WEB 情報に基づく訪日中国人観光客の味覚に関する嗜好性分析

范　敏 [a,*]　野中　尋史 [a]　Alemán Carreón Elisa Claire [a]　中井　堅誠 [a]　邊土名　朝飛 [a]

Abstract

近年のインターネットの急速な発展に伴い，訪日中国人観光客の消費行動は，SNS 等の eWOM（電子的口コミ）により精緻な分析が可能となっている．本研究では，中国人向けレストランポータルサイト「Dianping」から発信されている「食に関する消費動向」情報を着目し，エントロピー，SVM 及び統計的因果探索手法を用いて，訪日中国人観光客の味覚嗜好性に関するセンチメント分析を行った．結果として，中国人観光客は濃厚な味の店でサービス、雰囲気が良いと感じ，フグの刺身などの生ものや塩辛いものがあまり好んでいないことを示された．

キーワード：感情分析，観光情報，機械学習，因果関係，中国語

1. はじめに

日本を訪れる外国人観光客は年々増加している．日本政府は訪日キャンペーンの展開や訪日客に向けたサービスの向上を推進しており，今後も持続してインバウンド消費が拡大していく．観光産業にとって，如何にしてこのインバウンド需要を取り込むかが経営課題の一つとなっている．

2018 年日本政府観光局（JNTO）の訪日外客数の発表統計[1]等の観光統計が示すように，訪日外国人の中で中国人観光客は大きなボリュームゾーンとなっており，中国人観光客の消費額は訪日観光消費額全体の約 60%を占めるまでになっている．このような背景の中，訪日中国人観光客の消費行動の実態やニーズを把握し，より質の高い製品およびサービスを提供していくことは多くの企業にとって重要な課題となっている．一方，近年国や地方自治体による観光戦略の中で「食および食文化」による観光活性化が唱えられている．日本政府観光局のデータ「訪日前後の日本のイメージの変化」が示唆するように，訪日後に「日本の食・食文化」のイメージが大きくプラス方向に変化すること[2]などがその背景である．

マーケティングにおけるクチコミの影響は高い関心を集めてきたが，近年その重要性や効果に注目が集まって来ている[3]．近年ソーシャルメディアサービス（SNS）の急速な発展に伴い，消費者が容易に膨大なクチコミ情報を検索，利用できる．こうしたことから，訪日中国人観光客の消費行動は，SNS 等の eWOM（電子的口コミ）により精緻な分析が可能となっている．そのため，訪日中国人観光客の食に関する消費動向を把握するために，中国人向けレストランポータルサイト「Dianping」[4]で発信されているクチコミ情報の分析は非常に有意義である．

本研究では中国で多くのユーザーが利用している「Dianping」から発信されている「食に関する消費動向」情報を着目し，訪日中国人観光客が投稿しているクチコミ情報を取り上げ，クチコミから中国人観光客の味覚嗜好性に関するセンチメント分析を行う．本研究の実現により，飲食店側は様々な意見や評価を知ることが可能になりメニューやサービスの改善に生かすことができる．

2. 従来研究

従来の多くの観光市場研究は，中国人観光客を対象としたものが多い．Kau らは，シンガポールの中国人観光客に対して，アンケート調査を行い，観光についての動機，実際の行動，満足度について分析を行った上で，再訪する可能性を示した[5]．一方で，観光分野にもテキストマイニング手法を活用する研究が多数行われている．Li らは，北朝鮮を訪れる中国人観光客を対象として，ショッピング体験に着目して分析を行った[6]．結果として，親戚や友人にお土産を贈ることは中国人観光客にとって一番重要な購入動機であることを示した．ショッピング以外にも宿泊施設や観光地に関する

* E-mail: s185045@stn.nagaokaut.ac.jp
a 長岡技術科学大学
〒940-2188　新潟県長岡市上富岡町 1603－1

研究が多い．Berezina らは，旅行情報サイト・トリップアドバイザーに投稿された米国フロリダ州のホテルに関するレビュー2,510 件を，PASW モデルで分析した結果，利用者はホテルの立地・施設・サービスを重視していることを示した[7]．Viriya らは，トリップアドバイザーで掲載されているタイのヒット観光地に関するレビューを収集し，トピックモデルLDAと単純ベイズの二つの手法を用いて，タイのビーチ，島，歩行者天国などの観光地についての分析結果を示した[8]．

