時空間的な投稿数を考慮した密度に基づく適応的な時空間クラスタリング手法

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　〒 731-3194 広島県広島市安佐南区大塚東 3-4-1
E-mail: †da65003@e.hiroshima-cu.ac.jp, ††{ktamura,kitakami}@hiroshima-cu.ac.jp

あらまし　ソーシャルメディア上に投稿される位置情報付きのデータを用いて実世界で注目を集めているトピックの時空間的な変遷を分析する研究が行われている．我々は先行研究において，密度に基づく時空間クラスタリングを用いてトピックの時空間的な変遷をモニタリングする手法を提案している．先行研究のモニタリング手法で使用していた密度に基づく時空間クラスタリングでは，投稿数の多い地域と少ない地域，また投稿数の多い時間帯と少ない時間帯がある場合に，適切に時空間クラスタを抽出できないため，トピックの発生を正確に把握することができなかった．本論文では，この問題点を解決するために，密度に基づく適応的な時空間クラスタリングを提案する．提案手法は，時空間クラスタの抽出の基準となる閾値を各地域と各時間帯において適応的に変化させることで，投稿数が多い地域と投稿数が少ない地域，また時間帯を区別することなく，時空間クラスタを抽出することができる．提案手法をモニタリング手法に導入し評価実験を行った結果，提案手法は従来手法と比較して，より高性能にトピックの発生を捉えることができた．

キーワード　時空間クラスタリング，密度に基づくクラスタリング，Twitter，ジオタグ付きツイート，ソーシャルメディア

1. はじめに

ビッグデータへの関心の高まりとともに，ソーシャルメディア上に投稿されるデータを用いて実世界で注目を集めているトピックを分析する研究が行われている．また，GPS 付きスマートフォンの普及により，位置情報付きのデータがソーシャルメディア上に盛んに投稿されており，位置情報付きのデータを利用することで，トピックの時間変化だけでなく，空間的な変遷も分析が可能となってきている [1]．例えば，Twitter 上では，台風，大雨や大雪などの自然災害発生時にそれらの状況を伝えるジオタグ付きツイートが投稿されている [2], [3], [4]．このジオタグ付きツイートを用いることで，自然災害が発生している地域と当該事象の時間変化把握することができる．

我々は，先行研究 [5] において，Twitter 上に投稿されるジオタグ付きツイートを対象として，密度に基づく時空間クラスタリングを用いてトピックの時空間的な変遷をモニタリングする手法を提案している．この手法では，モニタリングの対象となっているトピックを含むツイートをナイーブベイズ分類器を用いて取り出す．そして，当該トピックに関連するツイートが盛んに投稿されている地域を，密度に基づく時空間クラスタリングを用いて時空間クラスタとして抽出することで，当該トピックの時空間的な変遷をモニタリングすることができる．

ここで，トピックの時空間的な変遷とは，ある地域においてトピックが発生，変化と消滅することを指す．例えば，トピックを大雨としたとき，大雨の降り始めをトピックの発生，大雨の降っている地域の拡大や移動などをトピックの変化，大雨が止むことをトピックの消滅とする．先行研究では，時空間クラスタの抽出と更新を行うことでトピックの時空間的な変遷を捉えることができる．ある地域で大雨が降り始めるとその地域では人々が盛んに大雨に関する投稿を始める．時空間的に密に投稿されたジオタグ付きツイート集合を時空間クラスタとして抽出することで，トピックの発生を捉えることができる．また，トピックの変化に合わせて，抽出される時空間クラスタの大きさが変化する．そして，時空間クラスタが抽出されなくなった地域は大雨に関するトピックが消滅したことを示す．

先行研究において評価実験を行った結果，大雨と大雪に関するトピックの時空間的な変遷をモニタリングできることを確認した．しかしながら，地域や時間帯によって投稿数に差異がある場合に，時空間クラスタの抽出の基準となる閾値を適切に設定することが困難になるという問題が生じていた．図 1 に問題が起こる例を示す．図 1 の左側は普段の投稿数が多い地域であり，右側は普段の投稿数が少ない地域であるとする．ここで，データ集合 A または B に合わせて閾値を設定すると，データ集合 A と B が時空間クラスタとして抽出される．閾値を下げ，データ集合 C または D に合わせて閾値を設定すると，データ集合 A，B，C と D が時空間クラスタとして抽出される．データ集合 C と D は同じ程度でデータが密集しているが，データ集合 C は普段の投稿数と比べると高密度であり，データ集合 D は普段の投稿数と比べると低密度であるという違いがある．このように地域や時間帯によって普段の投稿数を考慮した相対的

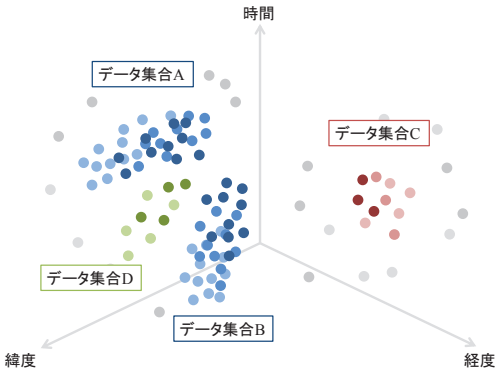


図 1　先行研究の問題点

な密度には違いがあり，データ集合 A，B と C が時空間クラスタとして抽出されるような閾値を設定する方法を考える必要がある．

本論文では，この問題点を解決するために，密度に基づく適応的な時空間クラスタリングを提案する．密度に基づく適応的な時空間クラスタリングは，各地域，各時間帯における過去の投稿数を考慮し，基準となる閾値を適応的に変化させる．密度に基づく適応的な時空間クラスタリングを用いることで，投稿数が多い地域と投稿数が少ない地域，また時間帯を区別することなく，時空間クラスタを抽出することができる．

