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

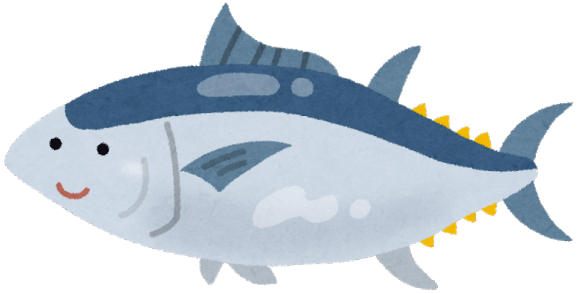
自然言語処理 on the WEB研究室
指導教員　山本幹雄・乾孝司・津川翔
201813550 斎藤明子

研究背景 (1/3)

- 観光情報をSNSで共有するユーザの増加
  - 9割が観光情報の取得にICTを活用(観光庁 2014)

- 旅行計画を立てる上で観光資源の旬は重要な情報

桜の見頃



魚の食べ頃

研究背景 (2/3)

ガイドブック

課題

「桜の旬は3月下旬」などと
記載されており，大まかな
旬期間しかわからない

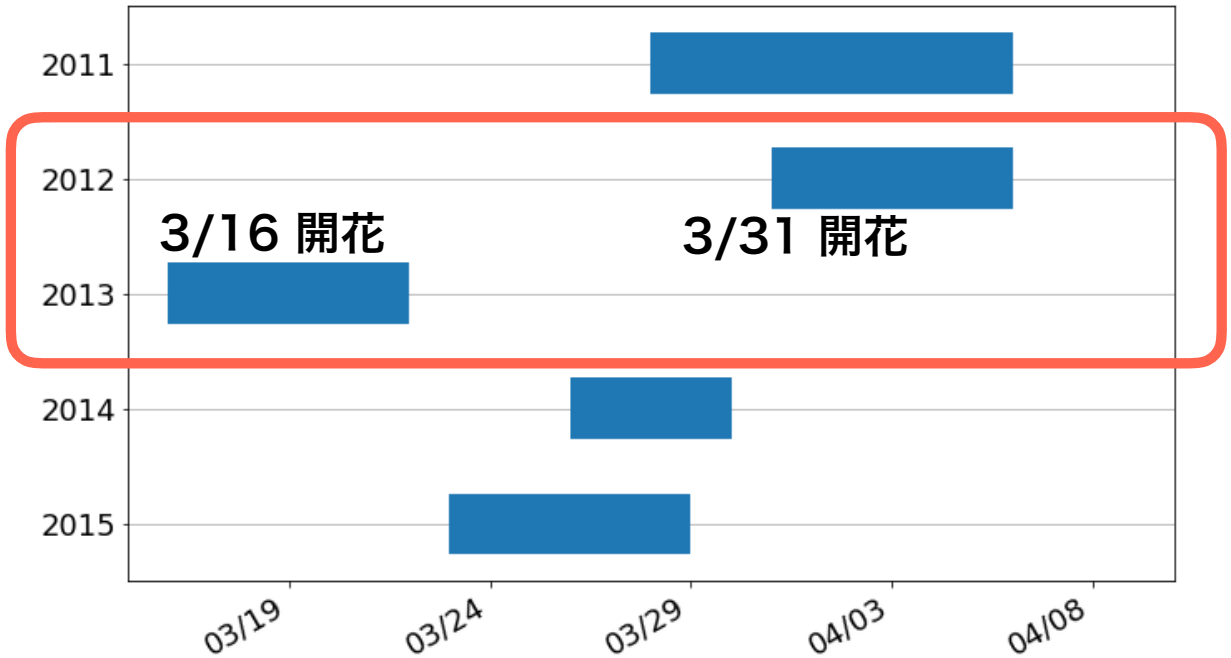
SNS・Webサイト

課題

リアルタイム情報はあるが，
情報がまとまっていないため，
検索などで獲得するのは困難

研究背景 (3/3)

- 自然資源の旬は年によってズレが生じる
  - 例）桜や紅葉などの観光資源



東京における桜の開花日～満開日(2011-2015)

