密度に基づく時空間クラスタリング手法を用いた話題の地域分析アプリケーション

Topical-Area Analysis Application using Density-based Spatiotemporal Clustering Method

酒井 達弘
Tatsuhiro Sakai
広島市立大学大学院
情報科学研究科
E-mail: my67011@edu.ipc.hiroshima-cu.ac.jp

田村 慶一
Keiichi Tamura
広島市立大学大学院
情報科学研究科
E-mail: ktamura@hiroshima-cu.ac.jp

***Abstract*—Nowadays, with the increasing attention being paid to social media, a huge number of georeferenced documents, which include location information, are posted on social media sites. People transmit and collect information over the Internet through these georeferenced documents. Georeferenced documents are usually related to not only personal topics but also local topics and events. Therefore, extracting local topics and events from georeferenced documents is one of the most important challenges in different application domains. We have been developping density-based spatiotemporal algorithms for extracting topical area in georeferenced documents. In this paper, a prototype of topical areas analysis application using the density-based spatiotemporal algorithm for geo-local emergency topics.**

I. はじめに

ソーシャルメディアへの関心の高まりとともに，インターネット上のユーザは，日々の事柄だけでなく，社会的な話題やイベントに関するデータをソーシャルメディアサイト（TwitterやFacebookなど）に投稿し，情報発信を行うようになってきている．近年，GPS付きスマートフォンの普及によって，位置情報が付与されたデータが次々に生成され，位置に関連した地域的な話題やイベントに関わる情報発信が盛んに行われており [1], [2]，これらのデータは情報源としての新たなメディアを形成しつつある．そこで，地域的な話題や出来事を位置情報が付与されたデータから抽出すること [3] は，動向分析，マーケティング，観光情報をはじめとして，情報検索や情報推薦などにとって重要な課題のひとつとなっている．

我々は，ソーシャルメディアサイト上に投稿される位置情報が付与された文書データを対象とし，密度に基づくクラスタリング手法を応用して，地域的な話題を取り出す手法の開発を行っている [4], [5]．キーワードを含む文書データが盛んに投稿されている地域は，そのキーワードに関連したトピックやイベントに関連している可能性が高い．例えば，Flickr上で投稿されるジオタグが付与された画像データから，ランドマークや観光スポットを抽出する研究 [6]，Twitter上の地震に関するつぶやきから震源地を特定し揺れが予測される地域を特定する研究 [7] が行われている．

本論文では，あるキーワードを含むトピックが話題となっている地域を (ϵ, τ)-密度に基づく時空間クラスタリング手法 [4] と分類器を用いて分析するためのアプリケーションを提案する．提案するアプリケーションでは，Webサーバ側で位置情報が付与された文書データとしてTwitter上のジオタグ付きツイートを収集する．次に，分類器を利用して特定のキーワードを含むトピックに関連するツイートを抽出する．次に，(ϵ, τ)-密度に基づく時空間クラスタリング手法を用いて (ϵ, τ)-密度に基づく時空間クラスタをリアルタイムに抽出する．最後に，Webブラウザもしくは，Androidタブレット端末上のアプリケーションからWebサービスを通してデータを送信し，地図上に時空間クラスタの表示を行う．

Twitter上において「雪」が話題となっている地域をマップ上で表示し動向分析を行うための「雪マップ」をWebアプリケーションとして試作し，評価実験を行った．評価実験の結果,「雪」について話題となっている地域をリアルタイムに検出することができ，地図上に表示することで，話題となっている地域を容易に把握し分析することが可能であることを確認できた．

本論文の構成は以下の通りである．第 2 章では，(ϵ, τ)-密度に基づく時空間クラスタリング手法とそのインクリメンタルなアルゴリズムについて説明する．第 3 章では，(ϵ, τ)-密度に基づく時空間クラスタリング手法を用いた話題の地域分析アプリケーションの詳細について述べる．第 4 章で評価実験の結果を示し，第 5 章で本論文のまとめを行う．