提案手法をモニタリング手法に導入し，評価実験を行った．評価実験の結果，提案手法は従来手法と比較して，より高性能にトピックの発生を捉えることができた．

本論文の構成は以下の通りである．第 2 章では，関連研究を述べる．第 3 章では，先行研究である (ϵ, τ)-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モニタリング手法とその問題点を示す．第 4 章では，提案手法である (ϵ, τ)-密度に基づく適応的な時空間クラスタリングについて述べる．第 5 章では，評価実験の実験結果を示し，第 6 章で本論文をまとめる．

2. 関 連 研 究

近年，ソーシャルメディアサイトの発展とスマートフォンの普及により，ソーシャルメディアサイト上の位置情報付きのデータは急速に増加している [6]．その中でも，最も普及しているソーシャルメディアサイトの一つである Twitter 上のジオタグ付きツイートを用いて，実世界で起こっているトピックを分析する研究が行われている [1]．例えば，Sakaki ら [7] は，位置情報付きのツイートから台風の軌道と地震の震源地を予測する手法を提案し，Ozdikis ら [8] も，Twitter 上に投稿される地震に関する情報から地震の場所を推定する研究を行った．これらのように，ジオタグ付きのツイートを用いることで地震や台風などの自然災害をはじめとした緊急性のあるトピックの分析が可能となる [9], [10], [11].

Murakami ら [12] は，2011 年に発生した東日本大震災時に投稿されたツイートを分析した．Vieweg ら [3] は，火災や洪水などの緊急的な状況に関する情報が Twitter 上に投稿されることを示した．Karimi ら [13] は，自然災害に関する内容のツイートの分類手法を提案している．しかしながら，これらの研究はツイートの分析や分類のみに焦点を当てている．本研究では，ツイート分類と時空間クラスタリングを組み合わせることで，トピックの時空間的な変遷のモニタリングを可能にする．

また，Thom ら [14] は，ジオタグ付きツイートから異常を検出し，対話型のクラウドシステム上で可視化することで，発生した地震などの自然災害がどのくらい影響があるのか提示するシステムを開発した．Aramaki ら [15] は，ツイートにサポートベクターマシン（SVM）やナイーブベイズなどの分類器を適用し，インフルエンザの流行を検出するための手法を提案した．Aramaki らの手法では，各地域の投稿数の増減を求めることで，インフルエンザの影響度を提示している．Aramaki らと Thom らの研究の研究は，我々の研究に最も近い研究であるが，彼らのシステムはトピックがどういう状況なのかということを把握することができない．本研究では，時空間クラスタリングによってトピックの変遷を確認でき，ジオタグ付きツイートの本文からトピックの状況を把握することができる．

Avvenuti ら [16] は，各地域の地震による損害を把握するための EARS（Earthquake alert and report system）というシステムを開発している．Kim ら [17] は，トピックの時空間的な変化を可視化する mTrend というシステムを提案した．Kumar ら [18] は，道路上で発生した危険な状況を Twitter 上のユーザを用いて検出した．しかしながら，これらの手法は地域や時間帯によって投稿数が異なることを考慮できていない．

また，密度に基づくクラスタリングのパラメータの変化によるクラスタリング結果の違いを分析するための手法，OPTICS が提案されている [19]．OPTICS は，密度に基づくクラスタリングの距離の基準となるパラメータについて，様々な値を設定した際に各パラメータ値でどのようなクラスタリング結果が得られるのか，分析することができる．OPTICS を用いることで，データセットごとにパラメータを適切な値に設定することができる．しかしながら，OPTICS では，クラスタリングを行う際には分析結果から一つのパラメータ値を決めてクラスタリングを行う．本論文の提案手法では，一つの値を定めるが，そのパラメータの値を各地域，各時間帯によって適応的に変化させることで，データセットの中で適切な時空間クラスタを抽出することができる．

3. (ϵ, τ)-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モニタリング手法

本章では，(ϵ, τ)-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モニタリング手法 [5] について説明する．

3.1 モニタリング手法の概要

図 2 に (ϵ, τ)-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モニタリング手法の概要図を示す．モニタリングでは，以降の処理を一定時間毎に実行する．ジオタグ付きツイートクローラを用いて Twitter からジオタグ付きツイートを収集し，ジオタグ付きツイートデータベースに保存する．ナイーブベイズ分類器 [20] を用いてモニタリング対象のトピックに関連する