➡ リアルタイム情報を用いた見頃の推定が求められる

先行研究

- ツイートを用いた桜や紅葉の見頃推定 (遠藤 2016)
  - 位置情報付きツイートの投稿数の移動平均を利用

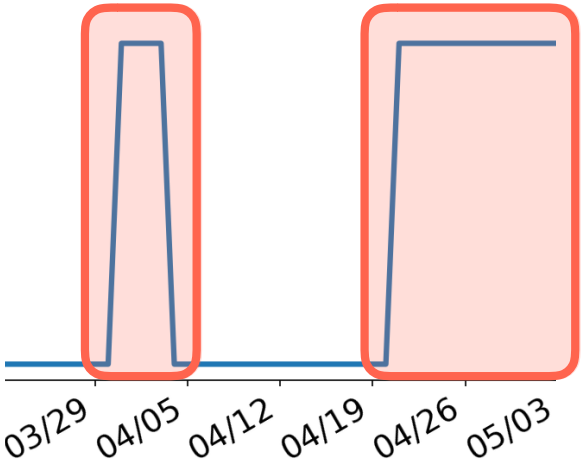
- バースト検知手法を用いた桜の見頃推定 (下園 2019)
  - 本来の見頃から外れたバーストの誤検知が課題

研究目的

- Twitterデータを用いた桜の見頃推定手法の改善
- 過去の見頃情報（見頃履歴）を利用し，
  下園手法で提起された問題点の解決を目指す
  - 入力：ツイートデータ
  - 出力：推定される見頃期間