II. (ϵ,τ)-密度に基づく時空間クラスタリング手法

本章では，リアルタイムで時空間クラスタを抽出するためのインクリメンタルな時空間クラスタリング手法を述べる．

A. 諸定義

密度に基づくクラスタリング手法では2つのデータ間の距離を定め，ある文書データ dp からの距離が ϵ 以内に存在する文書データ dq を ϵ-近傍 $N_\epsilon(dp)$ と定義する．(ϵ,τ)-密度に基づく時空間クラスタリング手法では，ϵ-近傍の定義を次のように再定義する．

定義 1 ((ϵ,τ)-近傍 $N_{(\epsilon,\tau)}(dp)$) 文書データ dp の (ϵ,τ)-**近傍**を $N_{(\epsilon,\tau)}(dp)$ と表記し，以下のように定義する．

$$N_{(\epsilon,\tau)}(dp) = \{dq \in D | dist(dp,dq) \leq \epsilon \ and \ alt(dp,dq) \leq \tau\} \quad (1)$$

関数 $dist$ は経度・緯度など座標値を使って，文書データ間の空間上の距離を求める関数，関数 alt は文書データ dp と文書データ dq 間の投稿間隔を求める関数である．

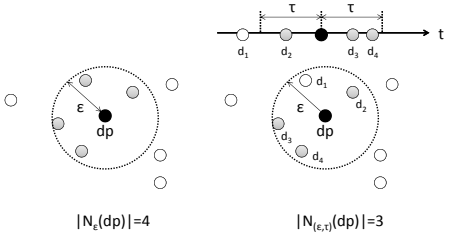


図 1. 定義 1 の例

図 1 の左は，ϵ-近傍の例であり，この例では，文書データは4つ存在し，$|N_\epsilon| = 4$ となる．一方，図 1 の右に (ϵ,τ)-近傍の例を示す．文書データ dp の (ϵ,τ)-近傍は，半径 ϵ 以内に存在する文書データでかつ，文書データ dp との投稿間隔が τ 以内の文書データである．この例では，文書データは3つ存在し，$N_{(\epsilon,\tau)} = \{d_2, d_3, d_4\}$ である．

定義 2 (核文書データ，周辺文書データ) 文書データ dp の (ϵ,τ)-近傍 $N_{(\epsilon,\tau)}(dp)$ について，$|N_{(\epsilon,\tau)}(dp)| \geq MinDoc$ を満たす文書データ dp を**核文書データ**，$|N_{(\epsilon,\tau)}(dp)| \leq MinDoc$ である文書データを**周辺文書データ**と呼ぶ．

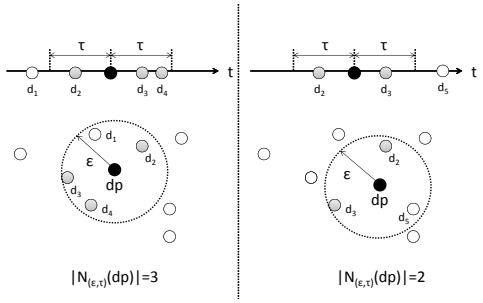


図 2. 定義 2 と定義 3 の例

(ϵ,τ)-密度に基づくクラスタリング手法では，核文書データの集合がクラスタの核データとなる．図 2 に例を示す．$MinDoc = 3$ とすると，図 2 の左では，文書データ dp は核文書データであり，図 2 の右では，文書データ dp は周辺文書データである．

定義 3 ((ϵ,τ)-密度的に直接到達可能) 文書データ dq が文書データ dp の (ϵ,τ)-近傍であり，$|N_{(\epsilon,\tau)}(dp)| \geq MinDoc$ を満たす時，文書データ dq は文書データ dp から (ϵ,τ)-**密度的に直接到達可能**であると表現する．

図 2 の例を使って，例を示す．図 2 の左では，文書データ dp は核文書データである．つまり，$|N_{(\epsilon,\tau)}(dp)| \geq MinDoc$ を満たす．このとき，文書データ d_2，d_3 と d_4 とは文書データ dp の (ϵ,τ)-近傍であり，文書データ dp から (ϵ,τ)-密度的に直接到達可能である．

