広島市立大学審査博士学位論文

ソーシャルメディアに対する
時空間データマイニングに関する研究

2018年9月

酒井　達弘

ソーシャルメディアに対する
時空間データマイニングに関する研究

酒井　達弘

要旨

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から有益な知識を発見することが注目されている．また，GPS 付きスマートフォンの普及により，ユーザは時間情報だけでなく位置情報をデータに付与し，ソーシャルメディア上に盛んに投稿するようになってきている．ソーシャルメディア上の位置情報付きのデータには個人的な話題だけでなく，ユーザが目にした事象や話題を含んでおり，位置情報付きのデータから実世界のトピックの分析や抽出を行うことは重要な研究課題の一つである．

ソーシャルメディア上の位置情報付きのデータを利用することで，実世界のトピックの時間変化だけでなく，空間上における変化も分析が可能となる．例えば，代表的なソーシャルメディアである 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位置情報付きのデータ，ジオタグ付きツイートが投稿されている．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の分析が可能となる．

そこで，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対象にして，実世界で注目されているトピックの分析を行う研究が盛んに行われている．その多くはデータに付与された時間情報と位置情報に着目し，データが盛んに投稿されている時間帯または領域には何かしらの有益な知識があるという考えに基づいている．しかしながら，これらの情報とともに投稿されるテキストと画像データの内容を考慮した時空間データマイニング手法は確立しているとはいえない．また，日々増加していくデータを効率的に処理するための高速化手法の開発が不可欠である．

ソーシャルメディア上の位置情報付きのデータを用いた既存の研究には，以下の五つの問題点がある．

（1）ジオタグ付きツイートに対して，分類器と時空間クラスタリングを組み合わせてトピックをリアルタイムに時空間分析するための手法が提案されていない．本研究における時空間分析とは，対象のトピックが注目されている地域の発生，その変化と消滅のモニタリングを行うことを示す．先行研究では，ジオタグ付きツイートに対してキーワード検索を行い，密度に基づく時空間クラスタリングを用いてジオタグ付きツイートが時空間上で密集している領域を抽出することで，トピックの分析を行ってい

る．しかしながら，対象のトピックに関連しないジオタグ付きツイートも時空間クラスタリングの対象としてしまうため，分類器を用いて対象のトピックに関連するジオタグ付きツイートとそうでないものを分類する必要がある．また，インクリメンタルに時空間クラスタリングを行えないため，リアルタイムに時空間分析を行うことができない．そして，トピックが注目されている地域を閲覧するためのユーザインタフェースが開発されていない．

（2）ジオタグ付きツイートに対して密度に基づく時空間クラスタリングを行う場合，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない時間帯が存在しているときに，適切に時空間クラスタの抽出ができない．これは，地域や時間帯によって，時空間クラスタを抽出する際に用いられる閾値を手動で設定することが困難なためである．

（3）ソーシャルメディア上に投稿される多種多様な画像データに対して，対象のトピックに関連する画像データのみを高性能に抽出するための画像分類が提案されていない．気象状況や自然災害などのトピックを正確にユーザへ伝えるためには，テキストを提示するよりも，画像データを提示する方がトピックを具体的に把握できるため，対象のトピックに関連する画像データのみを抽出するための高性能な画像分類は不可欠である．

（4）密度に基づくクラスタリングの代表的な手法である Density-based spatial clustering of applications with noise（DBSCAN）の高速化が十分に行われていない．最も高速とされる DBSCAN としてセルベースの DBSCAN が提案されているが，クラスタを形成するために行うセルの結合判定に多くの処理時間を要するという問題がある．

（5）ジオソーシャル画像データから，テキストと画像データの内容を考慮して各地域で注目されているトピックを抽出するための手法が提案されていない．本研究では，ソーシャルメディア上に投稿される，画像，テキストと位置情報を持つデータのことをジオソーシャル画像データと呼ぶ．先行研究では，位置情報に着目し，空間クラスタリングを用いてクラスタとしてトピックを抽出する手法が提案されているが，テキストと画像データの内容を考慮していないため，一つのクラスタに複数の異なるトピックが含まれてしまう．

本論文では，上記の五つの問題点を解決し，ソーシャルメディア上に投稿される時間情報と位置情報が付与されたテキストと画像データに対する時空間データマイニング手法の確立を目指す．具体的に，以下の五つの目的を達成する．

（1）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

Twitter 上に投稿されるジオタグ付きツイートを用いてトピックを時空間分析するための手法，密度に基づく時空間分析手法を提案する．提案手法は，対象となってい

るトピックの内容を含むジオタグ付きツイートを抽出するために，ナイーブベイズ分類器を用いてジオタグ付きツイートを分類する．次に，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを用いて，トピックが注目されている地域を時空間クラスタとしてリアルタイムに抽出する．そして，抽出された時空間クラスタについてその領域，ツイート内容と画像データを Web アプリケーション上に提示する．実際に Twitter 上からジオタグ付きツイートを収集し，トピックを「大雨」と設定して評価実験を行った結果，ツイート分類の交差検定における F 値として 0.78 を示した．また，トピックが注目されている地域を検出できたか評価を行った結果，検出率として 0.52 を示した．そして，抽出された時空間クラスタを Web アプリケーション上で確認することによって，本研究が目的とするトピックの時空間分析が可能であることを確認できた．

（2）密度に基づく適応的な時空間クラスタリング

密度に基づく時空間分析手法において，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない時間帯を区別することなく時空間クラスタを抽出するために，(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する．提案手法は，各地域，各時間帯における統計的な投稿数を用いて，時空間クラスタを抽出する基準となる閾値を適応的に変化させている．提案手法を密度に基づく時空間分析手法に導入し，トピック「大雨」について注目されている地域を検出できたか評価を行った結果，既存手法は閾値を変化させた場合，検出率が 0.73 から 0.32 まで落ちるのに対して，提案手法は閾値を変化させたとしても，0.80 から大きく変化することなく，高い検出率を示すことができた．

（3）密度に基づく時空間分析手法における画像分類

密度に基づく時空間分析手法において，対象となっているトピックに関連している画像データのみを抽出するための画像分類を提案する．提案する画像分類では，Bag-of-Features（BoF）または学習済み深層ネットワークを用いて画像データから特徴ベクトルを抽出し，Support Vector Machine（SVM）を用いて分類を行う．トピックを「大雨」と設定して行った評価実験の結果，提案手法は特徴ベクトル抽出手法として学習済み深層ネットワークである VGG-16 を用いた場合，交差検定における正解率として 0.89 示し，高性能に画像データを分類することができた．

（4）最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN

DBSCAN の高速化のために，最小外接矩形（MBR）とセルの再帰分割を用いたセルベースの DBSCAN を提案する．提案手法では，セルベースの DBSCAN のセルの結合判定について，セル中のデータを囲む MBR を作成し，MBR 間の距離を用いることで，条件を満たす場合に高速に判定することができる．また，セルを再帰的に分割し，計算の対象となるデータを減らしていくことで，高速にセルの結合判定ができ

る．人工データを用いて行った評価実験の結果，提案手法は既存手法よりも高速化ができた．特に，高次元のデータになるほど大幅な高速化ができた．

（5）密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出

ジオソーシャル画像データから各地域で注目されているトピックを抽出するために，密度に基づくマルチモーダル空間クラスタリングを用いたトピックの抽出手法を提案する．提案手法は，(ϵ, σ)-密度に基づく空間マルチモーダルクラスタリングを用いて，空間的また内容的にも類似したジオソーシャル画像データが密に投稿されている注目領域をマルチモーダル空間クラスタとして抽出する．また，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出し分かりやすくするために，ネットワークベースの重要度算出手法を用いて，代表画像データを抽出する．Twitter 上に投稿されるジオソーシャル画像データを用いて行った評価実験の結果，京都の「清水寺」,「渡月橋」や「金閣寺」などのトピックを抽出することができた．

本論文では，ソーシャルメディア上の位置情報付きのデータを用いた既存の研究にある五つの問題点を挙げ，ソーシャルメディア上のデータに対する時空間データマイニング手法の確立を目指した．五つの問題点を解決するために，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，密度に基づく適応的な時空間クラスタリング，密度に基づく時空間分析手法における画像分類，最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN と密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出の五つの目的を達成した．これらの研究成果は，ソーシャルメディアに対する時空間データマイニングの基盤となる重要な技術であり，本研究の目的とする時空間データマイニング手法の確立を達成できたといえる．今後の研究では，より高性能な時空間データマイニングを行うために，テキストや画像データに対して深層学習を用いた新しい手法を開発すること，ソーシャルメディア上のデータと気温や降雨量などの気象観測データを組み合わせた手法を開発すること，実用化を視野に入れ，手法の並列化による高速化を行うことが挙げられる．

キーワード：ソーシャルメディア，時空間データマイニング，密度に基づくクラスタリング，DBSCAN，深層学習，マルチモーダルデータ

A Study on Spatiotemporal Data Mining for Social Media

Tatsuhiro Sakai

Abstract

In the era of big data, the number of data posted on social media has been increasing rapidly, and the knowledge discovery from data on social media is paid the attention. Moreover, with the popularization of GPS-enabled smartphones, a significant amount of geo-annotated data, which include not only the posted time but also the posted location, are posted on social media. Geo-annotated data are typically related to not only personal topics but also social topics. Therefore, analyzing real-world topics using geo-annotated data is one of the most important challenges in many application domains.

