人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第22回)
SIG-AM-22-01

Twitterのバースト情報に基づく桜の見頃推定

Fine-grained Best Season Estimation for Cherry-blossom Viewing based on Burst Information on Twitter

下園良太[1*]　　乾孝司[1]
Ryota Shimozono[1]　Takashi Inui[1]

[1] 筑波大学大学院
[1] University of Tsukuba

Abstract: SNS services such as Twitter and Facebook have been recognized as useful tools for sightseeing. Tourists usually access to such services to collect local and timely information about their interesting areas. In this paper, we study an automatic estimation method of the best season for cherry-blossom viewing. We applied Kleinberg's burst detection algorithm to tweet data stream and compared with one of the previous estimation methods, the moving average method. From the experimental results, it suggests that the burst-based method can resolve the over-extraction problem with keeping sufficient precision.

1　はじめに

近年，SNS の普及に伴い電子コンテンツが急速に増加している．総務省の日本国内における 1500 人を対象にした調査によると，主要 SNS の利用者は 2012 年の 41.4 %から，2016 年には 71.2 %にまで増加しており，スマートフォンと合わせて SNS の利用が社会に定着してきたことが分かる [1]．これに伴い，SNS で扱われる動画や画像，テキストなど様々な形式のデータが増加している．

SNS として代表的な Twitter[2] ではリアルタイムなコミュニケーションを行うことが出来る．利用者は日常で思ったこと，体験したことをすぐに投稿することができ，利用者はリアルタイムに起こっている「今」について知ることが出来る．

さらに，近年では観光においても SNS などの ICT が活用されている．国土交通省の日本国内における 520 人を対象にした調査では観光者の 9 割以上の人が旅行を計画する際，情報通信機器を利用していることが分かった [3]．その際，観光情報を収集するためにウェブサイトが活用されたり，旅行体験を知るために SNS が利用されている．

観光情報の中でも観光資源の適時性は観光者にとって重要な情報となる．例えば桜や紅葉といった花の見頃や，カニやサンマなど魚介類の食べごろなど，観光資源の「旬」の情報を入手することが出来れば観光計画に役立てることができ，観光をより良いものにすることができる．

現状，旅行ガイドブックにも旬な情報が載っているが，1 ヶ月単位の情報（タラバガニの食べごろは 2 月〜4 月など）となっており旬期間の粒度が粗い問題がある．SNS にはより粒度の細かいリアルタイムな旬の情報があるが，投稿単体では情報不足であったり，過去データからの変遷がみえないなど，情報がまとまっておらず検索などで旬の情報を直接獲得することが難しくなっている．

本研究では観光資源の旬として桜の見頃に着目し，Twitter を用いた日単位の桜の見頃推定手法を検討する．Twitter の位置情報付きの投稿（ツイート）を活用することで各地域における桜の見頃を推定することを目指す．先行研究に移動平均による桜の見頃推定の手法を検討したものがあるが，以下のような課題がある [5]．

- 見頃推定期間が細かく分割される
- 生活周期に合わせて注目度が変化する
- キーワードのみで使用データを選択するとノイズが含まれる
- データ不足

このうち，本研究では「見頃推定期間が細かく分割される」という課題に着目した．これは見頃推定期間がまとまらずに細かくいくつも推定されてしまう問題である．

*連絡先：筑波大学大学院システム情報工学研究科 コンピュータサイエンス専攻
〒 305-0006 茨城県つくば市天王台 1 丁目 1 − 1
E-mail: shimozono@mibel.cs.tsukuba.ac.jp

例えば 3/27 - 4/4 が一つの見頃として推定されるとまとまりがあるが，3/27 - 3/29，3/31 - 4/4 として推定されると見頃推定期間が二つになってしまう．先行研究では見頃推定期間の中から桜に関する事前知識（開花日から満開日となる日数が約 5 日間である．など）を使用して人が直接判断して見頃期間を定めている．

