2019年度

筑波大学情報学群情報科学類

卒業研究論文

題目

見頃履歴を利用した
Twitterのバースト情報に基づく桜の見頃推定

主専攻　知能情報メディア主専攻

著者　斎藤 明子

指導教員　山本 幹雄・乾 孝司・津川 翔

要　旨

近年, 観光情報取得に ICT が活用されるようになっており, SNS 上ではユーザーによって観光情報が共有されている. 花見や紅葉のような自然を対象とした観光資源は年によって旬にずれが生じるため, SNS データを活用したリアルタイムな見頃の取得が求められている. このような背景のもと, 遠藤や下園によって Twitter データを活用した見頃推定手法が提案されてきた. 下園は Kleinberg のバースト検知手法を組み込んだ桜の見頃推定手法を提案しているが, バースト期間の誤検知が問題であった. 本研究では, 下園の手法を改善するため, ノイズツイート削除を目指した関連ツイート抽出手法を提案する. 提案手法では, 桜の見頃履歴を利用し, 過去の見頃期間のツイートから取得した関連キーワードによって絞りこみを行った. 下園の手法で用いられた対象キーワード照合法をベースライン手法として比較実験を行った. その結果, 提案手法では, 北海道で見られたバースト期間の誤検知が解消された. その他の地域の見頃推定への影響も小さかったが, データ数が少ない地域での推定精度が課題となった.

目 次

図 目 次

第1章　序論

1.1　研究背景と目的

近年, インターネットの普及に伴い, 観光情報の取得に ICT が活用されるようになった．同時に, 旅行中の体験をインターネットを介して共有するユーザーも多く存在する．観光庁による国内の旅行者 520 名を対象としたアンケート調査 [1] では, 90%以上が旅行計画時に情報通信機器を利用し, 旅行中の体験を共有した経験があるのは約 50%という結果が報告されている．

こうして共有された投稿の活用例として, 訪問予定地の桜や紅葉などの観光資源が旬を迎えているかという情報を取得することが考えられる. 観光客がこれから訪問する予定の場所の旬情報を把握しておくことで, 観光客の満足度向上が期待できる.

観光資源の大まかな旬情報は観光情報誌からも収集することができる. しかし, 観光資源は毎年おおよそ同じ時期に旬を迎えるものの, 気象条件によって時期の変動がある. そのため, 現在桜が旬を迎えているか, つまり見頃であるかどうかを知りたい場合, よりリアルタイムな情報源が必要となる. その点, Twitter をはじめとするマイクロブログはユーザーの実体験を即時に投稿できるため, ガイドブックなどの既存メディアよりもリアルタイムな観光情報が存在すると言える. Twitter には目的とする観光情報以外も大量に存在するため, 効率的に見頃情報を検索できるようなシステムが求められる. Twitter の情報を用いて, 桜をはじめとする生物の見頃情報を推定するという研究はすでに行われており, 遠藤ら [2] は移動平均を用いた生物の見頃推定手法を提案している. 移動平均による見頃推定手法は「見頃推定期間が分割される」という問題がある.

推定期間分割の問題を改善するため, 下園ら [3] は Kleinberg のバースト検知 [4] を用いた見頃推定手法を提案した. しかし, 依然として, 本来の見頃から外れた推定期間が存在する. 原因としては, 樹木の桜ではない「桜」という単語に言及した投稿がバースト検知時のノイズとなったためである.

本論文では, 桜の見頃推定におけるバースト検知時にノイズとなるツイートを削除する手法を提案する. これにより, バーストの誤検知を減らし, 見頃推定精度の向上を目指す. ノイズツイート削減のために, バースト検知に使用するツイートに対して検知対象と関連性の高い単語で絞り込みをかける.

評価実験では, 各地域での桜の見頃推定を行い, 提案手法の有効性を検証する. 見頃の正解期間と推定見頃期間を比較し, 正答率によって評価を行う. 関連する単語の使用割合による正答率の変化も評価する.

1.2 本論文の構成

2 章では関連研究について述べる. 3 章では過去の桜見頃データの比較を行う. 4 章では, バースト検出における関連ツイート抽出と, そのベースライン手法である対象キーワード照合法の問題点を述べる. 5 章では関連ツイート抽出処理を改善した提案手法について述べる. 6 章ではベースライン手法の追試, および提案手法の評価実験を行い, 結果と考察を述べる. 7 章ではまとめと今後の課題について述べる.

第2章　関連研究

2.1　バースト検出

事象やイベントの一時的な盛り上がりをバーストという. バースト検出は異常検知の一種であり, 主にトレンド分析に用いられる. Kleinberg は文章データストリームに対するバースト検出手法を提案しており [4], 検知対象に言及しているなどといった関連がある文章の出現頻度が一時的に上昇しバースト状態となっている期間を検出する. 検知対象に関連がある文章を関連文章と呼ぶ. Kleinderg のバーストは断続的に出現する関連文章発生時刻の時間間隔を元にバースト検出を行う連続型, 単位時間ごとに発生した関連文章数と総文章数を比較することによりバースト検出を行う列挙型がある.

2.2　ツイートデータを用いた見頃推定

2.2.1　移動平均を用いた手法

以前よりツイッターに代表されるマイクロブログの投稿を基にした植物の見頃推定手法は研究されており, 遠藤らの移動平均を用いた見頃推定手法もその 1 つである. 遠藤らは, 位置情報付きツイートの関連ツイート数から算出した移動平均を用いて, 桜やカエデの見頃を推定する手法を提案した. 桜の場合,「桜」,「さくら」,「サクラ」といった推定対象を表す単語を含むツイートを関連ツイートとして, 関連ツイート数を数える. 1 年移動平均, 7 日移動平均, 5 日移動平均をそれぞれ比較することで, 実際の見頃期間を含む見頃の推定を可能にした.

一方移動平均を用いた手法には以下のような問題点がある.

1. 見頃推定期間が細かく分割される
2. 生活周期に合わせて注目度が変化する
3. 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
4. データ不足

また, 各移動平均を比較して最終的には人手で判断する必要があった.