上記のような観光分野でテキストマイニングを用いたアプローチは，主にショッピングや宿泊に焦点を当てていた．一方で，観光客の食に関するニーズを分析する研究においてテキストデータを用いた研究は少なく，WEB 情報における詳細な投稿内容に着目した研究は筆者が知る限り存在しなかった．

3. 手法

3.1 前処理

クローリングの段階で「Dianping」サイトの URL の構造が店の ID 番号で決められているところに着目し，自動的にページを読み込み，HTML ファイルを収集した．次に，HTML ファイルからユーザーID，各レビューの文章，飲食店の味，雰囲気，サービスに関する評価点数などの情報を選出し，データベースで保存した．レビューの文章に対しては形態素解析を行った．本研究では中国語形態素解析器 Jieba を利用した[9]．形態素に文章を分割した後，中国語の表記を簡体字中国語に統一して，意味をもたないノイズとなる記号類，助動詞や接続詞などを除去した．

3.2 センチメント分析

3.2.1 エントロピーを基づいたキーワードを抽出

本研究では各単語のシャノンエントロピー（以下，エントロピーという）に基づいて特徴語の選択を行った[10]．情報理論におけるエントロピーとは，情報の期待値であり，ある事象の予測不確定性を量で表現することができる．これを自然言語処理の研究に適用することで，コーパス内で任意の単語の確率分布を観察することができる．

例えば，コーパス内で様々な文書に使用されている単語は，どの文書に出現するかを予測するのは困難であり，その単語が高いエントロピーを有することになる．それに対して，特定の文書にのみ使用され，コーパス内の他の文書に殆ど含まれない単語は，大きな文書偏在性を持つ，どの文書に含まれるかという予測の不確実性が減少し，エントロピーは 0 に近づく．この概念を以下の図 1 に示す．

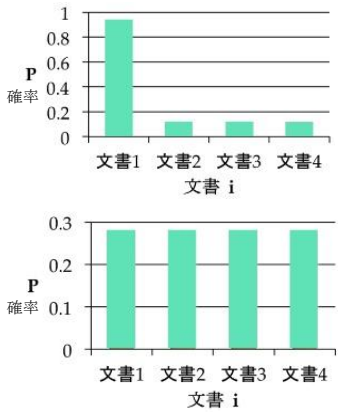

図 1　文書 i に含まれる単語 j の出現確率
上図：0 に近いエントロピーとなる場合
下図：高いエントロピーとなる場合

エントロピーの定義に基づき感情を持った単語を抽出する．ポジティブな文書内に高いエントロピーを持つ単語は，その単語の確率分布がネガティブな文書よりポジティブな文書で広がっている．つまりその単語はネガティブな文書よりもポジティブな文書で頻出する．これはネガティブに関連する単語でも同様である．

本研究では，研究室のメンバーが定義したセンチメント分析におけるエントロピーを用いた特徴語選択の方法により，分類されるものであればポジティブなキーワード，分類されなければネガティブなキーワードとして抽出した．

まずあらかじめタグを付いた教師データを利用し，ポジティブな文書集合に含まれる各文書jに各単語i が出現回数はN_{ijP}であり，一方で，ネガティブの文書集合に関してはN_{ijN}を計算する．次に以下の数式を用いて，各文書に出現する単語の出現確率を求めた．ポジティブがP_{ijP}式(1)，ネガティブがP_{ijN}式(2)である．

$$P_{ijP} = \frac{N_{ijP}}{\sum_{i=1}^{M} N_{ijP}} \tag{1}$$

$$P_{ijN} = \frac{N_{ijN}}{\sum_{i=1}^{M} N_{ijN}} \tag{2}$$

以上の値を次のエントロピーに基づいて定義した計算式の中に代入した．ポジティブの文書集合H_{Pj}(3)の各文書jにおける各単語iのエントロピーと，ネガティブの文書集合H_{Nj}(4)のエントロピーを計算する式を以下に示す．