ある期間において，ある事象が急激に増加する現象として「バースト」がある．これは，観光資源がある期間中に見頃となる現象に似ている．本研究では，Twitter のバースト期間を見頃期間として考えて，バースト検知手法を適用しその有効性を検証する．バースト検知手法としては Kleinberg のバースト [4] を用いる．これはバースト状態と平常状態の移り変わりが簡単に行われないため，まとまったバースト期間が現れる．これにより「見頃推定期間が細かく分割される」問題が解決できることを期待する．

本稿ではバースト検知手法を用いて桜の見頃推定を行った結果を報告する．2 章では関連研究，3 章では今回使用したバースト検知手法の説明，4 章では見頃推定とバースト検知の対応づけについて説明を行い，5 章で実験の方法と結果を報告する．最後に，6 章でまとめと今後について述べる．

2 関連研究

本研究の先行研究として，遠藤ら [5] の研究がある．遠藤らは桜や紅葉の見頃推定手法として，各地域での位置情報付きツイートの出現数に対する移動平均を活用した．桜の場合，対象語となる「さくら，桜，サクラ」のツイートの出現数を数え，その出現数の移動平均を一年，7 日間 (一週間)，5 日間 (生物の特性) の三つを定義し，その移動平均の比較を行い見頃推定を行う手法である．前の章で述べたように，この手法には課題点がいくつか存在する．

Kleinberg は，テキストの時系列データである document stream においてバーストを検知できる手法を示した [4]．Kleinberg は，2 種類のバースト解析手法を提案しており，1 つは連続時間で送られるドキュメントデータに対して，バーストしている期間の判定や，バーストの強さ毎のバースト期間の判定ができる．2 つ目は，Enumerating バーストとよばれ，離散時間で送られるドキュメントデータを一つのバッチとして考え，着目するキーワードがバーストしているか否かの判定を行うことができる．

3 Kleinberg のバースト検知

本研究では見頃推定に Kleinberg が提案するバースト検知手法を用いる．これはテキストの時系列データである document stream 内で急激にテキストの発生頻度が上昇しているようなバーストと呼ばれる期間を検知するアルゴリズムである．Kleinberg のバーストには 2 種類あり，時間軸に沿って断続的に発生する document stream の時刻を元にバースト検知を行うもの（連続型）と，単位時間毎に発生した document stream 内の関連文書を数える Enumerating バースト（列挙型）が定義されている．本研究では 1 日単位のテキスト発生に対する見頃推定を行うため，Enumerating バーストを使用する．

Enumerating バーストは，離散時間で送られる文書の集合に対して適用される．本稿では，各日ごとのツイート集合を一つの文書集合の単位とし，以下では単にツイート集合と呼ぶ．2 状態オートマトン $\mathcal{A}^2$ を定義し，2 つの状態を非バースト状態 q_0，バースト状態 q_1 とおく．入力に対して状態が遷移することにより 2 つの状態を切り替える．着目するツイートを「関連ツイート」，そうでないツイートを「非関連ツイート」とし，バーストか否かはツイート集合中の関連ツイートの割合によって決まる．

解析期間において，n 個のツイート集合 $B_1, ..., B_n$ が離散時間で送られてくる状況を考える．t 番目のツイート集合を B_t とし，そのツイート集合に含まれるツイートの数を d_t とおく．さらに，ツイート集合 B_t に含まれる関連ツイートの数を r_t とおく．解析期間における全てのツイートの数 D を $D = \sum_{t=1}^{n} d_t$，全ての関連ツイートの数を $R = \sum_{t=1}^{n} r_t$ と表す．

