4N-02

# 密度に基づく適応的な時空間クラスタリング手法を用いたトピックの時空間分析手法

酒井 達弘 † 田村 慶一 † 北上 始 †
† 広島市立大学大学院 情報科学研究科

## 1 はじめに

近年，ソーシャルメディア上において位置情報付きデータが盛んに投稿されており，位置情報付きデータは，地震，台風，洪水などの自然災害の状況分析への利用が期待されている．特に，Twitter 上のジオタグ付きツイートを利用し，自然災害発生時の状況分析を行う研究が盛んに行われている．我々は先行研究 [1] において，$(\epsilon,\tau)$-密度に基づく時空間クラスタリングを用いたトピックの時空間分析手法を提案している．提案されている手法を用いることで自然災害に関するトピックの時空間的な変化を捉えることができる．しかしながら，田舎などの投稿数が少ない地域では時空間クラスタを抽出することができないという課題があった．そこで，本研究では，この問題点を解決するために，時空間的な投稿数の差異を考慮した $(\epsilon,\tau)$-密度に基づく適応的な時空間クラスタリングを用いた，新しいトピックの時空間分析手法を提案する．

## 2 提案手法

本章では，提案手法について説明する．

### 2.1 データモデルと概要

ジオタグ付きツイートの中から，分析対象のトピック $(etp)$ を含むジオタグ付きツイート集合を，$RGTS^{(etp)} = \{rgt_1^{(etp)}, \cdots, rgt_m^{(etp)}\}$ と定義する．各 $rgt_i^{(etp)}$ は，投稿時間，位置情報，文書データで構成される．

提案手法の処理概要を次に示す．

(1) ジオタグ付きツイートをリアルタイムに収集し，ナイーブベイズ分類器を用いて，分析対象のトピックを含むジオタグ付きツイート $rgt_p^{(etp)}$ を抽出する．
(2) 新しい $rgt_p^{(etp)}$ が抽出される度に，$(\epsilon,\tau)$-密度に基づく適応的な時空間クラスタリングを行い，時空間クラスタをリアルタイムに抽出する．
(3) 地図上に抽出した時空間クラスタを表示する．

### 2.2 分類器

トピックを含むジオタグ付きツイートを抽出するために，最初に，分析したいキーワードを含むジオタグ付きツイートのみを抽出する．そして，教師あり学習のナイーブベイズを用いて “relevant” と “irrelevant” クラスに分類する．ナイーブベイズは，ベイズの定理に基づいた単純な確率的分類器である．“relevant” クラスに分類されたジオタグ付きツイートを $rgt_p^{(etp)}$ として抽出する．

### 2.3 $(\epsilon,\tau)$-密度に基づく適応的な時空間クラスタリング手法

密度に基づくクラスタリング手法は，データが高密度である領域をクラスタと定義する．先行研究では，距離が $\epsilon$ 以内であり，投稿時間が $\tau$ 以内のジオタグ付きツイートを $(\epsilon,\tau)$-密度に基づく近傍と定義し，時空間クラスタの核となるジオタグ付きツイートの $(\epsilon,\tau)$-密度に基づく近傍の数は閾値 $(MinRGT)$ を超えなければならないとしている．本研究では，時空間的な投稿数を考慮して閾値を適応的に変化させることで局所的に高密度な地域を時空間クラスタとして抽出可能にする．

#### 2.3.1 諸定義

$(\epsilon,\tau)$-密度に基づく近傍と適応的な閾値について説明する．

**定義 1 ($(\epsilon,\tau)$-密度に基づく近傍)**
ジオタグ付きツイート $rgt_p^{(etp)}$ の $(\epsilon,\tau)$-密度に基づく近傍は，$rgt_p^{(etp)}$ から距離が $\epsilon$ 以内であり，投稿時間間隔が $\tau$ 以内であるジオタグ付きツイート集合とし，$STN_{(\epsilon,\tau)}(rgt_p^{(etp)})$ と表記する．

**定義 2 (地域的な時空間投稿密度，適応的な閾値)**
ジオタグ付きツイート $rgt_p^{(etp)}$ の地域的な時空間投稿密度を $lstd(rgt_p^{(etp)})$ と表記し，$rgt_p^{(etp)}$ に対する適応的な閾値 $AT(rgt_p^{(etp)}, MinRGT)$ を次のように定義する．

$$AT(rgt_p^{(etp)}, MinRGT) = (MinRGT - 1) \times lstd(rgt_p^{(etp)}) + 1 \quad (1)$$

地域的な時空間投稿密度 $lstd(rgt_p^{(etp)})$ は，過去に投稿されたジオタグ付きツイートの投稿数から求める．対象時空間を 3 次元にグリッド分割を行い（分割数 $= div_{lng} \times div_{lat} \times div_{time}$），各時空間グリッドに含まれる投稿数を正規化した値を $lstd(rgt_p^{(etp)})$ とする．