Using geo-annotated data on social media enables the analysis of real-world topics with temporal and spatial variations. For example, users post geo-tagged tweet, which is geo-annotated text and image data on Twitter, related to what they are witnessing during natural disasters such as earthquakes, typhoons, and floods. Areas where natural disasters occur and the temporal variation in the natural disasters can be analyzed using these geo-annotated data.

Many researchers have been tackling the development of new data mining techniques for geo-annotated data to analyze real-world topics. Most of them are based on the idea that by focusing on the posted time and posted location, some useful information are available in the time periods and areas with many posts. However, the spatiotemporal data mining method considering the contents of text and image data together with these information has not been established. Moreover, it is necessary to develop a speed-up method to efficiently process the data that have been increasing.

The previous studies using geo-annotated data on social media have had the following problems:

(1) The method of spatiotemporal analysis of topics using a classifier and spatiotemporal clustering to geo-tagged tweets has not been proposed. The spatiotemporal analysis of this work is to identify the occurring, changing, and disappearing of target topics. In the previous work, the method, which extracts geo-tagged tweets related to a target topic using the keyword search, and the areas of interest of the target topic using the density-based spatiotemporal clustering, have been proposed. However, the previous method should

classify the geo-tagged tweets into tweets that are related to the target topic and tweets that are not related to it using the classifier, because the geo-tagged tweets are extracted if they include a keyword, not a target topic. In addition, the density-based spatiotemporal clustering does not support real-time extraction. Moreover, the interface has not been developed for browsing the areas of interest of the target topic.

(2) Appropriately extracting spatiotemporal clusters is difficult in the density-based spatiotemporal clustering when the number of geo-tagged tweets between areas or time periods is difference. This is because manually setting the threshold of spatiotemporal clustering is difficult for each area and time period.

(3) The image classification with high performance to extract images including the target topic from images of many variations on social media has not been proposed. When visualizing emergency topics such as natural disasters, images are better than texts, because it is easier to understand the content. Therefore, the image classification with high performance to extract images including the target topic is necessary.

(4) The acceleration of density-based spatial clustering of applications with noise (DBSCAN), which is a typical algorithm of density-based clustering, have not been sufficiently carried out. The cell-based DBSCAN is one of the fastest DBSCANs. The connecting-cells step, which is performed to form clusters, of the previous cell-based DBSCAN is time-consuming.

(5) The method considered text and image data to extract the topics of interest for each area from geo-social images has not been proposed. In this work, the geo-social image is data included text, image data, and the posted location on social media. In the previous work, the topic extraction method using the spatial clustering based on the posted location has been proposed. However, different topics are included in a cluster because the method does not consider the text and image data.

This paper aims to establish a spatiotemporal data mining method for the text and image data included the posted time and posted location on social media by solving the above problems. This study achieves the following purposes:

(1) Spatiotemporal analysis of topics using density-based clustering

To analyze topics using geo-tagged tweets, we propose a spatiotemporal analysis method termed the density-based spatiotemporal analysis method. First, the proposed method classifies geo-tagged tweets using the Naive Bayes classifier to extract geo-tagged tweets including the target topic. Next, the proposed method extracts the areas of interest of the target topic as spatiotemporal clusters in real-time using the incremental algorithm of the

(ϵ, τ)-density-based spatiotemporal clustering. Subsequently, we can observe the areas of interest of the target topic through a Web application. We conducted experiments using actual geo-tagged tweets, and set the target topic as "heavy rain." The experimental results of the tweet classification using cross-validation showed 0.78 as the F value. The experimental results of the spatiotemporal clustering showed 0.52 for the detection rates of heavy-rain areas. Moreover, the proposed method could identify the occurrence, change in, and disappearance of "heavy rain" through a Web application.

(2) Density-based adaptive spatiotemporal clustering

To extract spatiotemporal clusters without distinguishing the difference in the number of geo-tagged tweets between areas or time periods, we propose a (ϵ, τ)-density-based adaptive spatiotemporal clustering. The proposed method adaptively changes the threshold value used for spatiotemporal clustering using statistical data for each area and time period. We performed the experiments while changing the threshold value. The experimental results show that for detection rates of areas experiencing "heavy rain," the proposed method has an accuracy of 0.80, in contrast to the range of the previous method from 0.32–0.73.

(3) Image classification in density-based spatiotemporal analysis method

To extract images including the target topic in the density-based spatiotemporal analysis method, we propose an image classifier. First, the proposed image classifier extracts the feature vectors of the images by using the bag-of-features (BoF) or pre-trained deep network. Subsequently, the proposed image classifier classifies the images into ones that are related to a target topic or ones that are not related to it using a support vector machine (SVM). The experimental results using cross-validation show 0.89 as the classification accuracy of the images related to "heavy rain" when the VGG-16, which is pre-trained deep network, is used.

(4) Cell-based DBSCAN using minimum bounding rectangle and recursive cell partitioning

To speed up DBSCAN, we propose the cell-based DBSCAN using the minimum bounding rectangle (MBR) and recursive cell partitioning. The proposed method can accelerate the connecting-cells step using MBR and recursive cell partitioning when the condition is satisfied. The experimental results show that the proposed method outperforms the previous methods.

(5) Topic extraction using density-based multimodal spatial clustering

To extract the topics of interest from geo-social images, we propose a topic extraction method using a density-based multimodal spatial clustering. The proposed method extracts multimodal spatial clusters that are spatially and semantically separated from

other spatial clusters using (ϵ, σ)-density-based multimodal spatial clustering. Moreover, to present the primary topics for each multimodal spatial cluster, representative images are detected using network-based importance analysis. We conducted experiments using actual geo-tagged tweets that include photo images. The experimental results show that the proposed method can extract topics such as the "Kiyomizu-dera," "Togetsu bridge," and "Kinkaku-ji" in Kyoto.

In this study, to establish a spatiotemporal data mining method for the text and image data that include the posted time and posted location on social media, we achieved five purposes: spatiotemporal analysis of topics using density-based clustering, density-based adaptive spatiotemporal clustering, image classification in density-based spatiotemporal analysis method, cell-based DBSCAN using MBR, and recursive cell partitioning, and topic extraction using density-based multimodal spatial clustering. We could establish a spatiotemporal data mining method, which is the aim of this study, because these research results are important technologies form the foundation of spatiotemporal data mining for social media. In our future work, we intend to develop a method using deep learning for texts and images, and a method that combines the data on social media and meteorological observation data such as temperature and amount of rainfall, to perform spatiotemporal data mining with higher performance. Moreover, we intend to conduct the acceleration of the proposed methods by a parallelization method for a more practical realization.

