cheng. X, Yan. X, Lan. Y and Guo. J.：BTM:：Topic modeling over short texts, IEEE Transactions on Knowledge and Data Engineering, Vol.26, No.12, pp.2928–2941, 2014.
5) 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，Vol.68，No.5，pp.667-677，2012.
6) 矢部直人，倉田陽平：東京大都市圏における IC 乗車券を用いた訪日外国人の観光行動分析，GIS-理論と応用，Vol.21，pp.35-46，2013.
7) 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，Vol.72，No.5，pp.571-583，2016.
8) 三輪哲：潜在クラスモデル入門，https://www.jstage.jst.go.jp/article/ojjams/24/2/24_2_345/_pdf（2023.3.2 閲覧）
9) 辰巳嘉大，塚井誠人：トピックモデルを用いた訪日外国人周遊分析，運輸政策研究，Vol.23，pp.20-34，2021.
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，pp.507-517，2019.
11) 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分析手法の開発，運輸政策研究，Vol.20，pp.20-29，2018.
12) 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.
13) 岩田真治：トピックモデル，講談社，2015.
14) JNTO：月別・年別統計データ（訪日外国人・出国外国人），https://www.jnto.go.jp/jpn/statistics/visitor_t

rends/（2023.2.24 閲覧）

STUDY ON IDENTIFICATION OF TOUR OF FOREIGN VISITORS TO JAPAN WITH DATA ON RAILWAY USE

Wataru INABA, Shingo NAKAGAWA,
Takuya WATANABE and Noriko FUKASAWA

The purpose of this paper is to identify the tour of foreign visitors to Japan with data on railway use. Using data collected by automatic ticket gates, we analyzed the daily railway use of foreign visitors to Japan who tour around major cities. We assume the combination of destination stations as a tour, and we try to identify similar tour from the large data set by the topic model. Since the number of destination stations is small, we use BTM, an effective topic model for short texts. Based on the likelihood ratios, it is appropriate to set the 30 topics. Besides, we found different operators included in the same topic and the seasonal feature of topic. Using these results, we proposed the specific measures, such as the sale of excursion tickets and destination guides.