

# 観光に関するユーザーレビューデータを用いた観光客の話題分析と地域観光振興への活用の検討

野守 耕爾<sup>1\*</sup>, 神津 友武<sup>2</sup>

<sup>1</sup> 株式会社アナリティクスデザインラボ

<sup>2</sup> 有限責任監査法人トーマツ, デロイトアナリティクス

\* Corresponding Author: koji.nomori@analyticsdlab.co.jp

## Abstract

This paper describes an approach of tourism development from tourists' perspective by analyzing tourist review data. The approach enables us to understand characteristics of regions and tourists based on sightseeing themes extracted by applying text mining and PLSA to tourist review data. This study extracted 15 themes from about 170 thousand review data on tourist spots across Japan; scenery, historical buildings, sea views, shopping areas, flowers, shrines and so on. A regional tourism marketer can quantitatively understand characteristic themes of the region and target tourists who show interest in these themes. Some regions with similar themes can get a round trip effect by cooperating to establish a tour route experiencing the theme. Some regions with different themes can get new targets by cooperating to mutually send tourists showing interests in respective themes. The approach will be useful for tourism planning and regional developments to attract and pull in more tourists.

## Keywords

Tourist review data, Topic analysis, Text mining, PLSA

## 1 はじめに

地方創生が国の重要政策課題の一つに位置づけられている昨今、日本経済新聞社産業地域研究所が自治体向けに実施したアンケート調査によれば、政府に期待することとして約半数が「観光振興への支援策」と回答し、今後地域の戦略産業として育成したい分野には「観光レジャーなどの集客産業」が16.9%で最も多く回答されており（田辺 2014）、観光による地域の経済・産業の活性化が求められている。近年では訪日外国人観光客の伸び率が拡大しており、2017年では2,800万人を突破し（日本政府観光局 2017）、2020年までに4,000万人が目標に掲げられている。これに伴い観光活性化の検討もインバウンド戦略ばかりに注目が集まっているが、国内の旅行消費額で見ると、2017年では全体（日本人と訪日外国人の旅行消費額の合計）が約26.7兆円であるのに対し、訪日外国人の消費額は約4.4兆円であり、過去最高額ではあるものの全体の約16.5%であり、ほとんどが日本人の消費となっている（観光庁 2017）。インバウンド戦略も重要であるが、持続的な地域の観光活性化には日本人観光客の誘客促進を検討することが不可欠である。そこでは地域がその土地の観光資源の特徴と観光客のニーズを理解し、効果的な観光振興施策を検討することが求められるが、これまでの観光振興のアプローチは、自治体や旅行会社といったサービス提供者側の事情や都合を反映したものが多く、観光客の目線に立った取組みが疎かになっているということが指摘されている（才原 2015）。

この“観光客目線”でニーズを把握するには、実際に観光客が発するデータを分析し、そこから観光客が感じている価値や関心事を抽出することが有効であると考えられるが、その手段の一つとして、Web上に投稿されたユーザーレビューデータを活用することが挙げられる。ユーザーレビューは、従来のアンケート調査のように要求された設問に回答するのではなく、レビュー投稿者自らが発した生の声であり、また大量にデータを取得できる利点がある。またユーザーレビューは消費者の意思決定によく利用されていることも報告されており（NTTレゾナント 2012）、同じSNSの一つであるTwitterに関しては、実際にツイート数が販売実績に影響を与えている効果も確認されており（鶴見 2015）、ユーザーレビューから観光客のニーズを抽出して活用することは有効であると思われる。一方、個々のユーザーレビューには個人や製品の性質によって消費者の意見形成にバイアスが生じるため、実務上でユーザーレビューを活用する際には注意が必要であるが、個々のデータに着目するのではなく、例えばセグメント別に分析することによって、簡易的ではあるがそのバイアスのある程度解消し誤った解釈を避けることができる（大西 2015）。観光に関連するユーザーレビューデータの活用においても、個別の観光客、個別の観光資源を分析するのではなく、観光客の属性別や地域別といったセグメント単位で分析することで、またそれを大量のデータから行うことで有用な知見が得られることが期待できる。

本研究では、日本の観光資源を対象としたユーザーレビューのデータを分析することで、地域観光振興への活用の可能性を検討する。具体的には、観光客がレビューの中で話題にしている観光テーマを抽出し、そのテーマを軸とした地域の特徴と観光客の特徴をそれぞれ定量的に把握することで、各地域における誘客ターゲットの設定や効果的な連携地域の選定といった施策立案への活用を検討する。

## 2 分析方法

### 2.1 使用するデータ

本研究では、フォートラベル株式会社が運営する「旅行のクチコミと比較サイト フォートラベル」(<http://4travel.jp/>)における国内旅行の全国の観光資源に対するユーザーレビュー（フォートラベルではクチコミと呼ばれる）データを用いて分析する。フォートラベルのユーザーレビューで得られる情報は以下のように大別できる。

#### ①観光資源情報

施設の名称、施設カテゴリ、所在地域（都道府県単位、市郡単位、区町村単位）、施設住所

#### ②投稿者情報

性別、年代

#### ③レビュー情報

タイトル、コメント、評点、同行者、旅行年月

なお、観光資源情報の施設カテゴリとは、観光、グルメ・レストラン、ショッピング、交通、ホテルという5項目が設定されているが、本研究では「観光」カテゴリのレビューを対象とし、2004年7月27日から2014年8月15日までに投稿された日本全国の各観光資源の全ユーザーレビューのうち、レビュー件数が10件以上の観光資源（5,553箇所）におけるユーザーレビュー（169,407件）のデータを分析対象とする。

本レビューデータを分析対象とした理由としては、企業が運営するレビューサイトであるため、投稿者は会員登録した者に限られること、利用規約があること、不適切な投稿は削除されることなど、データがきちんと管理されており、他のSNSの匿名投稿と比較してデータが信頼できるということにある。企業が運営するレビューサイトの中でも特にフォートラベルを選択した理由としては、各投稿者のコメントの量が他のサイトよりも多いことにある。ほとんどの投稿者は写真まで投稿しており、単純に良かった、悪かったということではなく、その観光地でどのような体験をしてどのような印象を持ったのかということが事細かに記述されている。レビューのコメントから観光客目線による観光テーマを抽出する際、そのコメントの内容がより詳細であればあるほど、より具体的なテーマが抽出可能となるため、フォートラベルのユーザーレビューを使用することとした。

### 2.2 ユーザーレビュー分析の関連研究

SNSの普及に伴い、ユーザーレビューのコメントを対象にした分析も盛んに行われており、ユーザーの評

価やニーズを理解して製品やサービスの向上につなげるなどの応用が検討されている（奥村 2012）。例えばECサイトの化粧品のユーザーレビューを対象とした分析事例として、レビューのコメントを分析して消費者の知覚品質の手がかりとなるキーワードを抽出することでブランドマネジメントへの応用を検討したり（ジャムサランジャワバーサンフー 2008）、ヒットの背景にある要因をキーワードで探り、新たな視点のマーケティングを検討するものがある（上原 2011）。また携帯電話のレビューコメントを分析し、その評価語句の時系列変化や類似性を可視化して製品の特徴把握に活用するものや（打田ら 2010）、フィーチャーフォンとスマートフォンの出現ワードの特徴を比較して消費者の関心の変化を把握するものがある（鷲山 2011）。他にも野菜のインターネット通販のレビューコメントから品目ごとの評価の特徴を分析するものや（竹崎 2013）、レビューコメントの分析結果を応用した推薦システムを提案する研究として、学術本の難易度をそのレビューのキーワードから推定し、難易度に応じた学術本の推薦を検討する研究や（中山ら 2012）、映画のレビューの肯定的または否定的な評価表現と評価対象に基づいて、自身の投稿レビューから自身の嗜好に適した他者レビューを推薦する研究がある（林ら 2015）。宿泊や旅行に関しては、宿泊施設のレビューから評価の要因となるキーワードを抽出して宿泊施設の戦略導出を検討するものや（田邊ら 2008）、訪日外国人による観光地のレビューから出身国や観光地ごとの特徴的なイメージを分析するものがある（大久保 2014）。

ユーザーレビューのコメントを対象にした上記のような従来の分析において、そのアプローチの多くはそのコメント部分にテキストマイニングを実行し、単語あるいは係り受けの表現を抽出し、その頻度を集計したり、共起関係のネットワークを構築して全体像の傾向を把握したり、ある属性に特徴的に出現する単語をコレスポネンス分析（数量化Ⅲ類）やTF-IDFなどの指標を用いた特徴分析などが行なわれている。こうした分析の結果は、抽出した単語をベースに結果を解釈していくが、抽出した単語は数が多くアウトプットが複雑になりがちであり、また単語だけではそれが出現する背後の状況や文脈までは分からず、結果の特徴を解釈しづらいことが課題として挙げられる。また抽出された多くの単語から成るネットワークやマップを構築し、その位置関係から全体における話題や特徴の解釈と考察が行われているが、どうしても定性的な解釈になってしまっている。

コレスポネンス分析（数量化Ⅲ類）はテキストマイニングでもよく適用される手法であり、抽出した単語の出現に関する多次元データからそれを集約する軸を抽出し、寄与度の高い二つの軸による二次元平面上に単語や属性情報をマッピングしてその関係を解釈するが、その軸はより少ない次元で元のデータを近似するように抽出されたものである。特に第一成分軸は頻度の大きい単語に集中した軸が抽出されたり、当たり前の軸が抽出されることもあり、レビューコメント全体を構成する話題を分解して抽出する手法にはなっ

いない。またデータ件数が多く膨大な単語を分析対象とするときには、各成分軸の寄与率は小さくなりやすく、単純に二次元のマップで解釈することは困難となる。一方、書籍のレビューコメントから抽出した単語にPLSA（確率的潜在意味解析）を適用して次元圧縮し、その低次元データにSOM（自己組織化マップ）を実行することで、各レビューデータを二次元平面にマッピングし、そのマップをベースに各書籍の満足度の分布を可視化して比較するという研究もある（齊藤2014）。これによりレビューコメント全体を考慮した特徴の可視化ができるが、PLSAは抽出された多数の単語に基づく高次元データを低次元化するために利用されており、その圧縮された次元については解釈の対象とされていないため、どのような話題がその評価と関係しているのかという考察には至っていない。