Keywords: social media, spatiotemporal data mining, density-based clustering, DBSCAN, deep learning, multimodal data

目次

図目次

表目次

第 1 章　序論

本章では，研究の背景と目的，関連研究，本論文の構成について説明する．

1.1　研究の背景

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から有益な知識を発見することが注目されている．また，GPS 付きスマートフォンの普及により，ユーザは時間情報だけでなく位置情報をデータに付与し，ソーシャルメディア上に盛んに投稿するようになってきている [1, 2]．ソーシャルメディア上の位置情報付きのデータには個人的な話題だけでなく，ユーザが目にした事象や話題を含んでおり，位置情報付きのデータから実世界のトピックの分析や抽出を行うことは重要な研究課題の一つである [3, 4]．

ソーシャルメディア上の位置情報付きのデータを利用することで，実世界のトピックの時間変化だけでなく，空間上における変化も分析が可能となる．例えば，代表的なソーシャルメディアである 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位置情報付きのデータ，ジオタグ付きツイートが投稿されている [5, 6, 7, 8]．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の分析が可能となる．

そこで，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対象にして，実世界で注目されているトピックの分析を行う研究が盛んに行われている [9, 10]．その多くはデータに付与された時間情報と位置情報に着目し，データが盛んに投稿されている時間帯または領域には何かしらの有益な知識があるという考えに基づいている [11, 12]．しかしながら，これらの情報とともに投稿されるテキストと画像データの内容を考慮した時空間データマイニング手法は確立しているとはいえない．また，日々増加していくデータを効率的に処理するための高速化手法の開発が不可欠である．

ソーシャルメディア上の位置情報付きのデータを用いた既存の研究には，以下の五つの問題点がある．

（1）ジオタグ付きツイートに対して，分類器と時空間クラスタリングを組み合わせてトピックをリアルタイムに時空間分析するための手法が提案されていない．本研究における時空間分析とは，対象のトピックが注目されている地域の発生，その変化と消滅のモニタリングを行うことを示す．先行研究では，ジオタグ付きツイートに対してキーワード検索を行い，密度に基づく時空間クラスタリングを用いてジオタグ付きツイートが時空間上で密集している領域を抽出することで，トピックの分析を行っている [12]．しかしながら，対象のトピックに関連しないジオタグ付きツイートも時空間クラスタリングの対象としてしまうため，分類器を用いて対象のトピックに関連する

ジオタグ付きツイートとそうでないものを分類する必要がある．また，インクリメンタルに時空間クラスタリングを行えないため，リアルタイムに時空間分析を行うことができない．そして，トピックが注目されている地域を閲覧するためのユーザインタフェースが開発されていない．

（2）ジオタグ付きツイートに対して密度に基づく時空間クラスタリングを行う場合，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない時間帯が存在しているときに，適切に時空間クラスタの抽出ができない．これは，地域や時間帯によって，時空間クラスタを抽出する際に用いられる閾値を手動で設定することが困難なためである．投稿数の多い場合と少ない場合に分けて閾値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．

（3）ソーシャルメディア上に投稿される多種多様な画像データに対して，対象のトピックに関連する画像データのみを高性能に抽出するための画像分類が提案されていない．気象状況や自然災害などのトピックを正確にユーザへ伝えるためには，テキストを提示するよりも，画像データを提示する方がトピックを具体的に把握できるため，対象のトピックに関連する画像データのみを抽出するための高性能な画像分類は不可欠である．

（4）密度に基づくクラスタリングの代表的な手法である Density-based spatial clustering of applications with noise（DBSCAN）[13, 14] の高速化が十分に行われていない．最も高速とされる DBSCAN としてセルベースの DBSCAN が提案されている [15, 16]．セルベースの DBSCAN は，データセット全体を小さいセルに分割し，密度をセル単位で考えることによって，処理の高速化が実現できている．セルベースの DBSCAN では，二つのセル間に距離 ϵ 以内の任意のデータのペアがある場合にのみ，その二つのセルを結合してクラスタを形成する．このセルの結合判定は，単純な方法を用いると条件を満たすペアが見つからない場合に二つのセル間の全てのデータ間の距離計算を行わなければならないため，多くの処理時間を要する．

（5）ジオソーシャル画像データから，テキストと画像データの内容を考慮して各地域で注目されているトピックを抽出するための手法が提案されていない．本研究では，ソーシャルメディア上に投稿される，画像，テキストと位置情報を持つデータのことをジオソーシャル画像データと呼ぶ．先行研究では，位置情報に着目し，空間クラスタリングを用いてクラスタとしてトピックを抽出する手法 [11, 17] が提案されているが，テキストと画像データの内容を考慮していないため，一つのクラスタに複数の異なるトピックが含まれてしまう．

1.2　研究の目的

本論文では，1.1 節にて述べた五つの問題点を解決し，ソーシャルメディア上に投稿される時間情報と位置情報が付与されたテキストと画像データに対する時空間データマイニング手法の確立を目指す．具体的に，以下の五つの目的を達成する．

（1）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

Twitter 上に投稿されるジオタグ付きツイートを用いてトピックを時空間分析するための手法，密度に基づく時空間分析手法を提案する．提案手法は，対象となっているトピックの内容を含むジオタグ付きツイートを抽出するために，ナイーブベイズ分類器 [18] を用いてジオタグ付きツイートを分類する．次に，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを用いて，トピックが注目されている地域を時空間クラスタとしてリアルタイムに抽出する．そして，抽出された時空間クラスタについてその領域，ツイート内容と画像データを Web アプリケーション上に提示する．

（2）密度に基づく適応的な時空間クラスタリング

密度に基づく時空間分析手法において，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない時間帯を区別することなく時空間クラスタを抽出するために，(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する．提案手法は，各地域，各時間帯における統計的な投稿数を用いて，時空間クラスタを抽出する基準となる閾値を適応的に変化させている．よって，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない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．

（3）密度に基づく時空間分析手法における画像分類

密度に基づく時空間分析手法において，対象となっているトピックに関連している画像データのみを抽出するための画像分類を提案する．提案手法は，Bag-of-Features（BoF）[19] または学習済み深層ネットワークを用いて画像データから特徴ベクトルを抽出する．次に，抽出した画像データの特徴ベクトルを使用して Support Vector Machine（SVM）を学習させ，当該トピックに関連する画像データかどうか分類する．そして，当該トピックに関連する画像データのみを Web アプリケーション上に提示することで，密度に基づく時空間分析手法の有効性を向上することができる．

（4）最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN

DBSCAN の高速化のために，最小外接矩形（MBR）とセルの再帰分割を用いたセルベースの DBSCAN を提案する．提案手法では，セルベースの DBSCAN のセルの結合判定について，セル中のデータを囲む MBR を作成し，MBR 間の距離を用いる

ことで，条件を満たす場合に高速に判定することができる．また，セルを再帰的に分割し，計算の対象となるデータを減らしていくことで，高速にセルの結合判定ができる．

（5）密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出

ジオソーシャル画像データから各地域で注目されているトピックを抽出するために，密度に基づくマルチモーダル空間クラスタリングを用いたトピックの抽出手法を提案する．提案手法は，(ϵ, σ)-密度に基づく空間マルチモーダルクラスタリングを用いて，空間的また内容的にも類似したジオソーシャル画像データが密に投稿されている注目領域をマルチモーダル空間クラスタとして抽出する．ジオソーシャル画像データ間の類似度を正確に算出するために，マルチモーダル類似度を定めている．また，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出し分かりやすくするために，ネットワークベースの重要度算出手法を用いて，代表画像データを抽出する．

1.3　関連研究

ソーシャルメディア上に投稿されるデータを用いて，実世界で注目されているトピックの分析や抽出を行う研究が盛んに行われている．例えば，投稿されるテキストに着目し，分類器を用いて特定のトピックに関する情報を抽出する研究 [20, 21, 22, 23] や頻出している語句を抽出することで注目されているトピックを抽出する研究 [24, 25] が行われている．また，ソーシャルメディア上に投稿される画像データに着目した研究も行われている [26, 27, 28]．しかしながら，これらの研究はテキストや画像データとともに投稿される時間情報と位置情報を考慮していないため，本論文の目的とする時空間データマイニングとは異なる．

ソーシャルメディア上のデータに付与された位置情報に着目し，空間クラスタリングを用いてデータが盛んに投稿されている領域を抽出する研究が行われている [11, 29, 30, 31]．また，位置情報だけでなく時間情報を考慮することで，時空間上でトピックを分析する研究が行われている [12, 32, 33, 34]．しかしながら，これらの研究はテキストや画像データの内容を考慮していない，もしくはキーワード検索によって分析対象のトピックに関連するデータを取り出している．キーワード検索を行うのみでは，対象のトピックに関連しないデータも含まれてしまうため，分類器を用いてテキストや画像データを分類する必要がある．

ソーシャルメディア上に投稿されるテキストと位置情報に着目した研究が行われている．例えば，分類器を用いてテキスト分類を行い，テキストに付与された位置情報を用いて地図上にマッピングを行うことで，特定のトピックに関する情報の投稿分布を確認することができる [35, 36, 37, 38]．また，特定の地域でのみ頻出している語句を抽出することで，地域ごとに注目されているトピックを抽出することができる [39, 40]．さらに，画像データと位置情報を用いて，投稿された画像データを地図上に可視化する研究が行われている [41, 42, 43]．

これらの研究はテキストや画像データとともに投稿される位置時間を考慮しているが，時間情報を考慮していないため，時空間上でのトピックの分析を行うことはできない．

既存の研究では，ソーシャルメディア上に投稿されるデータに対して，テキストと画像データに着目して分類器を，時間情報と位置情報に着目して時空間クラスタリングを用いているが，それらを組み合わせた手法は提案されていない．本論文では，既存の研究では行われていない，時間情報と位置情報とともに投稿されるテキストと画像データの内容を考慮した時空間データマイニング手法の確立を目指す．

