Vol.2014-MPS-101 No.3
2014/12/9

地域的なトピック抽出のための密度に基づく適応的な空間クラスタリング手法

酒井 達弘[1,a)]　田村 慶一[1,b)]　北上 始[1,c)]　伊東 晴奈[2,d)]

概要：
ソーシャルメディアサイト上に投稿されている位置情報付き文書データは，地域的なトピックと関連しており，それらを抽出することは重要な研究課題の一つである．本研究では，新しい地域的なトピック抽出手法として，位置情報付き文書データから (ϵ, σ)-密度に基づく適応的な空間クラスタリング手法を用いて地域的なトピックが話題となっている地域を空間クラスタとして抽出し，さらに，空間クラスタから代表文書データを空間クラスタの中心的な内容として提示する手法を提案する．(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法は，空間クラスタを抽出する基準となる閾値を，各地域の投稿数によって適応的に変化させ，地域的な投稿密度に区別なく空間クラスタをシームレスに抽出可能な手法となっている．また，各空間クラスタからの代表文書データ抽出にはネットワークベースの重要文抽出手法を用いている．提案手法の有効性を確認するために，Twitter に投稿されたジオタグ付きツイートを用いて実験を行った．その結果，投稿数が少ない低密度な地域と投稿数が多い高密度な地域を区別することなく，地域的なトピックを抽出することができ，代表文書データを空間クラスタの中心的な内容として抽出することができた．

キーワード：密度に基づく空間クラスタリング，位置情報付き文書データ，地域的なトピック抽出，ソーシャルメディア，ネットワークグラフ

1. はじめに

近年，GPS 付きスマートフォンの普及とともに，位置情報付きデータが盛んに投稿され，ソーシャルメディア上のユーザは位置に関連した情報発信を行うようになってきている [1]．位置情報付きデータは個人的な話題だけでなく，地域的なトピックやイベントと結びついている可能性が高い．この中でも，位置情報付きの文書データから，地域で話題として取り上げられているトピックやイベントを抽出することは，社会的な動向分析，マーケティングや観光情報にとって重要な研究課題のひとつとなっている．

我々は，位置情報付き文書データから地域的なトピックが話題として取り上げられている地域を抽出するための新しい空間クラスタリング手法として，(ϵ, σ)-密度に基づく空間クラスタリング手法を提案している [2]．Twitter 上のジオタグ付きツイートを用いた評価実験の結果，空間クラスタとして地域的なトピックが話題として取り上げられている地域を取り出すことができることを確認できた．しかしながら，(ϵ, σ)-密度に基づく空間クラスタリング手法には，投稿数の違いがある地域が混在している場合，シームレスに空間クラスタを抽出できないことと，空間クラスタ内の文書データひとつひとつを閲覧しないと空間クラスタが持つトピックを把握できない点が課題として残っている．

本研究では，これらの課題を解決するために，(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法を用いてシームレスに空間クラスタを抽出し，さらに，空間クラスタから代表文書データを抽出する新しい地域的なトピック抽出手法を提案する．具体的には，空間クラスタを抽出する基準となる閾値を，各地域の投稿数によって適応的に変化させ，投稿数が多い高密度な地域と投稿数が少ない低密度な地域を区別なく空間クラスタをシームレスに抽出可能にする．そして，ネットワークベースの重要文抽出手法を用いて各空間クラスタの代表文書データを抽出する．代表文書データを閲覧することで各空間クラスタの中心的な内容を知ることができる．

[1] 広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan
[2] 広島市立大学情報科学部
Faculty of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan
a) my67011@e.hiroshima-cu.ac.jp
b) ktamura@hiroshima-cu.ac.jp
c) kitakami@hiroshima-cu.ac.jp
d) u20017@e.hiroshima-cu.ac.jp

提案手法を評価するために，Twitter 上のジオタグ付きツイートを用い，評価実験を行った．評価実験では，日本全国の地域的なトピックが話題として取り上げられている地域を抽出し，広島地区で抽出した地域に対し，代表文書データの抽出を行った．評価実験の結果，地域的なトピックが話題として取り上げられている地域をシームレスに抽出することができ，代表的な文書データを抽出することで各空間クラスタが持つトピックを容易に把握することができることを確認できた．