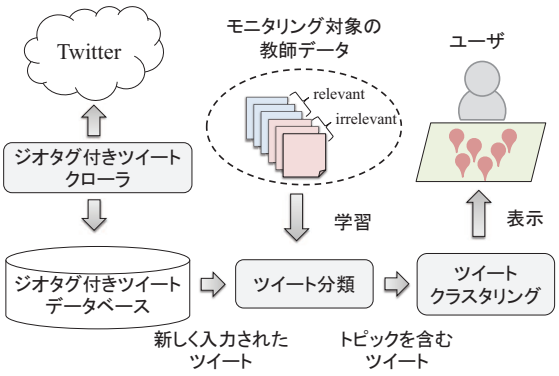


図 2　モニタリング手法の概要

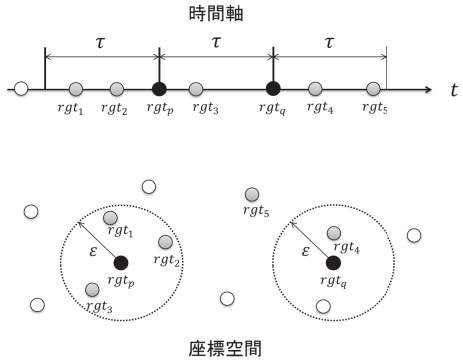


図 3　定義 1 と 3 の例

関連ジオタグ付きツイートとそれ以外のツイートに分類を行う．関連ジオタグ付きツイートとこれまでに抽出された時空間クラスタ集合を入力として，(ϵ,τ)-密度に基づく時空間クラスタリングを用いて新しく時空間クラスタ集合を抽出する．(ϵ,τ)-密度に基づく時空間クラスタリングについては，3.3 節で説明する．抽出された時空間クラスタの内容を地図上に表示し，ユーザへ提示する．

3.2　データ定義

ジオタグ付きツイートを gt_i と表記し，その集合を $GTS = \{gt_1, \cdots, gt_n\}$ とする．ここで，gt_i は文書データ $text_i$，投稿時間 pt_i と位置情報 pl_i の 3 つから構成される．本研究では位置情報として経度，緯度を用いる．また，モニタリングの対象としているトピックの内容を含むジオタグ付きツイートを関連ジオタグ付きツイート rgt_j（$= gt_{\phi(j)}$）と呼ぶ．関連ジオタグ付きツイート集合を $RGTS = \{rgt_1, \cdots, rgt_m\}$ とすると，GTS は $RGTS$ を包含しており（$RGTS \subset GTS$），次の単射で表現される．

$$\phi(j) : RGTS \to GTS;\ rgt_j \mapsto gt_{\phi(j)} \qquad (1)$$

3.3　(ϵ,τ)-密度に基づく時空間クラスタリング

密度に基づく空間クラスタリング [21], [22] では，データが密集している部分を空間クラスタ，密集していない部分を空間クラスタではないと定義し，空間クラスタを抽出する．(ϵ,τ)-密度に基づく時空間クラスタリング [23] は，密度に基づく空間クラスタリングを拡張し，ジオタグ付きツイートの位置情報と投稿時間に着目し，距離が ϵ 以内であり，投稿間隔が τ 以内であるジオタグ付きツイートを (ϵ,τ)-密度に基づく近傍と定義している．

［定義 1］（(ϵ,τ)-密度に基づく近傍）　関連ジオタグ付きツイート rgt_p の (ϵ,τ)-密度に基づく近傍を $STN_{(\epsilon,\tau)}(rgt_p)$ と表記し，以下のように定義する．

$$STN_{(\epsilon,\tau)}(rgt_p) = \{rgt_q \in RGTS \mid dist(rgt_p, rgt_q) \leqq \epsilon \ \ and \ \ iat(rgt_p, rgt_q) \leqq \tau\} \quad (2)$$

関数 $dist$ は経度・緯度などの座標値を使って，ジオタグ付きツイート間の空間上の距離を求める関数，関数 iat はジオタグ付きツイート間の投稿間隔を求める関数である．

図 3 に例を示す．関連ジオタグ付きツイート rgt_p から距離 ϵ 以内に，rgt_1，rgt_2 と rgt_3 の 3 つ，また，rgt_p から投稿間隔 τ 以内にも，rgt_1，rgt_2 と rgt_3 の 3 つの関連ジオタグ付きツイートが存在する．この時，rgt_p の (ϵ,τ)-密度に基づく近傍は，rgt_1，rgt_2 と rgt_3 の 3 つとなる．また，rgt_q の (ϵ,τ)-密度に基づく近傍は，rgt_4 の 1 つとなる．

ここで，時空間クラスタの核となる関連ジオタグ付きツイートの (ϵ,τ)-密度に基づく近傍には，$MinRGT$（ユーザパラメータ）以上の関連ジオタグ付きツイートが存在すると定義する．そして，(ϵ,τ)-密度に基づく時空間クラスタリングでは核となる関連ジオタグ付きツイートの (ϵ,τ)-密度に基づく近傍を結合していくことで，クラスタリングを行う．

3.4　先行研究の問題点

先行研究で用いられている (ϵ,τ)-密度に基づく時空間クラスタリングは，時空間的に高密度な時空間クラスタを抽出するために有効な手法である．しかしながら，データの分布によっては，時空間クラスタを抽出することが困難な場合がある．これは，地域や時間帯によって投稿数には差異があり，適切な $MinRGT$ を設定しなければならないからである．投稿数の多い場合と少ない場合に分けて $MinRGT$ を設定する方法が考えられるが，このような時空間的な投稿数の差異は三次元的に複雑に生じているため，手作業で行うのは困難である．例えば，高密度な地域に合わせて $MinRGT$ を設定すると，図 1 のデータ集合 C のような，高密度な地域から考えれば低密度であるが周辺の投稿数から考えると高密度な地域，つまり，局所的に高密度な時空間クラスタを抽出できない．反対に，低密度な地域に合わせて $MinRGT$ を設定すると，投稿数の多い地域では，図 1 のデータ集合 D のような，周辺の投稿数から考えると低密度な時空間クラスタが抽出されてしまう．