定義 4 ((ϵ,τ)-密度的に到達可能) 文書データ dp_i が文書データ dp_{i+1} について，文書データ dp_{i+1} が文書データ dp_i から (ϵ,τ)-密度的に直接到達可能である，文書データ列 $(dp_1, dp_2, \cdots, dp_n)$ を考える．この時，文書データ dp_1 と dp_n は，(ϵ,τ)-**密度的に到達可能**であると表現する．

定義 5 ((ϵ,τ)-密度的に接続) 文書データ dp と文書データ dq とが文書データ do と (ϵ,τ)-**密度的に到達可能**であり，文書データ do が $|N_{(\epsilon,\tau)}(do)| \geq MinDoc$ を満たす時，文書データ dp と文書データ dq とは (ϵ,τ)-**密度的に接続**していると表現する．

B. (ϵ,τ)-密度に基づく時空間クラスタ

(ϵ,τ)-密度に基づく時空間クラスタリング手法では，時空間的に密集している文書データを (ϵ,τ)-密度に基づく時空間クラスタと定義する．

定義 6 ((ϵ,τ)-密度に基づく時空間クラスタ) 位置情報が付与された文書データ集合 GD において，(ϵ,τ)-**密度に基づく時空間クラスタ** STC は以下の2つの条件を満たす部分文書データ集合 GD^+ である．

(1) 任意の文書データ $dp \in GD$ と $dq \in GD$ について，**時空間クラスタ** STC に文書データ dp が所属（$dp \in STC$）し，文書データ dq が文書データ dp から (ϵ,τ)-密度的に到達可能であれば，文書データ dq は**時空間クラスタ** STC に所属（$dq \in STC$）する．

(2) **時空間クラスタ** STC に所属する任意の文書データ $dp \in STC$ と $dq \in STC$ は，(ϵ,τ)-密度的に接続している．

トピックと関連する文書データ集合から，(ϵ,τ)-密度に基づく時空間クラスタを抽出することで，当該トピックが話題となっている地域を抽出することができる．また，文書データを可視化することでどのようなことが発生しているか分析可能となる．

input : gp - 新しく追加される文書, D - これまでの文書データ集合, $CSTC$ - 現在の時空間クラスタ集合, ϵ - ε 値, τ - τ 値, $MinDoc$ - 最小文書データ数
output: $NSTC$ - 更新された時空間クラス集合

```
NSTC ← CSTC;
RD ← GetRecentData(gp,τ,D);
for i ← 1 to |RD| do
    pd ← rd_i ∈ RD;
    N ← GetNeighborhood(pd,ϵ,τ);
    if |N| ≥ MinDoc then
        if IsClustered(pd) == false then
            stc ←MakeNewCluster(cid,pd);
        end
        else
            stc ← GetCluster(pd,NSTC);
        end
        EnQueue(Q,N);
        while Q is not empty do
            pq ← DeQueue(Q);
            if IsClustered(pq) == true then
                N ←
                GetNeighborhood(pq,ϵ,τ);
                if |N| ≥ MinDoc then
                    stc' ←
                    GetCluster(pq,NSTC);
                    stc ←
                    AppendClusters(stc,stc');
                end
            end
            else
                stc ← stc ∪ pq;
                N ←
                GetNeighborhood(pq,ϵ,τ);
                if |N| ≥ MinDoc then
                    EnNniqueQueue(Q,N);
                end
            end
        end
        NSTC ← NSTC ∪ stc;
    end
end
return NSTC;
```