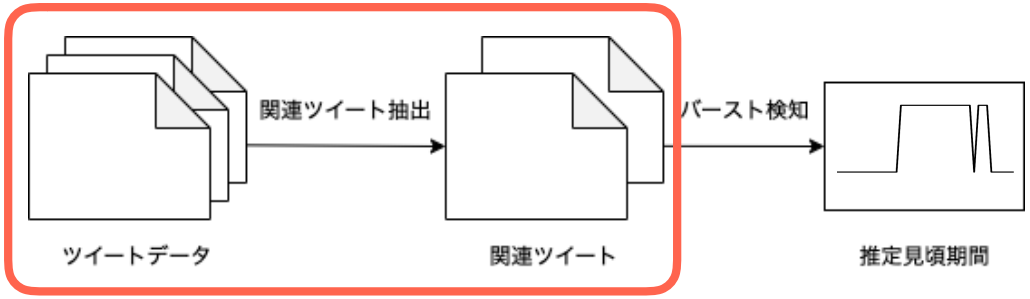
バースト検知による見頃推定

- バースト (Kleinberg 2002)
  - ある活動・事象の一時的な盛り上がり
    - 例）地震，音楽イベントなど



- バースト検知
  - 単位時間ごとの関連ツイート数と総文章数を比較
  - 検知対象に関連するツイートを抽出する必要がある

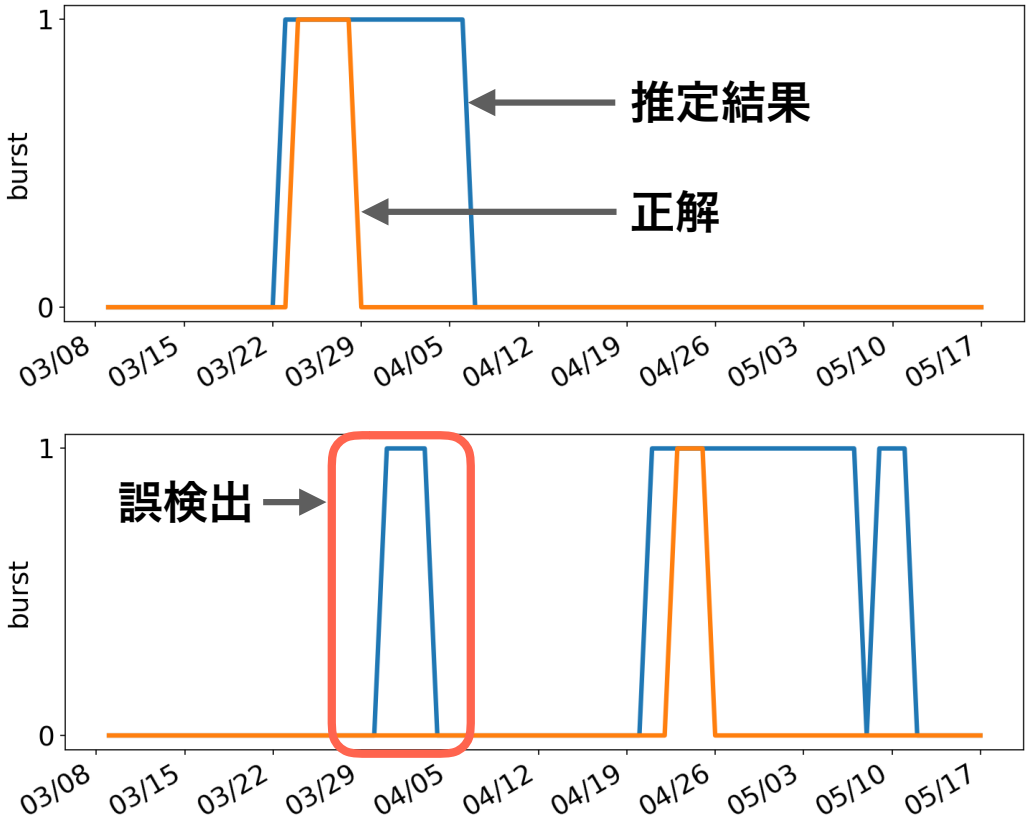
➡ **関連ツイート抽出**



バースト検知を利用した見頃推定の流れ図

下園(2019)の追試結果

（0: 非バースト，1: バースト）



- 東京都

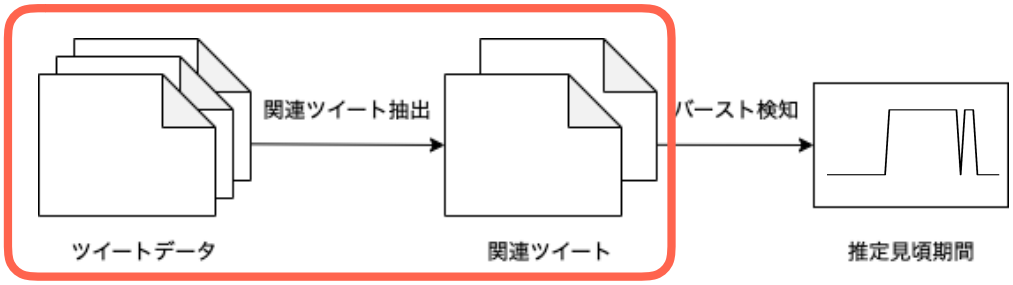
  推定した見頃期間と
  本来の見頃がほぼ一致

- 北海道

  本来の見頃から外れた
  見頃期間が推定された

先行手法の問題点

- 対象キーワードが含まれるものを関連ツイートとして抽出
  - 対象キーワードとして”桜”, ”さくら”, ”サクラ”を使用



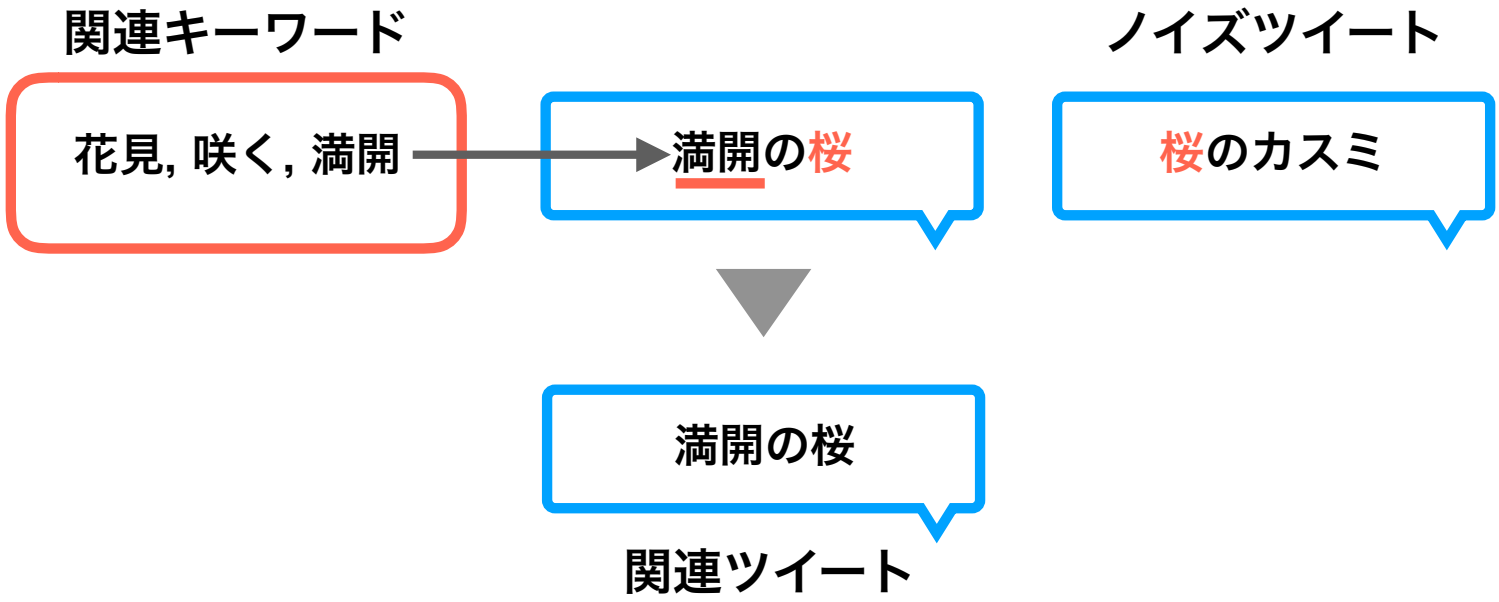
再掲：バースト検知を利用した見頃推定の流れ図

- 植物の桜と関連のないツイートがノイズとなり,
  期待しないバーストが検出される

➡ 関連ツイートからノイズツイートを削除する

関連キーワード照合法

- ツイートから桜と関連のある単語を抽出
- 対象キーワードと関連キーワードを併用して，絞り込み照合を行う