次に，オートマトンの 2 状態にそれぞれ期待値を割り当てる．初期状態である非バースト状態 q_0 には，解析期間全体から算出した期待値 $p_0 = R/D$ を割り当てる．バースト状態 q_1 には，p_0 にパラメータ s をかけた値である $p_1 = sp_0$ を割り当てる．ただし，$s > 1$ であり，$p_1 \leqq 1$ となるような s でなくてはならない．s はバーストの見なされやすさを表している値であり，この値が小さいほど，ツイート集合中の関連ツイートの割合が低くてもバーストと見なされやすくなる．バースト性の解析は，n 個のツイート集合が与えられたときの状態の系列を通るためのコスト計算によって行う．考えられる状態の系列のうち，最も系列のコストが小さいものが解となり，その系列の状態に応じて，バースト期間と非バースト期間が決定される．状態遷移は d_t と r_t によって定まる．状態の系列は $\boldsymbol{q} = (q_{i1}, ..., q_{in})$ と表され，q_{in} は，n 番目のツイート集合によって決定された状態 $q_i (i = 0, 1)$ である．ツイート集合中の関連ツイートが二項分布 $B(d_t, p_i)$ に従って現れるという考えに基づき，状態 q_i にいることに対してコストを与える関数 $\sigma\ (i, r_t, d_t)$ を以下のように定義する．

$$\sigma\ (i, r_t, d_t) = -\ln[\binom{d_t}{r_t} p_i^{r_t} (1 - p_i)^{d_t - r_t}] \qquad (1)$$

ただし，頻繁に状態遷移が起こると，バースト状態と非バースト状態が頻繁に切り替わることになる．そこで，現在の状態 q_iから次の状態 $q_j(j = 0, 1)$ へ，状態遷移を妨げるためのコスト関数 $\tau\ (i, j)$ を定義する．

$$\tau\ (i,j) = \begin{cases} \gamma\ \ln n & (j > i) \\ 0 & (j \leqq i) \end{cases} \tag{2}$$

τ は，パラメータ γ によって調節されるが，特に理由がない場合は $\gamma\ = 1$ とする．以上に述べた，ある状態 q にいることに対してコストを与える関数 σ と，状態遷移のコスト関数 τ を使って，状態の系列 $\boldsymbol{q}$ を通るためのコスト関数は以下となる．

$$c(\boldsymbol{q}|r_t, d_t) = \sum_{t=0}^{n-1} \tau\ (i_t, i_{t+1}) + \sum_{t=1}^{n} \sigma\ (i_t, r_t, d_t) \tag{3}$$

オートマトン $\mathcal{A}^2$ は二つのパラメータ s，γ によって決まることから，${\mathcal{A}^2}_{s,\ \gamma}$ と表記される．一般的に $s = 2$，$\gamma\ = 1$ と設定するので，本実験ではこれに従い ${\mathcal{A}^2}_{2,1}$ のオートマトンを用いる．

4 見頃推定とバースト検知の対応づけ

本研究では，桜に関するツイートに対して 1 日単位でバースト検知を行なう．そして，バースト検知で検出されたバースト期間を，桜の見頃推定期間とみなす．

5 実験

5.1 実験概要

本実験では見頃推定にバースト検知手法を位置情報付きツイートに適用し，その効果と先行研究の課題の一つである「見頃推定期間が細かく分割される」を改善できるかを確認するため先行研究の手法との比較実験を行った．

気象庁の観測官署において観測される春の植物であり，観光資源の一つである桜を対象とした．実験条件は先行研究と同じ条件として，実験対象地域は「東京都」・「石川県」・「北海道」とし，実際の見頃期間は気象庁の観測データ [6] による桜の開花日から満開日までの期間とした．

文字列「桜」・「さくら」・「サクラ」を含むツイートを関連ツイートとし，地域ごとにバースト検知を行った．実際の見頃期間に対して，先行研究 [5] による移動平均を使った見頃推定期間と Kleinberg のバーストによる見頃推定期間を比較した．