$$H_{Pj} = -\sum_{i=1}^{M}[P_{ijP} log_2(P_{ijP})] \quad (3)$$

$$H_{Nj} = -\sum_{i=1}^{M}[P_{ijN} log_2(P_{ijN})] \quad (4)$$

各単語のエントロピーにより感情分類を行った後，パラメータαの調整を行った．αは評価データの F 値から最高値を選択する．最適なαを確定する際に，ポジティブなキーワードの場合，式(5)を満たす単語を選択する．また式(6)を用いて，ネガティブなキーワードを選出した．

$$H_{Pj} > \alpha H_{Nj} \quad (5)$$

$$H_{Nj} > \alpha' H_{Pj} \quad (6)$$

3.2.2 SVM を用いたクチコミの分類

次に，機械学習を用いて収集したクチコミ情報を分類した．SVM は，2 値分類問題を解くために広く用いられる機械学習の手法である．線形カーネル法の SVM は以下の式(7)によって定義される．教師データの各点がベクトルに与える影響は，ウエイトベクトル w に含まれるウエイトw_nによって定義される．バイアス係数b は，超平面の位置を決定する．その後未知の新しいデータを分類する際に，以下の式(8)の条件を適用する．評価は K-fold Cross Validaiton 法を用いて F1 値により行った．

$$f(x) = w^T \cdot x + b \quad (7)$$

$$f(x) = \begin{cases} \geq 0, & y = +1 \\ < 0, & y = -1 \end{cases} \quad (8)$$

3.3 統計的因果探索

統計的因果探索（Causal Discovery）とは，データから因果関係を推測するための機械学習技術である．つまり「何かを変化させたときに，他のどの変数が変化するか」と言える．

因果探索の手法は主に二種類が挙げられる．一つは制約ベース（Constraint-based）で，もう一つはスコアベース（Score-based）である．制約ベースは前提条件を定める必要があり，仮定した前提条件によって多重テストの問題が発生する．それに対して，スコアベースはこの問題を克服することができる．スコア関数モデルを使用して，最適なスコアを持つ因果グラフを出力できる．しかし，実際の複雑なデータの中，例えば因果関係は非線形の場合，線形モデルを使用すれば，情報を失う可能性が高い．そこで，Huang らは，複雑な関係性を持つデータから因果関係を探索するため，再生核ヒルベルト空間を利用し，汎化性能も高い一般化スコア関数を提案した[11].

この手法は離散や連続などのデータ性質を依存しない交差検証尤度S_{CV}（Cross-Validated Likelihood）と限界尤度S_M（Marginal Likelihood）を提案している．交差検証尤度はデータが不足する場合に，汎化性能を高める手法である．以下の式(9)に交差検証尤度の定義式を示す．限界尤度は過学習を避けるために広く使われている手法で，式(10)によって定義される．本研究では，S_{CV}，S_Mを用いて因果関係を分析する．

$$S_{\mathrm{CV}}(X_i, PA_i^{\mathcal{G}_h}) = \frac{1}{Q}\sum_{q=1}^{Q} \ell(\hat{F}_i^{(q)} | D_{0,i}^{(q)}) \quad (9)$$

$$\begin{aligned} & S_M(X_i, PA_i^{\mathcal{G}_h}) \\ & = -\frac{1}{2}\mathrm{trace}\{\tilde{K}_{X_i}(\tilde{K}_{PA_i^{\mathcal{G}_h}} + \hat{\sigma}_i^2 I)^{-1}\tilde{K}_{X_i}\} \\ & - \frac{n}{2}\log|\tilde{K}_{PA_i^{\mathcal{G}_h}} + \hat{\sigma}_i^2 I| - \frac{n^2}{2}\log 2\pi. \end{aligned} \quad (10)$$

4. 実験・結果

4.1 データの処理

本研究では，「Dianping」より 154,970 件の訪日中国人観光客が投稿したクチコミ情報を収集し，その中から有用な情報を抽出して，データベースで保存した．また，併せてレビューの評価点数も収集した．

