

ジオタグ付きツイートをを用いた観光情報の推定

Estimation of tourist information using geo-tagged tweets

藤本 雅大[†] 佐藤 淳哉[†] 高森 勇佑[†] 大野 成義[†] 遠藤 雅樹[†] 石川 博[‡]

Masahiro Fujimoto Junya Sato Yusuke Takamori Ohno Sigeyoshi Masaki Endo Hiroshi Ishikawa

1. はじめに

近年、SNS を利用した観光情報収集は一般的な手法となっており、SNS に投稿されるビッグデータの分析は観光客に有益な情報をリアルタイムに提供できる可能性がある。SNS として代表的な Twitter は、ユーザによる投稿が容易なためリアルタイムでのコミュニケーションツールとしても活用でき、一般的に身の回りの出来事や趣味等の情報を投稿・閲覧するために利用される。Twitter に投稿されるテキストや画像等のことを「ツイート」と呼ぶ。Twitter ではユーザがツイートごとに現在の位置情報の公開の有無を指定でき、位置情報を公開するツイートのことを「ジオタグ付きツイート」と呼ぶ。Twitter では、ツイートの投稿によって観光情報に関する発信も広く行われており、近年、観光を予定している人が行き先を決める指標としても活用されている。これより、観光客がリアルタイムで地域の観光情報を取得及び推定するためのソーシャルセンサーとして役立つことが期待されている。現在、Twitter を用いた研究が多く行われている。SNS の個人利用状況は、総務省の令和 2 年通信利用動向調査[1]によると、6 歳以上では 73.8% であり、令和元年の 69.0% と比べると利用が拡大していることがわかる。SNS の一つである Twitter は、リアルタイムなコミュニケーションツールとして活用できることから日本国内での利用者も多く日々大量の情報発信が行われている。

本稿では、これまで我々が Twitter から収集したデータから紅葉に関するツイートを抽出し分析することにより紅葉の見頃推定を行い、気象庁の生物季節観測の情報と比較することで本手法の有効性を実験結果と共に報告する。本稿の構成は以下のとおりである。2 章では本稿に関連する研究について記述する。3 章では提案手法を示す。4 章では実験結果とその評価結果を報告する。5 章では本稿のまとめを記述する。

2. 関連研究

SNS などの普及に伴い今後も位置情報や画像、文字列など、様々な情報は飛躍的に増大することが予想され、この大量のデジタルデータの有効活用に関連した研究が数多く行われている。以下に、SNS や時系列予測に関する研究について述べる。

鳥海ら[2]はソーシャルメディア上における、新型コロナウイルスに関する投稿を収集して、情報を発信するユーザに変化とその感情に着目し分析を行った。データセットとして Twitter Search API で取得した新型コロナウイルスに関するツイートデータを分析に用いることで日本の新型コロナウイルスのツイートの傾向を確認した。

遠藤ら[3]はマイクロブログを用いて紅葉シーズンの国内の季節変化の推移を追跡した。データセットとして日本国内における位置情報付きツイートを利用し、対象語が含まれるツイートのトレンド推移を確認した。生物季節の対象語に関する気象庁の観測日とツイート数のピークが同等に観測可能であり、対象語に関するツイートの盛り上がりを実世界での季節変化の関連を明らかにした。その結果、生物に関連するツイートの観測から見頃(ピーク期)を推定できる可能性を示した。

Horikawa ら[4]は機械学習を用いて一定期間先の桜の見頃推定を行った。データセットとして 2015 年 2 月 1 日から蓄積してきた桜に関するツイート情報を学習データとして、機械学習を用いた日付に対するツイート数の時系列予測を実施した。SNS データによる時系列予測手法は従来手法と比べ、再現率の向上を確認した。

Takamori ら[5]は、東京都のオープンデータとして公開されているデータを元に、新規陽性者数の予測を行っている。機械学習による時系列予測を利用し、データセットとして新型コロナウイルス新規陽性者数や関連性のある変数を含む時系列データを用いることで、東京都の新規陽性者数の予測手法として利用できる可能性を示している。

小原ら[6]は、Twitter 本文から地域連想語やパターンマッチングを用いて観光情報の抽出やユーザの居住地の把握を行っている。取得した地域連想語の割合や観光情報をツイートしたユーザ数の推移のグラフ化し、ツイートされた場所や推定したユーザの居住地の分布を表示するシステムの構築を行っている。パターンマッチングを用いた観光情報抽出手法は 80.2% の適合率を達成した。