本論文の構成は次の通りである．第 2 章では，関連研究を述べる．第 3 章では，提案手法の処理手順を示す．第 4 章では，(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法について説明する．第 5 章では，代表文書データの抽出手法について述べる．第 6 章では，評価実験の実験結果を示し，第 7 章で本論文をまとめる．

2. 関連研究

位置情報付き文書データは，地理空間データとして扱うことができる．地理空間データの空間クラスタリング手法として最も有効的な手法として，密度に基づく空間クラスタリング手法が提案されている．密度に基づく空間クラスタリング手法は，空間データが多く存在する高密度な領域を，空間データが少ない低密度な領域と分離し，任意形状の空間クラスタとして抽出することができる．

Kisilevich ら [3] は，密度に基づく空間クラスタリング手法のひとつである DBSCAN[4] を拡張した空間クラスタリング手法として P-DBSCAN を提案している．データの近傍に存在するユーザ数に応じた新しい密度を定義し，ジオタグが付与された画像データを用いて，注目されているイベントや場所を分析することができる．田村ら [5] は，時空間クラスタを抽出することができる (ϵ, τ)-密度に基づく時空間クラスタリング手法を提案している．

Kisilevich らと田村らの研究は DBSCAN の拡張という点で本研究と似ているが，これらの研究ではデータの内容をクラスタリングの過程では考慮していない．一方，(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法はデータの内容を考慮している．我々の調査では，位置情報付き文書データを対象として適応的な空間クラスタリング手法を提案することは，本研究がはじめての試みである．提案手法では，適宜パラメータを調整する必要なく，局所的に高密度な領域の空間クラスタを抽出することができる．

3. 提案手法

位置情報付き文書データの集合を $GDS = \{gd_1, gd_2, \cdots, gd_n\}$ と表し，位置情報付き文書データ gd_i はその文書データ投稿された時刻 pt_i（投稿時刻），文書データ $text_i$（テキスト），位置情報 pl_i（経度と緯度）の三つの要素から構成され，$gd_i =< pt_i, text_i, pl_i >$ と表す．例えば，Twitter 上のジオタグ付きツイートであれば，時刻はツイートが投稿された時刻であり，テキストはツイート本文，また，位置情報はツイートに付与されたジオタグとなる．

提案手法の処理手順は次の通りである．

(1) 新しい位置情報付き文書データ gd_k を取得すると，(2) へ移る．

(2) 新しく取得した gd_k，これまでに投稿された位置情報付き文書データ集合 GDS，これまでに取得済みの空間クラスタ集合 $DASC$ を入力として，(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法を用いて (ϵ, σ)-密度に基づく適応的な空間クラスタ集合 $NDASC = \{ASC_1, ASC_2, \cdots, ASC_m\}$ を抽出する．(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法については，第 4 章で説明する．

(3) 各空間クラスタ ASC_i に対して，ネットワークベースの重要文抽出手法を用いて代表文書データを抽出する．代表文書データ抽出手法については，第 5 章で説明する．

(4) 抽出した空間クラスタ集合 $NDASC$ を地域的なトピックとして，代表文書データを中心的な内容としてユーザに提示する．(1) へ戻る．

4. (ϵ, σ)-密度に基づく適応的な空間クラスタリング手法

本章では，(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法について説明する．

4.1 概要

(ϵ, σ)-密度に基づく空間クラスタリング手法 [2] では，2 つの文書データ間の距離と類似度を定め，距離が ϵ 以内であり，類似度が σ 以内の位置情報付き文書データを (ϵ, σ)-近傍と定義している．そして，空間クラスタに存在する位置情報付き文書データは，$MinGdoc$（ユーザパラメータ）以上の位置情報付き文書データが (ϵ, σ)-近傍に存在する必要がある．(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法では，各地域の投稿数を各地域の投稿密度として考え，投稿密度により $MinGdoc$ を適応的に変化させる．投稿数の多い地域では $MinGdoc$ は高く，投稿数の少ない地域では $MinGdoc$ は低く設定される．

4.2 諸定義

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

定義 1 ((ϵ, σ)-近傍 $GN_{(\epsilon,\sigma)}(gdp)$) 位置情報付き文書データ gdp の (ϵ, σ)-近傍を $GN_{(\epsilon,\sigma)}(gdp)$ と表記し，次のように定義する．

