研究指導 中澤 真 教授

特徴語に着目したTwitter上の観光情報判別分析

堺 愛純

1. はじめに

近年，InstagramやTwitterなどを使用し有名な観光スポットや観光名物の情報を得てから，本格的に旅行の計画を立てるといった，SNSを旅行の情報収集源としている旅行者が増えている[1]．そのため，観光地域の事業者らはSNSを中心とした観光PRが必要不可欠である．SNSを中心とした観光PRをするにあたり，旅行者のニーズを把握することが重要となってくる．

しかし，「○○（観光地名） 観光」と検索しただけでは同業者の営利団体，事業者団体が発信したものや関係のないものなどがヒットしてしまい，本来の観光客のニーズを把握しづらい現状にある．これに対し，SNSの中でも速報性が高く，手軽に情報発信が可能なTwitterに着目し，ツイートの中から観光分析に有用なツイートのみを効率的に抽出するための研究が盛んにおこなわれている[2][3][4]．しかしツイートを収集する際に季節の偏りが見られたり，位置情報の指定により十分な観光ツイートが収集できていなかったりと問題点がある．

そこで本研究では，ツイートを十分に収集するために，位置情報を指定しない方法でツイートを取得する．また，ツイート抽出において季節に偏りがないよう月別に100件ずつ収集することとした．月別に収集したデータを4つの季節ごとで分析することで，観光情報を判別するための特徴語をこれまでよりも適切に抽出し，判別精度の向上を目指す．

2. Twitterを使用した観光情報分析

2.1 観光情報分析の現状

日本交通公社[1]のSNS・写真に対する意識調査によると，Z世代と呼ばれる1996-2003年に生まれた若い世代が旅行先の選択にSNSからの情報を最も重視しているという調査結果が示されている[1]．そのため，観光地域の事業者らはSNSを中心とした観光PRが必要になってくる．観光PRの手法として，まずSNSで観光客のニーズを把握し，そのニーズに合わせ事業者は情報発信していくことが必要である．

しかし，観光客のニーズを把握するためには，SNSで観光客が投稿したものを抽出しなければならないが，SNS上で一つ一つ目視しながら観光情報に関するツイートのみを，人手により多数抽出し分析することは難しい．そこで，近年SNSの中でも閲覧者数が多く，気軽に情報発信が可能なTwitterを用いてツイート文から観光情報のみを抽出する研究が行われている．Twitterで観光情報のみを抽出する方法としては，位置情報を使用し観光地で投稿されたツイートを分析する方法[2]や，投稿されたすべてのツイートを分類器にかけ観光ツイートだけを抽出して分析する方法[3]などが考案されている．

2.2 Twitterを使用した観光情報分析の先行研究

松本ら[2]は，下関地域を対象にTwitter API[1]を使用し，地図上の座標を用いてツイートを収集する地域を選択して，ツイートを取得している．さらに，観光客のツイートのみを抽出するために，投稿者のプロフィール欄に観光地名や観光地に関連するワードが見られた場合には，地元在住者によるツイートと判断してこれを除外した．これらのデータに基づいて抽出語リストや共起ネットワークを作成・分析しているが，指定範囲内の位置情報が記録されたツイートに限定して抽出しているため，観光ツイートを十分に得ることができなかった．

関谷ら[3]は，ツイート分類においてラベル付けの作業を除くために，すべてのツイートの中から観光情報が含まれるツイートだけを能動学習により抽出した．本来，ツイートから観光情報だけを抽出するというような機械学習における「分類」という作業は，膨大な量のツイートから人手により仕分ける必要がある．その作業を省いたものが能動学習である．学習アルゴリズムがユーザーや他の情報源に対話的に問い合わせることで，学習に有用なデータを優先して選択・生成しているため，正誤のラベル付きデータを使用しない点が特徴として挙げられる．

2.3 本研究の目的

松本ら[2]の問題点は，位置情報だけでツイートを取得しているため，GPS機能をオフにしている投稿者のツイートを抽出できないことである．また，観光客のツイートのみを抽出する際のフィルタリングが，プロフィール欄に対する観光地名などの記載の有無のみで取捨選択しているため，非観光ツイートが多く紛れ込んでしまっていることも問題の一つである．