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

C. アルゴリズム

Algorithm 1 に，リアルタイムに (ϵ,τ)-密度に基づく時空間クラスタを抽出するためのインクリメンタルなアルゴリズムを示す．新たに投稿された文書データ，これまでの文書データ集合と現時点での時空間クラスタ集合とパラメータを入力として，新しい時空間クラスタを出力するインクリメンタルなアルゴリズムとなっている．新たに文書データが追加された場合，当該文書データの (ϵ,τ)-近傍に存在する各文書データの (ϵ,τ)-近傍が変化する．反対に当該文書データの (ϵ,τ)-近傍ではない文書データについては，(ϵ,τ)-近傍の変化はない．そこで，新たに追加された文書データの (ϵ,τ)-近傍に存在する各文書データのみを再クラスタリングすることで，新しい時空間クラスタ集合を求めることができる．再クラスタリングの過程で他の時空間クラスタに辿り着いた場合は，2 つの時空間クラスタを併合し，1 つの時空間クラスタとする．

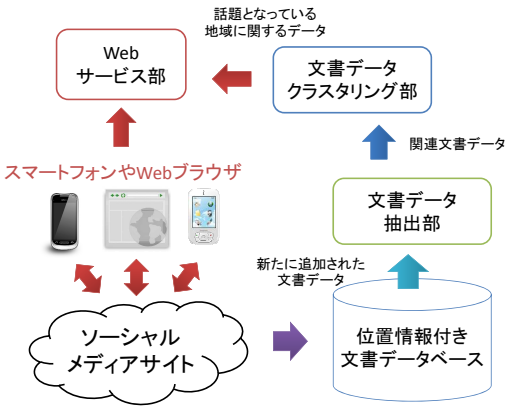


図 3. アプリケーションの概要

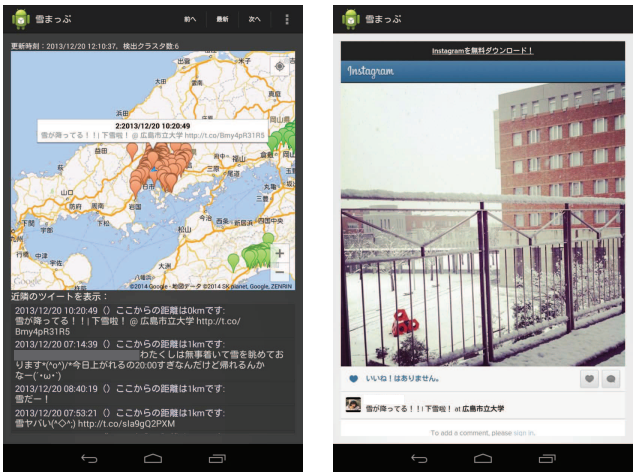
III. 話題の地域分析アプリケーション

図 3 に (ϵ,τ)-密度に基づく時空間クラスタリング手法を用い，話題の地域を分析するためのアプリケーションの概要を示す．

(1) ソーシャルメディアサイトから位置情報付き文書データを収集し，位置情報付き文書データベースに保存する．
(2) 文書データ抽出部では，分類器を使って監視を行っているトピックに適合する文書データを取り出し，文書データクラスタリング部に入力する．分類器として教師あり学習のナイーブベイズを用いる．監視を行うトピックの正例・負例によって事前に学習をさせ，監視を行うトピックを含む文書データを抽出する．
(3) 文書データクラスタリング部では，現時点の時空間クラスタ集合を保持するとともに，Algorithm 1 で示したアルゴリズムと新たに入力された位置情報付き文書データを用いて時空間クラスタ集合を更新する．
(4) Web サービス部では，時空間クラスタに関するデータを Web サービスとして提供する．
(5) Web サービス部を通して，スマートフォン，タブレット端末側では，Web ブラウザによる表示，もしくはアプリケーション上で話題となっている地域を確認できる．