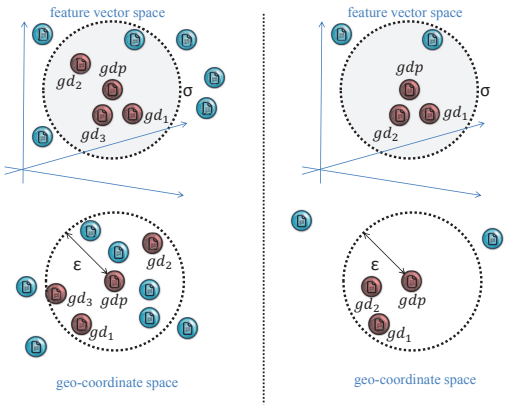

図 1 定義 3 の例
Fig. 1 Example of definition 3

$$GN_{(\epsilon,\sigma)}(gdp) = \{gdq \in GDS |\ dist(gdp, gdq) \leq \epsilon \quad and \quad sim(gdp, gdq) \geq \sigma\} \quad (1)$$

関数 $dist$ は経度・緯度など座標値を使って，位置情報付き文書データ gdp と gdq 間の空間上の距離を求める関数であり，本研究では Lambert-Andoyer の公式を用いて距離を返す．関数 sim は位置情報付き文書データ gdp と gdq 間の類似度を返す関数である．関数 sim は 4.3 節で説明する．

定義 2 (地域的な投稿密度，適応的な閾値) 位置情報付き文書データ gdp が存在する地域の投稿密度を $ld(gdp)$ と表記する．位置情報付き文書データ gdp の地域的な密度の適応的な閾値 AT を次のように定義する．

$$AT(gdp, MinGdoc) = (MinGdoc - 1) \times ld(gdp) + 1 \quad (2)$$

関数 $ld(gdp)$ は，位置情報付き文書データ gdp が存在する地域の投稿密度を返す関数である $(0 \leq ld(gdp) \leq 1.0)$．関数 $ld(gdp)$ は 4.3 節で説明する．

定義 3 (核文書データ，周辺文書データ) 位置情報付き文書データ gdp の (ϵ,σ)-近傍 $GN_{(\epsilon,\sigma)}(gdp)$ について，$|GN_{(\epsilon,\sigma)}(gdp)| \geq AT(gdp, MinGdoc)$ を満たす位置情報付き文書データ gdp を核文書データ，$|GN_{(\epsilon,\sigma)}(gdp)| < AT(gdp, MinGdoc)$ である位置情報付き文書データ gdp を周辺文書データと呼ぶ．

$MinGdoc$ はユーザパラメータである．**図 1** を使って，定義 3 の例を示す．$MinGdoc = 3$，$ld(gdp) = 0.8$ とすると，$AT(gdp, MinGdoc) = 2.6$ となる．つまり，図 1 の左では，gdp は核文書データであり，図 1 の右では，gdp は周辺文書データである．

定義 4 ((ϵ,σ)-密度に基づいて適応的に直接到達可能) 位置情報付き文書データ gdq が位置情報付き文書データ gdp の (ϵ,σ)-近傍であり，$|GN_{(\epsilon,\sigma)}(gdp)| \geq AT(gdp, MinGdoc)$ を満たす時，gdq は gdp から (ϵ,σ)-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,σ)-密度に基づいて適応的に到達可能) 位置情報付き文書データ gdp_{i+1} が位置情報付き文書データ gdp_i から (ϵ,σ)-密度に基づいて適応的に直接到達可能である，位置情報付き文書データ列 $(gdp_1, gdp_2, \cdots, gdp_n)$ を考える．この時，gdp_1 と gdp_n は，(ϵ,σ)-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,σ)-密度に基づいて適応的に接続) 位置情報付き文書データ gdp と位置情報付き文書データ gdq とが位置情報付き文書データ gdo と (ϵ,σ)-密度に基づいて適応的に到達可能であり，gdo が $|GN_{(\epsilon,\sigma)}(gdo)| \geq AT(gdo, MinGdoc)$ を満たす時，gdp と gdq とは (ϵ,σ)-密度に基づいて適応的に接続していると表現する．