## 2.3 本研究の分析アプローチ

本研究のアプローチの特徴は、各観光資源のレビューコメントにテキストマイニングを実行し、従来のようにそこで抽出された表現をベースに地域や観光客の特徴を分析・考察するのではなく、まず文脈上近い表現を集約することでユーザーレビューで話題にされている「テーマ」を抽出し、そのテーマを軸に地域や観光客の特徴を定量的に分析することにある。

約17万件あるレビューコメントを従来のアプローチのように表現ベースで分析してしまうと、出現する表現の数も膨大となるため結果が非常に複雑になり、それを解釈して考察することは容易ではない。またその複雑な結果を解釈したとしても、人によってその解釈は異なることが予想される。本研究のアプローチでは約17万件のレビューコメントを複数個のテーマに集約し、それを新たな量的変数としてデータに加えることで、地域や観光客の特徴を複数個のテーマを軸にシンプルにかつ定量的に解釈することができる。

本研究のアプローチでは以下四つのステップを実行する。これによって地域観光振興の施策立案を観光客目線で検討する。アプローチの全体像を図1に示す。

### 観光テーマの抽出

以下の手順により、日本全国の各観光資源のユーザーレビューのデータから観光客が話題にしているテーマを抽出する。

(1) レビューコメントの係り受け表現を抽出し、観光資源×係り受け表現の共起行列を作成する

レビューのタイトルとコメント本文を結合させたテキストデータに対し、テキストマイニングの係り受け分析を実行し、係り受け表現を抽出する。本研究では、観光における観光客の評価や体験に関連する表現を抽出するため、名詞と形容詞（形容動詞含む）、及び名詞と動詞（サ変接続名詞含む）という組み合わせの係り受け表現を抽出する。その係り受け表現の出現頻度を観光資源ごとに集計することで、観光資源と係り受け表現の共起行列（クロス集計表）を作成する。

なお本研究では、株式会社NTTデータ数理システムのText Mining Studio 5.0を使用しテキストマイニングを実行する。

(2) PLSAを実行して観光客が話題にする観光テーマを抽出する

作成した共起行列にPLSAを適用し、観光資源 $S_i$ と係り受け表現 $E_j$ の背後にある潜在的な観光テーマ $T_k$ を抽出する。PLSAについては後述する。PLSAの実行により、観光テーマ $T_k$ を条件とした観光資源 $S_i$ の条件付出現確率 $P(S_i|T_k)$ 、観光テーマ $T_k$ を条件とした係り受け表現 $E_j$ の条件付出現確率 $P(E_j|T_k)$ 、観光テーマ $T_k$ の出現確率 $P(T_k)$ が計算される。抽出された観光テーマ $T_k$ は、 $P(S_i|T_k)$ によって重み付けされた観光資源の名称や $P(E_j|T_k)$ によって重み付けされた係り受け表現からその意味を解釈する。

なお本研究では、国立研究開発法人産業技術総合研究所が開発したサービス店舗支援システムAPOSTOOL2（竹中ら2011）のPLSAプログラムを使用する。

### 地域別による話題の特徴分析

抽出した観光テーマを軸として各地域の特徴を定量的に把握する。

PLSAの実行によって観光テーマ $T_k$ ごとに計算された各観光資源 $S_i$ の条件付出現確率 $P(S_i|T_k)$ について、式(1)のようにその観光資源が存在する地域 $R_h$ で確率を集約し、観光テーマ $T_k$ を条件とした地域 $R_h$ の条件付出現確率 $P(R_h|T_k)$ を計算する。この $P(R_h|T_k)$ にベイズの定理を適用することで、式(2)のように各地域 $R_h$ を条件とした観光テーマ $T_k$ の条件付出現確率 $P(T_k|R_h)$ を計算する。これは各地域 $R_h$ における観光テーマ $T_k$ の構成割合を意味しており、各地域が有するテーマ性を定量的に把握することができる。なお式(2)における $P(T_k)$ はPLSAの実行によって得られている。この観光テーマ $T_k$ の構成割合が地域 $R_h$ にとって特徴的であるかどうかは、事後確率 $P(T_k|R_h)$ を事前確率 $P(T_k)$ で除した $P(T_k|R_h)/P(T_k)$ の大きさで確認する。この指標は、日本全国から抽出された観光テーマ $T_k$ の平均的な確率が、その地域 $R_h$ という条件を加えることで何倍に上昇するのかということの意味しており、1が基準となるが値が大きいテーマほどその地域でよく話題にされる特徴的なテーマといえる。なお、ここでの「地域」とは、フォートラベルのサイトで定義されている区町村単位の地域分類を想定している。

$$P(R_h|T_k) = \sum_i P(S_i|T_k) \quad (S_i \in R_h) \quad (1)$$

$$P(T_k|R_h) = \frac{P(R_h|T_k)P(T_k)}{\sum_k P(R_h|T_k)P(T_k)} \quad (2)$$

### 観光客の属性別による話題の特徴分析

観光客の属性別によく話題にされる観光テーマを定量的に把握する。

まず各レビューデータ $D$ において各テーマ $T_k$ がどれくらい話題にされているのかということをも $P(D|T_k)$ によって計算する。この条件付出現確率はそのテーマをよく話題にしているレビューほど値が高くなるが、これをテーマを構成する係り受け表現 $E_j$ に基づいて定義すると、 $P(D|T_k)$ は式(3)で計算される。係り受け表現 $E_j$ が

含まれるレビューのデータ件数を $N(E_j)$ とすると、 $P(D|E_j)$ は $N(E_j)$ の逆数として計算される。なお $P(E_j|T_k)$ はPLSAの実行によって得られている。

続いて、計算した $P(D|T_k)$ について、式(4)のようにそのレビューに付与された観光客の属性Aで確率を集約し、観光テーマ $T_k$ を条件とした観光客属性Aの条件付出現確率 $P(A|T_k)$ を計算する。この $P(A|T_k)$ にベイズの定理を適用することで、式(5)のように各観光客属性Aを条件とした観光テーマ $T_k$ の条件付出現確率 $P(T_k|A)$ を計算する。これは各観光客属性Aでみた観光テーマ $T_k$ の構成割合を意味しており、各観光客層における各テーマの話題の度合いを定量的に把握することができる。その観光テーマ $T_k$ の構成割合が観光客の属性Aにとって特徴的であるかどうかは、事後確率 $P(T_k|A)$ を事前確率 $P(T_k)$ で除した $P(T_k|A)/P(T_k)$ の大きさを確認する。この指標は、日本全国から抽出された観光テーマ $T_k$ の平均的な確率が、その観光客層Aという条件を加えることで何倍に上昇するのかということの意味しており、1が基準となるが値が大きいテーマほどその観光客層がよく話題にしているテーマといえる。

$$P(D|T_k) = \sum_j P(D|E_j)P(E_j|T_k) \quad (3)$$

$$P(A|T_k) = \sum_l P(D_l|T_k) \quad (D_l \in A) \quad (4)$$

$$P(T_k|A) = \frac{P(A|T_k)P(T_k)}{\sum_k P(A|T_k)P(T_k)} \quad (5)$$

### 地域観光振興の施策の検討

観光客がよく話題にする観光テーマを軸にして分析した地域別の特徴と観光客の属性別の特徴に基づいて、

各地域における観光振興の施策を検討する。例えば以下のような検討が考えられる。

(1) 自身の地域でよく話題にされるテーマに対し、それに関心の強い観光客層をターゲットに設定する

自身の地域でよく話題にされる特徴的なテーマと、そのテーマをよく話題にして関心の強い観光客の属性を照合することで、自身の地域に魅力を感じると考えられる観光客層を把握できる。特に周辺の地域においてよく話題にされるテーマとも比較し、自身の地域はどのようなテーマが特徴となっているのか把握し、その地域の独自性のある特徴を有効なターゲット層にPRして集客を図ることが考えられる。

(2) よく話題にされるテーマが類似する他の地域と連携して誘客の相乗効果を狙う

自身の地域及び周辺地域においてよく話題にされるテーマを把握し、自身の地域のテーマ性と類似する地域がある場合、その地域と連携して誘客施策を検討することで、そのテーマをよく話題にして関心の強い観光客層を地域間で回遊させるといった誘客の相乗効果や誘客コストの削減が見込まれる。

(3) 新たなターゲット層の開拓に効果的な連携地域を検討する

自身の地域でよく話題にされるテーマに対し、そのテーマをよく話題にしている観光客層ではないが、シニア層や家族旅行客など、特定の観光客層をターゲットとして設定したい場合、そのような観光客層がよく話題にしているテーマを特徴として有する地域が周辺に存在するかどうかを確認する。該当する地域があればその地域と連携することで新たなターゲット層の観光客を呼び込める可能性がある。

