文書データの類似度を考慮した 密度に基づくクラスタリングによる地域的なトピック抽出
Density-based Clustering Considering Document Similarity for Extracting Local Topics

酒井 達弘
Tatsuhiro Sakai
広島市立大学
情報科学部
E-mail: u20084@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

北上 始
Hajime Kitakami
広島市立大学大学院
情報科学研究科
E-mail: kitakami@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. In this paper, a novel spatiotemporal clustering algorithm, called the (ϵ, σ)-density-based spatial clustering algorithm, for extracting local topics and events from georeferenced documents is proposed. The proposed spatial clustering algorithm can recognize spatially-separated clusters by considering document similarity. To evaluate our proposed spatial clustering algorithm, geo-tagged tweets posted on the Twitter site are used. The experimental results show that the (ϵ, σ)-density-based spatial clustering algorithm can extract local topics and events.**

I. はじめに

近年，GPS 付きスマートフォンの普及とソーシャルメディアへの関心の高まりとともに，位置情報が付与された文書データがインターネット上において盛んに投稿され，位置に関連した情報発信が行われてきている [1], [2]．位置情報が付与された文書データは，位置に関連したローカルなトピックやイベントと結びついている可能性が高く，ローカルで話題となっているトピックやイベントを抽出すること [3] は，社会的な動向分析，マーケティング，観光情報をはじめとして，地域的な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．

そこで，位置情報に着目し，ローカルで話題となっているトピックやイベントを発見する手法が研究されている．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 [4]，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 [5] が行われている．

本研究では，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，ローカルで話題となっているトピックを取り出す手法を提案する．提案手法では，密度に基づくクラスタリング [6] を (ϵ, σ)-密度に基づくクラスタリングとして拡張する．そして，文書データから (ϵ, σ)-空間クラスタをローカルなトピックとして抽出する．(ϵ, σ)-密度に基づくクラスタリングを用いることで，どのようなトピックを含むか分からない未知の文書データから，地域的なトピックが抽出可能となる．

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，(ϵ, σ)-密度に基づくクラスタリングを提案する．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. 関連研究

近年，ソーシャルメディアサイト上ではユーザは位置情報を付与したデータを盛んに投稿するようになってきている．位置情報が付与されたデータは，ユーザをセンサと考えると，実世界のあらゆる事象を観測したセンサデータとして捉えることができ [7]，実世界の様々な事象が記録されているといえる．これらの位置情報が付与されたデータは，集団的な知識（集合知）を有するようになってきており，位置情報が付与されたデータから有益な知識を発見するための研究が盛んに行われている．

Jaffe ら [8] と Rattenbury ら [9] は，位置情報が付与された Flickr 上の投稿画像データを対象として，位置情報を用いてクラスタリングを行い，投稿画像データをクラスタとしてまとめ，トピックやイベントを検出する手法を提案している．Watanabe ら [10] は，投稿

される Twitter データから話題となっている地点を検出する手法を提案している．Lee ら [11] は，空間をボロノイ図で分割し，各ボロノイ区画の通常時の投稿数を求め，投稿数が急増したボロノイ区画を検出することで地域的なイベントを検出する手法を提案している．

このように位置情報を用いた空間的なクラスタリングを応用し，ソーシャルメディアサイト上のデータからトピックやイベントを抽出する研究は多数行われているが，本研究と最も関連するのが密度に基づくクラスタリングを応用した話題地域の抽出である．Kisilevich ら [12] は密度に基づくクラスタリング手法を拡張し，ジオタグが付与された投稿画像データから話題となっている地域を取り出す手法を提案している．本研究の手法は，Kisilevich ら [12] の研究と同様に話題となっている場所を取り出すが，ただ単に空間的に近いだけでなく，データ間の類似度を考慮してクラスタリングすることができる．