2.2.2 バースト検知を用いた見頃推定

下園らは遠藤の手法に Kleinberg の列挙型バースト検出を組み合わせ, 1 の「見頃期間が細かく分割されるという」問題の解決を図った [3].

Kleinberg の列挙型バースト検知では, 単位時間ごとの総文章数と検知対象の関連文章数を利用する. Kleinberg のバーストをツイートデータに適用するため, 検知対象に関するツイートを関連ツイートとし, 関連文章数の代わりに関連ツイート数を用いてバースト検知を行なっている. 下園の手法では, 桜を見頃推定の対象とした. 遠藤の手法と同じく, 「桜」, 「さくら」, 「サクラ」という推定対象を表す単語を含むツイートを関連ツイートとしている. 1 日ごとに関連ツイート数と総ツイート数を数え, バースト検知によりバーストと判定された期間, バースト期間を見頃期間とする.

見頃推定にバースト検知を利用することによって見頃期間の分割はある程度抑えられたが, 1 章で述べたとおり, 本来見頃ではない期間にバーストが検知されるという問題は解決していない.

該当期間のツイートより, 桜という名称を含んだイベントが SNS 上で開催されており, バーストが検出されてしまったことが原因にあると考えられる.

第3章　過去の桜見頃データ分析

観光資源は大きく人工資源と自然資源の 2 種類に分けられる. 人工資源は建物や文化などの人手によって生み出されたものを指し, 自然資源は桜やカエデなどの植物や魚, 動物など自然界に存在するものが該当する. 人工資源と自然資源のうち, 自然資源の旬は気象状況に左右される.

図 3.1 は 2011 年から 2015 年の東京都における桜の開花日から満開日の期間を表したグラフである [5]. このように, 桜の開花日・満開日には年によってばらつきがあり, 満開日は最大で 2 週間近くズレがある.

大まかな旬の時期をガイドブックなど従来の情報源から取得することは可能だが, 毎年変化する旬の期間を確実に把握することはできない. この例からも, リアルタイムな観光情報を取得できる SNS データを用いた見頃推定の必要性が確認できる.

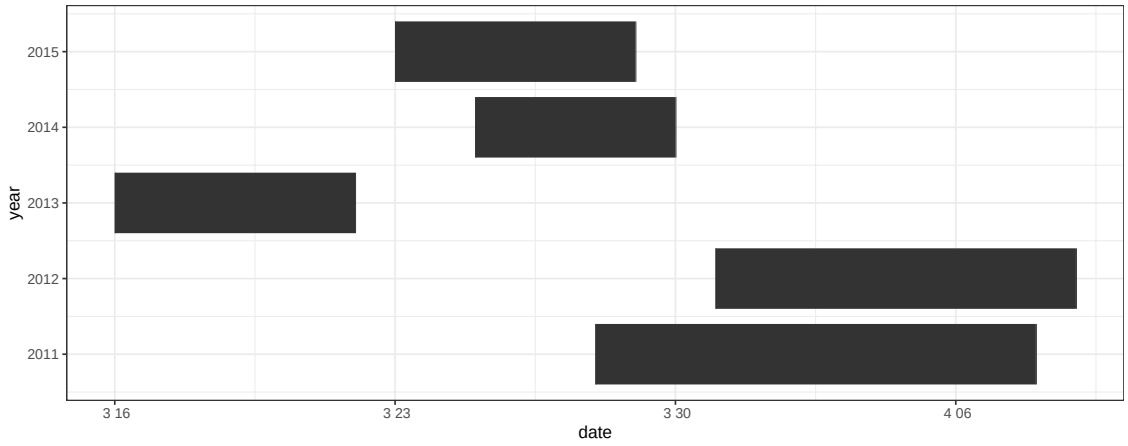


図 3.1: 2011 年から 2015 年までの桜の開花日と満開日 (東京)

第4章　関連ツイート抽出

4.1　バースト検出における関連ツイート抽出について

2 章で述べたように, Kleinberg らが提案したバースト検知手法をツイートデータに適用するためには, 検知対象の関連ツイートを抽出する必要がある. 図 4.1 にツイートデータに Kleinberg のバーストを適用した場合の処理の流れを示す. ツイートデータから関連ツイートを抽出する処理を関連ツイート抽出と呼ぶ.

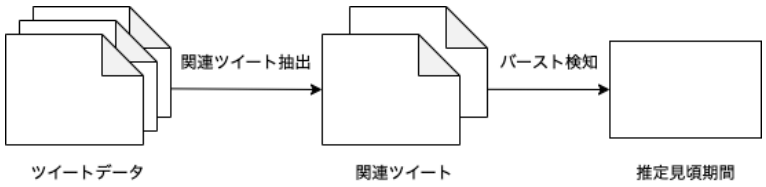


図 4.1: ツイートデータに Kleinberg のバーストを適用した場合の流れ図

下園らの手法ではツイートデータに対し Kleinberg の列挙型バースト検知を適用する. 列挙型バースト検知では関連文章の数を必要とするが, 代わりに検知対象に言及しているツイートを関連ツイートとして利用している.

4.2　下園手法における関連ツイート抽出とその問題点

下園らは桜の見頃を推定するため, 桜の関連ツイート抽出を行なっている. その際,「桜」,「さくら」,「サクラ」というキーワードのいずれかを含んだツイートを全て関連ツイートとして抽出している.

対象のキーワード群を利用して関連ツイートを抽出するため, 本稿では下園らの手法における関連ツイート抽出手法を対象キーワード照合法と呼ぶ.

対象キーワード照合法の場合, 対象キーワードが指し示す意味は植物の桜に限らないため, 対象キーワードのみで抽出した関連ツイートの言及している「桜」は人の名前や地名である可能性がある. このように, 対象キーワードを含んでいるが関連していないツイートはバースト検出結果に影響する. 後述の北海道のツイートデータに下園法を適用した結果から分かるように, 本来の見頃期間とは外れた 4 月初めにバーストが存在する. 当該期間には桜という名称を用いたキャンペーンが行われており, キャンペーンに関連したツイートが桜の関連ツイートとして抽出されたため, 推定見頃期間にノイズが生じたと考えられる.