以上、SNS や時系列分析に関する研究が行われている。紅葉の国内の季節変化の推移や桜の見頃推定の研究は報告されているが、紅葉の見頃推定に関する研究は報告されていない。そこで本研究では、時系列予測を用いた一定期間先の紅葉の見頃推定を行う。

3. 提案手法

本章では、データの収集の手法について説明する。

3.1 節に見頃推定に利用したデータセット、3.2 節に分析方法、3.3 節に本提案手法の推定条件を示した。

3.1 データセット

Twitter の投稿の中で位置情報の付与されたジオタグ付きツイートを 2015 年 3 月 1 日から 2018 年 12 月 31 日と 2020 年 1 月 1 日から 2020 年 12 月 31 日の期間を対象とした。

対象としたジオタグ付きツイートの中に対象語が記述されたツイートを抽出し分析を行った。まず対象語は気象庁において生物季節観測[7]が行われている「かえで」と「いちよう」、また我々は一般に秋の観光対象として用いられる表現である「こうよう」も対象語とした。更に対象語として、「コウヨウ」「紅葉」「黄葉」「カエデ」「楓」

[†] 職業能力開発総合大学校 Polytechnic University

[‡] 東京都立大学 Tokyo Metropolitan University

「イチョウ」「銀杏」を追加し、対象語の漢字・ひらがな・カタカナ表記を含むジオタグ付きツイート出現頻度と日付別推移を収集した。本稿において対象語を含むツイートを「紅葉」と記述する。

3.2 分析方法

我々が収集したツイートデータから紅葉の見頃推定を行う期間に対してツイート数の時系列予測を行う。

日本国内では「かえで」や「もみじ」は主に10月後半から12月前半までが紅葉日となるため、対象語の見頃推定期間は2021年10月16日から12月14日までの2か月間とする。

本手法の学習データは毎年秋にピークとなる対象語に関するツイートを収集しているため、学習データは季節性がある。よって、AWS(Amazon Web Services)のAmazon Forecast サービスから利用できる時系列予測アルゴリズムの中で、季節性のある学習データに適する Prophet を用いて見頃推定を行った。Prophet は、非線形傾向が年、週、及び日常の季節性に適合する加法モデルに基づく時系列予測アルゴリズムであり、強い季節効果と数シーズンの履歴データを持つ時系列において最も効果的である。

予測値は、加重分位損失で得られた50%の需要を満たす値を用いる。加重分位損失とは予測時のメトリクス的一种である。予測されたツイート数は小数点以下を切り捨てとし、ツイート数がマイナスの場合は0として扱う。機械学習によって予測した一定期間のツイート件数の推移により、対象語の見頃推定を行う。ツイート数の時系列予測に使用する学習データは2015年3月1日から2020年12月31日までのツイート情報を使用する。我々はシステムの都合により、2019年におけるツイート情報を取得できていないため、予測は2015年3月1日から2018年12月31日と2020年1月1日から2020年12月31日の期間におけるツイート情報を用いて行う。ツイート数の時系列予測を行う期間は2020年10月16日から12月14日の2か月間である。

3.3 推定条件

ツイートの推移による紅葉の見頃推定条件は、加重移動平均を用いる。加重移動平均とは、各値に重み付けした移動平均である。本研究において使用する手法では加重移動平均に用いる重みの中央値を1とし、中央値から±0.5の値をそれぞれ最小値及び最大値にする。

紅葉の見頃推定を行うにあたり、キーワードを含むジオタグ付きツイート出現頻度の分析を行う。結果を日付別に分析して、最適な視聴期間を推定する。対象語に関するツイートが増加した期間を、1年単純移動平均を求めた。加えて、気象庁は「いちょう」と「かえで」の紅葉日から落葉日までの平均日数を10日としている。そこで、10日加重移動平均 H_{avg10} を用いた。以下に使用する数式を示す。 x_y は、 y 日目のツイート数 x を示す。

$$H_{avg10} = (x_{10} \times 0.5 + x_9 \times 0.61 + x_8 \times 0.72 + x_7 \times 0.83 + x_6 \times 0.94 + x_5 \times 1.05 + x_4 \times 1.16 + x_3 \times 1.27 + x_2 \times 1.38 + x_1 \times 1.5) / 10$$
 (1)