関連キーワード抽出

- 過去の見頃期間のツイートより関連キーワード取得
  - 過去の見頃期間のツイートには
    ”桜”の関連ツイートが多数含まれるという仮定に基づく
- 対象キーワードとの関連度が正である単語を抽出
  - 関連度：2単語の共起頻度
  - 関連度はSoAにより算出

$$PMI(\#s, w) = \log 2 \frac{P(\#s, w)}{P(\#s) \times P(w)}$$

$$SoA(\#s, w) = PMI(\#s, w) - PMI(\neg\#s, w)$$

#s：対象キーワード
w：ツイート中の任意単語

関連キーワードの抽出例

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, . . .

抽出した関連キーワードの例

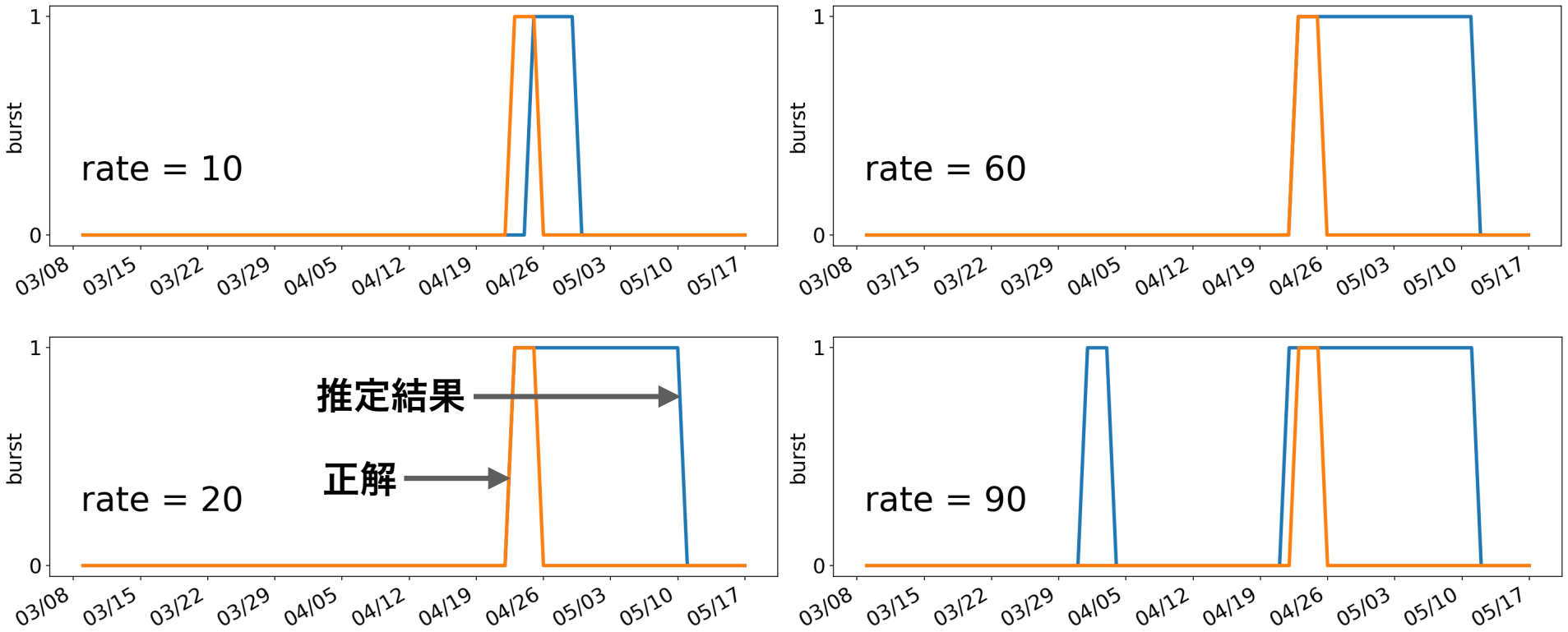
- 桜の開花場所や咲き具合を表す単語が抽出された

評価実験

- データセット
  - 位置情報付きの日本語ツイートデータ
    - 北海道・石川県・東京都のデータ
    - 関連キーワード取得：2014年の各地の桜の開花日-満開日
    - 見頃推定対象：2015/2/17 - 2015/12/31
- 関連キーワードの使用割合を変化させて実験
  - 関連度が上位のキーワードから10%ごと
- 開花日から満開日までを見頃の正解期間とする
  - 正解期間と推定結果からF-scoreを算出