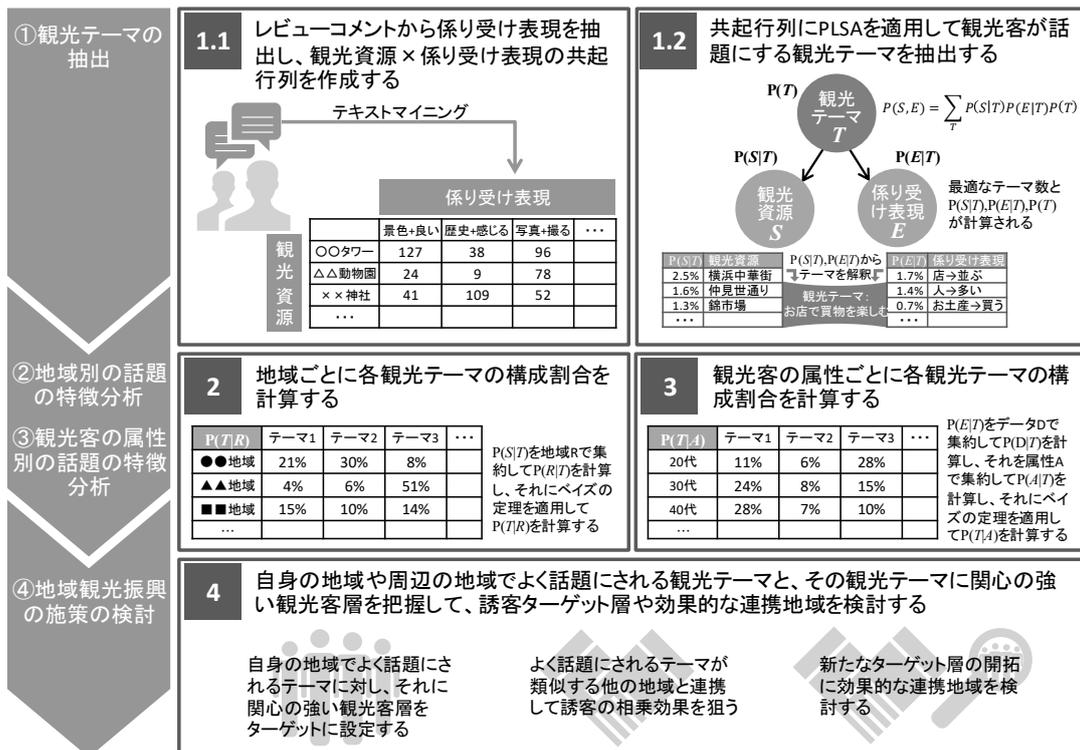


図1 観光レビューデータを地域観光振興に活用する分析アプローチ

## 2.4 PLSA (確率的潜在意味解析)

### PLSAの概要と特長

PLSA (Probabilistic Latent Semantic Analysis)は、文書分類のために開発された次元圧縮手法である (Hofmann 1999)。文書 $D$ とそこに出現する単語 $W$ の間には潜在的な意味クラス $C$ があることを想定し、各文書における単語の出現頻度が記録された「文書」×「単語」の共起行列データを学習し、文書と単語の共通のトピックとなるような特徴を見つける手法である。PLSAでは文書 $D_i$ と単語 $W_j$ の共起確率 $P(D_i, W_j)$ を潜在クラス $C_k$ を用いて式(6)のように分解して考える。ここで、文書 $D_i$ における単語 $W_j$ の出現回数を $N(D_i, W_j)$ とすると、式(7)の対数尤度を最大にする $P(D_i|C_k)$ ,  $P(W_j|C_k)$ ,  $P(C_k)$ をEMアルゴリズムを用いて式(8)のEステップと式(9)~(11)のMステップを計算することで最尤推定する。つまりPLSAの実行によって得られるアウトプットは3種類の確率変数 $P(D_i|C_k)$ ,  $P(W_j|C_k)$ ,  $P(C_k)$ の値である。これにより「文書」×「潜在クラス」という低次元データに変換することができ、クラスタリングの手法としても用いられる。

$$P(D_i, W_j) = \sum_k P(D_i|C_k) P(W_j|C_k) P(C_k) \quad (6)$$

$$L = \sum_i \sum_j N(D_i, W_j) \log P(D_i, W_j) \quad (7)$$

$$P(C_k|D_i, W_j) = \frac{P(D_i|C_k)P(W_j|C_k)P(C_k)}{\sum_k P(D_i|C_k)P(W_j|C_k)P(C_k)} \quad (8)$$

$$P(D_i|C_k) = \frac{\sum_j N(D_i, W_j)P(C_k|D_i, W_j)}{\sum_i \sum_j N(D_i, W_j)P(C_k|D_i, W_j)} \quad (9)$$

$$P(W_j|C_k) = \frac{\sum_i N(D_i, W_j)P(C_k|D_i, W_j)}{\sum_i \sum_j N(D_i, W_j)P(C_k|D_i, W_j)} \quad (10)$$

$$P(C_k) = \frac{\sum_i \sum_j N(D_i, W_j)P(C_k|D_i, W_j)}{\sum_i \sum_j \sum_k N(D_i, W_j)P(C_k|D_i, W_j)} \quad (11)$$

データクラスタリングという観点から、PLSAが他のクラスタリング手法と特に異なる点は主に以下の二つが挙げられる。

#### ①行と列を同時にクラスタリングする

一般的なクラスタリング手法は、列をベースに行をクラスタリングする、あるいは行をベースに列をクラスタリングするため、どちらか一方しかクラスタリングできない。一方PLSAでは、潜在クラスに対する行の要素の所属度合いと列の要素の所属度合いを式(9)(10)によって同時に計算する。これにより、抽出されるクラスは行と列の二つの軸から構成され、意味解釈もしやすくなる。

#### ②ソフトクラスタリングである

k-meansや階層的クラスタ分析などはハードクラスタリングと呼ばれ、各変数は必ず一つのグループに限定して所属する。一方PLSAはソフトクラスタリングと呼ばれ、全ての変数は全てのクラスにまたがって所属し、その所属度合いが条件付出現確率 $P(D_i|C_k)$ ,  $P(W_j|C_k)$ で与えられる。これにより複数の重要な意味を持つ変数がある場合でも柔軟なクラスタリングができる。

### 関連手法

文書分類の次元圧縮手法には他にLSA (Latent Semantic Analysis) (Deerwester 1990) やLDA (Latent Dirichlet Allocation) (Blei 2003) が知られている。LSAは特異値分解による次元圧縮手法であるが、テキスト分析でよく用いられる数量化Ⅲ類・コレスポネンス分析も特異値分解によって軸を抽出する手法であり、数学的には同様の手法といえる。LSAを確率的に処理し発展させたものがPLSAとなり、LSAにおける特異値分解の行列表記をPLSAでは確率 (aspectモデル) で表記しているが、数学的な考え方は同じである。LSAは入力する行列の成分をそのまま使用すると、大きな値をとりやすいベクトルに引っ張られて次元圧縮される傾向があるため、TF-IDFなどで重み付けされた行列を用いられることが多い。PLSAはそうした重み付けの事前処理をする必要なく次元圧縮を実行できる。

LDAはPLSAをさらに拡張させた手法として開発されている。個々の文書における各トピックの現れやすさを表す確率が、PLSAではあくまで学習させた観測データのみから定義されるが、LDAではディリクレ分布という確率分布を仮定して生成させる。PLSAでは、観測データに過剰に適合して他のデータの適合度が下がってしまうオーバーフィッティングが生じやすく、新しい文書におけるトピックの生成確率は定義されないが、LDAではこれを推定できる。情報検索の分野では、新しいデータがどのトピックに分類されるのかということが重要となるため、PLSAよりもLDAが適用されることが主流である。またLDAはPythonやMahoutなどのオープンソースライブラリとして公開されていることもあり、適用事例も多くPLSAよりも注目されているといえる。

本研究でPLSAを用いる理由は、ユーザーレビューのコメントで話題にされていることの現状を把握するためである。確かにPLSAは観測データにオーバーフィットし、新しいデータの対応が難しいが、観測データのみからその現状を示す潜在トピックを抽出できる。LDAではディリクレ分布を仮定していることで、オーバーフィッティングは回避しているが、その分抽象度が高い結果となりやすく、また純粋な観測情報から得られた結果とはいえない。本研究の取り組みは各観光資源のユーザーレビューの分析結果を用いて、現状における地域の特徴と観光客層の特徴を理解し、それらを総合して地域における観光振興の施策を検討することへの活用を想定したものである。したがって観光客の声の現状を率直に理解することが重要であると考え、本研究ではPLSAを採用する。

### PLSAのクラス数の決定

PLSAではクラス数をあらかじめ設定する必要があるが、AIC (Akaike 1973) などの情報量基準により数学的に最適なクラス数を決定することもできる。例えば、クラス数の異なる分析結果それぞれについてAICを計算し、AIC最小となるクラス数の結果を採用すればよい。またPLSAは初期値依存性があり、初期値によって結果が異なる。そこで、クラス数に範囲を持たせて複数設定し、初期値を変えてそれぞれのクラス数

でPLSAを複数回実行し、その全結果の中でAIC最小となる結果を採用するといった対応がとられている(石垣 2011)。

### 本研究におけるPLSAの適用方法

本来のPLSAの考え方と本研究におけるPLSAの適用の仕方を比較したものの図2に示す。本来のPLSAの実行では、「文書」×「単語」という共起行列をインプットとするが、行に設定された「文書」はつまりは文書IDであり、それ自体に意味を持たない情報であるため、抽出された潜在クラスの意味解釈には使用されない情報である。行と列を同時にクラスタリングできるPLSAでは、行と列は双方が十分な意味を持つ情報で構成すれば、抽出された潜在クラスの意味を二つの情報軸から解釈することができる。また「文書」×「単語」の共起行列は、基本的に0と1で構成されるデータであり、ほとんどが0のスパース(疎)なデータであるため、文書間・単語間で差が出にくく特徴的でクリアな潜在クラスが得られにくい。

本研究では、PLSAの本質は共起行列を構成する行と列の二つの軸の背後にある潜在的な要因を抽出する知識発見手法と捉え、各観光資源のレビューコメントで出現する係り受け表現の件数を記録した「観光資源」×「係り受け表現」の共起行列(クロス集計表)にPLSAを適用する。「観光資源」と「係り受け表現」はどちらも単独で意味を持つ情報となる。これにPLSAを適用することで、観光客が各観光資源を訪問するときに話題にしている潜在的な観光テーマを抽出できると考えられる。また、この共起行列は係り受け表現の出現有無に関する0か1のデータではなく、具体的な出現件数が値として入っているクロス集計型の行列であるため、スパース性の問題の影響を受けにくく、より解釈のしやすいクリアなテーマが抽出されることが期待できる。さらにその共起行列のサイズは本来のPLSAで用いる共起行列に比べて(特に行数において)とても小さくなっており、計算時間も大幅に削減できる。