定義 7 ((ϵ,τ)-密度に基づく適応的な時空間クラスタ) 文書データ集合 GDS において，(ϵ,σ)-密度に基づく適応的な空間クラスタ ASC 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) 任意の位置情報付き文書データ $gdp \in GDS$ と $gdq \in GDS$ について，(ϵ,σ)-密度に基づく適応的な空間クラスタ ASC に gdp が所属（$gdp \in ASC$）し，gdq が gdp から (ϵ,σ)-密度に基づいて適応的に到達可能であれば，gdq は (ϵ,σ)-密度に基づく適応的な空間クラスタ ASC に所属（$gdq \in ASC$）する．

(2) (ϵ,σ)-密度に基づく適応的な空間クラスタ ASC に所属する任意の位置情報付き文書データ $gdp \in ASC$ と $gdq \in ASC$ は，(ϵ,σ)-密度に基づいて適応的に接続している．

4.3 類似度算出関数

類似度算出関数 sim は，語句に基づくシンプソン係数 $wsim$ とキーワードに基づくシンプソン係数 $ksim$ とで構成される．

$$sim(gd_i, gd_j) = w_1 \times wsim(gd_i, gd_j) + w_2 \times ksim(gd_i, gd_j) \quad (3)$$

だたし，$w_1 + w_2 = 1.0$ とする．

テキスト $text_i$ に含まれる語句集合を，$dt_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,nw(i)}\}$，$w_{i,j} \in W$ とする．W は全ての語句集合とし，$nw(i)$ は dt_i に含まれる語句数とする．本研究では，形態素解析を行い，名詞，動詞，形容詞を語句として抽出する．語句に基づくシンプソン係数 $wsim$ を次のように定義する．

$$wsim(gd_i, gd_j) = \frac{|dt_i \cap dt_j|}{min(|dt_i|, |dt_j|)} \quad (4)$$

ここで，key_i を dt_i に含まれるキーワード集合，$key_i = \{k_{i,1}, k_{i,2}, \cdots, k_{i,nk(i)}\}$，$k_{i,j} \in K$ とする．ただし，K は W に含まれている全てのキーワードとし，$nk(i)$ は dt_i に含まれるキーワード数とする．キーワードに基づくシンプソン係数 $ksim$ を次のように定義する．

$$ksim(gd_i, gd_j) = \frac{|key_i \cap key_j|}{min(|key_i|, |key_j|)} \quad (5)$$

4.4 地域的な投稿密度

本研究では，投稿数の統計データを用いて地域的な投稿密度を算出する．日本の最西端である与那国島の緯度・経度 (24.4494,122.93361) と最北端である択捉島の緯度・経度 (45.5572,148.752) からなる矩形を空間分割の対象領域とする．対象領域を $1,000 \times 1,000$ の $1,000,000$ グリッドに空間分割する．各グリッドの投稿数を求め，正規化した値をグリッドの密度とする．グリッド i に含まれる位置情報付き文書データの数を gn_i とし，関数 $geo\_gid(gdp)$ をグリッド ID を求める関数とすると，位置情報文書データ gdp が位置する地域の投稿密度 $ld(gdp)$ は次の式で定義される．

$$ld(gdp) = \frac{gn_{(geo\_gid(gdp))} - gn_{min}}{gn_{max} - gn_{min}} \quad (6)$$

ただし，gn_{min} はグリッドに含まれる位置情報付き文書データの最大数であり，gn_{max} は最小数である．

4.5 アルゴリズム

空間クラスタの抽出手順を次に示す．新たに投稿された位置情報付き文書データ，これまでの位置情報付き文書データ集合，現在の空間クラスタ集合と各パラメータを入力として，更新された空間クラスタ集合を出力する．