4.　(ϵ,τ)-密度に基づく適応的な時空間クラスタリング

本章では，提案手法である (ϵ,τ)-密度に基づく適応的な時空間クラスタリングについて述べる．

4.1　概　　要

3.4 節であげた問題点を解決するためには，各地域，各時間

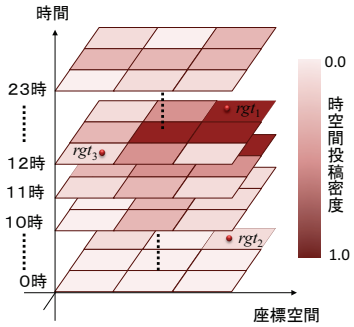


図 4　時空間投稿密度の例

帯によって適切な閾値を設定しなければならない．(ϵ, τ)-密度に基づく適応的な時空間クラスタリングでは，各地域，各時間帯によって適応的に変化する新しい閾値を定義する．具体的に，各地域，各時間帯の過去における投稿数を時空間投稿密度として定義し，時空間投稿密度を用いて閾値を適応的に変化させる．投稿数の多い地域，投稿数の多い時間帯では閾値は高く，投稿数の少ない地域，投稿数の少ない時間帯では閾値は低く設定される．したがって各地域，各時間帯にとって適切な閾値が設定することができる．

4.2　適応的な閾値

本節では，適応的な閾値について説明する．

［定義 2］（適応的な閾値）　関連ジオタグ付きツイート rgt_p の適応的な閾値を $AT(rgt_p, MaxMinRGT)$ と表記し，以下のように定義する．

$$AT(rgt_p, MaxMinRGT) = (MaxMinRGT - 1) \times lstd(rgt_p) + 1 \quad (3)$$

関数 $lstd(rgt_p)$ は rgt_p の時空間投稿密度を返す関数であり $(0 \leqq lstd(rgt_p) \leqq 1)$，$MaxMinRGT$ はユーザが与えるパラメータである．

時空間投稿密度は過去に投稿されたジオタグ付きツイートの統計量から算出する．まず，対象とする時空間領域を 3 次元の時空間グリッドに分割（$div_{lng} \times div_{lat} \times div_{time}$）する．次に，各時空間グリッドに含まれる過去に投稿されたジオタグ付きツイートの数をカウントする．関数 $lstd$ は次の式で正規化した値を返す．

$$lstd(rgt_p) = \frac{stnum(geo\_gid(rgt_p)) - stnum_{min}}{stnum_{max} - stnum_{min}} \quad (4)$$

ただし，関数 $stnum(i)$ は時空間グリッド i のジオタグ付きツイート数を返す関数である．$geo\_gid(rgt_p)$ は rgt_p が属する時空間グリッドの ID を返す関数である．$stnum_{max}$ と $stnum_{min}$ は最大数と最小数である．

図 4 に例を示す．この例では，全時空間が $3 \times 3 \times 24$ の時空間に分割され，各地域，また，各時間帯において時空間投稿密度が変化している．関連ジオタグ付きツイート rgt_1 と rgt_2 とは同一地域に存在しているが，時間帯が異なる．関連ジオタグ付きツイート rgt_1 は日中であり，時空間投稿密度が rgt_2 よりも高い．関連ジオタグ付きツイート rgt_3 は rgt_1 と同一時間帯であるが，地域が異なり，普段の投稿数が少ない地域であるため，時空間投稿密度が小さくなっている．

4.3　諸　定　義

本節では，(ϵ, τ)-密度に基づく適応的な時空間クラスタリングの諸定義について説明する．

［定義 3］（核/周辺 関連ジオタグ付きツイート）　関連ジオタグ付きツイート rgt_p について，$|STN_{(\epsilon,\tau)}(rgt_p)| > AT(rgt_p, MaxMinRGT)$ を満たすとき，rgt_p を核関連ジオタグ付きツイート，反対に，$|STN_{(\epsilon,\tau)}(rgt_p)| \leqq AT(rgt_p, MaxMinRGT)$ であるとき，rgt_p を周辺関連ジオタグ付きツイートと呼ぶ．

図 3 を使い，例を示す．$MaxMinRGT = 3$，$lstd(rgt_p) = 0.8$，$lstd(rgt_q) = 0.8$ とすると，$AT(rgt_p, MaxMinRGT) = 2.6$，$AT(rgt_q, MaxMinRGT) = 2.6$ となる．つまり，関連ジオタグ付きツイート rgt_p は核関連ジオタグ付きツイートであり，関連ジオタグ付きツイート rgt_q は周辺関連ジオタグ付きツイートである．

［定義 4］（(ϵ, τ)-密度に基づいて適応的に直接到達可能）　関連ジオタグ付きツイート rgt_q が rgt_p の (ϵ, τ)-密度に基づく近傍に存在し，$|STN_{(\epsilon,\tau)}(rgt_p)| > AT(rgt_p, MaxMinRGT)$ を満たす時，関連ジオタグ付きツイート rgt_q は rgt_p から (ϵ, τ)-密度に基づいて適応的に直接到達可能であると表現する．

［定義 5］（(ϵ, τ)-密度に基づいて適応的に到達可能）　関連ジオタグ付きツイート rgt_{p+1} が rgt_p から (ϵ, τ)-密度に基づいて適応的に直接到達可能である，関連ジオタグ付きツイート列 $(rgt_p, rgt_{(p+1)}, \cdots, rgt_{(p+l)})$ を考える．この時，関連ジオタグ付きツイート $rgt_{(p+l)}$ は rgt_p から，(ϵ, τ)-密度に基づいて適応的に到達可能であると表現する．