[1] https://help.twitter.com/ja/rules-and-policies/twitter-api

関谷ら[3]は能動学習自体の問題点として，導き出された正解の根拠が分からないというブラックボックス問題のほか，抽出したデータが1月上旬～5月上旬までのものであるため季節に偏りがあること，分類器に入力する特徴語が名詞だけであることから，抽出されたデータには偏りが生じていることなどが挙げられる．

以上の問題点より，本研究では，位置情報を限定せずに収集を行った．また，関谷ら[3]の問題点を解決する手段として，1年を季節で区切って特徴語を抽出する手法を考えた．例えば，冬であれば「雪まつり」や「寒い」など，その季節特有の特徴語が抽出されるようになる可能性があり，季節をまとめて特徴語を抽出した分類器よりも精度がよくなると考えられる．そのため季節ごとに特徴語を見つけて入力することを本研究の新規性とする．

3.　観光情報の分類に用いる特徴語

まず，ツイート分類においての特徴語を検討する．ツイートの収集から特徴語抽出までの流れを図 1に示す．詳細は以下のセクションで説明する．

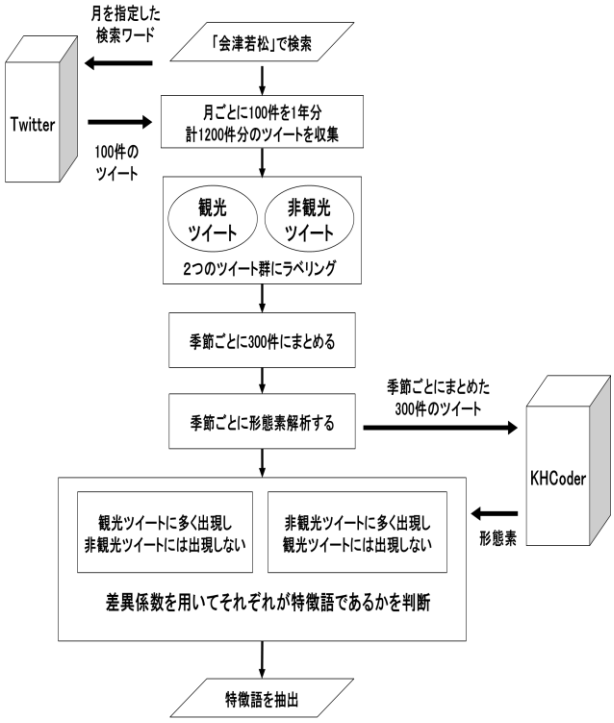

図 1　特徴語抽出の流れ

3.1 ツイートの収集とラベリング

今回の分析では観光都市である会津若松市を題材として検証する．まず，ツイート収集のための検索ワードは「会津若松」とし，ツールとしてGoogle Chromeの拡張機能である「ついすぽ[2]」を使用した．先行研究のような収集したツイートの季節の偏りを生じさせないため，ひと月ごとに100件ずつ1年分のツイートを収集した．また，リツイートやボットのツイートは非収集としている．理由としては，複数回にわたって同一のツイートを投稿している可能性があり，それによるデータの重複を防ぐためである．

また，観光ツイートと非観光ツイートのそれぞれの特徴を見つけるには，まず収集したツイートを2群に分ける必要がある．そこで本研究では，観光ツイートの判断基準として「観光地に関連する観光資源の名称を含むツイート[3]」と定義した．この際，観光資源[4]の名前のみを含むツイートや，ネガティブな内容を扱っているツイートも対象とした．定義にしたがって2群に分けた結果，ツイート1200件中302件が観光ツイート，898件が非観光ツイートとなった．

3.2 観光ツイートと非観光ツイートの特徴語

本研究では分析ツールとして，TMS[5]とKH Coder[5]を使用した．これらのツールを用いてラベリングした2つのツイート群を形態素解析した．2つのツイート群で出現傾向に差異がある語があれば，2つの群を判別するための特徴語となる可能性がある．よって，観光ツイートに多く出現し非観光ツイートにはあまり出現しない単語，また非観光ツイートに多く出現し観光ツイートにはあまり出現しない単語を抽出する．