評価実験｜北海道

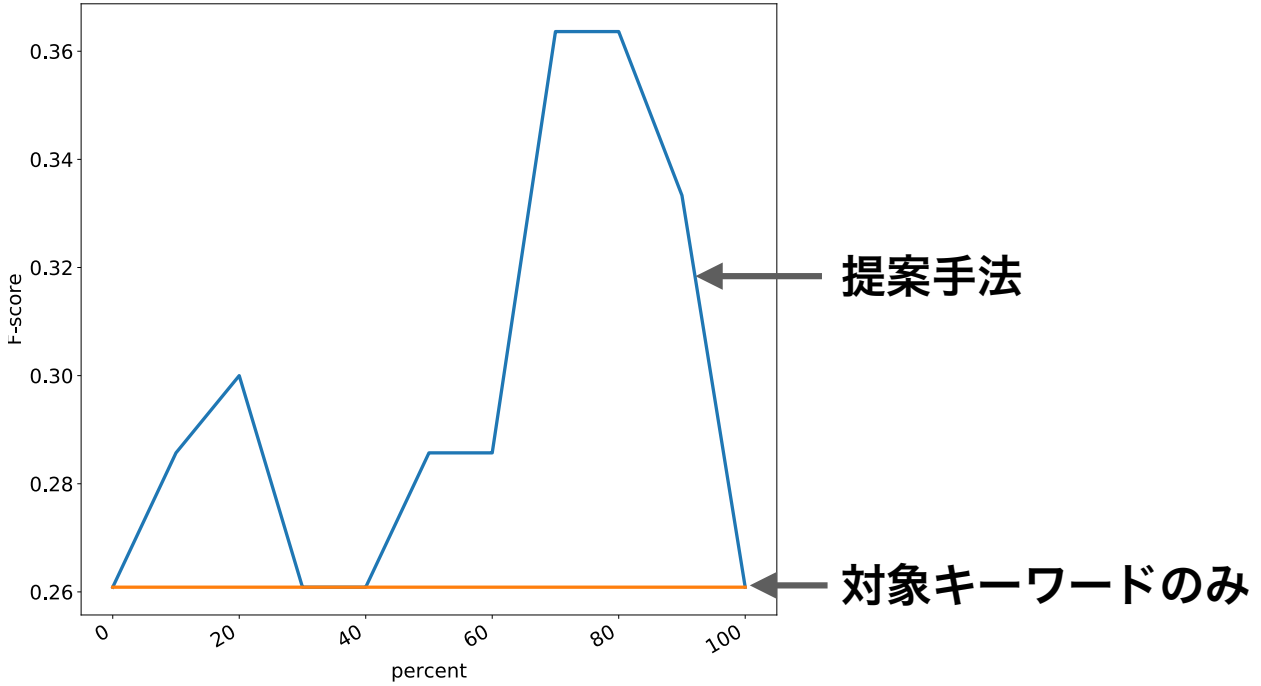
- 先行手法で見られたバーストの誤検知が解消した
  - 使用割合20%~60%ではバースト開始と実際の見頃期間開始が一致



北海道の見頃推定結果

評価実験｜北海道

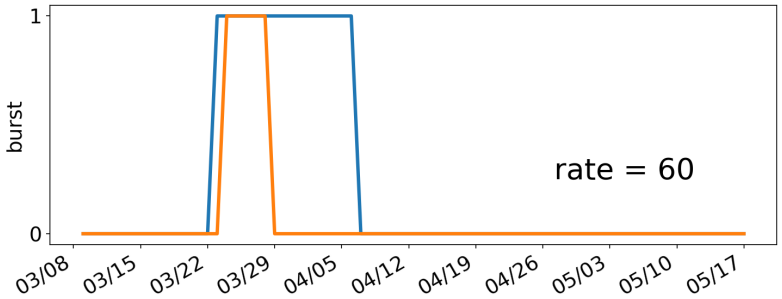
- 先行手法で見られたバーストの誤検知が解消した
  - 使用割合20%~60%ではバースト開始と実際の見頃期間開始が一致
- F-scoreは対象キーワードのみを使用している先行手法以上のスコア