［定義 6］（(ϵ, τ)-密度に基づいて適応的に接続）　関連ジオタグ付きツイート rgt_p と rgt_q とが rgt_o から (ϵ, τ)-密度に基づいて適応的に到達可能であり，rgt_o が $|STN_{(\epsilon,\tau)}(rgt_o)| > AT(rgt_o, MaxMinRGT)$ を満たす時，rgt_p と rgt_q とは (ϵ, τ)-密度に基づいて適応的に接続していると表現する．

4.4　(ϵ, τ)-密度に基づく適応的な時空間クラスタ

(ϵ, τ)-密度に基づく適応的な時空間クラスタリングでは，時空間的に密集している関連ジオタグ付きツイート集合を (ϵ, τ)-密度に基づく適応的な時空間クラスタと定義する．

［定義 7］（(ϵ, τ)-密度に基づく適応的な時空間クラスタ）　関連ジオタグ付きツイート集合 $RGTS$ において，(ϵ, τ)-密度に基づく適応的な時空間クラスタ $astc$ は次の 2 つの条件を満たす部分関連ジオタグ付きツイート集合である．

(1)　任意の関連ジオタグ付きツイート $rgt_p \in RGTS$ と $rgt_q \in RGTS$ について，(ϵ, τ)-密度に基づく適応的な時空間クラスタ $astc$ に関連ジオタグ付きツイート rgt_p が所属 $(rgt_p \in astc)$ し，関連ジオタグ付きツイート rgt_q が rgt_p から (ϵ, τ)-密度に基づいて適応的に到達可能であれば，rgt_q は $astc$ に所属（$rgt_q \in astc$）する．

(2)　(ϵ, τ)-密度に基づく適応的な時空間クラスタ $astc$ に所属する任意の関連ジオタグ付きツイート $rgt_p \in astc$ と

$rgt_q \in astc$ は，(ϵ, τ)-密度に基づいて適応的に接続している．

4.5 アルゴリズム

(ϵ, τ)-密度に基づく適応的な時空間クラスタリングのアルゴリズムを Algorithm1 に示す．新しく取得した関連ジオタグ付きツイート，これまでに取得した関連ジオタグ付きツイート集合，前回抽出された時空間クラスタ集合（$ASTCS$）を入力し，更新された時空間クラスタ集合（$ASTCS'$）を出力する．関連ジオタグ付きツイートを取得したとき，新たに取得した関連ジオタグ付きツイートの (ϵ, τ)-密度に基づく近傍に存在する関連ジオタグ付きツイートのみに影響があるため，それらの関連ジオタグ付きツイートを対象に再クラスタリングを行う．また，再クラスタリングの際に，二つの (ϵ, τ)-密度に基づく適応的な時空間クラスタが結合し，一つとなることもある．

Algorithm1 を詳しく説明する．

（1） 新しく取得した関連ジオタグ付きツイート rgt の (ϵ, τ)-密度に基づく近傍を取得し，再クラスタリングの対象として $NRGT$ に挿入する．

（2） $NRGT$ から関連ジオタグ付きツイート $nrgt$ を 1 つ取り出す．ただし，$NRGT$ が空ならば（9）へ進む．

（3） $nrgt$ が核関連ジオタグ付きツイートであるかチェックする．もし，核関連ジオタグ付きツイートであれば（4）を実行し，核関連ジオタグ付きツイートでなければ（2）に戻る．

（4） $nrgt$ が時空間クラスタに所属していれば，$nrgt$ の所属する時空間クラスタを取得し，$astc$ とする．時空間クラスタに所属していなければ，新たに時空間クラスタ $astc$ を作成する．

（5） $nrgt$ の (ϵ, τ)-密度に基づく近傍を関連ジオタグ付きツイート集合としてキュー Q に挿入する．

（6） Q から関連ジオタグ付きツイート $qrgt$ を取り出し，次の処理を行う．

（a） $qrgt$ が核関連ジオタグ付きツイートであるかチェックし，核関連ジオタグ付きツイートであれば，(b) へ移る．核関連ジオタグ付きツイートでなければ（7）へ進む．

（b） $qrgt$ が時空間クラスタに所属していれば（i）を，時空間クラスタに所属していなければ，(ii) を実行する．

i. $qrgt$ の所属する時空間クラスタ $astc'$ と $astc$ を結合する．

ii. $qrgt$ を $astc$ に挿入する．$qrgt$ の (ϵ, τ)-密度に基づく近傍から，キュー Q に存在しない関連ジオタグ付きツイートを Q に挿入する．

（7） キュー Q が空であれば（8）へ移る．空でなければ（6）へ戻る．

（8） 時空間クラスタ集合 $ASTCS$ に $astc$ を加え，(2) へ戻る．

（9） $ASTCS$ から，現在時刻から τ 前までに更新がない時空間クラスタを削除し，$ASTCS'$ とする．$ASTCS'$ を更新された時空間クラスタ集合として出力する．