### 3 観光テーマの抽出

ユーザーレビューのコメントにテキストマイニングの係り受け分析を実行し、「観光資源」×「係り受け表現」の共起行列(クロス集計表)を作成してPLSAを適用することで、観光資源と係り受け表現の背後にある観光テーマを抽出した。

#### 3.1 係り受け表現の抽出

全169,407件のユーザーレビューのタイトルとコメント本文を結合させたテキストデータに対して、テキストマイニングの係り受け分析を実行し係り受け表現を抽出した。特に出現頻度が30件以上の係り受け表現に限定したところ、6,834件の表現が抽出された。なお出現頻度の1件とは、その表現が出現したレビューの件数であり、1件のレビューの中である表現が複数回出現しても重複はカウントせず、その表現の出現回数は1件とカウントした。

#### 3.2 PLSA適用による観光テーマの抽出

全5,553箇所の観光資源のユーザーレビューで出現する6,834種の係り受け表現の件数を記録した「観光資源」×「係り受け表現」の共起行列(クロス集計表)を作成し、PLSAの学習データとした。PLSAの実行では、クラス数を5から20まで1刻みで変化させ、それぞれに対してPLSAを5回ずつ初期値を変えて実行し、AICを計算した。その結果、クラス数に対して下に凸のカーブを描き、クラス数15の実行結果の一つがAIC最小となったのでこの結果を採用することとした。なお、PLSAでは先述の通り初期値依存性があり、初期値によって局所最適化が起り、結果が変動する可能性がある。そのため同じクラス数15の解でも異なる初期値ではまたクラスの内容も変わる可能性もあり、その意味ではPLSAの解はロバストではないともいえる。本研究ではAICという数学的な指標を目安に使用し、複数出力された解の中から一つを採用したが、これは絶対的な正解ではなく、あくまで解の一つ、つまりユーザーレビューで話題にされる観光テーマの軸の一つのパターンとして捉えるものであり、クラス数を変えればまたそれぞれ異なる軸が抽出される。本研究ではユーザーレビューを分析することで、地域における観光振興の施策検討へ活用することを想定しているが、その施策検討を支援する軸を提供することが重要であり、その軸は絶対的な正解である必要はないと考える。つまり、クラス数は15でなくても10でも20でも良いが、一つのパターンの結果に定める目安としてAICという基準を使用しクラス数15の解を採用した。

本研究におけるPLSAでは、観光資源Sと係り受け表現Eの背後にある観光テーマ(潜在クラス)Tを想定したため、PLSAのアウトプットはP(S|T), P(E|T), P(T)が計算された。抽出された15個のクラスについて、所属確率の高い観光資源と係り受け表現から、そのクラスを意味する観光テーマを解釈し、各クラスの名称を付けた。各クラスの名称および解釈した観光テーマと、それに所属する観光資源および係り受け表現をまとめたものを表1に示す。なお表1において、同じ行に

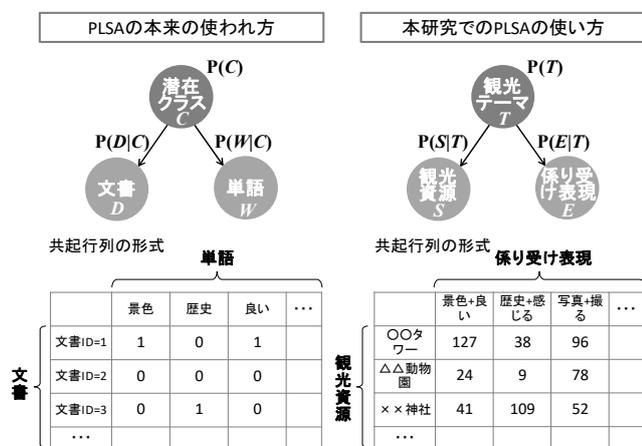


図2 PLSAの本来の使い方と本研究での使い方

ある観光資源と係り受け表現は互いに対応しているわけではなく、各クラスに所属する観光資源と係り受け表現について、所属確率の高い上位10件をそれぞれ列挙したものである。

本分析より、日本全国の各観光資源において観光客は多種多様な話題をしているが、AICという数学的な基準においては、歴史や自然、寺社、テーマパーク、買物などの15個の観光テーマに集約されることが分かった。なお、テーマ13は「温泉に入ること」と「車やバスで訪問すること」という二つのテーマが複合したテーマと解釈している。これは「車で移動して温泉に入った」というようなレビューコメントがされた観光

資源がある程度存在し、この共起関係が媒介して「温泉に入る」テーマと「車やバスでの訪問」という二つのテーマが一つのテーマとして抽出されたことが考えられる。そのため所属確率の高い観光資源は必ずしも温泉地ではない。PLSAに限らず次元圧縮・クラスタリングの手法では、元々のデータを複数個のクラスに集約するため、少なくとも一つは複数の意味が混在した「その他」のクラスが抽出されることが多い。しかし、今回のPLSAによる結果は全体的にクリアな構成となっており、テーマ13は「その他」と解釈するほど意味は混在していないため「温泉に入ること・車やバスで訪問すること」という複合的なテーマとした。

表1 抽出された15個の観光テーマと所属する観光資源および係り受け表現

1 歴史資料展示		2 展望景色		3 動物園・水族館			
資料が展示された施設で歴史を学ぶこと		高い場所からきれいな景色を展望すること		動物園や水族館で家族で楽しむこと			
P(s t)	観光資源名	P(e t)	係り受け表現	P(s t)	観光資源名	P(e t)	係り受け表現
1.9%	首里城公園(首里城)	1.2%	資料→展示	11.1%	沖縄美ら海水族館	1.2%	大人→楽しむ+できる
1.8%	熊本城	0.8%	場所→ある	6.4%	旭川市旭山動物園	1.1%	動物→いる
1.8%	名古屋城	0.7%	歴史→展示	2.5%	東京都恩賜上野動物園	0.9%	間近→見る+できる
1.4%	原爆ドーム	1.2%	中→入る	2.3%	アドベンチャーワールド	0.8%	子供→楽しむ+できる
1.3%	国宝松本城	0.7%	歴史→学ぶ+できる	2.2%	海遊館	0.6%	泳ぐ→姿
0.9%	首里城正殿	0.7%	歴史→わかる	1.3%	鳥羽水族館	0.8%	ある→水族館
0.9%	松山城	0.6%	石垣→残る	1.3%	京都水族館	0.5%	動物→触れ合う+できる
0.9%	松江城(千鳥城)	0.5%	展示→見る	1.3%	横浜 八景島シーパラダイス	0.5%	子供→遊ぶ
0.8%	鶴ヶ城(若松城)	0.5%	場所→思う	1.3%	名古屋港水族館	0.5%	ショー→見る
0.8%	大坂城	0.5%	写真→展示	1.3%	すみだ水族館	0.5%	餌→あげる
4 桜・公園		5 歴史的建物		6 海			
桜のきれいな広大な公園で自然を楽しむこと		歴史的な建物で歴史を感じること		青くきれいな海を眺めること			
P(s t)	観光資源名	P(s t)	観光資源名	P(s t)	観光資源名	P(e t)	係り受け表現
1.7%	上野恩賜公園	1.1%	札幌時計台	2.7%	川平湾	1.6%	天気→良い
1.3%	大坂城公園	0.8%	北海道庁旧本庁舎	2.3%	古宇利大橋	1.4%	海→きれい
1.1%	千鳥ヶ淵	0.7%	大浦天主堂	2.2%	古宇利島	1.3%	海→見る+できる
1.0%	日比谷公園	0.7%	道後温泉本館	2.1%	万寿毛	1.3%	青い→海
0.9%	井の頭恩賜公園	0.6%	旧函館区公会堂	1.6%	厳島神社(広島)	1.2%	きれいな→海
0.9%	新宿御苑	0.8%	首里城公園(首里城)	1.3%	コンドイビーチ	1.1%	海→眺める
0.9%	大通公園	0.7%	函館ハリストス正教会	1.2%	平久保崎	1.1%	海→見る
0.9%	代々木公園	0.7%	大阪市中央公会堂	1.2%	砂山	0.9%	風→強い
0.8%	円山公園(京都)	0.4%	東京駅 赤レンガ駅舎	1.0%	与那覇前浜	0.8%	人→いる+ない
0.8%	山下公園	0.4%	グラバー園	1.0%	沖縄美ら海水族館	0.6%	晴れる→日
7 買物街		8 紅葉・寺		9 景勝地			
多くの店で賑わう場所で買物を楽しむこと		お寺などのきれいな紅葉を楽しむこと		雄大な景勝地の景色を眺めること			
P(s t)	観光資源名	P(s t)	観光資源名	P(s t)	観光資源名	P(e t)	係り受け表現
2.5%	横浜中華街	3.8%	清水寺(京都)	1.6%	大涌谷	1.4%	天気→良い
2.4%	国際通り	2.8%	鹿苑寺(金閣寺)	1.2%	高尾山	1.2%	富士山→見る+できる
2.4%	浅草寺	1.7%	東福寺	1.1%	オジンコシンの滝	0.7%	紅葉→時期
1.7%	高山の古い町並み	1.6%	東大寺	1.1%	富士山	0.7%	滝→見る+できる
1.6%	仲見世通り(東京)	1.4%	天龍寺(京都)	0.9%	華厳滝	0.6%	海→行く
1.5%	南町	1.2%	銀閣寺(慈照寺)	0.9%	袋田の滝	0.6%	きれいな→見る+できる
1.5%	おかげ横丁	1.1%	高徳院(鎌倉大仏)	0.8%	高千穂峡	0.6%	滝→見る
1.3%	錦市場	1.1%	龍安寺	0.8%	九重「夢大吊橋」	0.5%	景色→良い
1.3%	倉敷美観地区	1.1%	南禅寺	0.8%	竹田城跡	0.5%	人→いる
1.3%	道頓堀	1.1%	高台寺	0.7%	阿蘇山	0.5%	標高→高い
10 テーマパーク		11 花		12 散策			
テーマパークで家族で楽しむこと		広大な敷地できれいに咲く花を楽しむこと		散策しながらきれいな自然を堪能すること			
P(s t)	観光資源名	P(s t)	観光資源名	P(s t)	観光資源名	P(e t)	係り受け表現
4.1%	東京ディズニーシー	2.2%	ファーム富田	1.9%	美瑛 青い池	0.9%	遊歩道→整備
4.0%	東京ディズニーランド	1.8%	兼六園	1.4%	白川郷合掌造り集落	0.7%	気持ち→良い
3.4%	ユニバーサルスタジオジャパン	1.2%	四季彩の丘	1.1%	高千穂峡	0.7%	写真→撮る
2.5%	東京スカイツリー	1.1%	足立美術館	1.0%	奥入瀬渓流	0.6%	流れる→川
1.6%	ハウステンボス	1.0%	国営ひたち海浜公園	0.9%	忍野八海	0.6%	遊歩道→歩く
1.1%	なばなの里 ウィンターイルミネーション	0.9%	金沢21世紀美術館	0.8%	小樽運河	0.6%	川→流れる
1.1%	白い恋人パーク	0.8%	グラバー園	0.8%	斎場御嶽	0.6%	水→きれい
0.9%	カップヌードルミュージアム	0.8%	岡山後楽園	0.7%	金鱗湖	0.6%	天気→良い
0.8%	横浜アンパシオンこどもミュージアム	0.8%	ハイツテンボス	0.7%	東尋坊	0.6%	水→流れる
0.7%	三鷹の森ジブリ美術館	0.8%	栗林公園	0.7%	十和田湖	0.4%	遊覧船→乗る
13 温泉・アクセス		14 写真スポット		15 神社			
温泉に入ること・車やバスで訪問すること		橋や建造物などの写真を撮ること		神社を参拝して厳かな雰囲気を感じる			
P(s t)	観光資源名	P(s t)	観光資源名	P(s t)	観光資源名	P(e t)	係り受け表現
1.4%	道後温泉本館	2.4%	札幌時計台	3.2%	伊勢神宮 内宮	2.3%	ある→神社
0.9%	羊ヶ丘展望台	1.6%	東京駅 赤レンガ駅舎	2.3%	出雲大社	1.1%	祀る→神社
0.6%	血の池地獄	1.6%	東京タワー	2.1%	伏見稲荷大社	1.1%	鳥居→くぐる
0.6%	ファーム富田	1.2%	渡月橋	1.8%	伊勢神宮外宮	0.9%	人→多い
0.6%	湯畑	1.1%	東京スカイツリー	1.5%	大宰府天満宮	0.9%	場所→ある
0.5%	砂むし会館砂楽	1.1%	さっぽろテレビ塔	1.5%	明治神宮	0.9%	有名→神社
0.5%	竹瓦温泉	1.0%	大通公園	1.4%	厳島神社(広島)	0.8%	境内→ある
0.5%	別府の地獄	1.0%	レイニーブリッジ	1.2%	八坂神社(京都)	0.6%	厳か→雰囲気
0.4%	宗谷岬	1.0%	赤レンガパーク	1.2%	熱田神宮	0.5%	参道→歩く
0.4%	ニッカウキスキー北海道工場	0.9%	ほりまや橋	1.2%	鶴岡八幡宮	0.5%	初詣で→行く