1.4　本論文の構成

第 2 章では，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について説明する．また，Twitter 上のジオタグ付きツイートを収集し，モニタリングの対象とするトピックを「大雨」と「大雪」と設定して行った評価実験の結果を示す．第 3 章では，密度に基づく適応的な時空間クラスタリングについて説明し，密度に基づく時空間分析手法に導入して行った評価実験の結果を示す．第 4 章では，密度に基づく時空間分析手法における画像分類を提案し，密度に基づく時空間分析手法に導入して行った評価実験の結果を示す．第 5 章では，最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN について説明する．また，人工データと実データを用いて行った評価実験の結果を示す．第 6 章では，密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出について説明する．また，Twitter 上に投稿された画像データを持つジオタグ付きツイートを用いて行った評価実験の結果を示す．第 7 章では，本論文をまとめ，今後の課題を述べる．

第 2 章　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

本章では，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について説明する．

2.1　はじめに

近年，インターネット上のユーザはソーシャルメディアを通して位置情報付きのデータを盛んに発信するようになってきている．ソーシャルメディア上に投稿される位置情報付きデータの中には，ユーザの個人的な話題だけでなく，ユーザの身の回りで発生した出来事やイベントの状況をリアルタイムに伝えるデータが存在しており，それらのデータから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．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位置情報付きのデータ，ジオタグ付きツイートが投稿されている．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の分析が可能となる．

位置情報付きデータは地理空間データとして扱うことができるため，空間クラスタリングを用いて空間クラスタを発見することで，トピックに関するデータが投稿されている地域を見つけることができる．地理空間データの空間クラスタリング手法の中で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 [13, 14] が提案されている．密度に基づく空間クラスタリングは，空間データが密集している高密度な領域を，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．よって，密度に基づく空間クラスタリングを用いることで盛んに投稿が行われている地域，つまり，ある事象が注目されている地域を取り出すことができる．

密度に基づく空間クラスタリングを応用して，ジオタグ付きツイートを用いてトピックの時空間分析をする研究が行われている [12]．この先行研究では，密度に基づく空間クラスタリングを密度に基づく時空間クラスタリングへと拡張し，(ϵ, τ)-密度に基づく時空間クラスタリングを提案している．(ϵ, τ)-密度に基づく時空間クラスタリングは，モニタリングの対象となっているトピックを含むジオタグ付きツイートをキーワード検索で求め，ジオタグ付きツイートの投稿が空間的かつ時間的に高密度な地域を時空間クラスタとして抽出する．モニタリングの対象となっているトピックのキーワードを含むジオタグ付きツイートが時空間的に高密度な地域を抽出することで，モニタリングの対象となっているトピックが注目されている地域を抽出することができる．

本研究では，Twitter 上に投稿されるジオタグ付きツイートを用いて対象となっているト

ピックをリアルタイムに時空間分析するための手法の開発を目指す．本研究が目的とする時空間分析とは，対象とするトピックが注目されている地域の発生，その変化と消滅のモニタリングを行うことを示す．ある地域で特定のトピックが発生した時には，そのトピックに関連する内容のジオタグ付きツイートが盛んに投稿される．対象とするトピックに関連するジオタグ付きツイートの投稿を時空間上で捉え，投稿されるジオタグ付きツイートを閲覧することで，時空間分析を可能にする．

先行研究で提案されている手法を用いて本研究の目的である時空間分析を行うには，以下の問題点がある．

- モニタリングの対象となっているトピックを含むジオタグ付きツイートをキーワード検索のみで求めているため，トピックに関連しないジオタグ付きツイートも時空間クラスタリングの対象としてしまう．この問題は，キーワードを含むツイートが必ずしもトピックには関連するとは限らないために発生する．
- リアルタイムに処理を行うことができないため，注目されている地域と各地域に投稿されたジオタグ付きツイートの内容をリアルタイムに確認できない．
- トピックが注目されている地域を閲覧するためのユーザインタフェースが開発されていない．

本研究では，これらの問題点を解決した密度に基づく時空間分析手法を提案する．提案手法の特長は以下の 3 点である．

- モニタリングの対象となっているトピックに関する内容を含むジオタグ付きツイートを抽出するために，ナイーブベイズ分類器 [18] を用いてジオタグ付きツイートの分類を行う．ジオタグ付きツイートをモニタリング対象のトピックに関連するものとそれ以外に分類し，トピックに関連するジオタグ付きツイートのみを時空間クラスタリングの対象とする．
- トピックが注目されている地域を抽出するために，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを用いて時空間クラスタを抽出する．インクリメンタルなアルゴリズムを用いることで，注目されている地域を時空間クラスタとしてリアルタイムに抽出できる．
- 抽出された時空間クラスタについてその領域，ツイート内容と画像データを Web アプリケーション上に提示する．Web アプリケーション上では時空間クラスタに含まれるツイートと画像データがリアルタイムに地図上に表示される．抽出した時空間クラスタは時空間上に保存してあり，時間を前後することで時空間クラスタの発生，その変化と消滅を確認することができる．

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収集し，評価実験を行っ

た．評価実験の結果，提案手法はトピック「大雨」と「大雪」について，ツイート分類と時空間クラスタリングについて有効性を示すことができた．また，作成した Web アプリケーション上でトピックの時空間クラスタの発生と消滅をリアルタイムに確認することができた．

本章の構成は以下の通りである．2.2 節では，関連研究を述べる．2.3 節では，(ϵ, τ)-密度に基づく時空間クラスタリングについて説明する．2.4 節では，提案手法である密度に基づく時空間分析手法について述べる．2.5 節では，評価実験の結果を示し，2.6 節で本章をまとめる．

2.2　関連研究

近年，ソーシャルメディア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上の位置情報付きのデータは急速に増加している．その中でも，最も普及しているソーシャルメディアの一つである 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の時空間上における変化を分析する研究が盛んに行われている．例えば，Sakaki ら [35] 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 [44] 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．また，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に対しバースト検出手法を用いることで，大雨や大雪などのトピックの注目度をバーストとして捉える手法が提案されている [45, 46, 47, 48, 49, 50]．このように，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 [51, 52, 53].

Murakami ら [24] 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，Vieweg ら [5] 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 [54] 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．Hwang ら [55] は，ツイートに対して時空間分析を行うことで，インフルエンザの流行の分析を行った．また，Marcus ら [56] は，Twitinfo という可視化システムを開発している．Twitinfo は，グローバルなトピックとその時間変化を分析するのに適している．しかしながら，これらの研究は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，トピックの時空間上における変化のモニタリングを可能にする．

また，Thom ら [57] 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 [21] 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注目されている詳細な地域を把握することができない．本研究の提案手法では，時空間クラスタリングによって対象とするトピックが注目されている地域の発生，その変化と消滅を把握することができる．

Avvenuti ら [58] 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 [59] は，トピックの時空間的な変化を可視化する mTrend と呼ばれるシステムを提案した．Kumar ら [38] 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．このように多くの研究者が Twitter 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の抽出，分析やモニタリングのために様々なアプローチを行っている．本研究の先行研究である (ϵ, τ)-密度に基づく時空間クラスタリング [12] は，モニタリングの対象となっているトピックを含むジオタグ付きツイートをキーワード検索で求め，時空間的に高密度に投稿されている地域を抽出している．(ϵ, τ)-密度に基づく時空間クラスタリングは，距離 ϵ と投稿間隔 τ を基準として時空間クラスタリングを行い，大雨や大雪などのトピックが注目されている地域を時空間クラスタとして抽出できる．

2.3　(ϵ, τ)-密度に基づく時空間クラスタリング

本節では，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 [12] について説明する．

2.3.1　概要

密度に基づく空間クラスタリング [13, 14] では，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，クラスタリングを行う．この手法は，クラスタ内の各データは定義された近傍内に，$MinPts$（ユーザパラメータ）以上のデータが存在しなければならないということに基づいている．密度に基づいているため，図 1 のように円状ではないクラスタを抽出することが可能である．よって，密度に基づく空間クラスタリングは地理空間データのクラスタリングとして広く用いられている．

(ϵ, τ)-密度に基づく時空間クラスタリングでは，密度に基づく空間クラスタリングを拡張し，時空間的に密集しているジオタグ付きツイート集合を時空間クラスタとして定義している．そのために，(ϵ, τ)-密度に基づく時空間クラスタリングでは，近傍の定義を空間上における距離 ϵ 以内かつ時間上における投稿間隔 τ 以内に存在するジオタグ付きツイート集合としている．そして，この近傍に存在するジオタグ付きツイートの数が $MinGT$（ユーザパラ