5. 評価実験

提案手法を評価するために，評価実験を行った．本章では，評価実験の結果を示す．

input : rgt - 新しく取得した関連ジオタグ付きツイート, $RGTS$ - これまでに取得した関連ジオタグ付きツイート集合, $ASTCS$ - 前回抽出した時空間クラスタ集合, ϵ, τ, $MaxMinRGT$ - ユーザパラメータ
output: $ASTCS'$ - 更新された時空間クラスタ集合

```
NRGT ← GetNeighborhood(rgt,ε,τ);
for i ← 1 to |NRGT| do
    nrgt ← nrgt_i ∈ NRGT;
    STN ← GetNeighborhood(nrgt,ε,τ);
    if |STN| > AT(nrgt, MaxMinRGT) then
        if IsClustered(nrgt) == false then
            astc ← MakeNewCluster(nrgt);
        end
        else
            astc ← GetCluster(nrgt,ASTCS);
        end
        EnQueue(Q,STN);
        while Q is not empty do
            qrgt ← DeQueue(Q);
            STN ← GetNeighborhood(qrgt,ε,τ);
            if |STN| >AT(qrgt, MaxMinRGT) then
                if IsClustered(qrgt) == true then
                    astc' ← GetCluster(qrgt,ASTCS);
                    astc ← AppendClusters(astc,astc');
                end
                else
                    astc ← astc ∪ qrgt;
                    EnNniqueQueue(Q,STN);
                end
            end
        end
        ASTCS ← ASTCS ∪ astc;
    end
end
ASTCS' ← DeleteOldCluster(ASTCS);
return ASTCS';
```

Algorithm 1: (ϵ, τ)-密度に基づく適応的な時空間クラスタリングアルゴリズム

5.1 データセットと実験内容

評価実験では，モニタリング手法の時空間クラスタリング部分について，従来手法 [5] と提案手法を用いた場合で比較を行う．モニタリングの対象とするトピックを「大雨」と「大雪」とし，2014 年 6 月と 7 月に投稿されたジオタグ付きツイートを用いてトピック「大雨」を，2014 年 1 月と 2 月に投稿されたジオタグ付きツイートを用いてトピック「大雪」の評価を行う．

パラメータとしては，$\epsilon = 5km$，$\tau = 3600sec$ を用いた．従来手法のパラメータ $MinRGT$ と提案手法のパラメータ $MaxMinRGT$ はそれぞれ 2 から 10 まで変化させて実験を行った．時空間投稿密度を求める対象領域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) からなる矩形とし，パラメータは $div_{lng} = 1,000$,

表 1　大雨が観測された 16 日間に収集したツイート数

日付	ツイート数	関連ツイート数
2014/6/4	325095	2249
2014/6/6	312145	6401
2014/6/7	330540	4433
2014/6/13	346507	2589
2014/6/16	340675	750
2014/6/22	411863	4172
2014/6/23	355384	700
2014/6/25	393441	2331
2014/6/29	441959	4838
2014/7/3	341770	4738
2014/7/7	376734	4173
2014/7/8	366887	1405
2014/7/9	374707	4704
2014/7/10	395061	4803
2014/7/11	383704	1763
2014/7/19	412403	5369

表 2　大雪が観測された 11 日間に収集したツイート数

日付	ツイート数	関連ツイート数
2014/1/10	282370	2665
2014/1/14	284215	981
2014/1/17	283809	995
2014/2/6	284065	2821
2014/2/8	350867	27823
2014/2/11	289628	3564
2014/2/13	306106	3953
2014/2/14	378368	21834
2014/2/15	256378	10060
2014/2/16	307708	5121
2014/2/18	262145	2325

$div_{lat} = 1{,}000$，$div_{time} = 24$ を用いた．また，統計データとして 2013 年 12 月 13 日から 23 日の間に投稿された 3,301,605 件のジオタグ付きツイートを用いて，時空間投稿密度を算出した．

5.2　再現率の比較

まず，投稿数の少ない地域や時間帯で時空間クラスタを抽出することができたかについて，新聞報道に基づき評価を行う．実験期間に日本で大雨が観測された 16 日間，大雪が観測された 11 日間に注目し，「大雨」と「大雪」に関するトピックが報道されている新聞記事 (注1) から，記事に記載されている地域（市町村）を抽出した．新聞記事から抽出した各地域から，本実験で用いたデータセット中でジオタグ付きツイートが 1 件も投稿されていない地域は取り除いた．その結果，6 月と 7 月に大雨と報道されていた地域は 77，1 月と 2 月に大雪と報道されていた地域は 89 となった．なお，同じ地域であっても別々の日に報道があれば別々にカウントしている．

表 1 と表 2 に実験期間の 16 日間と 11 日間に収集したジオタグ付きツイート数とツイート分類によって抽出された関連ジオタグ付きツイート数を示す．また，各 $MinRGT$ と $MaxMinRGT$ において抽出された時空間クラスタ数を表 3 と表 4 に示す．表 3 と表 4 には，各 $MinRGT$ と $MaxMinRGT$ において，各日付で抽出された時空間クラスタ数を合計した値を示している．

「大雨」,「大雪」と報道された地域を検出できたかどうかの判定は，対象の地域において時空間クラスタが抽出され，ツイートの内容が当該トピックの内容と一致していれば検出できたと判定した．抽出された時空間クラスタとツイートの内容の確認は人手によって行った．

(注1)：2014 年 1 月，2 月，6 月と 7 月に発刊された朝日新聞の朝刊と夕刊

表 3　従来手法の抽出クラスタ数

$MinRGT$	「大雨」のクラスタ数	「大雪」のクラスタ数
2	1325	1762
3	928	1150
4	670	809
5	534	633
6	420	497
7	342	437
8	304	375
9	246	320
10	224	280