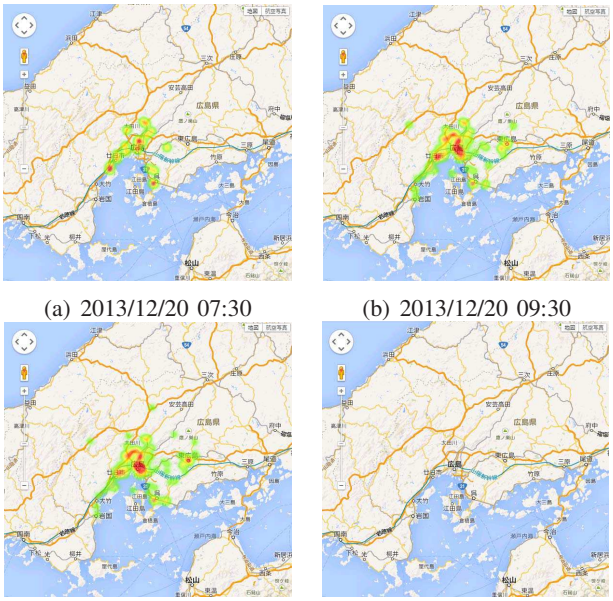
IV. 評価実験

提案アプリケーションを評価するために，Twitter 上において「雪」が話題となっている地域を分析する「雪



(a) Android アプリケーションの画面 1　(b) Android アプリケーションの画面 2

図 4.　2013 年 1 2 月 20 日の雪に関する話題を表示する「雪マップ」のアプリケーションキャプチャ画面



(a) 2013/12/20 07:30　(b) 2013/12/20 09:30

(c) 2013/12/20 11:30　(d) 2013/12/20 13:30

図 5.　2013 年 1 2 月 20 日の雪に関する話題の地域的な遷移

マップ」を試作し，評価実験を行った．Twitter 上で取得したジオタグ（経度，緯度）が付与されたツイートをリアルタイムに収集し，話題地域をリアルタイムに検出した．試作を行った「雪マップ」は Android アプリケーションで実装を行った．また，分類器では，「雪」についての正例 791 件，負例 2764 件を用いた．図 4 に「雪マップ」のスクリーンキャプチャを示す．図 4(a) に示すように話題地域のツイート一覧を地図上に示し，また，下部ではユーザの現在位置から近い順に「雪」関するツイートが表示される．また，現在位置だけでなく，時間でソートして表示も可能となっている．ツイート一覧をタッチすると当該ツイートの位置に移動し，関連するツイートのみを閲覧可能でき，長押しをすると画像ツイートが添付されている場合は画像を閲覧することができる．また，Web ブラウザによるインタフェースでは，話題の地域をヒートマップを使って可視化を行い，話題となっている地域の時間変化を観測することができる（図 5）

V. まとめ

本論文では，あるキーワードを含むトピックが話題となっている地域を (ϵ, τ)-密度に基づく時空間クラスタリング手法と分類器を用いて分析するためのアプリケーションについて述べた．Twitter 上において「雪」が話題となっている地域を地図上で表示し動向分析を行うための「雪マップ」を試作し，評価実験を行った．「雪」について話題となっている地域をリアルタイムに検出することができ，分かりやすく地図上に表示することができ，話題となっている地域を容易に把握することが可能となった．これからの課題として，内容を掌握するためのツイート要約機能の作成などがあげられる．

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] J. Chon and H. Cha, “Lifemap: A smartphone-based context provider for location-based services,” *IEEE Pervasive Computing*, vol. 10, pp. 58–67, Apr. 2011.
[2] M. Naaman, “Geographic information from georeferenced social media data,” *SIGSPATIAL Special*, vol. 3, pp. 54–61, July 2011.
[3] H. Yang, S. Chen, M. R. Lyu, and I. King, “Location-based topic evolution,” in *Proceedings of MLBS ’11*, pp. 89–98, 2011.
[4] K. Tamura and T. Ichimura, “Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents,” in *Proceedings of The 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, pp. 2079–2084, 2013.
[5] T. Sakai, K. Tamura, and H. Kitakami, “A new density-based spatial clustering algorithm for extracting attractive local regions in georeferenced documents,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I*, pp. 360–365, 2014.
[6] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, “Mapping the world’s photos,” in *Proceedings of WWW ’09*, pp. 761–770, 2009.
[7] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in *Proceedings of WWW ’10*, pp. 851–860, 2010.

問い合わせ先
〒 731-3194
広島市安佐南区大塚東 3-4-1
広島市立大学大学院情報科学研究科
酒井 達弘