5.2 データセット

データセットは，先行研究で行われたものと同じ条件にしたいため，同じ期間における同じ位置の位置情報付きツイートを使用した．2015 年 2 月 17 日から 12 月 31 日までの期間で，日本の領土を含む範囲である緯度経度が 10 進法表記で 120.0 ≦ 経度 ≦ 154.0 かつ 20.0 ≦ 緯度 ≦ 47.0 の位置情報付きデータとする．データ数は，約 5,000 万件であった．今回対象地域の東京都では約 850 万件，石川県では約 34 万件，北海道では約 190 万件となった．

5.3 実験結果

見頃推定の実験結果を図 1 から図 3 に示す．桜の開花時期の 2 月から 5 月の間における，気象庁の開花日と満開日の観測データ，先行研究の移動平均手法 (AVE, Average）の見頃推定期間と Kleinberg のバースト検知手法（Burst）の見頃推定期間をまとめた．また，正解期間である，開花日から満開日の期間を，図中の黒い横線で表している．

図 1 の東京都の見頃推定結果では，実際の見頃 3/23 - 3/29 に対して，AVE では 3/24 - 4/4 を含む 6 つの見頃期間が推定され，Burst では 3/23 - 4/6 の 1 つの見頃期間が推定された．図 2 の石川県の見頃推定結果では，実際の見頃 3/31 - 4/4 に対して，AVE では 4/1 - 4/9 を含む 6 つの見頃期間が推定され，Burst では 4/1 - 4/14 の 1 つの見頃期間が推定された．図 3 の北海道の見頃推定結果では，実際の見頃 4/22 - 4/26 に対して，AVE では 4/22 - 5/1 を含む 8 つの見頃期間が推定され，Burst では 4/23 - 5/11 を含む 2 つの見頃期間が推定された．

5.4 考察

AVE と Burst を比較すると，どの地域においても AVE による推定期間数よりも Burst による推定期間数が少なくなっており，大きく削減された．これは，Kleinberg のバースト検知の際の，状態遷移コストによるバースト期間の分割を抑える効果のためではないかと考える．推定期間の数や推定期間の長さを考えていく上で，パラメータ s と γ の値を設定することが重要である．パラメータ s と γ はバーストの検出具合に関わるパラメータである．s は状態 q_iであることに対してのコストの調整を行い，γ は状態遷移のコストの調整を行うものである．今後最適な設定をもとめる実験をしていく必要がある．