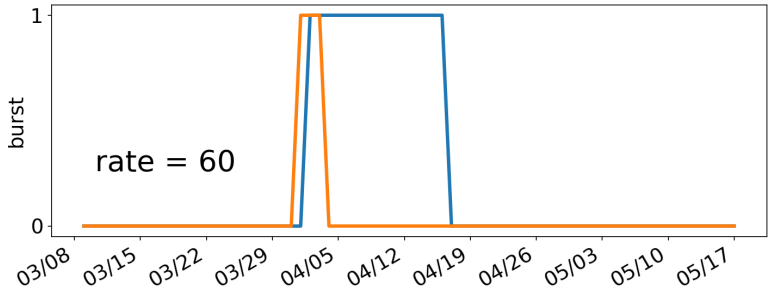
関連キーワード使用割合ごとのF-score：北海道

評価実験｜東京都・石川県

- 推定見頃期間の開始が正解期間とほぼ一致

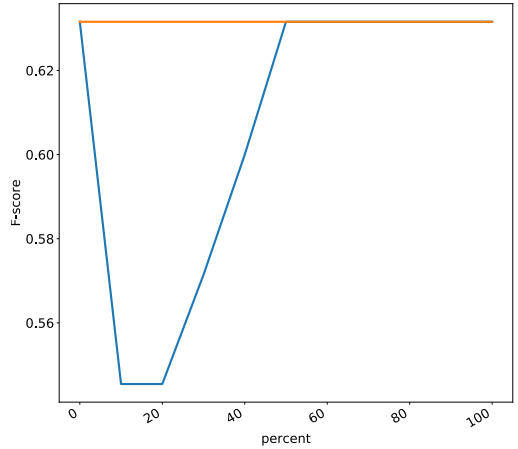


東京都の見頃推定結果

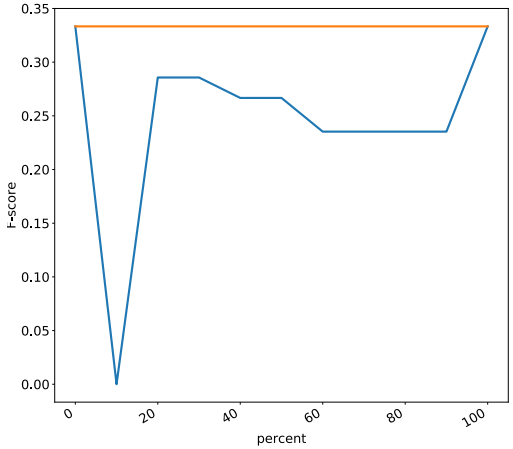


石川県の見頃推定結果

- 提案手法に変更してもF-scoreへの影響が小さい



東京都のF-score



石川県のF-score

考察

- 関連キーワード照合法により,
  ノイズによるバーストの誤検知は解消可能
  - 関連キーワードの使用割合80%以下の場合
- 他の地域の見頃推定精度への影響は小さい
  - データ数不足により適切な関連キーワードを
    取得できない可能性がある
- 関連キーワードの使用割合は50%~80%が適切

今後の課題

- 推定結果と正解期間の終了日に差が生じた
  - F-scoreの低下の一因だと考えられる
  - 正解期間の設定など，より適した評価方法を検討

- 分類モデルを使用した関連ツイート抽出の検討

使用データについて

- 推定対象のツイート数
  - 北海道：約200万個
  - 東京都：約850万個
  - 石川県：約33万個

- 関連キーワード取得対象のツイート数
  - 北海道：約4万個
  - 東京都：約36万個
  - 石川県：約2万個

使用データについて

- 関連キーワード取得対象のツイート数
  - 北海道：約4万個
  - 東京都：約36万個
  - 石川県：約2万個

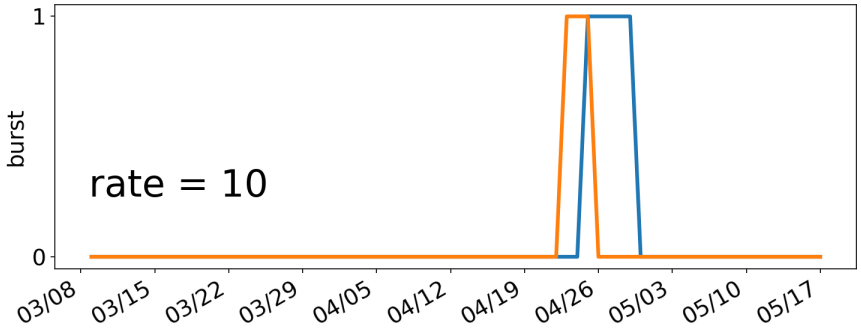
    → データ数の不足により適切な関連キーワードが抽出できなかった可能性がある
- 関連キーワード数
  - 北海道：490個
  - 東京都：6484個
  - 石川県：700個