人の目での選別はその尺度があいまいになってしまうため，2つの群の各単語の出現差異を測る評価尺度として差異係数[6]を使用する．また，特徴語の抽出条件を，差異係数の絶対値が0.7以上かつ2つの群どちらかの出現頻度が5以上の語とした．品詞は形容詞，形容動詞，動詞，名詞を対象品詞とした．

$$\text{差異係数} = \frac{\left(\frac{\text{非観光ツイート群の対象単語出現数}}{\text{非観光ツイート数}}\right) - \left(\frac{\text{観光ツイート群の対象単語出現数}}{\text{観光ツイート数}}\right)}{\left(\frac{\text{非観光ツイート群の対象単語出現数}}{\text{非観光ツイート数}}\right) + \left(\frac{\text{観光ツイート群の対象単語出現数}}{\text{観光ツイート数}}\right)}$$

季節ごとの特徴語を抽出するために収集した1200件を，3～5月，6～8月，9～11月，12～2月の4つの季節300件ずつに分割する．これらのデータを差異係数に基づき抽出した語の一覧を以下の表に示す．なお，語は差異係数の絶対値の降順で並べてある．

[2] https://tilde.afonomics.com/TweetExport/

[3] 地名の名前のみのもの・利益団体や事業者団体が独自で発信しているもの・画像でしか観光資源と判断できないものは含めない．

[4] 特定のエリア周辺でのみ催されるイベントや，飲食・購入可能なものを観光資源とする．

[5] https://jdream3.com/service/analysis-partner/tmstudio.html

表 1 3～5 月の特徴語

語	出現頻度（非観光）	出現頻度（観光）	差異係数
御朱印	0	7	-1.00
新型	18	0	1.00
多い	7	0	1.00
ラーメン	8	32	-0.83
コロナ	25	1	0.80
食べる	4	12	-0.78

表 2 6～8 月の特徴語

語	出現頻度（非観光）	出現頻度（観光）	差異係数
路線	28	0	1.00
鶴	1	12	-0.96
美味しい	1	10	-0.95
温泉	4	12	-0.85
買う	4	11	-0.84

表 3 9～11 月の特徴語

語	出現頻度（非観光）	出現頻度（観光）	差異係数
ネット	11	0	1.00
情報	9	0	1.00
配信	9	0	1.00
紅葉	1	6	-0.86
食べる	2	8	-0.80
駅前	4	10	-0.70

表 4 12～2 月の特徴語

語	出現頻度（非観光）	出現頻度（観光）	差異係数
温泉	0	7	-1.00
感染	16	0	1.00
確認	13	0	1.00
カツ	0	5	-1.00
暖かい	7	0	1.00

次に，季節別に抽出した特徴語と比較検証するために，季節ごとに分けない通年のツイートに対する特徴語の抽出を試みる．各季節の300件のツイートからランダムに75件ずつ選び出したツイートを統合し，季節を考慮しないツイート群として合計300件のツイートを収集した．この場合の特徴語と差異係数を表 5に示す．

表 5 季節を考慮しない場合の特徴語

語	出現頻度（非観光）	出現頻度（観光）	差異係数
コロナ	13	0	1.00
御朱印	0	7	-1.00
ラーメン	4	21	-0.87
食べる	2	8	-0.84
ネット	18	1	0.72
新型	17	1	0.70

表 1～表 4の季節ごとに分類した結果を見ると，観光ツイートに「鶴」や「温泉」，「紅葉」などが抽出されている．これらの特徴語は季節特有の可能性がある．また，観光ツイートに出現する単語の特徴として，いずれも観光資源が上位に出現していることが挙げられる．その原因として，ラベリングする際に「観光地に関連する観光資源の名称を含むツイート」としていることが考えられる．

それに対し，表 5の季節を考慮しない場合の結果を見ると，「ラーメン」や「御朱印」などが抽出されている．これは特徴語が季節特有のものではなく，年間を通じて入手可能なものであるためと考えられる．また，非観光ツイートに「新型」や「コロナ」，「ネット」などが抽出されており，これは新型コロナウイルス関連のネットニュースによるツイートが，1年を通じて投稿されていたことによると考えられる．

本研究では以上のような季節特有の特徴語を使用し，判別的中率を上げることを目指す．なお，季節を考慮しない場合の特徴語は1年を通して出現しているため，季節ごとの特徴語と重複する部分もあるが，このことについて特別な処理は施さない．