第5章　提案手法

5.1　見頃履歴の取得

気象庁では, 年毎に植物の開花日や動物の初見日を観測する生物季節観測が行われている. 桜をはじめとする植物の開花日や満開日, 動物の鳴き声が初めて聞こえた日などが全国指定の各地点で観測されている [5]. 観測結果から推察される, 季節ごとに発生する生物特有の現象を生物季節特性と呼ぶ. 遠藤の手法では, 生物季節特性を生物ごとの移動平均の日数を決定するのに活用している. 本稿では, 過去の見頃情報である見頃履歴を求めるため, 生物季節特性を活用する.

ただし, 見頃履歴を取得できる生物は限られている. バラは春から秋にかけて様々な品種が入れ替わりで開花する. バラのように開花のピークを複数回迎える生物の場合, 過去の見頃を一定期間に定めることは困難である. 一方, 桜は年に一度, 春頃に開花するという特性がある. したがって, 桜の場合は生物季節特性と実際の観測記録より, 過去の見頃期間情報である見頃履歴を取得することができる.

本稿では, 見頃期間に投稿されたツイートを旬ツイートと定義する. ある地点で生物が見頃を迎えると, 付近のユーザーは平常時よりも生物に言及したツイートを投稿する機会が増加する. 旬ツイートには, その生物の関連ツイートが多数含まれていると仮定できる. この仮定に基づき, 見頃履歴を利用した関連ツイート抽出手法を提案する.

5.2　関連キーワード照合法

これまで述べてきた通り, バースト検知におけるノイズ削減のためには関連ツイート抽出手法を改善する必要がある. 対象キーワード照合法では, 対象キーワードのみで絞り込んだ関連ツイートへのノイズ混入が課題であった. この課題に対し, 見頃履歴を元に取得した過去の旬ツイートを使用して解決を図る. 過去の旬ツイートより, 桜と同時に出現しやすいキーワードが取得できる. これらの対象キーワードと共起度が高いキーワードを関連キーワードと呼ぶ.

関連キーワードとして, 過去の旬ツイート内に出現する全ての名詞・動詞と, 対象キーワード「桜」,「さくら」,「サクラ」のいずれかのキーワードと共起した単語のうち, 関連度が閾値以上の単語を抽出する.

関連度の導出には PMI と SoA を用いる. PMI は Pointwise Mutual Information のことであり, ともに出現する 2 単語の関連度の強さを表す尺度である. 共起する単語 x,y について, それぞれ出現する確率が $P(x)$,$P(y)$, 同時に出現する確率が $P(x,y)$ とする. この時, x と y の PMI は

以下の式で求められる.

$$PMI(x,y) = \log 2 \frac{P(x,y)}{P(x)P(y)} \tag{5.1}$$

$PMI(x,y) > 0$ の場合, x と y は共起しやすい傾向があり, 一方, $PMI(x,y) < 0$ の場合, x と y は共起しづらい傾向があることを示す. $|PMI(x,y)|$ が大きいほど傾向が強い. なお, $PMI(x,y) = 0$ の場合は x と y は独立である.

ある単語 x と共起する特徴的な単語を求める場合, 全ての単語と関連度が高い単語と, x との関連度のみが高い単語を PMI の値で見分けることができない. Strength of Association(SoA) は単語との非関連度を考慮することで PMI の欠点を補う. 単語 x と y の SoA は 5.1 式を用いて, 以下の式のように求めることができる.

$$SoA(x,y) = PMI(x,y) - PMI(\neg x, y) \tag{5.2}$$

5.2 式で $\neg x$ は, x を除いたツイートデータ中に含まれる全ての単語を表す. SoA では, 単語 x と共起しない単語 y の頻度, つまり $PMI(\neg x, y)$ を非関連度として, x, y の PMI 値より除算する.

対象キーワード群 $\#s$ と共起した単語 w の PMI と SOA は式 (5.1), および式 (5.2) より以下の式で算出できる.

$$\begin{aligned} PMI(\#s,w) &= \log 2 \frac{P(\#s,w)}{P(\#s) \times P(w)} \\ &= \log 2 \frac{\frac{|\#s,w|}{N}}{\frac{|\#s|}{N} \times \frac{|w|}{N}} \\ &= \log 2 \frac{|\#s,w| \times N}{|\#s| \times |w|} \end{aligned} \tag{5.3}$$

$$SOA(\#s,w) = PMI(\#s,w) - PMI(\neg\#s,w) \tag{5.4}$$

今回のように複数の対象キーワードに対して関連度を計算する際, 本来は 1 つの単語に対し各キーワード全ての関連度を求め, 平均値を関連度とする必要がある. あらかじめ対象キーワードを全て「桜」に置き換え, ツイート群に含まれる全単語と「桜」の PMI を計算することで処理を簡単化した. 関連キーワードの抽出アルゴリズムを Algorithm5.2.1 に示す.

関連ツイート抽出の条件として対象キーワードと関連キーワードを併用することで, さらに絞り込んだ照合を行う. この関連ツイート抽出手法を関連キーワード照合法と呼ぶ. 関連キーワード照合法のアルゴリズムを Algorithm5.2.2 に示す. 関連キーワード照合法では Algorithm5.2.1 によって抽出した関連キーワードと対象キーワード, および推定対象期間のツイートデータを用いる. バースト検知ではツイート数に着目し, ツイート内容は使用しないため, 関連ツイート数を返り値とする. Algorithm5.2.2 によって得られた日毎の関連ツイート数と, 日毎の総ツイート数を, バースト検知の入力とする.