見頃推定期間の中から最終的に人が見頃期間を定める先行研究に対して，Burst では見頃推定期間の数が大

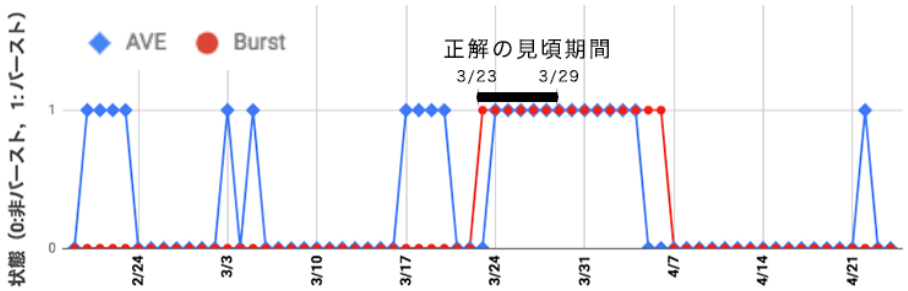

図 1: 東京都の見頃推定結果

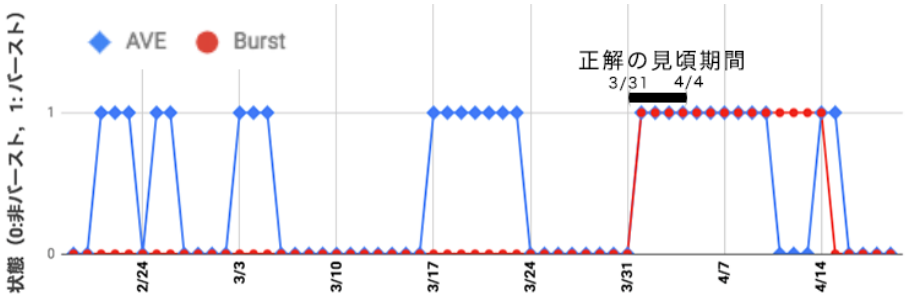

図 2: 石川県の見頃推定結果

きく削減されることから，見頃検出の自動化に近づいていると言える．しかし，図 3（北海道）の Burst の結果を見てみると Burst の見頃推定期間が二つ存在する．図 3 の Burst の 4/1 - 4/3 のツイート内容を確認すると，生物の桜とは関係のない，デジタルコンテンツの桜に関連するイベントについての内容であった．これは意図している対象についての内容ではなく，1 章で述べた課題の一つの「キーワードのみで使用するデータを選択するとノイズが含まれる」に該当する．今後見頃推定の精度を高めるために残りの課題について知見を深め，解決していく必要があると考える．

6　まとめ・今後について

本稿では，観光資源の旬の日単位での推定を行うことを目的として，バースト検知手法を用いた桜の見頃推定を行った．先行研究の移動平均による手法の課題をいくつかあげ，その内の「見頃推定期間が細かく分割される」という問題の解決の期待と，これまでに見頃推定にバースト検知手法が使用されたことが無いことからバースト検知手法の一つである Kleinberg のバースト検知アルゴリズムを適用した．この手法と先行研究の手法の比較実験を行なった．結果，無駄な見頃推定期間が削減され，期待していた問題の解決が出来ることが示唆された．しかし，見頃推定期間が未だ複数出現することから，見頃推定の精度を上げるために，残りの課題を解決する必要があることが分かった．今後はその課題の対応，バースト検知手法のパラメータの調整と，手法の評価を深めるため他の地域や観光資源における見頃推定の実験を行うことを考えている．

謝辞

実験データの収集にあたり，豊橋技術科学大学の吉田光男氏に多大な協力をいただきました．氏に深く感謝いたします．

参考文献

[1] 総務省 平成 29 年情報通信メディアの利用時間と情報行動に関する調査 報告書, http://www.soumu.go.jp/main_content/000564530.pdf, 2018.

[2] Twitter 公式サイト, https://twitter.com.

[3] 観光庁 ICT 活用による観光振興サービスガイド, http://www.mlit.go.jp/common/001080544.pdf, 2014.

[4] J. Kleinberg. Bursty and hierarchical structure in streams. *In Proc. 8th SIGKDD*, pp. 91-101, 2002.

[5] 遠藤雅樹，三富恵佑，佐伯圭介，江原遥，廣田雅春，大野成義，石川博：ツイートを用いた生物季

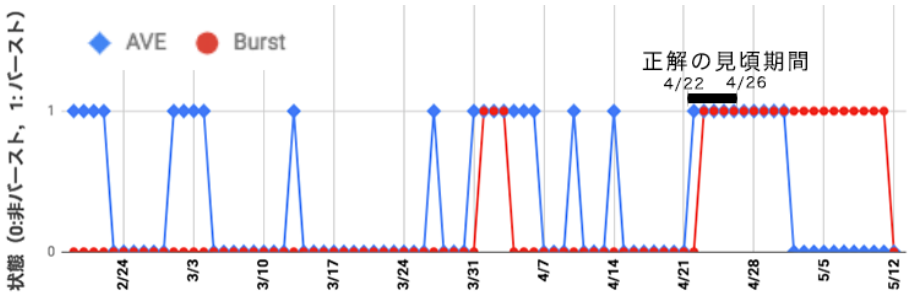

図 3: 北海道の見頃推定結果

節観測の見頃推定手法による情報提供の検討, 観光情報学会誌 12(1), 47-60, 2016.

[6] 気象庁:さくらの観測, 入手先 https://www.data.jma.go.jp/sakura/data/index.html, (閲覧日: 2018/10/02).