#### 4 地域別による話題の特徴分析

抽出した15個のテーマを軸として、全国の各地域がどのようなテーマの話題をされているのかということを定量的に分析した。

PLSAの実行によって観光テーマTごとに計算された各観光資源Sの条件付出現確率 $P(S|T)$ について、前述の式(1)のようにその観光資源が存在する地域Rで確率を集約し、観光テーマTを条件とした地域Rの条件付出現確率 $P(R|T)$ を計算し、さらに式(2)のようにベイズの定理を適用して各地域Rを条件とした観光テーマTの条件付出現確率 $P(T|R)$ を計算した。この値は各地域における15個の観光テーマの構成割合を意味しており、各地域における観光資源の特徴を定量的に把握することができる。特に事後確率 $P(T|R)$ を事前確率 $P(T)$ で除した $P(T|R)/P(T)$ の値は、元々の観光テーマTの確率が、その地域Rという条件を加えることで何倍に上昇するのかということの意味しており、1が基準となるが値が大きいテーマほどその地域の特徴的なテーマといえる。

例えば、北海道の主要な地域（レビュー総数が上位20箇所の地域）について $P(T|R)$ を計算した結果を表2に示す。表2において、各地域の観光資源数とは10件以上のレビューがある観光資源の数であり、レビュー総数とはそれらの観光資源に対するレビュー件数の総和である。地域の並びはなるべく近い地域でまとめて並べている。また $P(T|R)/P(T)$ を計算したとき、値が1.5以上及び2.0以上のものを網掛けで示し、特に値の大きなものを強調することで特徴を見やすくした。なお地域の分類はフォートラベルのサイトで利用されている区町村単位の地域分類を使用した。

また、それぞれの地域の特徴的なテーマは具体的にどのような観光資源が影響を与えているのかということを確認するため、前述の式(2)のベイズの定理をPLSAで計算された $P(S|T)$ に適用して、観光資源Sを条件とした観光テーマTの条件付出現確率 $P(T|S)$ を算出した。この値は各観光資源における15個の観光テーマの構成割合を意味している。また各観光資源において特徴的なテーマを把握するため、事後確率 $P(T|S)$ を事前確率 $P(T)$ で除した $P(T|S)/P(T)$ も同様に算出した。表2においてレビュー総数の多かった、「札幌駅周辺・大通り」「札幌すすきの」「札幌市南西部」「小樽」

「函館」を対象に、それぞれの地域でレビュー件数が上位5箇所の観光資源を例に算出した結果を表3に示す。表2と同様に $P(T|S)/P(T)$ が1.5以上及び2.0以上のものを網掛けで示し、特に値の大きなものを強調することで特徴を見やすくした。なお観光資源の名称はフォートラベルのサイトで利用されている施設名称を使用した。

表2, 3より、例えば札幌の駅周辺や大通りは、時計台や北海道庁旧本庁舎、テレビ塔や大通公園などがあるので、歴史的建物や写真スポットというテーマの話題が多い地域であること、札幌のすすきの、全国的にも有名な繁華街で飲食店も多いが、それだけでなく中島公園や創成川公園などもあり、公園や買物街というテーマの話題が多い地域であること、札幌市の南西部にはジャンプ競技場やもいわ山展望台、白い恋人パークや円山動物園などがあるので、展望景色、テーマパーク、動物園・水族館といったテーマの話題が多い地域であることが分かる。小樽は運河やノスタルジックな町並みで有名であるが、建物が歴史的で、散策しながら買物が楽しめる地域であること、函館は夜景で有名であるが、函館山や五稜郭タワーからの展望景色が綺麗であることがよく話題にされている地域であると分かる。また富良野と美瑛は互いに近い地域であり、両地域ともお花畑が綺麗で有名だが、花以外のテーマについては、富良野は車などでの移動や景勝地のテーマが特徴的で、美瑛は散策や展望景色のテーマが特徴的であることが分かる。ウトロは世界自然遺産の知床があるため、景勝地と散策のテーマの話題が多く、千歳は千歳空港のターミナルビルが商業施設化しており、ミュージアムや体験施設が充実していることから、テーマパークや資料展示のテーマの話題が多く出ていると思われる。また、道東に位置する釧路、摩周・弟子屈、阿寒は、散策をしながら展望景色や景勝地を楽しめる地域であることが想像できる。特に阿寒は北海道最大のアイヌ民族の集落であるアイヌコタンがあり、アイヌの民芸品やアイヌ料理などが楽しめる地域であり、買物街としてのテーマも特徴的である。