図 5.1 に北海道の 2014 年のツイートデータから抽出した桜の関連キーワードの例を示す. 「咲き」, 「7 分」などの桜の開花に関する単語や, 「札幌市資料館」, 「山の上」といった桜の咲いている地点を示した単語が関連キーワードとして抽出されていることがわかる. 「cherrytree」

Algorithm 5.2.1 関連キーワードの抽出アルゴリズム

Input: $\#s$ { 対象キーワード }
Input: $season\_tweets$ { 過去の旬ツイートデータ }
  $word = []$
  $word\_relevance = []$
  for each $tweet \in season\_tweets$ **do**
    $temp\_word+ = separated\_words(tweet)$
    for each $s\_word \in separeted\_words$ **do**
      if s_word not in $words$ **then**
        $words$append.(s_words)
      end if
    end for
  end for
  for each $word \in words$ **do**
    $relevance =$ calc_relevance$(\#s, word)$
    if $relevance > 0$ **then**
      $word\_relevances$.append$(\{"word" = word, "relevance" = relevance\})$
    end if
  end for
  sort$(word_relevances)$
Output: $word\_relevances$ { ソートされた単語, 関連度のペアのリスト }

Algorithm 5.2.2 関連キーワード照合法のアルゴリズム

Input: $\#s$ { 対象キーワード }
Input: $word\_relevances$ { 関連度でソートされた単語, 関連度のペアのリスト }
Input: $tweets$ { 任意の 1 日間に投稿されたツイートデータ }
Input: $rate$ { 関連キーワードの使用割合 }
  $word = []$
  $word\_relevance = []$
  $related\_tweet\_count = 0$
  $limit = len(word\_relevance) * (rate/100)$ $word\_relevances$ =word_relevances$[0 : limit]$
  for each $word\_relevance \in word\_relevances$ **do**
    $related\_words.append(word\_relevances.word)$
  end for
  for each $tweet \in tweets$ **do**
    for each $related\_word \in related\_words$ **do**
      if $tweet$ has (#s **and** related_word) **then**
        $related\_tweet\_count+ = 1$
        break
      end if
    end for
  end for
Output: $related\_tweet\_countt$ { 任意の 1 日間に投稿された関連ツイート数 }

や「instaplace」という英単語が抽出されているが, これは instagram[6] の投稿を Twitter に共有した際, 投稿に含まれたハッシュタグだと考えられる. 「ょんとみたんだよなぁ」という単語は形態素解析に失敗した結果である可能性が高く, 出現回数も少ないと推察される. こうした出現頻度が稀な関連キーワードが含まれていても, 絞り込み検索を行う上では影響がほとんどないため, キーワードから除外しない.

咲き, 札幌市資料館, instaplace, instaplaceapp, 構内, 知行, 長年, 3940, 映える, 山の上,

ジンパ, つぼむ, Creative, Inter, CROSS, サク, ねぷた, エルム, 杜, 7 分, ツツジ, 見頃,

北海道豊富町大通り, ょんとみたんだよなぁ, 80 円, いわた, 大福, luckypierrot, cherrytree,

ハン, 匿名希望, ヘイタイサン, SAITA, アサヒビール, KKR 札幌医療センター, ...

図 5.1: 北海道のツイートデータから抽出した桜の関連キーワードの例

第6章　評価実験

6.1　データセット

見頃推定対象として, 2015 年 2 月 17 日から 2015 年 12 月 30 日に投稿された日本語ツイートデータを使用する. 各ツイートには投稿地点となった都道府県情報が付与されている. 本研究では先行研究と同じ北海道, 東京都, 石川県を見頃推定の対象とする. 見頃履歴を元にした関連キーワード抽出には, 2014 年 1 月 1 日から 12 月 31 日までのツイートデータを使用する. 気象庁が 2014 年に観測した桜の開花日から満開日までを実際の見頃期間として, 各都道府県ごとに 2014 年の旬ツイートを取得する.

関連キーワード照合法に用いる関連キーワードは, 各都道府県ごとに抽出したキーワードのうち関連度が正であるものについて, 関連度が上位のものから関連キーワードとして使用する割合を変化させて実験を行う. 上位 10%から 10%刻みで 100%までの 10 段階であり, 使用率を 100%とした場合は関連度が正であるキーワード全てを使用する.

ベースライン手法には対象キーワード照合法を用いた.

6.2　評価手法

気象庁の記録にある開花日から満開日までを桜の見頃の正解期間として評価する. 推定された見頃期間のうち, 正解期間である日数を数え, F-score を算出する. 推定した見頃と実際の見頃の一番早い日付から一番遅い日付までの期間について正答率を算出する. 表 6.1 に示す, 気象庁の観測による 2015 年の桜の実際の見頃データを正解データとして使用する [5]. 北海道は観測地点が複数存在するため, 本研究では札幌の観測記録を使用する.

表 6.1: 2015 年の正解見頃期間

	開花日	満開日
北海道 (札幌)	4/22	4/26
石川県	3/31	4/4
東京	3/23	3/29

6.3 実験結果

6.3.1 ベースライン手法の追試結果

ベースラインには対象キーワード照合法を用いる. 図 6.1a, 図 6.1b はそれぞれ東京都, 石川県のツイートデータに対し対象キーワード照合法を適用した結果である. バーストが検知された日付の値が 1 に, 検知されなかった日付の値が 0 になっている. どちらの都市の結果も正解期間とバースト期間の開始がほぼ同じであり, 正しく見頃を推定できていると言える. 正解期間後もバーストが検知されているが, これは正解期間を開花から満開までとしているためであり, 満開後も桜の見頃はある程度続くと考えられるため問題としない. 次に, 2015 年の北海道のツイートデータに対し対象キーワード照合法を適用した結果を 6.1c に表す. 2015 年の北海道における実際の見頃期間は, 表 6.1 に示す通り 4 月 22 日からだったが, 本来の見頃期間と被る推定見頃期間の他に 4 月初旬にバーストが検知されている. この誤検知は 4.3 章で述べたように植物の桜以外に言及したツイートがノイズとなった結果だと確認した.

6.3.2 関連キーワード照合法の実験結果

抽出した関連キーワードの使用割合毎のバースト検知結果を図 6.2, 図 6.4, 図 6.4 に示す. SoA により関連度を算出して抽出した関連キーワードを使用している. 関連キーワードの使用割合が 10%から 80%の時, 正解期間とバースト期間はほぼ一致していることがわかる. さらに, 使用割合が 20%から 60%の時は正解期間とバースト期間の開始が一致した. しかし, 使用割合を増加させ 90%以上になると正解期間から大きく外れたバースト期間が出現した.