III. (ϵ, σ)-密度に基づくクラスタリング

本章では，密度に基づくクラスタリングの諸定義を拡張し，(ϵ, σ)-密度に基づくクラスタリングを提案する．

A. 諸定義

密度に基づくクラスタリング手法 [6] は，データが密集している部分をクラスタ，密集していない部分をクラスタではないと定義し，クラスタを抽出する．密度に基づいているため，円状ではない（つまりクラスタ内のデータ間の距離が近いとは限らない）クラスタを抽出することが可能である．よって，この密度に基づくクラスタリング手法は空間データのクラスタリングとして広く使用されている．

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．本研究では，まず，ϵ-近傍の定義を以下のように拡張する．

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

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

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 sim は文書データ dp と文書データ dq 間の類似度を返す関数である．

図 1 に定義 1 の例を示す．図 1 の左は，ϵ-近傍の例であり，文書データ dp の ϵ-近傍は半径 ϵ 以内に存在する文書データ集合である．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, σ)-近傍の例を示す．文書データ dp の (ϵ, σ)-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 との類似度が σ 以上の文書データである．この例では，

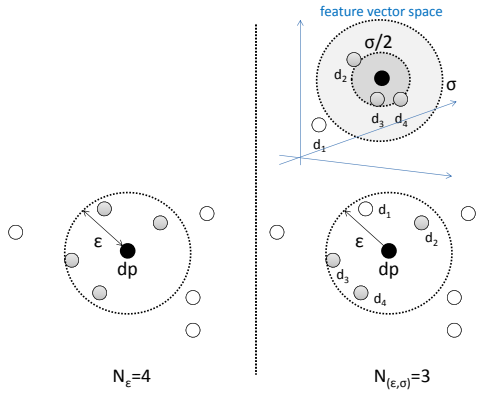

図 1. 定義 1 の例

文書データは 3 つ存在し，$N_{(\epsilon,\sigma)} = \{d_2, d_3, d_4\}$．文書データ d_1 は ϵ 以内にあるが，類似度が，(ϵ, σ)-近傍に入らない．

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

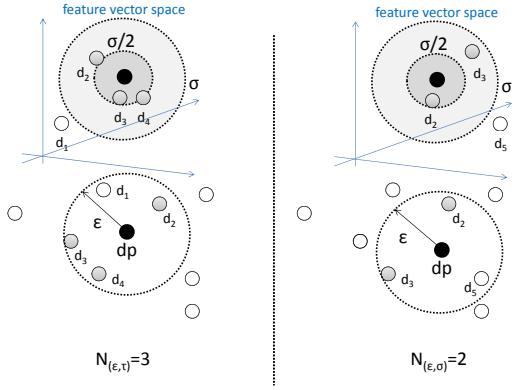

図 2. 定義 2 と定義 3 の例

(ϵ, σ)-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, σ)-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, σ)-近傍であり，$N_{(\epsilon,\sigma)}(dp) \geq MinDoc$ を満たす時，文書データ dq は文書データ dp から (ϵ, σ)-**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，$N_{(\epsilon,\sigma)}(dp) \geq MinDoc$ 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, σ)-近傍であり，文書データ dp から (ϵ, σ)-密度的に直接到達可能である．

定義 4 ((ϵ, σ)-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,σ)-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,σ)-**密度的に到達可能**であると表現する．

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

B. (ϵ,σ)-空間クラスタ

密度に基づくクラスタリングでは近距離に密集している文書データを空間クラスタとして定義している．一方，(ϵ,σ)-密度に基づくクラスタリングでは，類似度を考慮して近距離に密集している文書データを (ϵ,σ)-空間クラスタと定義する．

定義 6 ((ϵ,σ)-空間クラスタ) 文書データ集合 D において，(ϵ,σ)-**空間クラスタ** SC は以下の 2 つの条件を満たす部分文書データ集合である．