4. 季節を考慮した特徴語の検証と考察

4.1 季節を考慮した特徴語による判別の精度

先に述べた特徴語を用いて，観光情報か否かを判別する際の精度についてここでは述べる．まず，得られた観光ツイート群は非観光ツイート群と比較するとサンプルサイズが小さいことから，分析ではこれを揃えるために非観光ツイート群からランダム抽出して，観光ツイート群とサンプルサイズを揃えた．

分析の手法としては，SVM[7]と判別分析[8]を使用した．なお，この分析では特徴語の出現頻度を説明変数とし，観光ツイートか否かを目的変数としてい

る. 季節ごとの判別的中率の結果を表 6に示す.

表 6 通年および季節別の判別的中率

	季節の考慮なし(通年)	3～5月	6～8月	9～11月	12～2月
SVM	62.9%	63.7%	70.4%	54.3%	61.1%
判別分析	62.9%	63.1%	70.4%	54.4%	61.1%

結果として, 3～5月, 6～8月においては, SVM, 判別分析どちらの判別的中率も季節を考慮しない場合より数値を上回ることができたが, 9～11月, 12～2月においては上回ることができなかった.

4.2 部分的な精度改善に対する考察

部分的な精度改善にとどまったことについて, 2つの原因が考えられる. 1つ目の原因は, 収集したツイート数が不足していたことである. 本研究では1200件ほどツイートを収集したが, 1季節につき300件のツイートではサンプルサイズ的に, 適切な特徴語が現れにくい可能性がある. そのため, 特徴語を抽出するためのツイート数をひと月あたり1000件以上収集することで, 判別に有効な特徴語が出現するようになるのではないかと思われる.

2つ目の原因として, 祭事などのイベントは特定の季節に催されることが多く, 季節によってツイートを収集した地域の観光資源の量が変動したのではないかと考えられる. 観光資源の量が乏しければ, 特徴語の出現頻度も減少するため, 収集したツイートに含まれる特徴語の量に季節ごとの差があったのではないかと思われる.

5. むすび

本研究では, Twitterから観光ツイート, 非観光ツイートの収集を行い, 季節ごとに特徴語を抽出することで二つのツイート群の判別的中率の精度向上を目指した. その結果, 3～5月, 6～8月の季節では, 季節別にしない通年での判別的中率を上回る結果となった. しかし, その他の9～11月, 12～2月では, 上回ることができなかった. 収集したツイート数が1200件であったが, 季節ごとに分割すると300件ずつと少なく, これにより適切な特徴語が出現しにくい状況になってしまったことが原因として考えられる.

今後の課題の一つは, 十分な数の季節ごとの特徴語を抽出して判別精度を高めることである. このために, ひと月あたり1000件以上のツイートを収集して新たな分析に取り組む必要がある. また今回は, 検索キーワードを「会津若松」としたため, 検索キーワードを他地域などに変更した場合についても検証をしていく.

参考文献

[1] 公益財団法人, 日本交通公社, 内旅行におけるSNS・写真に対する意識／実態, https://www.jtb.or.jp/research/statistics-tourist-sns-pictures2022/, (参照 2023-02-07).
[2] 松本ほか, “下関地域における Twitter を利用した観光情報分析”, バイオメディカル・ファジィ・システム学会誌, vol.22, No.2, pp.59-66, 2020.
[3] 関谷ほか,“Tweet の分類による観光情報の取得”, 第 19 回情報科学技術フォーラム, D-012, pp105-106, 2020.
[4] 小原ほか, “Twitter 本文を用いた観光情報抽出及び分析システムの構築”, 第 29 回人工知能学会全国大会論文集, pp.1-3, 2015.
[5] 末吉美喜, テキストマイニング入門:Excel と KH Coder で分かるデータ分析, オーム社, 2019.
[6] 小林雄一郎, ことばのデータサイエンス, 朝倉書店, 2020.
[7] 加藤公一 監修, 秋庭伸也, 杉山阿聖, 寺田学, 見て試して分かる機械学習アルゴリズムの仕組み機械学習図鑑, 翔泳社, 2019.
[8] 栗原伸一, 入門統計学-検定から多変量解析・実験計画法まで-, オーム社, 2014.