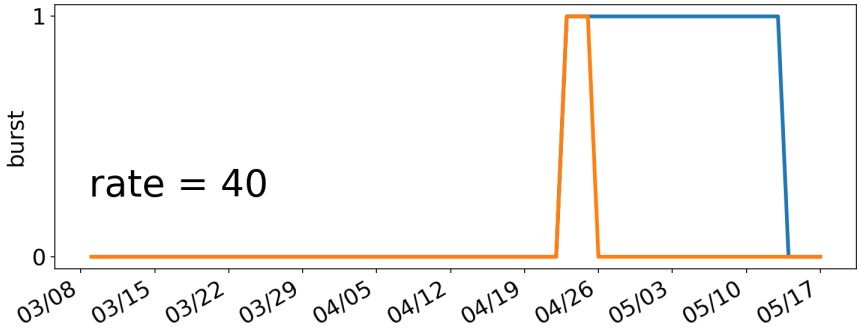
使用データについて

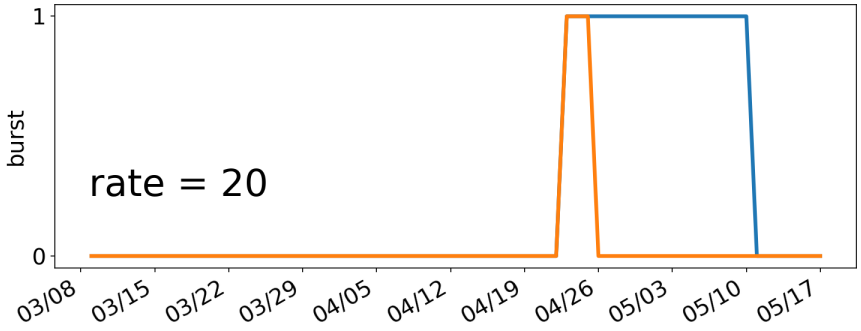
- 使用割合ごとの関連キーワード数(北海道)

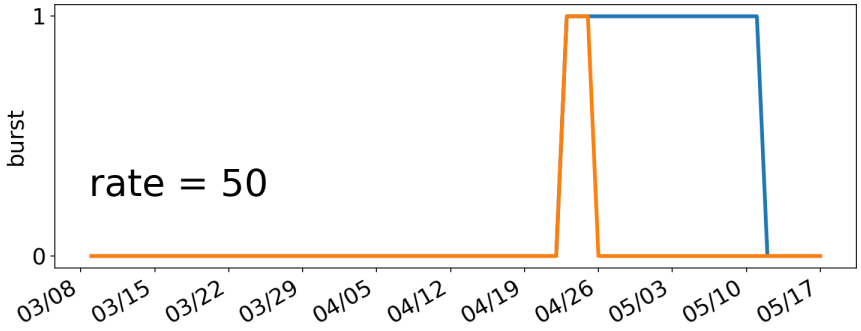
使用割合	関連キーワード数	使用割合	関連キーワード数
上位10%	71	60%	295
20%	108	70%	344
30%	148	80%	392
40%	198	90%	441
50%	248	100%	490