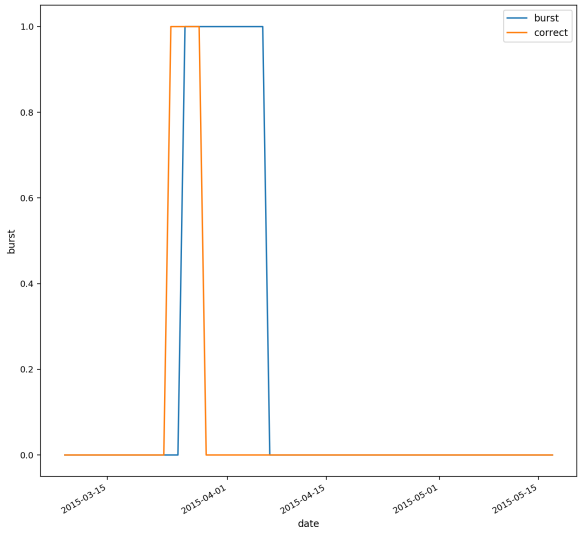
関連キーワード照合法の関連キーワード使用割合と F-score のグラフを図 6.5, 図 6.7, 図 6.6 に示す. PMI, SoA はそれぞれ関連キーワードの関連度を PMI, SoA それぞれで算出した場合の結果である. Target は同じデータで対象キーワード照合法を使用した場合の F-score を示している. 図 6.5 より, 北海道のデータにおいて関連キーワードの使用割合が 40%以上になると, 関連キーワード照合法における F-score が対象キーワード照合法を上回った. 関連キーワード使用割合が 70%の時, F-score は最大 0.1 ポイントほど改善した. PMI で関連度を算出した場合, 上位の関連キーワードのみを使用すると F-score は 0 ポイント近くまで落ち込んだ. 一方, SoA を使用すると, 関連キーワード使用割合が 30%以下の場合でも F-score は対象キーワード照合法を上回った. 図 6.7 に示す, 東京都の見頃推定結果を見ると, PMI は北海道と同じく大きく落ち込んだが, SoA 関連キーワード照合法を使用した場合は, 使用割合が 50%を超えたあたりから対象キーワード照合法と同程度の F-score を記録した. 一方, 図 6.7 に示す, 石川県の見頃推定結果は PMI, SoA 共に落ち込みが見られ, 使用割合を増加させても対象キーワード照合法を 0.3 から 1 ポイントほど下回る結果となった.

6.4 考察

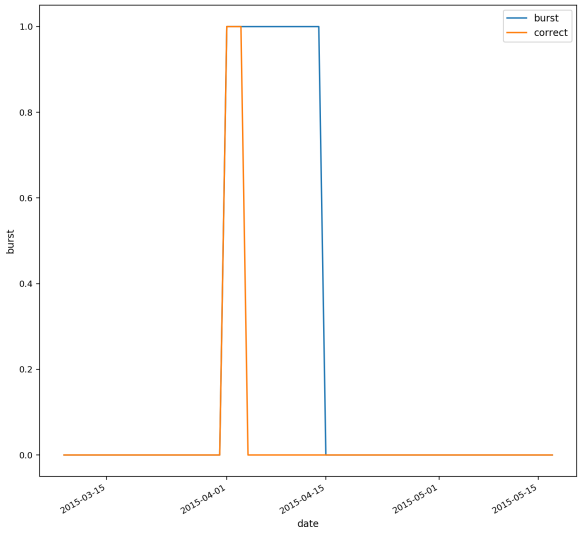
前節で述べたように, 提案手法である関連キーワード照合法における関連キーワードの使用割合を 80%以下に抑えた場合, 北海道の見頃推定におけるバーストの誤検知は解消が可能である.

また, 先行手法で問題なく見頃推定が行えていた東京都の見頃推定結果への影響も少ないと言える. 石川県の見頃推定結果は対象キーワード照合法より F-score が低下したが, これは石川県のツイートデータ数が不十分であったため, 適切な関連キーワードが抽出できていない可能性がある.

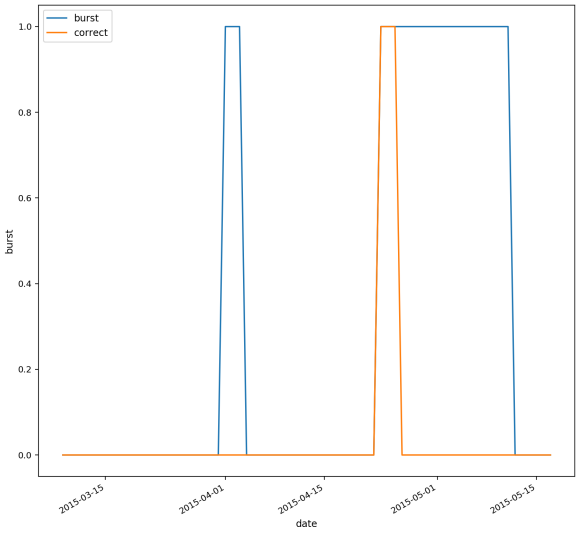
関連度算出には非関連度の考慮された SoA を用いることで, 関連キーワードの使用割合を抑えられる. 関連キーワードの使用割合を抑えることで照合すべき単語数が減少するため, 見頃推定の高速化につながる. 北海道, 東京都, 石川県の各見頃推定結果より, 関連キーワード照合法における関連キーワードの使用割合は 50%から 80%が適切であると考えられる.