(1) 新たに投稿された位置情報付き文書データ gdp の (ϵ, σ)-近傍を取得し，(ϵ, σ)-近傍と gdp を文書データ集合として CGD に挿入する．
(2) CGD の各位置情報付き文書データ $pgd = cgd_i$ に対して次の処理を行う．
(3) pgd が核文書データであるかチェックし，キュー CQ に (ϵ, σ)-近傍を挿入する．そうでなければ，CGD の次の文書データに移り，(2) へ戻る．
(4) pgd が空間クラスタに属しているかチェックを行う．空間クラスタに属していなければ，新たに空間クラスタ asc を作成する．空間クラスタに属していれば，その空間クラスタに所属する文書データを asc に保存する．
(5) キュー CQ が空になるまで，CQ から位置情報付き文書データ pgq を取り出し，pgq がすでに空間クラスタに属しているかチェックを行い，次の処理を繰り返す．
  (a) 空間クラスタに所属していれば，pgq が核文書データであるかチェックする．もし，核文書データであれば，pgq が所属する空間クラスタの文書データ集合を取得する．そして，その空間クラスタと asc を結合する．
  (b) 空間クラスタに属していなければ，その位置情報付き文書データを asc に挿入する．次に，pgq が核文書データであるかチェックする．核文書データであれば，pgq の (ϵ, σ)-近傍に挿入する．挿入では，CQ に存在せず，また，他の空間クラスタに属していない位置情報付き文書データのみを CQ に挿入する．
(6) 空間クラスタ集合 $NDASC$ に，空間クラスタ asc を加える．
(7) CGD の次の文書データに移り，(2) へ戻る．
(8) $NDASC$ を更新された空間クラスタ集合として返す．

5. 代表文書データ抽出

代表文書データ抽出ではネットワークベースの重要文抽出手法を用いて各空間クラスタの代表文書データを抽出する．各空間クラスタ ASC_i について，空間クラスタ ASC_i を構成する文書データを形態素解析し，名詞・動詞・形容詞を取り出す．次に，文書データ間の類似度を語句集合に基づくシンプソン係数により算出する．文書データをノード，また，類似度が α（ユーザパラメータ）以上である文書データ間に辺を挿入した，類似度グラフ NG_i を作成し，各ノードの重要度を PageRank アルゴリズムもしくは媒介中心性により算出する．最後に，重要度上位 β 件を ASC_i の代表文書データとして抽出する．

6. 評価実験

提案手法を評価するために，評価実験を行った．

6.1 データセットと実験環境

評価実験では，Twetter streaming API で取得した（2011 年 11 月から 2012 年 2 月まで）392,912 件のジオタグ付きツイートを位置情報付き文書データとして扱い実験を行う．(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法のパラメータは，$\epsilon = 500m$，$\sigma = 0.7$，$MinGdoc = 5$，$w_1 = 0.5$，$w_2 = 0.5$ を，代表文書データ抽出手法のパラメータは，$\alpha = 0.5$，$\beta = 3$ を用いた．評価実験では，先行研究 [2] の手法（DBSC と表記する）と提案手法（DBASC と表記する）の比較を行う．

6.2 (ϵ, σ)-密度に基づく適応的な空間クラスタリング手法の評価実験

表 1 に DBSC と DBASC のそれぞれで抽出された空間クラスタの総数を示す．表 1 より，DBASC によって新たに抽出された空間クラスタがあるのが分かる．新たに抽出された空間クラスタを数えたところ，その数は 1,311 クラスタであった．この 1,311 クラスタのジオタグ付きツイー

表 1 抽出空間クラスタ総数
Table 1 Number of extracted spatial clusters

	抽出空間クラスタ総数
DBSC	2109
DBASC	3403

表 2　検出率
Table 2　Detection rates

対象データ	DBSC の検出率 (%)	DBASC の検出率 (%)
日本百景	11	33
新日本観光地 100 選	42	60

表 3　データセットに存在しない観光地を除いた検出率
Table 3　Detection rates removing areas with no posted geotagged tweets

対象データ	DBSC の検出率 (%)	DBASC の検出率 (%)
日本百景	19.3	57.9
新日本観光地 100 選	47.2	67.4

トの内容を手作業で確認し，地域的なトピック（お寺などの観光地やローカルな飲食店など）を含んでいるか確認した．トピックを含む空間クラスタは，1,027 クラスタであった．よって DBASC の方が DBSC と比較して多くの地域的なトピックを含む新しい空間クラスタを抽出できたといえる．

DBASC の有効性を示すために日本百景と新日本観光地 100 選について，それぞれ 100 件に対し，DBSC と DBASC とで対象地を空間クラスタとして抽出できたかを判定する．判定基準は，対象の観光地などが存在する周辺で抽出された空間クラスタのジオタグ付きツイートの本文を確認し，その内容が対象の観光地などと一致していれば検出できたとする．検出率を**表 2** に示す．表 2 より，DBASC の方が DBSC より高検出率なのが分かるが，全体的に低検出率となっている．理由としては，2 つの対象データセットに本実験で用いたデータセットに含まれてない観光地などが含まれているからである．