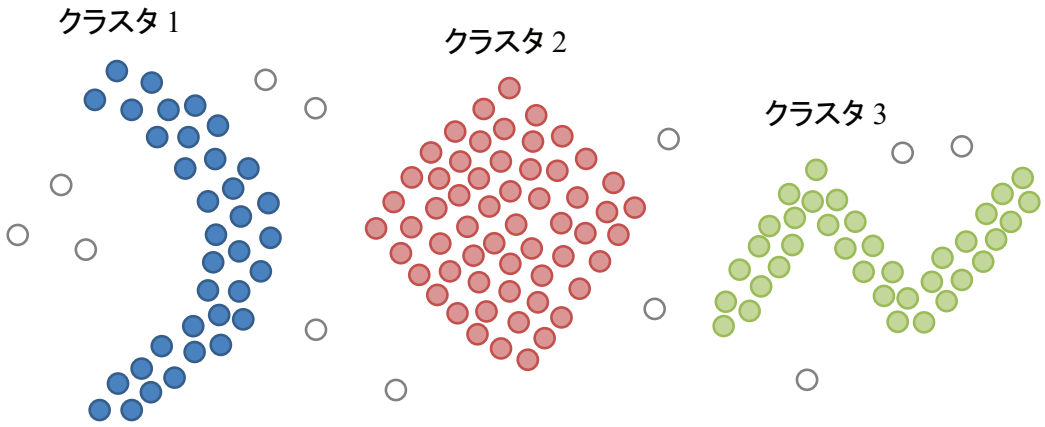


図 1: 密度に基づく空間クラスタリング

メータ）以上であるジオタグ付きツイートを核として時空間クラスタを作成する．

2.3.2　諸定義

ジオタグ付きツイート集合を GT とし，ジオタグ付きツイート gt_p の (ϵ, τ)-密度に基づく近傍を次のように定義する．

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

$$N_{(\epsilon,\tau)}(gt_p) = \{gt_q \in GT \mid dist(gt_p, gt_q) \leq \epsilon \ \ and \ \ iat(gt_p, gt_q) \leq \tau\} \tag{1}$$

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 2 に定義 1 の例を示す．ジオタグ付きツイート gt_p から距離 ϵ 以内に，gt_2，gt_3 と gt_4 の 3 つ，また，gt_p から投稿間隔 τ 以内にも，gt_2，gt_3 と gt_4 の 3 つのジオタグ付きツイートが存在する．このとき，gt_p の (ϵ, τ)-密度に基づく近傍は，gt_2，gt_3 と gt_4 の 3 つとなる．一方，gt_q の距離 ϵ 以内には，gt_2，gt_3 と gt_5 の 3 つのジオタグ付きツイートが存在するが，そのうち投稿間隔 τ 以内に存在するのは gt_2 と gt_3 であるため，gt_q の (ϵ, τ)-密度に基づく近傍は gt_2 と gt_3 の 2 つとなる．

定義 2（核ジオタグ付きツイート，周辺ジオタグ付きツイート）　ジオタグ付きツイート gt_p の (ϵ, τ)-密度に基づく近傍 $N_{(\epsilon,\tau)}(gt_p)$ について，$|N_{(\epsilon,\tau)}(gt_p)| \geq MinGT$ を満たすジオタ

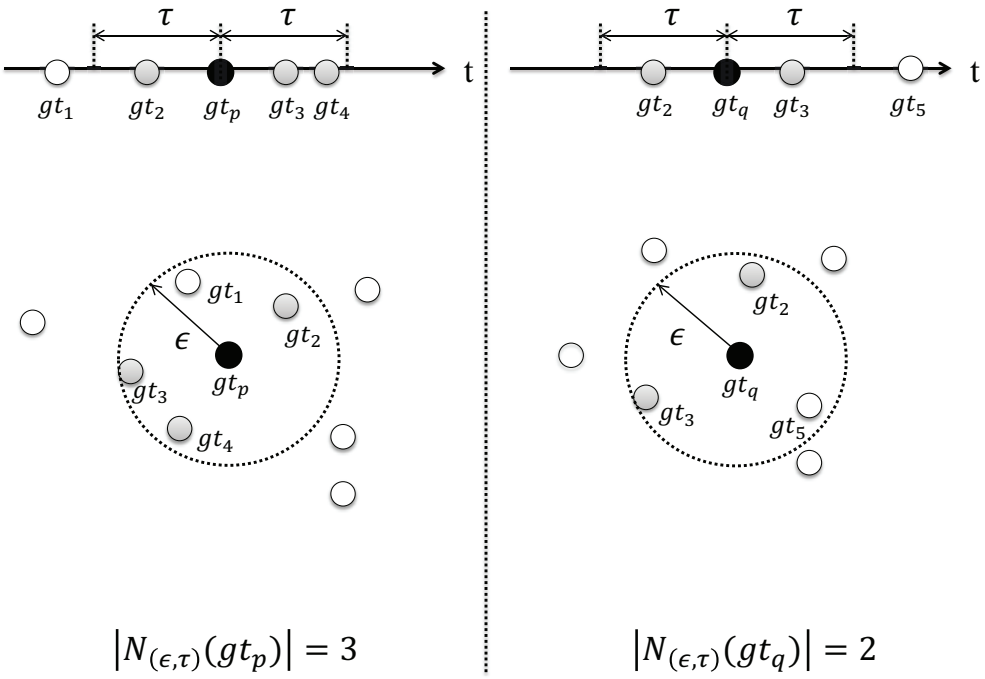


図 2: 定義 1，2 と 3 の例

グ付きツイート gt_p を核ジオタグ付きツイート，$|N_{(\epsilon,\tau)}(gt_p)| < MinGT$ であるジオタグ付きツイートを周辺ジオタグ付きツイートと呼ぶ．

$MinGT$ はユーザが与えるパラメータである．(ϵ,τ)-密度に基づく時空間クラスタリングでは，核ジオタグ付きツイートの集合が時空間クラスタの核となる．

図 2 を使って定義 2 の例を示す．$MinGT = 3$ とすると，ジオタグ付きツイート gt_p は $|N_{(\epsilon,\tau)}(gt_p)| \geq MinGT$ を満たすため，核ジオタグ付きツイートである．一方，ジオタグ付きツイート gt_q は $|N_{(\epsilon,\tau)}(gt_q)| \geq MinGT$ を満たさないため，周辺ジオタグ付きツイートとなる．

定義 3（(ϵ,τ)-密度的に直接到達可能） ジオタグ付きツイート gt_q がジオタグ付きツイート gt_p の (ϵ,τ)-密度に基づく近傍に存在し，$|N_{(\epsilon,\tau)}(gt_p)| \geq MinGT$ を満たす時，ジオタグ付きツイート gt_q はジオタグ付きツイート gt_p から (ϵ,τ)-密度的に直接到達可能であると表現する．

図 2 を使って定義 3 の例を示す．ジオタグ付きツイート gt_p は核ジオタグ付きツイートであるため，gt_2，gt_3 と gt_4 は gt_p 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） ジオタグ付きツイート gt_{p+1} がジオタグ付きツ

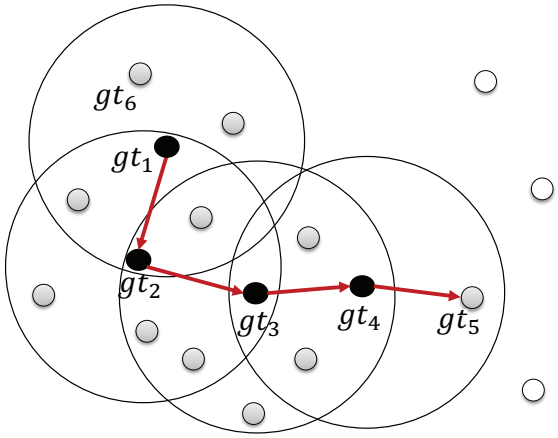


図 3: 定義 4 の例

イート gt_p から (ϵ, τ)-密度的に直接到達可能である，ジオタグ付きツイート列 $(gt_p, gt_{p+1}, \cdots, gt_{p+n})$ を考える．この時，ジオタグ付きツイート gt_p から gt_{p+l} へ，(ϵ, τ)-密度的に到達可能であると表現する．

図 3 に定義 4 の例を示す．ジオタグ付きツイート列 $(gt_1, gt_2, \cdots, gt_5)$ は gt_{p+1} がジオタグ付きツイート gt_p から (ϵ, τ)-密度的に直接到達可能である．つまり，ジオタグ付きツイート gt_5 は gt_1 から (ϵ, τ)-密度的に到達可能である．