(a) 東京都

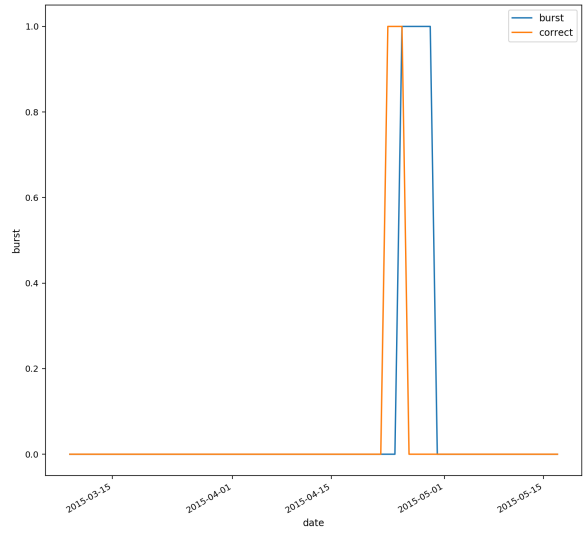


(b) 石川県

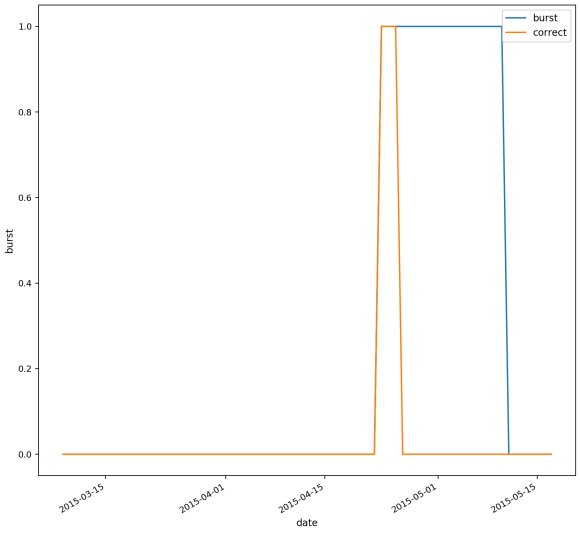


(c) 北海道

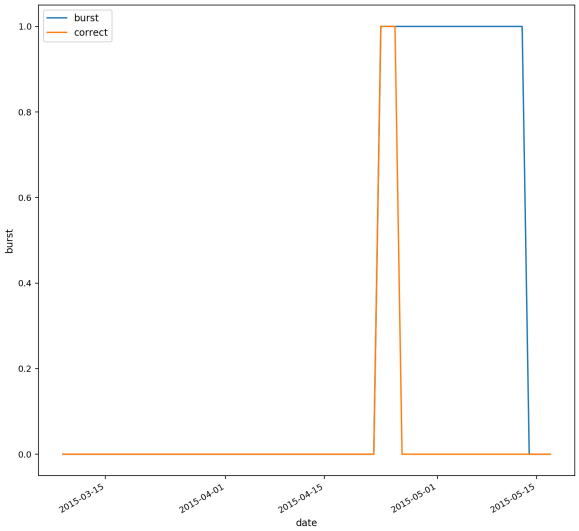
図 6.1: 対象キーワード照合法を適用した場合のバースト結果



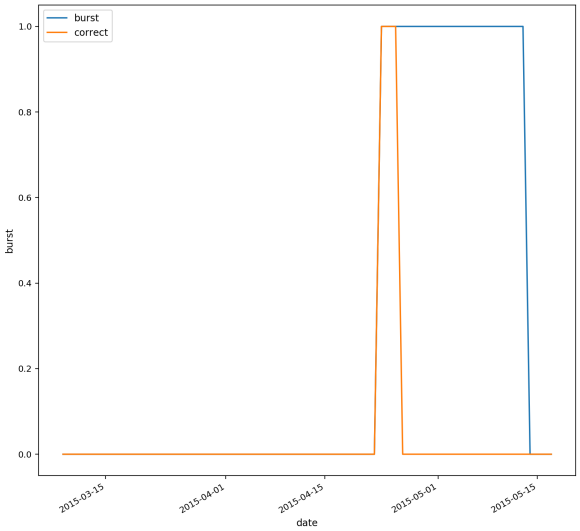
(a) 関連キーワード上位 10%使用



(b) 関連キーワード上位 20%使用

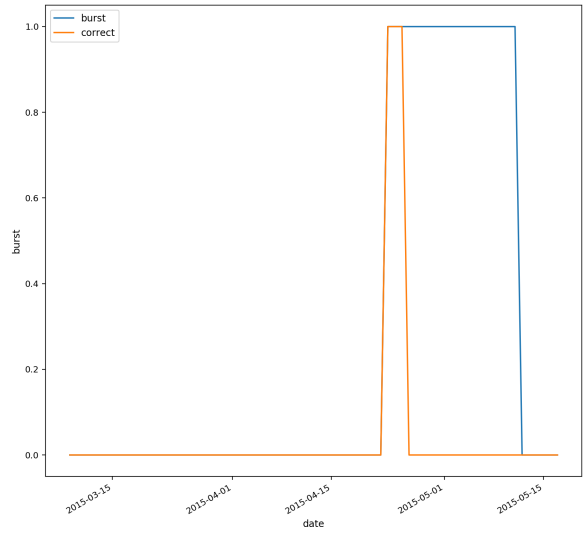


(c) 関連キーワード上位 30%使用

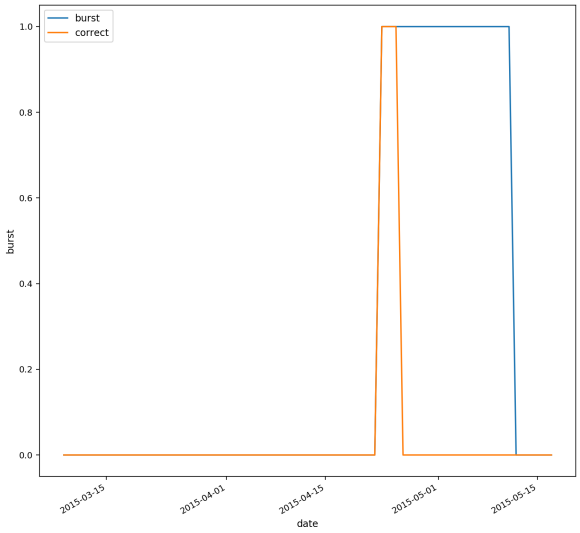


(d) 関連キーワード上位 40%使用

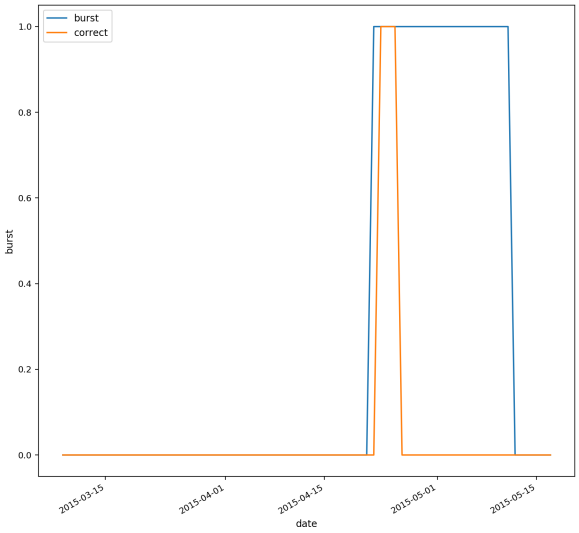
図 6.2: 関連キーワード照合法における関連キーワードの使用割合ごとのバースト検知結果 (10%-40%)