4.2 感情分類の評価実験

教師データを作成するため，無作為に 2500 件のクチコミの文章を選び，感情を含む文書に対して正解ラベルを付加した．正解ラベルの付与は研究室内のメンバーと協力して行った．作成した教師データに前処理を実行し，次にエントロピー理論を適用して，キーワードの抽出を行った．

教師データに対して，各単語のエントロピー値を計算した後，最適な α の値を求めるため 0.5 から 3.0 まで，0.25 の刻み幅で評価し，最大の F1 値を持つ素性を選択した．SVM を識別器として SVC を使用した，損失の大きさをどれくらい考慮するかを決めるパラメータCは 1.0 とするとき，分離超平面の誤差を最小化し，最適化プロセスに影響を与える．SVM での学習後，評価データに対して 5-Fold Cross Validation(K=5)を実施し，F1 値が最大となったαとα'に基づきポジティブキーワードリストとネガティブなキーワードリストを選択し

た．両方を組み合わせた素性を用いて最も高い精度で分類が可能なモデルを作成した．以下の表 1 は分類精度の結果を示す．

表 1　SVM の分類精度結果

	適合率	再現率	F1
Positive(α=2.0) Negative(α'=0.75)	0.86	0.94	0.90

4.3　統計的因果探索による関連性分析

どのような要素がユーザーの評価を影響するか原因を分析するため，SVM の分類により店ごとにクチコミ文書の中に頻出するキーワードの出現頻度とユーザーからの点数評価を組み合わせた調査を行った．ユーザーからの評価点数について，「味」，「サービス」，「雰囲気」，「一人当たり値段」の 4 種類のデータを取得した．その後，店ごとのコメント中にキーワードの出現頻度と評価点数の情報をベクトル化して，一般化スコア関数を用いて因果探索を行った．ポジティブな視点からの結果を以下の図 2 に示す．黒いノードは SVM の結果であり，赤いノードはユーザーがお店に付けた評価点数である．破線は交差検証尤度の結果を示し，実線は限界尤度の結果を表示する．赤い点線は両方の共通結果を示した．