次にデータセットに存在しない観光地などを，日本百景では 43 件，新日本観光地 100 選では 11 件を除いた検出率を**表 3** に示す．表 3 より，DBASC では 2 つの対象データでそれぞれ 50%以上の検出率を示している．ただし，検出できていない観光地も存在している．理由としては，島（「伊豆大島」や「九十九島」など）や地域そのもの（「八戸」や「釧路」など）は範囲が広いために，その観光地に存在するジオタグ付きツイート間の距離がパラメータ ϵ 以上離れていたためである．

抽出された空間クラスタの例として，DBSC と DBASC を用いて広島で抽出された空間クラスタを，**図 2** と**図 3** にそれぞれ示す．DBSC では 18 件，DBASC では 34 件の空間クラスタが抽出され，新たに抽出された空間クラスタは 17 件であった．図 2 と図 3 では，ジオタグ付きツイートのマーカーを作成し空間クラスタごとに色分けをして，Google Map 上にマッピングしている．

広島は，図 2 と図 3 の中央付近の地域が中心部であり，投稿密度が高い地域である．図 2 と図 3 を比較すると，

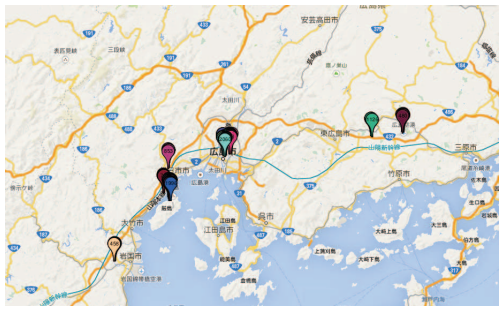

図 2　広島で抽出された空間クラスタ（DBSC）
Fig. 2　Extracted spatial clusters in Hiroshima (DBSC)

図 3　広島で抽出された空間クラスタ（DBASC）
Fig. 3　Extracted spatial clusters in Hiroshima (DBASC)

表 4　DBASC で新たに抽出された空間クラスタの例（広島）
Table 4　Example of newly extracted spatial clusters using DBASC (Hiroshima)

クラスタ No	ツイート本文
2149	恐羅漢についたどーーーーッ！！ヽ(*´▽)ノ♪ 恐羅漢楽し過ぎ！でもまだお NEW 板には乗れてない。修行と筋トレが必要。 恐羅漢 恐羅漢 ここへ来たからにはこれ。 (@ 恐羅漢)
3089	朝日山なー！ 朝日山の車どめ 朝日山なう人生初の「聖地巡礼」コンプリート！
3321	歴史のみえる丘公園に到着 歴史のみえる丘公園到着。下から登るの疲れた…

DBSC に比べて，DBASC では投稿密度が低い地域で空間クラスタを抽出できた．DBASC で新たに抽出された空間クラスタの例を**表 4** に示す．広島の観光地である「恐羅漢」，「朝日山」，「歴史のみえる丘公園」などの地域的なトピックが抽出できている．広島で新たに抽出された 17 件中 15 件の空間クラスタに地域的なトピックを含んでおり，この結果から，DBASC は DBSC と比較しておいて地域的なトピックを多く抽出できたといえる．

6.3　代表文書データ抽出の評価実験

代表文書データ抽出の評価実験では，DBASC によって広島で抽出された空間クラスタからそれぞれ投稿数が多い上位 3 位の空間クラスタで実験を行った．アンケート調査を行い，各代表文書データが空間クラスタの中心的な内容を表しているか，5 段階（1 から 5 で高い数字ほど代表文書データにふさわしい）で評価をしてもらい，代表文書デー

表 5 広島において PageRank アルゴリズムで抽出された代表文書データ
Table 5 Extracted representative documents using PageRank algorithm in Hiroshima