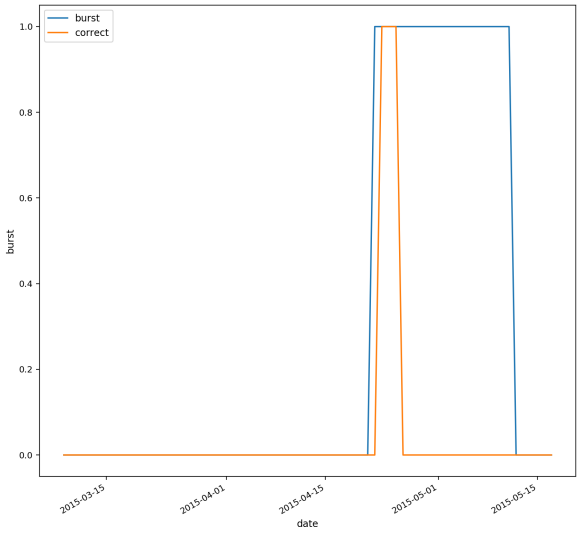
(a) 関連キーワード上位 50%使用



(b) 関連キーワード上位 60%使用

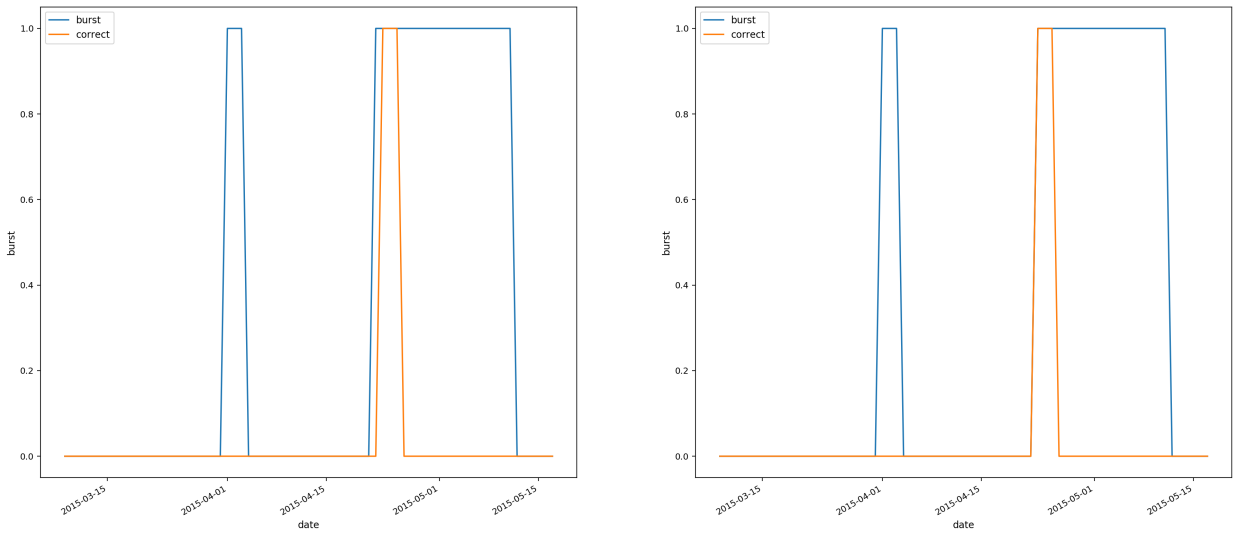


(c) 関連キーワード上位 70%使用



(d) 関連キーワード上位 80%使用

図 6.3: 関連キーワード照合法における関連キーワードの使用割合ごとのバースト検知結果 (50%-80%)



(a) 関連キーワード上位 90%使用　　(b) 関連キーワード全て使用

図 6.4: 関連キーワード照合法における関連キーワードの使用割合ごとのバースト検知結果 (90%, 100%)