表 4　提案手法の抽出クラスタ数

$MaxMinRGT$	「大雨」のクラスタ数	「大雪」のクラスタ数
2	2163	2879
3	2125	2813
4	1880	2424
5	1714	2291
6	1598	2039
7	1485	1897
8	1372	1750
9	1306	1720
10	1202	1652

図 5 と図 6 にトピック「大雨」とトピック「大雪」の再現率をそれぞれ示す．図 5 と図 6 は，各 $MinRGT$，各 $MaxMinRGT$ における実験結果を，横軸を抽出クラスタ数，縦軸を再現率とした散布図で示している．図 5 より，提案手法を用いた場合のトピック「大雨」の再現率は $MaxMinRGT$ を 2 から 10 まで変化させても約 80%から大きな変化がないのが分かる．一方，従来手法を用いた場合，トピック「大雨」の再現率は約 70%から約 30%まで落ちている．また，従来手法で抽出クラスタ数 1325，提案手法で抽出クラスタ数 1306 の場合を比べると，提案手法が抽出クラスタ数が少ないにもかかわらず，再現率は高くなっている．図 6 より，トピック「大雪」についても，トピック「大雨」ほどの差はないが，提案手法が従来手法よりも高再現率であることが分かる．以上の結果より，従来手法で $MinRGT$ を増加させると高密度な地域のみが抽出されるが，提案手法では適応的に閾値が地域，また，時間帯によって変化するため，再現率の低下を防ぐことができたといえる．

しかしながら，検出できていない地域も存在する．検出できなかった理由としては，対象の地域に複数のジオタグ付きツイートが投稿されていたとしてもジオタグ付きツイート間の距離や投稿間隔が，パラメータ $\epsilon = 5km$ と $\tau = 3600sec$ より離れていることがあり，時空間クラスタを抽出できなかった．今後，パラメータ ϵ と τ を適応的に変化させる方法を導入する必要がある．

5.3　抽出された時空間クラスタの評価

抽出された時空間クラスタを地図上で確認し，抽出地域とツイート内容の評価を行う．なお，この実験では $MinRGT = 5$，$MaxMinRGT = 5$ として行う．

図 7(a) と図 7(b) に，7 月 3 日午前 7 時に北九州にて抽出された時空間クラスタを地図上に示す．地図上の時空間クラスタは，中心点のみを傘マークまた雪マークのアイコンで示している．赤丸で示した時空間クラスタは提案手法でのみ抽出され，7 月 3 日の夕刊と 7 月 4 日の朝刊において，7 月 3 日の午前中

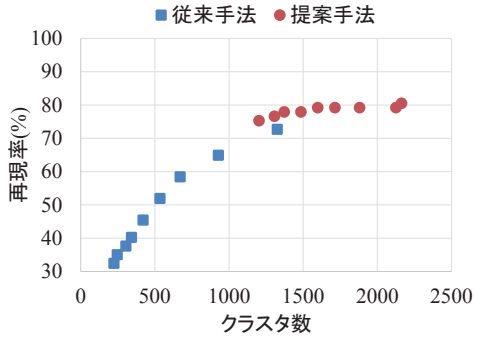


図 5　トピック「大雨」の再現率

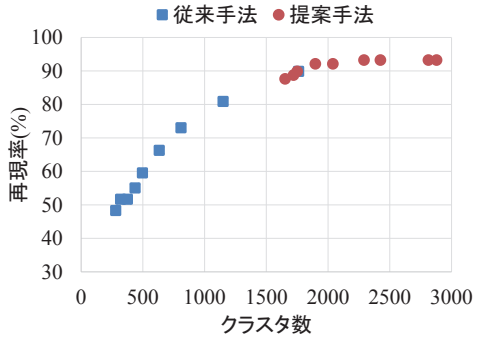


図 6　トピック「大雪」の再現率

(a) 従来手法を用いた結果　　(b) 提案手法を用いた結果

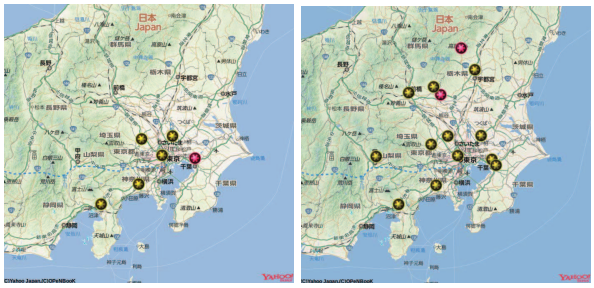
図 7　7 月 3 日に九州地方で抽出された時空間クラスタ

に大雨であったと報道されていた地域（福岡県朝倉市と大分県中津市）である．表 5 に，この二つの時空間クラスタに含まれていた全てのジオタグ付きツイートを示す．表 5 より，「大雨」である状況を伝えているジオタグ付きツイートを抽出できているのが分かる．

図 8(a) と図 8(b) に，2 月 8 日午前 4 時に関東地方にて抽出された時空間クラスタを地図上に示す．2 月 8 日と 9 日の朝刊では，2 月 8 日の未明から関東地方で大雪が観測されたと報道されている．図 8(a) と図 8(b) より，深夜の投稿が少ない時間帯において提案手法は従来手法と比較して多くの時空間クラスタが抽出できており，トピック「大雪」の発生を捉えることができた．