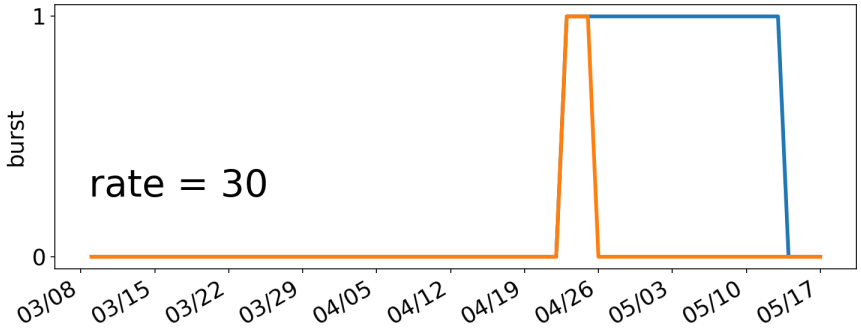
見頃推定結果（北海道）

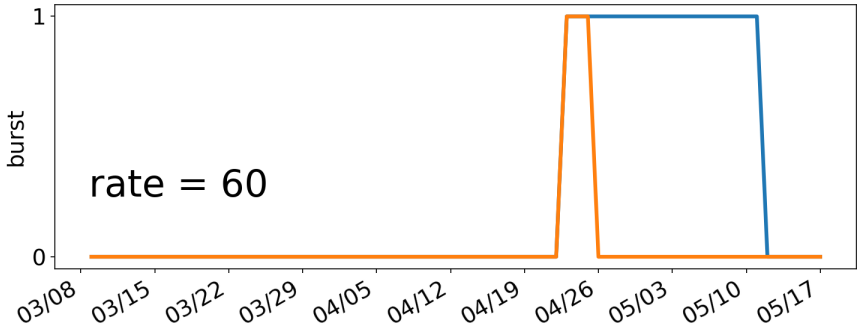

rate = 10
burst
1
0
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17


rate = 40
burst
1
0
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17

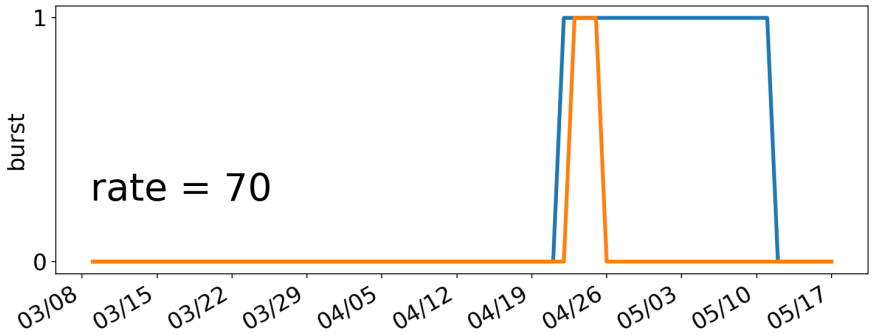

rate = 20
burst
1
0
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17

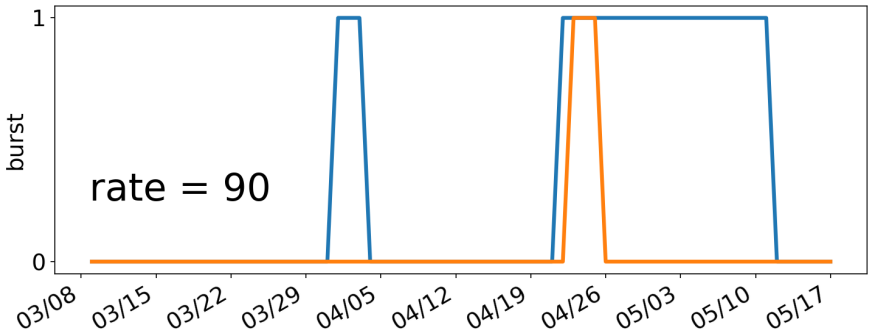

rate = 50
burst
1
0
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17

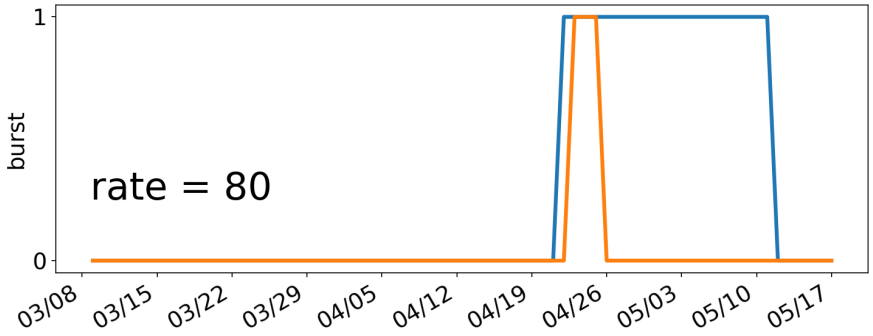

rate = 30
burst
1
0
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17

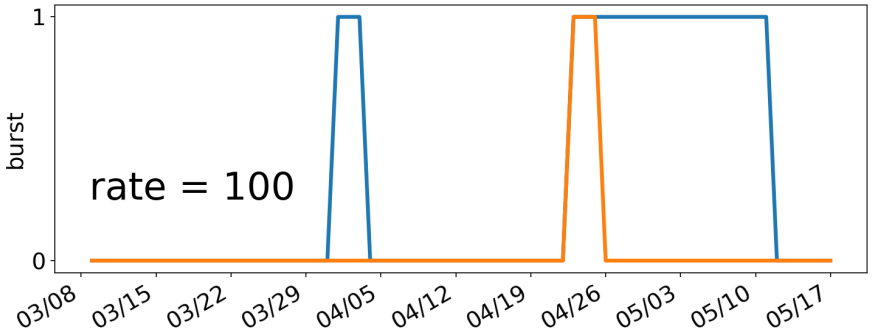

rate = 60
burst
1
0
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17

見頃推定結果（北海道）


burst
1
0
rate = 70
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17


burst
1
0
rate = 90
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17


burst
1
0
rate = 80
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17


burst
1
0
rate = 100
03/08
03/15
03/22
03/29
04/05
04/12
04/19
04/26
05/03
05/10
05/17