更に曜日によるツイート数の差の影響を少なくするため、7日加重移動平均を求めた。以下に使用する数式を示す。 x_y は、 y 日目のツイート数 x を示す。

$$H_{avg7} = (x_7 \times 0.5 + x_6 \times 0.67 + x_5 \times 0.83 + x_4 \times 1 + x_3 \times 0.17 + x_2 \times 1.33 + x_1 \times 1.5) / 7$$
 (2)

これらの推定基準を用いて、次の2つの条件を満たした場合に、紅葉の見頃と判定することとした。

- 1: ツイート数が1年単純移動平均よりも多い
- 2: 3日連続でツイート数の7日加重移動平均が10日加重移動平均よりも多い

これらの最適な時間推定基準を使用して、各観光スポットを見るのに最適な期間を推定した。

ここで、気象庁が判定する「かえで」や「いちょう」の紅葉観測条件を述べる。「かえでの紅葉日」[8]とは、標本木全体を眺めたときに、大部分の葉の色が紅色に変わった最初の日を指す。また、「いちょうの黄葉日」[9]とは、標本木全体を眺めたときに、大部分の葉が黄色に変わった最初の日を指す。「かえで」と「いちょう」のそれぞれの落葉日[10]は、標本木の約80%が落葉したと見えた日となる。風の有無は問わない。

本稿では「かえで」と「いちょう」の紅葉日から落葉日までを紅葉の見頃であるとし、気象庁が観測している紅葉日から落葉日の実測期間と我々が見頃と推定した期間を比較することで本手法の精度を評価する。

4. 実験結果

本章では、提案手法による対象語の分析の判定結果を示す。まず、時系列予測で得たツイート数から紅葉の見頃推定を行った結果を示す。2020年の京都府、大分県、栃木県における紅葉の見頃結果を図1、図2、図3に示す。それぞれの図の黒点線は1年移動平均、黒実線は10日移動平均、黒破線は7日移動平均、集合縦棒は時系列予測を行ったツイート数、灰色部分は見頃推定期間を示す。

ピーク予測の精度は、見頃推定結果と実測期間を比較し、再現率と適合率の値で評価する。実測期間は表1に示す通り気象庁が観測した紅葉日から落葉日までの期間とする。

本手法の予測したツイート数を用いて行った紅葉の見頃推定と実測期間の再現率と適合率を表2に示す。

図1の京都府は11月5日から11月30日を見頃と推定し、気象庁が観測した紅葉日から落葉日までの11月20日から11月30日と近似した見頃推定ができた。京都府における本手法での見頃推定再現率は100%、適合率は42.3%である。図2の大分県は11月9日から11月28日を見頃と推定し、気象庁が観測した紅葉日から落葉日の11月27日から12月11日と近似した見頃推定ができた。

表1 気象庁 紅葉日・落葉日(2020年)

| 対象地域 | 紅葉日 | 落葉日 |
|------|--------|--------|
| 京都 | 11月20日 | 11月30日 |
| 大分 | 11月27日 | 12月11日 |
| 栃木 | 12月7日 | 12月14日 |

表2 予測ツイート数の見頃推定手法の再現率と適合率

| 対象地域 | 再現率 | 適合率 |
|------|-------|-------|
| 京都 | 100% | 42.3% |
| 大分 | 13.3% | 10.0% |
| 栃木 | 0.0% | 0.0% |