定義 5（(ϵ, τ)-密度的に接続） ジオタグ付きツイート gt_p とジオタグ付きツイート gt_q とがジオタグ付きツイート gt_o から (ϵ, τ)-密度的に到達可能であり，ジオタグ付きツイート gt_o が $|N_{(\epsilon,\tau)}(gt_o)| \geq MinGT$ を満たす時，ジオタグ付きツイート gt_p とジオタグ付きツイート gt_q は (ϵ, τ)-密度的に接続していると表現する．

図 4 に定義 5 の例を示す．ジオタグ付きツイート gt_6 は gt_2 から (ϵ, τ)-密度的に到達可能である．また，ジオタグ付きツイート gt_5 も gt_2 から (ϵ, τ)-密度的に到達可能である．つまり，ジオタグ付きツイート gt_6 と gt_5 は (ϵ, τ)-密度的に接続している．

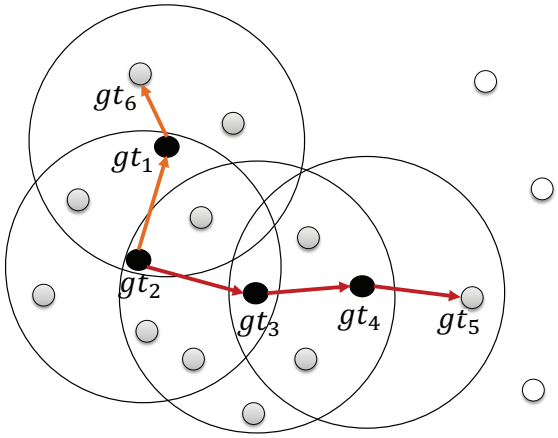


図 4: 定義 5 の例

2.3.3　(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では，密集している核ジオタグ付きツイートとその周辺ジオタグ付きツイートを時空間クラスタとして定義している．(ϵ, τ)-密度に基づく時空間クラスタを次のように定義する．

定義 6（(ϵ, τ)-密度に基づく時空間クラスタ） ジオタグ付きツイート集合 GT 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ジオタグ付きツイート集合である．

(1)　任意のジオタグ付きツイート $gt_p \in GT$ と $gt_q \in GT$ について，(ϵ, τ)-密度に基づく時空間クラスタ stc に gt_p が所属（$gt_p \in stc$）し，gt_q が gt_p から (ϵ, τ)-密度的に到達可能であれば，gt_q は stc に所属（$gt_q \in stc$）する．

(2)　(ϵ, τ)-密度に基づく時空間クラスタ stc に所属する任意のジオタグ付きツイート $gt_p \in stc$ と $gt_q \in stc$ は，(ϵ, τ)-密度的に接続している．

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

Input: GT, ϵ, τ, $MinGT$

Output: STC

```
1: STC ← φ
2: for i ← 1 to |GT| do
3:    gt_p ← gt_i ∈ GT
4:    if IsClustered(gt_p) == false then
5:       N_(ε,τ) ← GetNeighborhood(gt_p, ε, τ)
6:       if |N_(ε,τ)| ≥ MinGT then
7:          stc ←MakeNewCluster(gt_p)
8:          Q ← φ
9:          EnQueue(Q, N_(ε,τ))
10:         while Q is not empty do
11:            gt_q ← DeQueue(Q)
12:            N_(ε,τ) ← GetNeighborhood(gt_q, ε, τ)
13:            if |N_(ε,τ)| ≥ MinGT then
14:               EnUniqueQueue(Q, N_(ε,τ))
15:            end if
16:            stc ← stc ∪ gt_q
17:         end while
18:         STC ← STC ∪ stc
19:      end if
20:   end if
21: end for
22: return STC
```

2.3.4　アルゴリズム

Algorithm1 に (ϵ, τ)-密度に基づく時空間クラスタリングのアルゴリズムを示す．Algorithm1 は，キーワード検索によって得られたジオタグ付きツイート集合 GT とパラメータ ϵ，τ と $MinGT$ を入力として，時空間クラスタ集合 STC を出力する．Algorithm1 の内容を詳しく説明する．

（1）ジオタグ付きツイート集合 GT からジオタグ付きツイート gt_p を 1 つ取り出す．ただし，GT が空ならば，（8）へ進む．

（2）関数 IsClustered を用いて，gt_p が時空間クラスタに所属しているかチェックする．時空間クラスタに所属していなければ，関数 GetNeighborhood を用いて gt_p の (ϵ, τ)-密度に基づく近傍を取得する．
（3）もし，gt_p が核ジオタグ付きツイートであれば，（4）へ進む．gt_p が核ジオタグ付きツイートでなければ，（1）へ戻る．
（4）関数 MakeNewCluster を用いて新たに時空間クラスタ stc を作成する．
（5）ここで，gt_p の (ϵ, τ)-密度に基づく近傍に存在するジオタグ付きツイート集合を関数 EnQueue を用いてキュー Q に挿入する．
（6）キュー Q からジオタグ付きツイート gt_q を取り出し，次の処理を行う．
　（a）ジオタグ付きツイート gt_q の (ϵ, τ)-密度に基づく近傍を取得する．もし，gt_q が核ジオタグ付きツイートであれば，gt_q の (ϵ, τ)-密度に基づく近傍を関数 EnUniqueQueue を用いてキュー Q に挿入する．挿入では，Q に存在せず，他の時空間クラスタに所属していないジオタグ付きツイートのみを Q に挿入する．
　（b）ジオタグ付きツイート gt_q を時空間クラスタ stc に挿入する．キュー Q が空であれば，（7）へ進む．空でなければ（6）へ戻る．
（7）時空間クラスタ集合 STC に時空間クラスタ stc を加え，（1）に戻る．
（8）時空間クラスタ集合 STC を出力する．

2.3.5　問題点

(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が，いくつか問題点がある．まず，トピックと関連するジオタグ付きツイート集合をキーワード検索のみで求めているため，トピックに関連しないジオタグ付きツイートも時空間クラスタリングの対象としてしまう点である．キーワードを含むジオタグ付きツイートが必ずしもトピックに関連するとは限らない．また，Algorithm1 はジオタグ付きツイート集合を一度に入力して処理を行うバッチ処理であり，逐次発生するジオタグ付きツイートに対してクラスタリングを行うことができない．よってリアルタイムに時空間クラスタの発生，その変化と消滅を捉えることができない．

2.4　提案手法

本節では，提案手法である密度に基づく時空間分析手法について説明する．

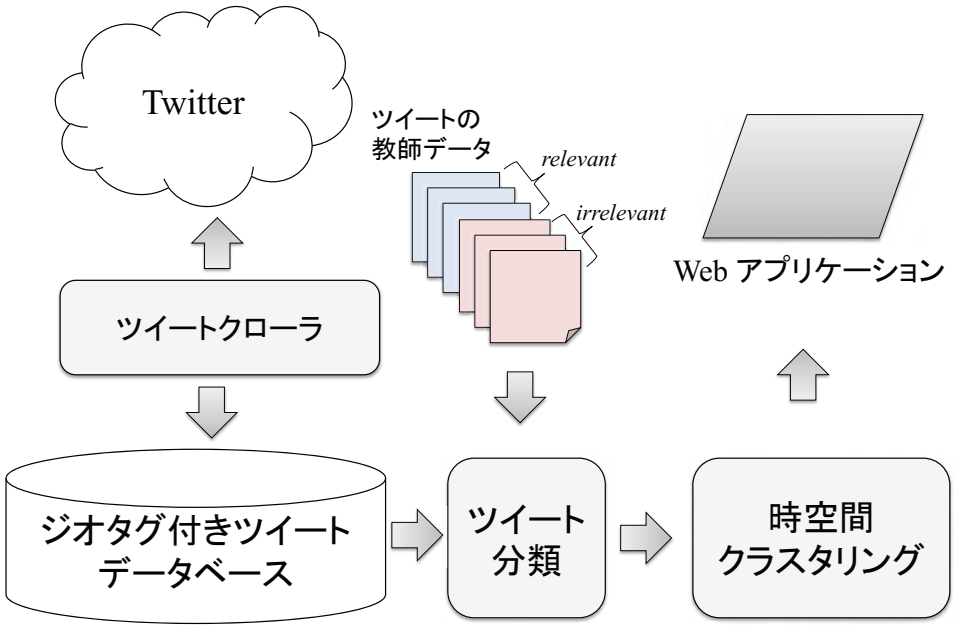


図 5: 密度に基づく時空間分析手法の処理手順

2.4.1　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 $GT = \{gt_1, \cdots, gt_n\}$ とする．ここで，$gt_i =< text_i, pt_i, pl_i, api_i >$ であり，$text_i$ は文書データ，pt_i は投稿時間，pl_i は位置情報，api_i は画像データである．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 mtp 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 $rgt_j^{(mtp)}$（$= gt_{\phi^{(mtp)}(j)}$）と呼ぶ．関連ジオタグ付きツイート集合を $RGT^{(mtp)} = \{rgt_1^{(mtp)}, \cdots, rgt_m^{(mtp)}\}$ とすると，GT は $RGT^{(mtp)}$ を包含しており（$RGT^{(mtp)} \subset GT$），次の単射で表現される．

$$\phi^{(mtp)}(j) : RGT^{(mtp)} \to GT;\ rgt_j^{(mtp)} \mapsto gt_{\phi^{(mtp)}(j)} \tag{2}$$