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

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

C. アルゴリズム

Algorithm1 に (ϵ,σ)-密度に基づくクラスタリングのアルゴリズムを示す．各文書データ d_i について，もし，その文書データがまだどこのクラスタにも所属していなければ，次の処理を行う．文書データ d_i の (ϵ,σ)-近傍を求める．もし，(ϵ,σ)-近傍の数がユーザが指定した $MinDoc$ 以上であれば，新しくその空間クラスタ stc_{cid} を作成し，その文書データを空間クラスタ stc_{cid} に所属させる．次に，(ϵ,σ)-近傍をキュー Q に挿入する．キュー Q が空になるまで，Q から文書データ pd を取り出し，次の処理を繰り返す．文書データ pd の (ϵ,σ)-近傍を求める．もし，(ϵ,σ)-近傍の数がユーザが指定した $MinDoc$ 以上であれば，その文書データを空間クラスタ stc_{cid} に加える．次に，(ϵ,σ)-近傍の中でクラスタに含まれていない文書データのみをキューに加える．

input : D - dataset with coordinates, ϵ - neighborhood radius, σ - similarity rate, $MinDoc$ is threshold value
output: SC - set of clusters

```
cid ← 1;
SC ← φ;
for i ← 1 to |D| do
    pd ← d_i ∈ D;
    if IsClustered(pd) == false then
        N ← GetNeighbors(pd,ϵ,σ);
        if |N| ≥ MinDoc then
            stc_cid ←MakeNewCluster(cid,pd);
            cid ← cid + 1;
            EnQueue(Q,N);
            while Q is not empty do
                pq ← DeQueue(Q);
                N ← GetNeighbors(pq,ϵ,σ);
                if |N| ≥ MinDoc then
                    EnNniqueQueue(Q,N);
                    stc_cid ← stc_cid ∪ pq
                end
            end
            SC ← SC ∪ stc_cid;
        end
    end
end
return SC;
```

Algorithm 1: (ϵ,σ)-密度に基づくクラスタリングアルゴリズム

IV. 評価実験

提案手法を評価するために，Twitter 上で取得したジオタグ（経度，緯度）が付与された 480,000 ツイート (2011 年 11 月~ 2012 年 2 月，言語を日本で設定しているユーザ) を用いて評価実験を行った．広島県庁から 50km に存在するツイートを取り出し，評価を行う．

(ϵ,σ)-密度に基づくクラスタリングアルゴリズムを適用するにあたり，文書データ間の類似度を計る必要がある．各ツイートは，形態素解析で形態素に分け，語句として名詞，形容詞と動詞を取り出した．各ツイート間の類似度の算出には，シンプソン係数を用いた．

表 I に抽出された (ϵ,σ)-空間クラスタの一覧を示す．パラメータとして，$\epsilon = 0.5(= 500m)$，$\sigma = 0.5$，$MinDoc = 5$ を用いた．15 クラスタ抽出できた．ツイート数で上位 8 位までのクラスタを示している．表では各クラスタのツイート数，経度・緯度の範囲，また，クラスタに含まれている頻出語句上位 5 件を示している．ただし，広島，市，区などの住所に関連する語句は省いている．

1 番目のクラスタは原爆ドームという語句が上位に来ているように，原爆ドームとその周辺の繁華街に関するツイートがクラスタとして抽出された．2 番目のクラスタは，駅，新幹線とあるように広島駅とその周辺のトピックが抽出されている．ただし，1 番目と 2 番目のクラスタの中心のトピックは原爆ドームと広島駅であるが，その周辺も繁華街であるため明確に 1 つのトピックとして抽出はできなかった．原因として，住所を併記しているツイートが多いためトピックが異な

表 I
クラスター一覧

No	ツイート数	経度	緯度	頻出語句上位 5 位
1	63	132.4535349 - 132.465292	34.3893519 - 34.40117185	Atom，原爆ドーム，店，相生通り
2	57	132.46988297 - 132.47762918	34.39416213 - 34.399713	駅，新幹線，ホーム，店，松原
3	30	132.31812559 - 132.32402913	34.29420941 - 34.30074774	厳島神社，宮島，大鳥居，本殿，藤
4	16	132.42674139 - 132.42702243	34.37271835 - 34.37327164	ハード，セミ，トースト，焼き，おはようございます
5	10	132.30438744 - 132.305471	34.35376995 - 34.354314	たこ，ボール，中，営業，宮園
6	10	132.45104849 - 132.45255053	34.39181602 - 34.39298908	Peace，平和記念公園，駅伝，男子，出走
7	10	132.31602073 - 132.31844813	34.36297389 - 34.36704947	宮島 SA，下り，山陽自動車道，店，Starbucks
8	9	132.45691723 - 132.45915413	34.40035934 - 34.40379812	城，Cast，中，たま，自転車