このように観光客の生の声から抽出された15個のテーマを軸にして各地域の特徴を観光客目線で考察することができる。

表2 北海道の主要な地域における15個のテーマの構成割合

地域名称	観光資源数	レビュー総数	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
			歴史資料展示	展望景色	動物園・水族館	桜公園	歴史的建物	海	買物街	紅葉・寺	景勝地	テーマパーク	花	散策	温泉・アクセス	写真スポット	神社
札幌駅周辺・大通り	74	3,002	6.6%	8.8%	0.2%	11.9%	14.7%	0.0%	6.8%	0.8%	0.4%	4.3%	6.6%	3.1%	9.4%	25.9%	0.6%
札幌すすきの	18	513	1.7%	0.9%	0.2%	32.9%	13.3%	0.0%	22.1%	1.6%	0.2%	0.3%	1.3%	6.0%	6.7%	9.2%	3.5%
札幌市南西部	16	717	3.0%	18.5%	14.0%	10.0%	1.0%	0.2%	0.2%	0.1%	9.6%	18.3%	5.9%	2.3%	10.4%	0.7%	5.7%
札幌市南東部	17	457	9.1%	11.4%	6.2%	12.0%	1.9%	0.0%	1.0%	0.2%	0.4%	8.1%	7.8%	0.0%	36.9%	4.9%	0.0%
札幌市北部	16	330	9.9%	4.4%	4.3%	26.3%	10.0%	0.6%	0.9%	0.4%	6.5%	7.9%	9.4%	2.0%	15.7%	1.8%	0.0%
小樽	29	1,001	8.7%	3.2%	5.6%	1.7%	15.3%	1.7%	24.8%	0.1%	0.4%	2.0%	1.7%	16.6%	7.2%	11.0%	0.1%
函館	39	1,818	7.5%	31.8%	0.3%	9.1%	14.0%	3.6%	10.1%	1.1%	5.5%	2.8%	1.6%	3.0%	4.9%	3.6%	1.0%
富良野	24	691	3.8%	5.7%	0.0%	0.8%	4.6%	1.5%	2.2%	0.4%	14.9%	2.9%	40.2%	10.1%	12.6%	0.4%	0.0%
美瑛	15	659	2.0%	14.1%	0.9%	0.7%	0.7%	2.6%	0.5%	0.1%	6.0%	0.3%	27.3%	37.0%	5.8%	2.2%	0.0%
旭岳・天人峡・白金	7	272	0.0%	2.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.5%	22.0%	0.0%	0.0%	65.9%	8.0%	1.4%	0.0%
旭川	6	531	3.7%	0.2%	87.5%	0.2%	0.2%	0.0%	0.9%	0.4%	0.4%	0.0%	2.9%	2.4%	1.0%	0.3%	0.0%
ウトロ	13	552	0.1%	8.8%	0.8%	0.0%	0.6%	9.7%	0.0%	0.0%	55.4%	0.8%	0.4%	19.4%	3.1%	0.9%	0.0%
網走	7	226	31.0%	9.5%	3.7%	0.0%	5.2%	14.7%	0.0%	0.0%	4.8%	1.5%	5.2%	11.1%	12.2%	1.0%	0.0%
千歳	12	333	24.2%	7.9%	9.0%	0.0%	0.5%	0.0%	1.2%	0.0%	0.0%	42.3%	0.0%	0.4%	11.9%	2.1%	0.5%
登別	9	247	1.0%	7.4%	9.8%	1.9%	0.9%	0.0%	2.2%	0.0%	10.8%	4.3%	0.0%	29.3%	32.5%	0.0%	0.0%
釧路	8	273	1.5%	25.1%	3.6%	0.2%	1.5%	2.9%	0.4%	0.0%	18.7%	0.3%	0.5%	26.9%	8.1%	10.3%	0.0%
摩周・弟子屈	4	193	0.0%	23.5%	0.6%	0.0%	0.0%	1.6%	0.3%	0.0%	34.5%	0.0%	0.9%	28.1%	10.4%	0.0%	0.0%
阿寒	6	191	3.7%	5.5%	3.4%	0.0%	0.8%	8.7%	20.3%	0.0%	1.7%	0.8%	0.2%	45.3%	7.8%	1.7%	0.0%
稚内	9	266	1.6%	12.9%	6.2%	6.1%	7.3%	35.8%	2.3%	0.0%	1.4%	0.1%	1.5%	2.4%	21.4%	0.9%	0.0%
帯広	7	202	11.7%	0.0%	9.1%	9.8%	1.4%	0.0%	3.3%	0.0%	0.0%	3.6%	23.6%	1.4%	34.0%	1.5%	0.7%
P(T)	5,553	169,407	6.7%	7.4%	5.1%	7.3%	7.1%	7.5%	6.7%	8.0%	7.5%	6.5%	5.2%	7.5%	5.3%	5.8%	6.5%

※   P(T)R/P(T)が1.5以上   P(T)R/P(T)が2.0以上

表3 北海道の主要な観光資源における15個のテーマの構成割合

観光資源名称	所在地域	レビュー件数	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
			歴史資料展示	展望景色	動物園・水族館	桜公園	歴史的建物	海	買物街	紅葉・寺	景勝地	テーマパーク	花	散策	温泉・アクセス	写真スポット	神社
札幌時計台	札幌駅周辺・大通り	361	8.4%	0.0%	0.0%	0.0%	30.8%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	4.9%	55.7%	0.0%
さっぽろテレビ塔	札幌駅周辺・大通り	278	0.0%	54.9%	0.0%	1.7%	0.1%	0.0%	0.0%	0.0%	0.0%	2.5%	0.6%	0.0%	0.6%	39.7%	0.0%
大通公園	札幌駅周辺・大通り	271	0.0%	0.0%	0.0%	41.7%	0.0%	0.0%	2.8%	0.0%	0.0%	3.5%	11.9%	0.5%	2.0%	37.4%	0.0%
北海道庁日本庁舎	札幌駅周辺・大通り	189	29.3%	0.0%	0.2%	5.1%	37.3%	0.0%	0.0%	0.1%	0.0%	0.0%	11.6%	1.6%	4.3%	10.5%	0.0%
さっぽろ雪まつり	札幌駅周辺・大通り	141	0.0%	0.0%	1.8%	1.3%	0.0%	0.0%	5.1%	0.2%	5.2%	40.3%	0.0%	0.0%	7.7%	38.3%	0.0%
すすきの	札幌すすきの	112	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	65.7%	0.0%	0.0%	0.7%	0.0%	0.0%	9.8%	23.7%	0.0%
中島公園	札幌すすきの	79	0.0%	0.0%	0.0%	91.1%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.1%	2.4%	1.7%	3.8%	0.0%
元祖さっぽろラーメン横丁	札幌すすきの	53	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
創成川公園	札幌すすきの	41	0.4%	0.0%	0.0%	60.5%	0.0%	0.0%	5.5%	0.0%	0.0%	0.0%	8.3%	17.0%	2.9%	5.3%	0.0%
豊平館	札幌すすきの	41	0.0%	0.0%	0.0%	10.9%	64.6%	0.0%	0.0%	4.0%	0.0%	0.0%	0.0%	10.9%	0.0%	9.6%	0.0%
白い恋人パーク	札幌市南西部	163	10.0%	0.0%	0.0%	0.0%	2.5%	0.0%	0.0%	0.0%	0.0%	63.7%	13.5%	0.0%	10.3%	0.0%	0.0%
円山動物園	札幌市南西部	133	0.0%	0.0%	68.8%	10.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	2.5%	17.2%	0.0%	0.0%	0.0%
大倉山ジャンプ競技場	札幌市南西部	101	0.0%	54.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	29.1%	8.7%	0.0%	0.0%	7.4%	0.0%	0.0%
北海道神宮	札幌市南西部	66	0.0%	0.0%	1.4%	22.4%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	3.9%	18.0%	0.0%	54.1%
もいわ山展望台	札幌市南西部	64	0.0%	82.2%	0.0%	0.0%	0.0%	0.0%	0.6%	0.0%	24.7%	4.9%	0.0%	0.0%	7.6%	0.0%	0.0%
小樽運河	小樽	299	0.4%	0.0%	0.0%	0.0%	2.5%	1.9%	44.5%	0.0%	0.0%	0.2%	0.0%	33.3%	0.2%	16.9%	0.0%
JR小樽駅	小樽	98	0.0%	1.0%	0.0%	0.0%	8.3%	1.0%	33.0%	0.0%	0.0%	0.0%	0.0%	0.0%	41.7%	15.0%	0.0%
旧手宮線跡	小樽	74	4.2%	0.0%	0.0%	18.4%	14.0%	0.0%	5.2%	0.0%	0.0%	0.0%	0.0%	50.7%	0.2%	7.4%	0.0%
おたる水族館	小樽	46	0.0%	2.5%	82.1%	0.0%	0.0%	6.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8.9%	0.0%	0.0%
日本銀行旧小樽支店金融資料館	小樽	45	32.9%	0.0%	0.0%	0.0%	65.5%	0.0%	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.9%	0.0%
五稜郭タワー	函館	204	17.0%	63.6%	0.0%	12.5%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%	0.0%	0.2%	4.2%	1.0%	0.0%
函館山	函館	173	0.0%	68.2%	0.0%	0.0%	0.0%	0.1%	1.7%	0.1%	21.9%	5.7%	0.0%	0.0%	2.3%	0.0%	0.0%
函館山展望台	函館	152	0.0%	65.0%	0.0%	0.0%	0.0%	0.3%	0.2%	0.0%	22.1%	11.3%	0.0%	0.0%	0.2%	0.8%	0.0%
金森赤レンガ倉庫	函館	126	1.3%	0.0%	0.0%	0.0%	4.0%	4.3%	62.4%	0.0%	0.0%	1.4%	0.5%	0.0%	1.8%	24.3%	0.0%
函館朝市	函館	120	0.0%	0.0%	2.1%	0.0%	0.0%	0.0%	91.7%	0.0%	0.0%	0.0%	0.0%	0.0%	6.2%	0.0%	0.0%
P(T)		169,407	6.7%	7.4%	5.1%	7.3%	7.1%	7.5%	6.7%	8.0%	7.5%	6.5%	5.2%	7.5%	5.3%	5.8%	6.5%

※   P(T)S/P(T)が1.5以上   P(T)S/P(T)が2.0以上

### 5 観光客の属性別による話題の特徴分析

各観光客属性はどのようなテーマの話題をする傾向にあるのかということを定量的に分析した。

PLSAの実行によって観光テーマTごとに計算された各係り受け表現Eの条件付出現確率P(E|T)に基づいて、前述の式(3)のように観光テーマTを条件とした各

レビューデータDの条件付出現確率P(D|T)を計算し、式(4)のようにP(D|T)を観光客の属性Aで集約し、さらに式(5)のようにベイズの定理を適用して各観光客属性Aを条件とした観光テーマTの条件付出現確率P(T|A)を計算した。この値は各観光客属性が15個の観光テーマのうちどのテーマの話題を多くしているのかという

構成割合を意味しており、各観光客属性における各テーマの話題の度合いを定量的に把握することができる。第4章と同様に事後確率P(T|A)を事前確率P(T)で除したP(T|A)/P(T)を求めたとき、1を基準に値が大きいテーマほどその観光客層にとってよく話題にし、関心の強いテーマといえる。

例えば、性別と年代及び性別と同行者という組み合わせの観光客属性について、P(T|A)を計算した結果をそれぞれ表4、5に示す。表4、5において、P(T|A)/P(T)を計算したとき、値が1.1以上及び1.2以上のものを網掛けで示しているが、地域の特徴分析と比べ値が小さくなっている。これは各属性カテゴリの件数が地域のカテゴリと比べて多く、P(T|A)がP(T)に近づくためであるが、その中でも相対的に値が高いものを網掛けで強調することで特徴を見やすくした。観光客属性のカテゴリはフォートラベルのサイトで利用されている投稿者属性を使用した。