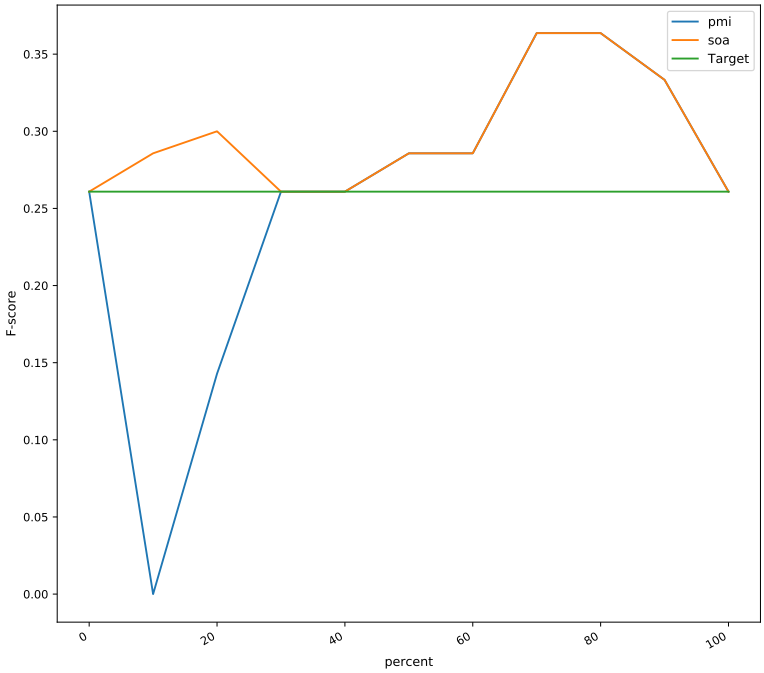


図 6.5: 関連キーワード使用割合ごとの F-score：北海道

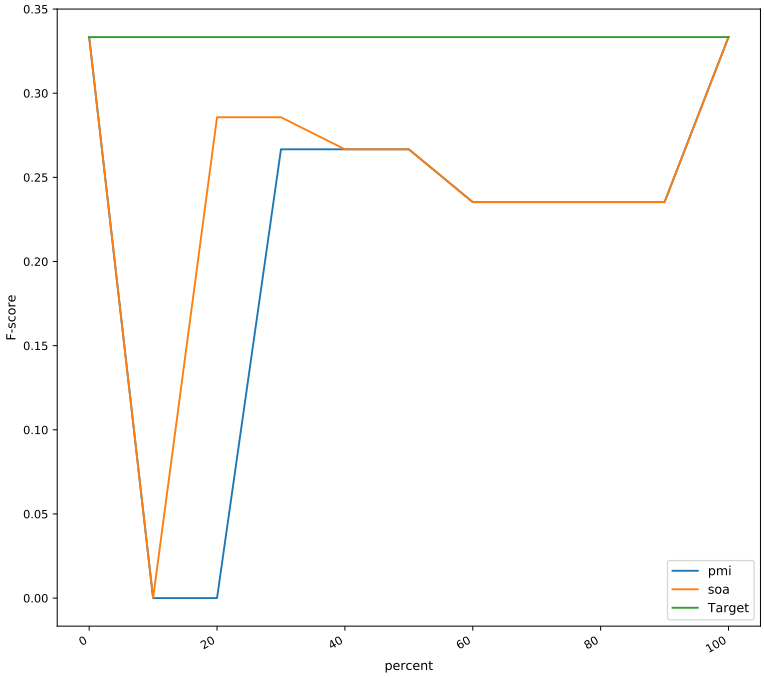


図 6.6: 関連キーワード使用割合ごとの F-score：石川県

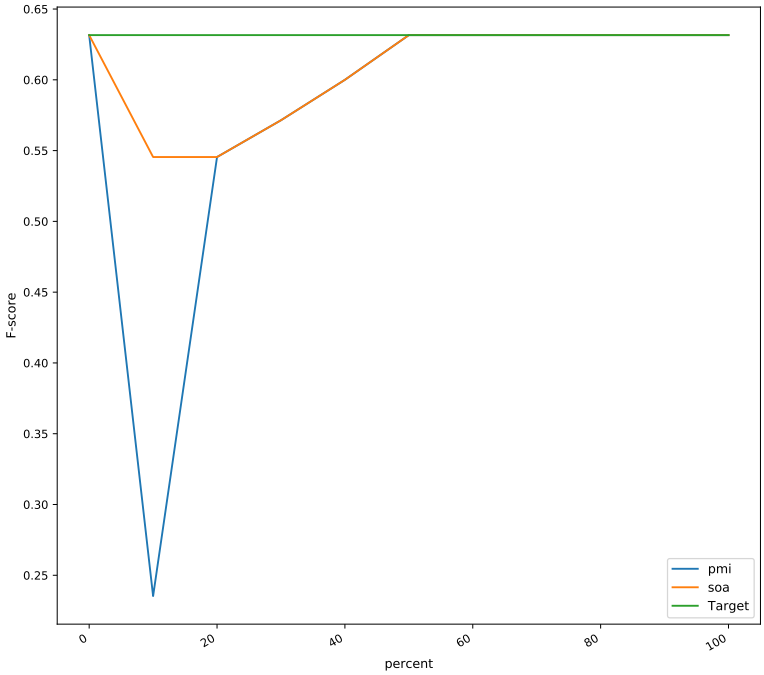


図 6.7: 関連キーワード使用割合ごとの F-score：東京都

第7章　結論

7.1　まとめ

本稿では, 下園 [3] が提案した Kleinberg のバースト検知 [4] を用いた桜の見頃推定手法における, バーストの誤検知という問題に着目した. この問題の解決のため, 関連ツイート抽出処理時に発生するノイズツイートの削除を目指した. 見頃履歴を利用して得た過去の旬ツイートから関連キーワードを取得し, 対象キーワードと関連キーワードを併用してツイートの絞り込みを行う関連キーワード照合法を提案した.

評価実験の結果, 先行手法で発生していた北海道の見頃推定におけるバーストの誤検知が, 関連キーワード照合法では解消された. 他の地域である東京都や石川県での見頃推定精度も保たれる結果となった.

7.2　今後の課題

今回は, 各単語の分類によるナイーブな手法によって改善を図った. 今後は, 過去の旬ツイートに対し関連キーワード照合法を適用した結果から作成した, ロジスティック回帰による関連ツイートの分類モデルを使用した関連ツイート抽出手法について取り組む予定である.

また, 実際の見頃期間と見頃推定期間の終了時期に差があり, F-score の低下につながっている. したがって, 今後は更に適した評価手法を考案する必要がある.

謝辞

本研究を進めるにあたり, 指導教官である筑波大学システム情報系情報工学域乾孝司准教授には丁寧にご指導いただきました. 心より感謝を申し上げます. 更に, 様々な助言をくださった知能情報・生体工学研究と評判グループの皆様に厚く御礼を申し上げます.

参考文献

[1] 観光庁. ICT 活用による観光振興サービスガイド. 情報通信技術を活用した観光振興策に関する調査業務 報告書, 第 I 部, p. 4, 2014.

[2] 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, Vol. 68, No. 12(1), pp. 47–60, 2016.

[3] 下園良太, 乾孝司. Twitter のバースト情報に基づく桜の見頃推定. 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), 2019.

[4] Jon Kleinberg. Bursty and hierarchical structure in streams. In *Proceeding of The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1–25, 2002.

[5] さくらの観測. ”https://www.data.jma.go.jp/sakura/data/index.html”. 閲覧日：2019/12/20.

[6] Instagram. ”https://www.instagram.com”.