#### 2.3.2 クラスタ抽出アルゴリズム

新たに抽出されたトピックを含むジオタグ付きツイート $rgt_p^{(etp)}$ と，これまでのジオタグ付きツイー

---

Spatiotemporal Analysis Method for Topics using Density-based Adaptive Spatiotemporal Clustering Algorithm

† Tatsuhiro Sakai, Keiichi Tamura and Hajime Kitakami Graduate School of Information Sciences, Hiroshima City University (†)

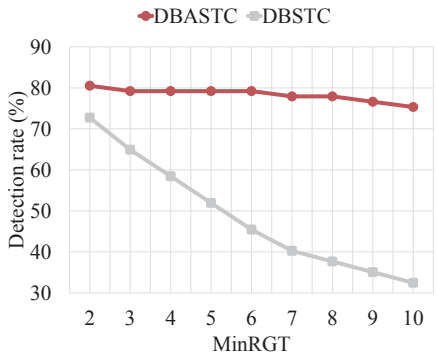

図 1: 検出率

ト集合 $RGTS^{(etp)}$ と現時点での時空間クラスタ集合を入力として，更新された時空間クラスタ集合を出力する．近傍の数が適応的な閾値以上である (つまり，$|STN_{(\epsilon,\tau)}(rgt_p^{(etp)})| \geq AT(rgt_p^{(etp)}, MinRGT)$) ツイートを結合していくことで，ひとつの時空間クラスタを抽出する．また，新しい $rgt_p^{(etp)}$ の入力がある度に $rgt_p^{(etp)}$ の近傍にあるツイートのみを再クラスタリングを行うことで，時空間クラスタをリアルタイムに抽出することができる．

## 3 評価実験

提案手法を評価するために,「雨」に関するトピックについての検出率を提案手法 (DBASTC と表記する) と先行研究の手法 (DBSTC と表記する) で比較する．比較では 2014 年 6 月と 7 月において大雨として新聞報道された 75 地域を検出できたかを比較する．パラメータは，$\epsilon = 5km$，$\tau = 3600sec$ を用いた．$MinRGT$ は 2 から 10 まで変更した場合について実験を行った．

地域的な時空間投稿密度を求めるための分割数は，日本の最西端の緯度，経度 (24.4494, 122.93361) と最北端の緯度，経度 (45.5572, 148.752) からなる矩形を対象とし，$div_{lng} = 1,000$，$div_{lat} = 1,000$，$div_{time} = 24$ の 24,000,000 分割を行った．各時空間グリッドについて，2013 年 12 月 13 日から 23 日の 3,301,605 件のジオタグ付きツイートの投稿数を数えた．

分類器の教師データは，6 月 4 日に投稿されたキーワードとして「雨」を含む，“relevant” クラスの 1,458 件と “irrelevant” クラスの 1,097 件，計 2,555 件のジオタグ付きツイートを用いた．教師データについて，交差検定（分割数は 5，10，20，25，50 分割）を行った結果，精度は約 75%，再現率は約 83%，正解率は約 74%を示した．

次に，DBASTC と DBSTC で検出率の比較を行う．6 月と 7 月に新聞で報道されていた大雨であった 75 地域を時空間クラスタとして検出できるか評価を行った．図 1 に検出率を示す．図 1 より，DBASTC は DBSTC よりも高検出率であることを示した．


(a) DBSTC


(b) DBASTC

図 2: 検出された地域

図 2 に，7 月 3 日午前 8 時 30 分に九州地方で抽出された時空間クラスタを地図上に示す．丸で示した 2 つの時空間クラスタ（福岡県朝倉市と中津市）は報道されていた大雨であった地域であり，DBASTC でのみ抽出することができた．

## 4 まとめ

本論文では，$(\epsilon, \tau)$-密度に基づく適応的な時空間クラスタリング手法を用いたトピックの時空間分析手法を提案した．評価実験において先行研究の手法と比較を行った結果，提案手法は高検出率で大雨であった地域を検出することができた．これからの課題として，検出した地域をユーザへ通知するための手法などが考えられる．

## 謝辞

本研究の一部は，JSPS 科研費 26330139 と広島市立大学・特定研究費（一般研究，研究課題名「時空間文書ストリーム上におけるバースト領域の抽出手法」）の支援により行われた．

## 参考文献

[1] Tatsuhiro Sakai and Keiichi Tamura. Identifying bursty areas of emergency topics in geotagged tweets using density-based spatiotemporal clustering algorithm. In *proceedings of IWCIA 2014*, pp. 95–100, Nov 2014.