重要度	代表文書データ	アンケートの平均値	アンケートの標準偏差
1.00	(1) 広島に到着しました。新幹線がいっぱいです！ (@ JR 広島駅 (Hiroshima Sta.))	3.95	0.86
0.93	(2) 到着?。 (@ JR 広島駅 (Hiroshima Sta.))	3.30	1.18
0.89	(3) 広島に来てます。	3.05	0.92
1.00	(1) 原爆ドームなう	4.15	1.10
0.67	(2) ドーム！ (@ 原爆ドーム (Atomic Bomb Dome))	3.05	1.11
0.67	(3) pray (@ 原爆ドーム (Atomic Bomb Dome))	2.65	1.15
1.00	(1) 宮島なう	4.10	0.88
0.45	(2) フェリーなう♪遠くに見えるのは宮島の大鳥居！	3.60	0.96
0.31	(3) 大鳥居	3.05	1.11

表 6 広島において媒介中心性で抽出された代表文書データ
Table 6 Extracted representative documents using betweeness centrality in Hiroshima

重要度	代表文書データ	アンケートの平均値	アンケートの標準偏差
1.00	(1) 広島に到着しました。新幹線がいっぱいです！ (@ JR 広島駅 (Hiroshima Sta.))	3.95	0.86
0.41	(2) ようやく広島に到着！	3.65	0.90
0.27	(3) 広島じゃけえ。ちょっと雪降ってるんですけど?。 (@ JR 広島駅 (Hiroshima Sta.))	3.05	1.20
1.00	(1) 夜の原爆ドーム。なんか福島の原発もオーバーラップして、ちょいと考えちゃいます。明日、資料館で勉強っす。	3.35	0.85
0.85	(2) 原爆ドームなう。	4.15	1.10
0.81	(3) 今日の雨は黒くない。 (@ 原爆ドーム (Atomic Bomb Dome))	2.40	1.15
1.00	(1) フェリーなう♪遠くに見えるのは宮島の大鳥居！	3.60	0.96
0.87	(2) I'm at 宮島	3.60	0.96
0.10	(3) 大鳥居	3.05	1.11

タとして適切であるか確認した．アンケートは実施期間は 2014 年 10 月 29 日から 31 日で，20 人の被験者（大学生 20 名）に対して実施した．

表 5 と**表 6** に上位 3 位の空間クラスタの代表文書データを示す．表 5 と表 6 は PageRank アルゴリズムと媒介中心性を用いて抽出された代表文書データであり，1 位，2 位と 3 位の空間クラスタ別に，重要度，代表文書データとアンケートの結果の平均値をそれぞれ示している．

上位 3 位の空間クラスタはそれぞれ，広島駅，原爆ドーム，宮島周辺にあるが，実際に各空間クラスタに所属する文書データを閲覧しないとその内容が分からない．しかしながら，表 5 と表 6 に示すように代表文書データを見ることで，各空間クラスタが，広島駅，原爆ドーム，宮島とその大鳥居をトピックとして扱っていることが分かる．また，アンケート結果では，それぞれ高い評価が得られており，代表文書データとして妥当であると言える．

7. まとめ

本論文では，位置情報付き文書データから地域的なトピックを抽出する手法を提案した．提案手法は，(ϵ, σ)-密度に基づく適応的な空間クラスタリング手法を用いてシームレスに空間クラスタを抽出し，さらに，空間クラスタから代表文書データを抽出する新しい地域的なトピック抽出手法となっている．評価実験の結果，従来手法と比較して提案手法の方が地域的な話題を多く検出でき，また，代表文書データを抽出することで各空間クラスタの内容を把握しやすくなることを確認できた．

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] Naaman, M.: Geographic information from georeferenced social media data, *SIGSPATIAL Special*, Vol. 3, No. 2, pp. 54–61 (2011).
[2] Sakai, T., Tamura, K. and Kitakami, H.: Extracting Attractive Local-Area Topics in Georeferenced Documents using a New Density-based Spatial clustering Algorithm, *IAENG International Journal of Computer Science*, Vol. 41, pp. 131–140 (2014).
[3] Kisilevich, S., Mansmann, F. and Keim, D.: P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos, *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, COM.Geo '10, pp. 38:1–38:4 (2010).
[4] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Second International Conference on Knowledge Discovery and Data Mining* (Simoudis, E., Han, J. and Fayyad, U. M., eds.), AAAI Press, pp. 226–231 (1996).
[5] Tamura, K. and Ichimura, T.: Density-Based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Georeferenced Documents, *Proceedings of the IEEE International Conference on System, Man, and Cybernetics, SMC 2013*, pp. 2079–2084 (2013).