表4より、例えば50代～70代の年配の男性は歴史的な建物や神社の話題が、10代～30代の若い女性は動物

園・水族館やテーマパークの話題が、50代～70代の年配の女性は花の話題が多い傾向にあることが分かる。また表5より、性別に関わらず家族旅行（乳幼児連れを含む）は動物園・水族館やテーマパークの話題が、一人旅の男性は歴史的な建物の話題が、社員・団体旅行の女性は海や景勝地、散策の話題が多い傾向にあることが分かる。

また各テーマについて見ると、表4,5より展望景色、買物街などのテーマは観光客の属性によって特徴があまり見られないテーマである。本研究では観光客の代表的な属性として、性別、年代、同行者を取り上げたが、こうした属性の分け方では話題の度合いに差が生まれにくいテーマも地域観光にはあるということが示唆される。つまり地域における観光振興では、観光のテーマだけでなく観光客の属性を分ける効果的な軸の設計も検討の余地があるといえる。

表4 性別×年代における15個のテーマの構成割合

性別	年代	レビュー総数	T1 歴史資料展示	T2 展望景色	T3 動物園・水族館	T4 桜・公園	T5 歴史的建物	T6 海	T7 買物街	T8 紅葉・寺	T9 景勝地	T10 テーマパーク	T11 花	T12 散策	T13 温泉・アクセス	T14 写真スポット	T15 神社
男性	10代	724	7.7%	8.3%	6.7%	5.4%	5.9%	7.7%	6.5%	7.8%	7.7%	8.5%	4.1%	6.5%	6.4%	5.8%	5.0%
	20代	8,707	7.4%	7.7%	5.0%	6.6%	7.9%	7.4%	7.0%	8.1%	7.3%	6.3%	4.5%	7.2%	5.5%	5.7%	6.5%
	30代	26,480	7.0%	7.4%	4.9%	7.4%	7.8%	7.4%	7.1%	7.9%	7.4%	5.9%	4.7%	7.3%	5.4%	6.0%	6.4%
	40代	26,275	7.2%	7.6%	5.0%	7.4%	7.3%	7.5%	6.8%	7.6%	7.5%	6.3%	4.7%	7.2%	5.4%	6.2%	6.4%
	50代	23,464	7.3%	7.0%	3.7%	7.8%	9.3%	6.6%	6.7%	8.4%	7.2%	4.9%	5.0%	7.4%	5.1%	6.4%	7.4%
	60代	9,071	7.9%	7.5%	3.5%	7.3%	8.8%	7.5%	6.3%	8.4%	7.7%	4.1%	5.3%	7.8%	4.9%	5.3%	7.4%
	70代	1,682	6.5%	6.2%	4.1%	9.5%	8.3%	5.7%	5.9%	8.8%	7.4%	4.6%	6.8%	7.1%	4.9%	7.0%	7.2%
女性	10代	428	6.4%	7.3%	8.4%	5.7%	4.8%	6.4%	6.9%	6.7%	6.8%	12.8%	5.1%	6.6%	5.7%	6.1%	4.3%
	20代	12,557	5.9%	7.7%	6.6%	6.6%	5.1%	8.0%	7.0%	7.8%	7.6%	8.3%	5.5%	7.5%	5.3%	5.3%	5.8%
	30代	27,272	6.0%	7.4%	6.3%	7.0%	5.5%	7.9%	6.6%	7.6%	7.6%	8.1%	5.5%	7.5%	5.3%	5.4%	6.2%
	40代	17,711	5.9%	7.2%	5.3%	7.4%	6.2%	7.9%	6.8%	8.0%	7.6%	7.3%	5.7%	7.7%	5.2%	5.5%	6.4%
	50代	10,728	6.5%	7.5%	4.4%	7.5%	7.5%	7.1%	6.4%	8.5%	7.9%	5.6%	5.9%	7.9%	5.1%	5.7%	6.5%
	60代	3,200	6.3%	7.2%	3.9%	8.2%	7.7%	7.6%	6.2%	8.9%	8.4%	4.6%	6.1%	8.5%	4.6%	5.4%	6.4%
	70代	201	5.4%	5.2%	4.4%	6.7%	7.5%	5.0%	6.6%	9.8%	6.3%	5.7%	7.4%	6.9%	3.7%	5.0%	14.5%
P(T)		169,407	6.7%	7.4%	5.1%	7.3%	7.1%	7.5%	6.7%	8.0%	7.5%	6.5%	5.2%	7.5%	5.3%	5.8%	6.5%

※  P(T|A)/P(T)が1.1以上  P(T|A)/P(T)が1.2以上

表5 性別×同行者における15個のテーマの構成割合

性別	同行者	レビュー総数	T1 歴史資料展示	T2 展望景色	T3 動物園・水族館	T4 桜・公園	T5 歴史的建物	T6 海	T7 買物街	T8 紅葉・寺	T9 景勝地	T10 テーマパーク	T11 花	T12 散策	T13 温泉・アクセス	T14 写真スポット	T15 神社
男性	一人旅	33,099	7.8%	7.0%	3.6%	8.2%	9.6%	6.3%	7.1%	8.6%	6.6%	4.3%	4.9%	7.0%	5.2%	6.6%	7.3%
	友人	9,246	6.7%	7.8%	4.5%	7.1%	6.8%	7.4%	7.1%	8.0%	8.1%	5.9%	4.6%	7.6%	5.9%	6.0%	6.5%
	カップル・夫婦	19,239	6.6%	7.8%	4.5%	6.9%	7.4%	7.8%	7.1%	8.1%	8.0%	5.6%	4.9%	7.7%	5.2%	5.9%	6.5%
	カップル・夫婦(シニア)	5,212	7.2%	7.8%	3.8%	6.9%	8.3%	8.5%	6.0%	8.4%	8.6%	4.2%	5.1%	8.4%	5.1%	4.8%	6.9%
	乳幼児連れ家族旅行	2,935	5.7%	6.6%	10.6%	7.3%	4.9%	8.0%	5.9%	6.2%	6.7%	12.0%	4.7%	6.3%	5.2%	5.0%	4.8%
	家族旅行	11,389	6.6%	7.5%	6.6%	6.4%	5.7%	8.1%	6.2%	7.2%	8.1%	8.8%	4.8%	7.5%	5.6%	4.9%	6.0%
	社員・団体旅行	1,165	8.0%	8.5%	4.9%	5.2%	6.7%	7.5%	6.5%	8.1%	9.0%	5.8%	4.3%	8.5%	5.6%	5.2%	6.3%
女性	一人旅	11,140	6.8%	6.9%	3.9%	8.2%	7.7%	7.1%	6.9%	9.1%	6.7%	4.8%	6.1%	7.5%	4.9%	5.9%	7.4%
	友人	12,427	5.8%	7.5%	5.5%	6.8%	5.4%	7.6%	7.1%	8.1%	7.7%	7.9%	5.7%	7.6%	5.5%	5.7%	6.3%
	カップル・夫婦	17,174	5.9%	7.8%	5.3%	6.9%	6.2%	8.3%	6.6%	7.9%	8.1%	6.5%	5.4%	8.0%	5.2%	5.5%	6.3%
	カップル・夫婦(シニア)	2,695	6.1%	7.7%	4.8%	7.3%	6.0%	8.2%	6.4%	8.1%	8.4%	6.3%	5.8%	8.4%	5.1%	5.3%	6.0%
	乳幼児連れ家族旅行	4,766	5.5%	6.3%	12.4%	7.1%	4.0%	7.5%	5.4%	5.4%	6.6%	14.6%	5.4%	6.0%	5.1%	4.4%	4.1%
	家族旅行	12,874	5.9%	7.6%	6.4%	6.6%	5.1%	7.9%	6.4%	7.5%	8.2%	8.7%	5.6%	7.9%	5.4%	4.9%	5.9%
	社員・団体旅行	1,094	7.2%	8.3%	5.3%	6.4%	5.7%	9.0%	5.8%	7.8%	9.7%	5.3%	5.2%	9.1%	5.0%	4.6%	5.4%
P(T)		169,407	6.7%	7.4%	5.1%	7.3%	7.1%	7.5%	6.7%	8.0%	7.5%	6.5%	5.2%	7.5%	5.3%	5.8%	6.5%

※  P(T|A)/P(T)が1.1以上  P(T|A)/P(T)が1.2以上

## 6 地域観光振興の施策の検討

第4章で分析した地域の特徴と第5章で分析した観光客の特徴に基づいて、地域における観光振興の施策の検討を試みる。本研究では、表2, 4, 5の結果を用いて、北海道の地域を例に効果的な誘客ターゲット層の設定や連携地域について、以下の三つの観点から施策化の可能性を考察した。

### 6.1 自身の地域でよく話題にされるテーマに対し、それに関心の強い観光客層をターゲットに設定する

例えば小樽を例にした場合、表2より、よく話題にされるテーマは「T5:歴史的建物」、「T7:買物街」、「T12:散策」である。「T5:歴史的建物」をよく話題にする観光客層は、表4, 5より50代~70代の年配の男性、一人旅の男性、シニアの夫婦旅行の男性などである。「T7:買物街」については、先述の通り観光客の属性で話題の度合いに大きな差はみられない。「T12:散策」をよく話題にする観光客層は、表4, 5より60代の女性や性別を問わずシニアの夫婦旅行及び社員・団体旅行の観光客である。特に小樽は札幌と近い位置にある観光地であるが、札幌周辺の観光地と比較して、「T12:散策」は小樽に特徴的なテーマであり、周辺地域に対して差別化できる可能性がある。

以上から、小樽でよく話題にされるテーマに対し、それをよく話題にして関心のある観光客層は、シニア層や社員・団体旅行者、一人旅の男性などが挙げられ、これらの観光客層をターゲットに設定して小樽の観光資源をPRすることで、誘客促進につながることが期待できる。