2.4.2　概要

図 5 に密度に基づく時空間分析手法の処理手順を示す．密度に基づく時空間分析手法では，予めモニタリングの対象とするトピックを設定し，次の処理を一定時間毎に実行する．

（1）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．ツイートの収集には Twitter Streaming API を

図 6: Web アプリケーション画面（アイコン）

使用している．

（2）ナイーブベイズ分類器を用いて，新たに追加されたジオタグ付きツイートの分類を行う．ここでは，モニタリングの対象となっているトピックに分類されたジオタグ付きツイートを関連ジオタグ付きツイートとして抽出する．ツイートの分類方法については，2.4.3 項で説明する．

（3）新たに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と前回抽出された時空間クラスタ集合を入力し，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを実行して新しく時空間クラスタ集合を抽出する．ここで，新しく抽出された時空間クラスタ集合と前回の時空間クラスタ集合を比較し，新しく追加された時空間クラスタは新しく話題として取り上げられている地域の発生を示している．また，どちらにも存在し，更新があった時空間クラスタは，話題として取り上げられている地域の変化を示しており，新しく抽出された時空間クラスタ集合に含まれなくなった時空間クラスタがあった場合は，話題として取り上げられている地域の消滅を意味している．インクリメンタルなアルゴリズ

図 7: Web アプリケーション画面（ツイートと画像データ）

ムについては，2.4.4 項で説明する．

（4）抽出された時空間クラスタの内容を地図上に表示し，Web アプリケーションから閲覧可能にする．作成した Web アプリケーションのスクリーンショットを図 6 と 7 に示す．時空間クラスタ集合は時系列データとして保存しており，時間を戻ることや進めることで時空間クラスタの発生，その変化と消滅を確認できる．地図上の時空間クラスタは中心点のみをアイコンで表示しており（図 6），このアイコンの発生と消滅が，モニタリング対象のトピックが注目されている地域の発生と消滅を表している．アイコンを選択すると時空間クラスタに含まれるすべての関連ジオタグ付きツイートと付与されている画像データが地図上に表示され，内容を確認することができる（図 7）．

2.4.3　ナイーブベイズ分類器を用いたツイート分類

ツイート分類では，ジオタグ付きツイートの文書データに含まれる語句に基づいて，ナイーブベイズ分類器を用いてツイートを分類する．モニタリング対象のトピックを「大雨」

表 1: ツイート分類の教師データの例

relevant クラス	*irrelevant* クラス
今、すごい大雨	昨日雨だった
雨降り始めた	今日雨の予報だったのに
ほんとすごい雨	明日は雨だろう
雨強すぎる	昨日はあんなに大雨だったのに

としたときの教師データの例を表 1 に示す．表 1 のように，つぶやかれたときに大雨が降っていることが分かるツイートを $relevant$ クラスに分類し，「大雨」に関する内容でもつぶやかれたときに大雨が降っていないツイートや「大雨」に関係の無いツイートを $irrelevant$ 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．クラス集合を $CLASS^{(mtp)} = \{\text{“}relevant,\text{”}\ \text{“}irrelevant\text{”}\}$ とし，モニタリングの対象となっているトピックを含むジオタグ付きツイートを $relevant$ クラス，それ以外のジオタグ付きツイートを $irrelevant$ クラスとして 2 つに分類する．

教師データは各クラスの内容を含むジオタグ付きツイート集合で，教師データ $TGT^{(mtp)}$ を $TGT^{(mtp)} = \{(tgt_1^{(mtp)}, c_1), (tgt_2^{(mtp)}, c_2), \cdots, (tgt_l^{(mtp)}, c_l)\}$ とする．ここで，$c_i = \{\text{“}relevant\text{” or “}irrelevant\text{”}\} \in CLASS^{(mtp)}$ となる．クラス $class \in CLASS^{(mtp)}$ に属する全ての語句を，$CW_{class}^{(mtp)} = \{cw_1^{(mtp)}, cw_2^{(mtp)}, \cdots, cw_{|CW_{class}^{(mtp)}|}^{(mtp)}\}$ とする．

ここで，ジオタグ付きツイート gt_p の語句集合を $GW_p = \{gw_1, gw_2, \cdots, gw_{|GW_p|}\}$ と表現する．ジオタグ付きツイート gt_p が所属するクラスの事後確率 $Pr(class|gt_p)$ は次の式で表される．

$$Pr(class|gt_p) = \frac{Pr(class)Pr(gt_p|class)}{Pr(gt_p)} \propto Pr(class)Pr(gt_p|class) \tag{3}$$

$Pr(class)$ は $class$ の事前確率であり，$Pr(gt_p|class)$ は尤度である．尤度は次の式で求める．

$$Pr(gt_p|class) = Pr(gw_1 \wedge gw_2 \wedge \cdots \wedge gw_k|class) = \prod_{i}^{|GW_p|} Pr(gw_i|class) \tag{4}$$

このとき，$Pr(gw_i|class)$ はクラス $class$ での語句 gw_i の発生頻度であり，次の式で求めることができる．

$$Pr(gw_i|class) = \frac{OW^{(mtp)}(gw_i, class) + 1}{\sum_{j=1}^{|CW_{class}^{(mtp)}|}(OW^{(mtp)}(cw_j^{(mtp)}, class) + 1)} \tag{5}$$

$OW^{(mtp)}(gw_i, class)$ はクラス $class \in CLASS^{(mtp)}$ での語句 gw_i の出現回数である．

そして，ジオタグ付きツイート gt_p は事後確率 $Pr(class|gt_p)$ を最大化する $class$ に分類される．

$$\begin{aligned} c_{gt_p} &= \arg\max_{class} Pr(class|gt_p) \\ &= \arg\max_{class}\left(Pr(class) \prod_{i}^{|GW_p|} Pr(gw_i|class) \right) \end{aligned} \tag{6}$$

2.4.4　(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズム

時空間クラスタリングでは，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを実行する．Algorithm2 に (ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを示す．新しい関連ジオタグ付きツイート $nrgt^{(mtp)}$ を取得する度に，これまでに取得した関連ジオタグ付きツイート集合 $RGT^{(mtp)}$，前回抽出された時空間クラスタ集合 $CSTC^{(mtp)}$，パラメータ ϵ，τ と $MinRGT$ を入力し，更新された時空間クラスタ集合 $NSTC^{(mtp)}$ 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)-密度に基づく時空間クラスタが結合し，一つとなることもある．

Algorithm2 を詳しく説明する．

（1）前回抽出された時空間クラスタ集合 $CSTC^{(mtp)}$ を $NSTC^{(mtp)}$ に挿入する．

（2）関数 GetNeighborhood を用いて，新たに取得した関連ジオタグ付きツイート $nrgt^{(mtp)}$ 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 $RRGT^{(mtp)}$ に挿入する．

（3）$RRGT^{(mtp)}$ から関連ジオタグ付きツイート $rgt_p^{(mtp)}$ を 1 つ取り出す．ただし，$RRGT^{(mtp)}$ が空ならば（11）へ進む．

（4）$rgt_p^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍を取得し，$N_{(\epsilon,\tau)}$ に挿入する．

（5）$rgt_p^{(mtp)}$ が核ジオタグ付きツイートであれば，（6）へ進む．$rgt_p^{(mtp)}$ が核ジオタグ付きツイートでなければ（3）へ戻る．

（6）関数 IsClustered を用いて，$rgt_p^{(mtp)}$ が時空間クラスタに所属しているかチェックする．時空間クラスタに所属していなければ，関数 MakeNewCluster を用いて新しく時空間クラスタ $stc^{(mtp)}$ を作成する．時空間クラスタに所属していれば，関数 GetCluster を用いて $rgt_p^{(mtp)}$ が所属している時空間クラスタを取得し，$stc^{(mtp)}$ と