6. ま と め

本論文では，(ϵ, τ)-密度に基づく適応的な時空間クラスタリングを提案した．(ϵ, τ)-密度に基づく適応的な時空間クラスタリングでは各地域，また各時間の統計的な投稿密度を用いることで，時空間クラスタを抽出するときに基準となる閾値を適応的に変化させている．提案手法を用いることで，投稿数が多い地域と少ない地域や時間帯を区別することなく時空間クラスタを抽出することができる．提案手法を実際に実装し，Twitter 上に投稿されたジオタグ付きツイートを用いて，トピックを「大雨」と「大雪」と設定し，評価実験を行った．評価実験の結果，提案手法は従来手法と比較して，より高性能にトピックの発生を捉えることができた．



(a) 従来手法を用いた結果　　(b) 提案手法を用いた結果

図 8　2 月 8 日に関東地方で抽出された時空間クラスタ

これからの課題として，普段の投稿数が多い地域または時間帯において，抽出されるべきではない時空間クラスタを削除することができたかについて評価することがあげられる．評価実験において，再現率を求めることにより普段の投稿数が少ない地域または時間帯において，トピックが取り上げられている地域を時空間クラスタとして検出できることを評価したが，精度については評価を行えていないため，今後，行う必要がある．

謝　　辞

本研究の一部は，JSPS 科研究費 26330139 と広島市立大学・特定研究費の支援により行われた．

文　　献

[1] Jie Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *Intelligent Systems, IEEE*, Vol. 27, No. 6, pp. 52–59, Nov 2012.

[2] Krishna Y. Kamath, James Caverlee, Kyumin Lee, and Zhiyuan Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proc. 22nd International Conference on World Wide Web*, WWW '13, pp. 667–678, 2013.

[3] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 1079–1088, 2010.

[4] Mai Miyabe, Asako Miura, and Eiji Aramaki. Use trend analysis of twitter after the great east japan earthquake. In *Proc. ACM 2012 Conference on Computer Supported Cooperative Work Companion*, CSCW '12, pp. 175–178, 2012.

[5] Tatsuhiro Sakai and Keiichi Tamura. Real-time analysis application for identifying bursty local areas related to emergency topics. *SpringerPlus*, Vol. 4, No. 162, 2015.

[6] Mor Naaman. Geographic information from georeferenced

表 5　7 月 3 日の 7 時 47 分に朝倉市と中津市で抽出されたジオタグ付きツイート

ID	ツイート本文
1-1	雨やばー学校行くの怖ー
1-2	雨やばし (笑) 昼から結婚式の打ち合わせ (((o(*▽*)o))) 昼からやむって言いよるけど本当にやむとかな (￣Д￣)
1-3	雨ヤバイから学校休みにして~ お願い~
2-1	さて雨の中仕事いやだー たまにはゆとりてー
2-2	さて、仕事や！(;o;) しかも雨やー……朝からテンション下がるわー m(。≧Д≦。)m おなかすいた (笑) 今日頑張れば明日休みだ！頑張ろーっと (´・ω・‘)
2-3	もう、雨で止まるくらいやったら豊肥本線爆発せんかな笑
2-4	おはょん笑　大分は雨だよー

social media data. *SIGSPATIAL Special*, Vol. 3, No. 2, pp. 54–61, 2011.

[7] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proc. 19th International Conference on World Wide Web*, WWW '10, pp. 851–860, 2010.

[8] Ozer Ozdikis, Halit Oguztuzun, and Pinar Karagoz. Evidential location estimation for events detected in twitter. In *Proc. 7th Workshop on Geographic Information Retrieval*, GIR '13, pp. 9–16, 2013.

[9] Karl Kreiner, Aapo Immonen, and Hanna Suominen. Crisis management knowledge from social media. In *Proc. 18th Australasian Document Computing Symposium*, ADCS '13, pp. 105–108, 2013.

[10] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proc. First Workshop on Social Media Analytics*, SOMA '10, pp. 71–79, 2010.

[11] Cindy Hui, Yulia Tyshchuk, William A. Wallace, Malik Magdon-Ismail, and Mark Goldberg. Information cascades in social media in response to a crisis: A preliminary model and a case study. In *Proc. 21st International Conference Companion on WWW*, pp. 653–656, 2012.

[12] Akiko Murakami and Tetsuya Nasukawa. Tweeting about the tsunami?: Mining twitter for information on the tohoku earthquake and tsunami. In *Proc. 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pp. 709–710, 2012.

[13] Sarvnaz Karimi, Jie Yin, and Cecile Paris. Classifying microblogs for disasters. In *Proc. 18th Australasian Document Computing Symposium*, ADCS '13, pp. 26–33, 2013.

[14] D. Thom, H. Bosch, S. Koch, M. Worner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Proc. 2012 IEEE Pacific Visualization Symposium*, pp. 41–48, 2012.

[15] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proc. Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1568–1576, 2011.

[16] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. Ears (earthquake alert and report system): A real time decision support system for earthquake crisis management. In *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 1749–1758, 2014.

[17] Kyoung-Sook Kim, Ryong Lee, and Koji Zettsu. mtrend: discovery of topic movements on geo-microblogging messages. In *Proc. 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, pp. 529–532, 2011.

[18] Avinash Kumar, Miao Jiang, and Yi Fang. Where not to go?: Detecting road hazards using twitter. In *Proc. 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1223–1226, 2014.

[19] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. In *Proc. 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pp. 49–60, 1999.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 169–194, 1998.

[22] Martin Ester, Hans-Peter Kriegel, Joerg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[23] Keiichi Tamura and Takumi Ichimura. Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents. In *Proc. 2013 IEEE International Conference on Systems, Man, and Cybernetics*, SMC 2013, pp. 2079–2084, 2013.