### 6.2 よく話題にされるテーマが類似する他の地域と連携して誘客の相乗効果を狙う

表2より、札幌（駅周辺・大通り・すすきの）、小樽、函館は「T5:歴史的建物」と「T7:買物街」のテーマがよく話題にされていることが共通しているといえ、これらのテーマをよく話題にして関心の高い観光客層は先述の通り、50代~70代の年配の男性、一人旅の男性、シニアの夫婦旅行の男性などであった。この二つのテーマに加え札幌では「T4:桜・公園」や「T14:写真スポット」、小樽では「T12:散策」、函館では「T2:展望景色」がよく話題にされる特徴的なテーマとなっている。

例えば、この三つの地域が一つの広域観光ルートとして連携し、「T5:歴史的建物」と「T7:買物街」を中心テーマにして各地の特徴的なテーマとともにPRする活動に取り組むことで、この「T5:歴史的建物」と「T7:買物街」をよく話題にして関心のある観光客を呼び込み、また各地域で異なる特徴テーマに触れさせることで地域間の回遊を促進させ、相互送客といった相乗効果が得られることや誘客コストの削減にもつながることが期待できる。

### 6.3 新たなターゲット層の開拓に効果的な連携地域を検討する

先述の通り、小樽でよく話題にされるテーマではシ

ニア層や社員・団体旅行者、一人旅の観光客などが効果的なターゲット層として考えられたが、例えば他に家族旅行者をターゲット層として設定したいと考えた場合、家族旅行者を取り込むのに効果的な連携地域を検討する。

家族旅行者がよく話題にする関心テーマは、表5より「T3:動物園・水族館」と「T10:テーマパーク」である。このテーマの話題がよくされる地域は、表2より「札幌市南西部」や「旭川」、「千歳」である。特に札幌市南西部は小樽と近い位置にあり、円山動物園や白い恋人パークなど、「T3:動物園・水族館」と「T10:テーマパーク」の両方で強みを持つ地域であり、効果的な連携先となりうる。例えば、札幌市南西部の地域と連携することで、多くの家族旅行者を小樽に呼び込める可能性が期待できる。

以上の検討は、本研究の分析結果を地域観光振興に活用することを分かりやすく示すため、観光地として有名な北海道の地域を例に取り上げて考察したものである。あくまで観光客のレビューコメントから得られた15個の観光テーマを軸として、表2の地域の特徴と表4, 5の観光客の特徴を照合し考察した一例であり、これを施策として実行して効果を検証したものではない。実際に地域における観光振興の施策を検討する際には、データの分析結果だけでなく検討されるのではなく、地域の事情や担当者の感覚など情報を総合的に扱い検討することになると思われるが、これまで得られなかった観光客の生の声の分析結果はその施策に観光客の目線を取り入れるものであり、新たな示唆を与えたり様々なアイデアの発想を支援する有用な情報であると考えられる。

## 7 まとめ

本研究では、ユーザーレビューという観光客が実際に発した大量の生の声のデータを分析することで、観光客の目線に基づいた地域における観光振興の施策立案への活用の可能性を検討した。

本研究における分析アプローチでは日本全国の各観光資源のユーザーレビューを用いて、観光客のコメントが記されたテキストデータにテキストマイニングとPLSAを適用することで、国内観光客がよく話題にする15個の観光テーマを抽出し、そのテーマを軸とした地域の特徴と観光客層の特徴を定量的に分析可能にした。この分析結果を用いることで、各地域の観光テーマに合致した誘客ターゲット層の設定や効果的な連携地域の選定などを考察でき、各地域における観光客目線の観光振興の施策検討に資するものと期待できる。

一方、ユーザーレビューには偏りがあり、必ずしもすべて信頼できるものではないかもしれないという注意はユーザーレビューの利用において意識しておくべきことである。ただ、観光客の生の声を大量に得られるデータというのは大変貴重であり、そこから得られる情報は多く、観光客の価値観の気づきや施策の手がかりを得るうえでは有用と考えられる。従来のアンケート調査では回答者の偏りを調整したデータ収集が可能かもしれないが、ユーザーレビューでは、膨大な量

の声を把握することができる。したがって、ユーザーレビューのデータも従来の調査手法と組み合わせて活用することで、より有用で効果的な施策検討を実行することが可能になるものと思われる。

また今回はユーザーレビューの活用という観点のみから地域の観光振興の施策立案への活用を検討したが、こうした客観的なデータだけでなく自治体関係者や地元の方の豊富な経験や感覚も大変貴重な情報源であり、実際の地域における観光振興では、地域側の事情や思惑もある。そのため、データのみから施策を検討するのではなく、データから得られることは重要な知見として活用しながら、その地域に関わる様々な関係者の考えも取り入れ、総合的に検討していくことが望ましい観光振興アプローチだと思われる。また地域観光振興も施策を立案するだけでは不十分であり、それを実際のアクションとして実行し、またその成果検証も適切なKPIを設定してモニタリングし、PDCAサイクルを回していく必要がある。筆者も本研究の分析アプローチを活用しながら実際に自治体と共に地域の観光振興の取り組みをしており、データ分析は強力な武器になると考えられるが、分析に留まるのではなく、観光振興に携わる関係者が皆高い意識を持ってその武器を使いこなし、施策を検討・実行・評価していくように、観光振興を共創していくことが重要であると思われる。今後はこうしたデータ分析に留まらない現場と共に地域振興を考えるアクションリサーチを研究の課題としたい。

なお、本研究では観光に関するユーザーレビューを分析し、それを地域の観光振興の施策立案に活用する可能性を検討したが、これは地域観光に限らず他の領域においても適用可能と思われる。例えば宿泊施設のユーザーレビューに適用し、周辺の宿泊施設と差別化する宿泊者目線のテーマを把握してPR方法を検討したり、飲食店のユーザーレビューに適用し、利用者目線でテーマが類似する他店と連携した誘客効果を高める企画を検討することも考えられる。またECサイトの商品のユーザーレビューに適用し、消費者目線による新たな評価軸を抽出して既存商品の評価・改善や新商品の企画を検討したり、消費者目線による商品の類型化によりユーザフレンドリーなECサイトを設計したり、消費者が話題にする評価テーマを軸にして関心が高いと思われる商品を推薦するといった推薦システムの構築も考えられる。このように本研究のアプローチはユーザーレビューを活用することで顧客と共に顧客価値を高める企業や社会のサービス活動に展開が期待できるものであり、それぞれの領域に応じた適用の検討やブラッシュアップは今後の課題となる。

## 8 謝辞

本研究で用いたユーザーレビューデータはフォートラベル株式会社の許可を得て利用させていただいた。

## 9 参考文献

- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. 2nd International Symposium on Information Theory, 267-281.
- Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.
- Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proc. of Uncertainty in Artificial Intelligence*, 289-296.
- 石垣司, 竹中毅, 本村陽一 (2011). 百貨店ID付きPOSデータからのカテゴリ別状況依存的変数間関係の自動抽出法. *オペレーションズ・リサーチ*, 56(2), 77-83.
- 上原義子 (2011). 伝統的商品に対する消費者の視点とマーケティング椿油のロコミを事例に、消費者の視点を探索的に求める一. *日本経営診断学会論集*, 11, 164-169.
- 打田裕樹, 吉川大弘, 古橋武, 平尾英司, 井口浩人 (2010). Webユーザーレビューにおける評価情報の時系列変化の可視化. *知能と情報*, 22(3), 377-389.
- NTTレゾナント (2012). 購買行動におけるクチコミの影響に関する調査. *gooリサーチ*, 207.
- 大久保立樹, 室町泰徳 (2014). 旅行ガイドブックとロコミの言語解析による訪日外国人の観光地イメージに関する研究. *都市計画論文集*, 49(3), 573-578.
- 大西浩志 (2015). レビュー:ソーシャルメディアとマーケティング研究(その2)ー市場の理解とソーシャルメディア・データのバイアス. *マーケティングジャーナル*, 34(3), 58-68.
- 奥村学 (2012). ソーシャルメディアを対象としたテキストマイニング. *電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review*, 6(4), 285-293.
- 観光庁 (2017). 旅行・観光消費動向調査および訪日外国人消費動向調査.
- 齊藤史哲 (2014). Webカスタマーレビュー文の理解支援を目的とした自己組織化マップによる評価分布の可視化法. *日本経営工学会論文誌*, 65(3), 180-190.
- 才原清一郎 (2015). 観光客視点からの着地型観光の課題の考察. *日本国際観光学会論文集*, 22, 21-28.
- ジャムサラランジャワバーサンフー (2008). Web上のクチコミデータに基づくブランド・マネジメント. *日本経営診断学会論集*, 8, 209-214.
- 竹崎あかね, 木浦卓治, 林武司 (2013). インターネット通販の商品レビューの基づく野菜購入嗜好性の推定. *電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション*, 113(213), 89-92.
- 竹中毅, 石垣司, 本村陽一 (2011). 生活者行動に着目したサービス需要予測技術の検討. *人工知能学会全国大会論文集*, 25, 1B1-2.
- 田辺省二, 佐藤隆二 (2014). 観光おもてなし最新事情ーリサーチ地方景況[上]. *日経グローバル*, 253, 10-13.
- 田邊亘, 後藤正幸 (2008). 宿泊施設の戦略構築を支援するユーザーレビュー分析に関する一考察. *武蔵工業大学環境情報学部情報メディアセンタージャーナル*, 9, 91-101.
- 鶴見裕之, 増田純也, 中山厚徳 (2015). マーケティングにおけるSNS上のテキスト・データ活用の可能性と限界. *マーケティングジャーナル*, 35(2), 38-54.
- 中山祐輝, 南保英孝, 木村春彦 (2012). レビュー情報を用いた学術本の難易度推定. *人工知能学会論文誌*, 27(3), 213-222.
- 日本政策観光局 (2017). 訪日外客統計.
- 林貴宏, 尾内理紀夫 (2015). Web上のレビューを利用した映画推薦システム. *人工知能学会論文誌*, 30(1), 102-111.
- 鷲山健人, 三友仁志 (2011). クチコミを用いた携帯電話の評価分析と消費者の選択行動に関する研究ーCGMへのテキストマイニングの適用一. *情報通信学会第28回学会大会*.