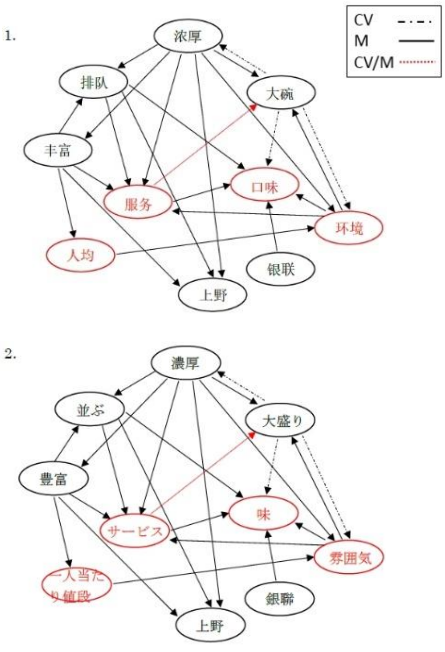

図 2　ポジティブな視点からの因果グラフ
1.実際の結果　2.日本語訳

以下の図 3 はネガティブな文書で頻出する単語とユーザー評価間の因果グラフを示す．上記の表示方法と同様である．

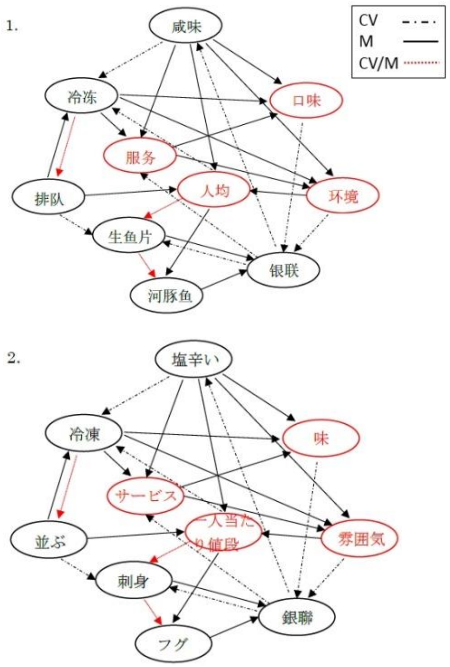

図 3　ネガティブな視点からの因果グラフ
1.実際の結果　2.日本語訳

5．考察

実際のレビュー内容を考察したところで，ウナギ屋やラーメン屋で濃厚な味を好んでいる中国人観光客が多い．調味ソースやラーメンの汁が好評になると考えられる．また中国人観光客は濃厚な味の店でサービス，雰囲気が良いと感じている．それに対して，塩辛いシーフードへの評価が低くなっていることを示した．実際のレビュー内容によって，具体的にはカニみそが好みではないことがわかった．フグの刺身などの生ものも好みではなく，冷凍食材を使うことはマイナスポイントになる．また近年，日本における中国の決済手段の導入によって中国人観光客の支払い満足度は向上したが，未だ不満を持つ人もいることが実際のレビュー文より考えられる．

6．おわりに

本研究では，中国人向けレストランポータルサイト「Dianping」から発信されている「食に関する消費動向」情報を着目し，エントロピー理論，SVM 及び統計的因果探索手法を用いて，訪日中国人観光客の味覚嗜好性に関するセンチメント分析を行った．センチメント分析にあたっては分類性能が高い（F1=0.90）手法を構築し，因果関係を探索するため汎化性能の高い一般化スコア関数を利用した．今後は得られた結果を踏まえて，多言語の

情報を利用し，全面的なユーザーの意見と意思決定の要因分析を行い，総合的な分析手法の開発を行う．

一方で，和食がユネスコ無形文化遺産に登録された際に，挙げられた特徴の一つとして「自然の美しさや季節の移ろいの表現」がある[12]．外国人観光客にとって季節ごとに違うものが食べられるという点は，日本の食の評価を高くしている一つのポイントだと考えている．そこで，季節ごとに食に関する情報を取得して時系列で分析する．この分析の実現によって訪日客は日本で食事に旬のものを見つけることが可能となり，季節ごとに再訪する観光客の増加が期待できる．

文献

[1] 2018 Visitor Arrivals & Japanese Overseas Travelers. Japan National Tourism Organization (JNTO),pp.2-4,2019. https://www.jnto.go.jp/jpn/statistics/data_info_listing/pdf/190116_monthly.pdf, (参照 2019-11-20).

[2] 鈴木 勝. 食文化を活用した国際ツーリズム振興. 大阪観光大学紀要,Vol.7, No.3, pp.16-18, 2007.

[3] Tonita Perea y Monsuwe, Benedict G.C. Dellaert, Ko de Ruyter.What drives consumers to shop online: A literature review. International Journal of Service Industry Management. International Journal of Service Industry Management, vol.15, No.1, pp.102-103, 2014.

[4] Dianping http://www.dianping.com/

[5] Ah Keng Kau, Pei Shan Lim. Clustering of Chinese Tourists to Singapore: An Analysis of Their Motivations, Values and Satisfaction. Int. J. Tourism Res. 7, 231–248, pp.231-233, 2005.

[6] Fangxuan (Sam)Li, Chris Ryan. Souvenir shopping experiences: A case study of Chinese tourists in North Korea. Tourism Management vol.64, pp.142-153, 2018.

[7] Katerina Berezina, Anil Bilgihan, Cihan Cobanoglu, Fevzi Okumus. Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. Journal of Hospitality Marketing & Management, vol.25, pp.1-24, 2016.

[8] Viriya Taecharungroj, Boonyanit Mathayomchan. Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. Tourism Management vol.75, pp.550–568, 2019.

[9] Jieba https://github.com/fxsjy/jieba

[10] Shannon E Claude, A Mathematical Theory of Communication. Bell System Technical Journal, vol.27, No.3, pp. 379-423, 1948.

[11] Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, Clark Glymour. Generalized Score Functions for Causal Discovery. Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 1551-1560, 2018.

[12] 無形文化遺産 | 文化庁 https://www.bunka.go.jp/seisaku/bunkazai/shokai/mukei_bunka_isan/