るが類似度が大きくなったことがあげられる．

また，３番目のクラスタは宮島であり，４番目と５番目のクラスタはローカルのパン屋とたこ焼き屋が抽出された．ただし，4 番目のパン屋はパン屋自体が投稿したツイートであり，ユーザ数によるフィルタリングを行い，このように宣伝のような投稿の重要度を低くする必要がある．6 番目のクラスタは原爆ドームの近隣にある平和記念公園や資料館を扱っているクラスタである．原爆ドームと平和記念公園とは隣接しているがうまく分離できている．また，7 番目のクラスタは高速道路の SA エリア，8 番目のクラスタは広島城について扱っているクラスタである．

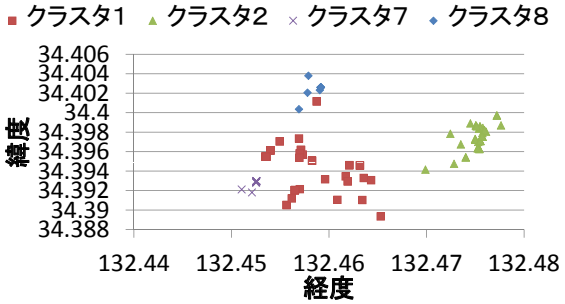

図 3.　近接クラスタ

図 3 に近接した空間クラスタを図示する．図 3 は，広島市内で抽出された 4 つの空間クラスタを経度緯度でプロットした図である．クラスタ 1，クラスタ 7 とクラスタ 8 は近接しているが，ツイートの内容が異なるため 3 つのクラスタとして分離することができている．

V. まとめ

本研究では，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，ローカルで話題となっているトピックを取り出す手法を提案した．提案手法では，密度に基づくクラスタリング手法を (ϵ, σ)-密度に基づくクラスタリング手法として拡張した．(ϵ, σ)-密度に基づくクラスタリング手法を用いることで，どのようなトピックを含むか分からない未知の文書データから，地域的なトピックが抽出できる．提案手法を実際に実装し，Twitter から取得したデータを用いて評価実験を行った．評価実験の結果，(ϵ, σ)-密度に基づくクラスタリング手法を用いることで文書データの内容を区別してトピックを形成している空間クラスタを取り出すことができた．

謝辞

本研究の一部は，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] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, “Mapping the world’s photos,” in *Proceedings of WWW ’09*, pp. 761–770, 2009.

[5] 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.

[6] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu, “Density-based clustering in spatial databases: The algorithm gdbscan and its applications,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 169–194, June 1998.

[7] M. F. Goodchild, “Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0,” *International Journal of Spatial Data Infrastructures Research*, vol. 2, pp. 24–32, 2007.

[8] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, “Generating summaries and visualization for large collections of geo-referenced photographs,” in *Proceedings of MIR ’06*, pp. 89–98.

[9] T. Rattenbury, N. Good, and M. Naaman, “Towards automatic extraction of event and place semantics from flickr tags,” in *Proceedings of SIGIR ’07*, pp. 103–110.

[10] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, “Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs,” in *Proceedings of CIKM ’11*, pp. 2541–2544, 2011.

[11] R. Lee and K. Sumiya, “Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection,” in *Proceedings of LBSN ’10*, pp. 1–10, 2010.

[12] S. Kisilevich, F. Mansmann, and D. Keim, “P-dbscan: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos,” in *Proceedings of COM.Geo ’10*, pp. 38:1–38:4, 2010.

問い合わせ先

〒 731-3194
広島市安佐南区大塚東 3-4-1
広島市立大学情報科学部
酒井 達弘