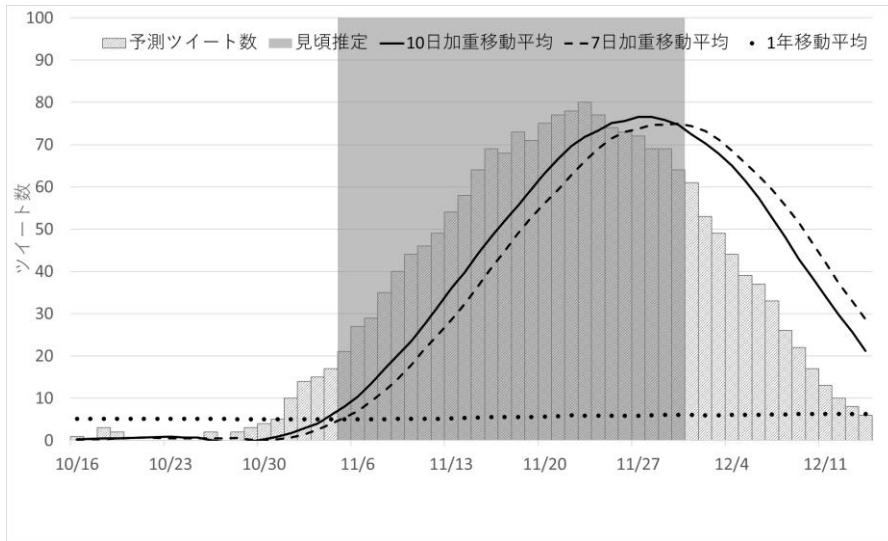


図 1 京都府「紅葉」時系列予測結果

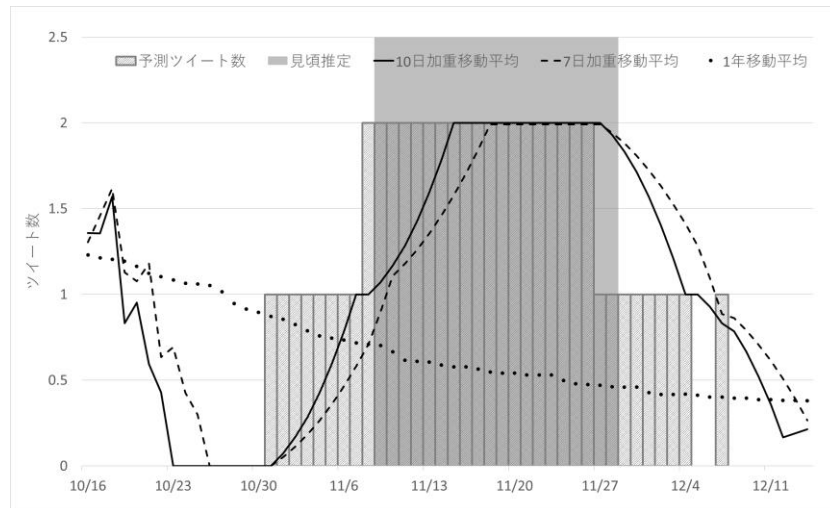


図 2 大分県「紅葉」時系列予測結果

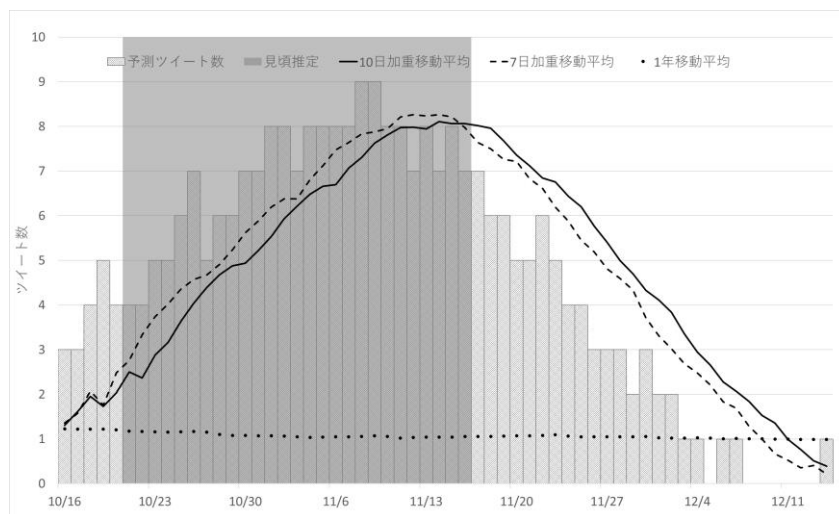


図 3 栃木県「紅葉」時系列予測結果

大分県における本手法での見頃推定再現率は 13.3%、適合率は 10%である。図 3 の栃木は 10 月 21 日から 11 月 16 日を見頃と推定し、気象庁が観測した紅葉日から落葉日の 12 月 7 日から 12 月 14 日と差異のある見頃推定となった。また十分な位置情報付きツイート量がないため対象語のみでの推定は困難であるという結果になった。栃木県における本手法での見頃推定再現率は 0%、適合率は 0%である。原因は気象台と観光名所の標高差による気温の違いが紅葉時期のずれとなったと考えられる。このため標本木のある気象台と観光名所の気温差が激しい地域では見頃に大きな差があると考えられる。

また京都府、大分県、栃木県のそれぞれの観光名所の見頃を確認した。京都観光オフィシャルサイト[12]によると 11 月 6 日から嵐山や常照皇寺などが見どころとなっており、我々が推定した時期とほとんど相違ないことがわかる。大分の中津邪馬溪環境協会[13]によると 11 月 1 日より見頃となっており、これも推定時期と相違ないことがわかる。さらに栃木県の日光観光ライブ情報局[14]によると日付は明記されていないが、各地域の見頃時期は 10 月 20 日前後に最も多いことがわかり、我々が推定した時期と同等である。このことから観光客が観光時期を決める際は本手法で推定できる可能性があることを確認した。

5.おわりに

本稿では、これまで我々が Twitter から収集したデータから紅葉に関するツイートの分析により、見頃推定を行った。気象庁の生物季節観測の結果と比較することで本手法の有効性を実験結果と共に報告した。その結果、京都府においては時系列予測による再現率と適合率は高く、紅葉に関するツイートの推移と現実世界での季節変化には関連があり、生物に関連するツイートの観測を行うことで見頃を推定可能であることを確認した。しかし大分県と栃木県に関しては標本木と紅葉名所の標高差が大きいことから高い再現率と適合率を得られることはできなかった。また、本稿で対象としたジオタグ付きツイートは、ツイート全体の 3%程度でありデータ量が少ない。このため、本稿では、地域ごとに十分なジオタグ付きツイートが集まらなかったことにより、提案手法を用いてジオタグ付きツイートの分析することで観光スポット別での推定は可能であることを確認したに留まった。今後は対象語の幅を広げることでデータ量を確保し、より高精度の予測を実現する方向で研究を進めていく予定である。

謝辞

本研究は、JSPS 科研費 20K12081、JSPS 科研費 22K13776、大川情報通信基金の研究助成により遂行されたものです。この場を借りて深く御礼申し上げます。

参考文献

- [1] “総務省の令和 2 年通信利用動向調査”, https://www.soumu.go.jp/main_content/000756018.pdf (2022 年 6 月 23 日閲覧)
- [2] 鳥海不二夫, 榊剛史, 吉田光男, “ソーシャルメディアを用いた新型コロナウイルス禍における感情変化の分析”, 人工知能学会論文誌, Vol.35, p. F-K45_1-7 (2020).
- [3] 遠藤雅樹, 廣田雅春, 倉田陽平, 横山昌平, 大野成義, “生物季節観測のための位置情報付きマイクロブログの分析手法の検討”, 情報科学技術フォーラム講演論文集, Vol.14 No.97-102 (2015).
- [4] T. Horikawa, M. Takahashi, M. Endo, S. Ohno, M. Hirota, and H. Ishikawa, “Estimating the Best Time to View Cherry Blossoms Using Time-Series Forecasting Method”, Vol.4, No.418-431 (2022).
- [5] Y. Takamori, R. Watanabe, D. Kato, M. Endo, and M. Ishikawa, “Numerical estimation of new COVID-19 positive cases using time series analysis by machine learning”, Proceedings of the 13th International Conference on Management of Digital Eco Systems, pp. 153-159 (2021).
- [6] 小原基季, 森田和宏, 泓田正雄, 青江順一, “Twitter 本文を用いた観光情報抽出及び分析システムの構築”, 人工知能学会全国大会論文集 第 29 回 p. 4M14-4M14 (2015).
- [7] “気象庁-生物季節観測の情報”, <https://www.data.jma.go.jp/sakura/data/index.html> (2022 年 6 月 23 日閲覧)
- [8] “気象庁-かえでの紅葉日”, <https://www.data.jma.go.jp/sakura/data/kaede2010.pdf> (2022 年 6 月 23 日閲覧)
- [9] “気象庁-いちょうの黄葉日”, <https://www.data.jma.go.jp/sakura/data/a/ichou2010.pdf> (2022 年 6 月 23 日閲覧)
- [10] “生物季節観測指針”, <https://www.data.jma.go.jp/sakura/data/kyu-shishin.pdf> (2022 年 6 月 23 日閲覧)
- [11] “日本植物生理学会”, https://jspp.org/hiroba/q_and_a/detail.html?id=1344 (2022 年 6 月 23 日閲覧)
- [12] “京都観光オフィシャルサイト”, <https://ja.kyoto.travel/doc/pdf/momiji2020.pdf> (2022 年 6 月 23 日閲覧)
- [13] “中津邪馬溪環境協会”, https://nakatsuyaba.com/?introduce=nakatsuyabakei_kouyou_2020 (2022 年 6 月 23 日閲覧)
- [14] “日光観光ライブ情報局”, <http://www.nikko-kankou.org/users/files/migoro.pdf> (2022 年 6 月 23 日閲覧)