Algorithm 2 (ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズム

Input: $nrgt^{(mtp)}$, $RGT^{(mtp)}$, $CSTC^{(mtp)}$, ϵ, τ, $MinRGT$
Output: $NSTC^{(mtp)}$

$NSTC^{(mtp)} \leftarrow CSTC^{(mtp)}$
$RRGT^{(mtp)} \leftarrow$ GetNeighborhood($nrgt^{(mtp)}$, ϵ, τ)
for $i \leftarrow 1$ to $|RRGT^{(mtp)}|$ **do**
  $rgt_p^{(mtp)} \leftarrow rrgt_i \in RRGT^{(mtp)}$
  $N_{(\epsilon,\tau)} \leftarrow$ GetNeighborhood($rgt_p^{(mtp)}$, ϵ, τ)
  if $|N_{(\epsilon,\tau)}| \geq MinRGT$ **then**
    if IsClustered($rgt_p^{(mtp)}$) $== false$ **then**
      $stc^{(mtp)} \leftarrow$ MakeNewCluster($rgt_p^{(mtp)}$)
    else
      $stc^{(mtp)} \leftarrow$ GetCluster($rgt_p^{(mtp)}$, $NSTC^{(mtp)}$)
    end if
    $Q \leftarrow \phi$
    EnQueue(Q, $N_{(\epsilon,\tau)}$)
    while Q is not empty **do**
      $rgt_q^{(mtp)} \leftarrow$ DeQueue(Q)
      if IsClustered($rgt_q^{(mtp)}$) $== true$ **then**
        $N_{(\epsilon,\tau)} \leftarrow$ GetNeighborhood($rgt_q^{(mtp)}$, ϵ, τ)
        if $|N_{(\epsilon,\tau)}| \geq MinRGT$ **then**
          $stc^{(mtp)'} \leftarrow$ GetCluster($rgt_q^{(mtp)}$, $NSTC^{(mtp)}$)
          $stc^{(mtp)} \leftarrow$ AppendClusters($stc^{(mtp)}$, $stc^{(mtp)'}$)
        end if
      else
        $N_{(\epsilon,\tau)} \leftarrow$ GetNeighborhood($rgt_q^{(mtp)}$, ϵ, τ)
        if $|N_{(\epsilon,\tau)}| \geq MinRGT$ **then**
          EnUniqueQueue(Q, $N_{(\epsilon,\tau)}$)
        end if
        $stc^{(mtp)} \leftarrow stc^{(mtp)} \cup rgt_q^{(mtp)}$
      end if
    end while
    $NSTC^{(mtp)} \leftarrow NSTC^{(mtp)} \cup stc^{(mtp)}$
  end if
end for
RemoveOldCluster($NSTC^{(mtp)}$, τ)
return($NSTC^{(mtp)}$)

する．

（7）$rgt_p^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍をキュー Q に挿入する．

（8）キュー Q から関連ジオタグ付きツイート $rgt_q^{(mtp)}$ を取り出し，次の処理を行う．

（a）$rgt_q^{(mtp)}$ が時空間クラスタに所属しているかチェックを行う．もし，時空間クラ

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

（i）$rgt_q^{(mtp)}$ が核ジオタグ付きツイートであるかチェックする．もし，核ジオタグ付きツイートであれば，関数 GetCluster を用いて $rgt_q^{(mtp)}$ が所属する時空間クラスタ $stc^{(mtp)'}$ 取得する．そして，関数 AppendClusters を用いて $stc^{(mtp)}$ と $stc^{(mtp)'}$ を結合する．

（ii）$rgt^{(mtp)_q}$ が核ジオタグ付きツイートであるかチェックする．もし，核ジオタグ付きツイートであれば，$rgt_q^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍を関数 EnUniqueQueue を用いてキュー Q に挿入する．挿入では，Q に存在せず，他の時空間クラスタに所属していない関連ジオタグ付きツイートのみを Q に挿入する．また，$rgt_q^{(mtp)}$ を $stc^{(mtp)}$ に挿入する．

（9）キュー Q が空であれば（10）へ進む．空でなければ（8）へ戻る．

（10）時空間クラスタ集合 $NSTC^{(mtp)}$ に，時空間クラスタ $stc^{(mtp)}$ を加え，（3）へ戻る．

（11）関数 RemoveOldCluster を用いて，$NSTC^{(mtp)}$ 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，$NSTC^{(mtp)}$ を更新された時空間クラスタ集合として出力する．

2.5　評価実験

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

2.5.1　実験内容

評価実験では，モニタリングの対象とするトピックを「大雨」と「大雪」とする．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．最初に，ツイート分類の評価実験を行う．ツイート分類では，トピック「大雨」と「大雪」についてそれぞれ教師データを作成し，交差検定によって評価を行う．次に，時空間クラスタリングの評価実験を行う．この実験では，新聞報道に基づきそれぞれのトピックが注目された地域を検出できたかを評価する．最後に，Web アプリケーション上で抽出された時空間クラスタを確認し，トピックの発生，その変化と消滅を捉えることができたか確認する．

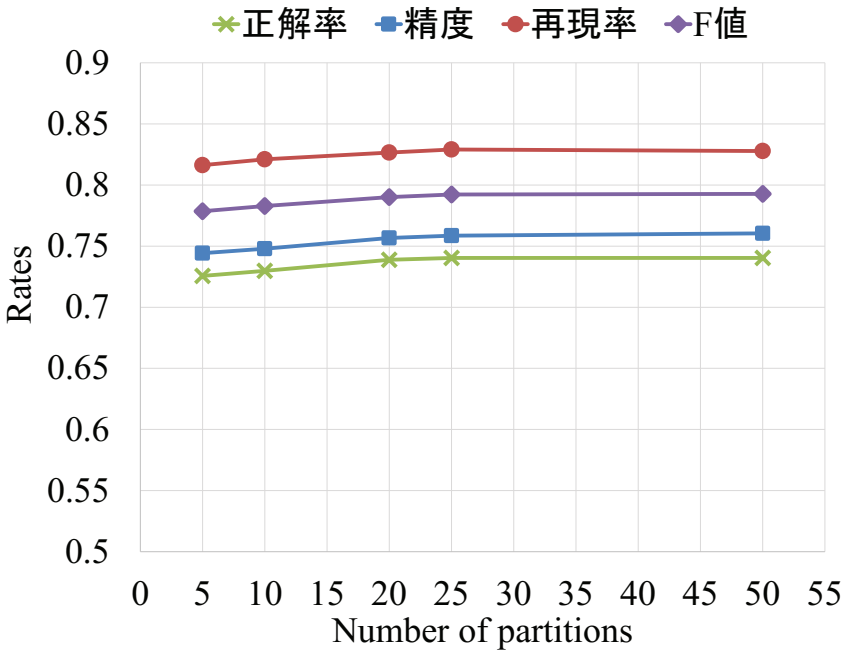


図 8: トピック「大雨」のツイート分類の交差検定

2.5.2　ツイート分類の評価実験

最初に，ナイーブベイズ分類器を用いたツイート分類の評価を行う．トピック「大雨」についての教師データとして，2014 年 6 月 4 日に投稿されたキーワードとして「雨」を含み，「大雨」に関するトピックを含む *relevant* クラス 1,458 件，「大雨」に関するトピックを含まない *irrelevant* クラス 1,097 件を用いた．また，トピック「大雪」についての教師データとして，2 月 8 日に投稿されたキーワードとして「雪」を含み，「大雪」に関するトピックを含む *relevant* クラス 1,648 件，「大雪」に関するトピックを含まない *irrelevant* クラス 852 件を用いた．

ツイート分類を評価するために交差検定を行った．交差検定の分割数には，5，10，20，25 と 50 分割を用いた．モニタリング対象のトピックを「大雨」としたときのツイート分類に対する交差検定の実験結果を図 8 に，モニタリング対象のトピックを「大雪」としたときのツイート分類に対する交差検定の実験結果を図 9 に示す．図 8 と図 9 には，各分割における正解率，精度，再現率と F 値を示している．トピック「大雨」の F 値として 0.78 から 0.79，トピック「大雪」の F 値は 0.81 をそれぞれ示した．ツイート分類を用いることで対象のトピックを含むツイートを抽出できることを確認できた．また，トピック「大雨」について，7 月 3 日に *relevant* クラスに分類された 4,738 件のジオタグ付きツイートがトピック「大雨」に関する内容であるかどうかを手作業で確認し，精度を計算したところ，0.93 の

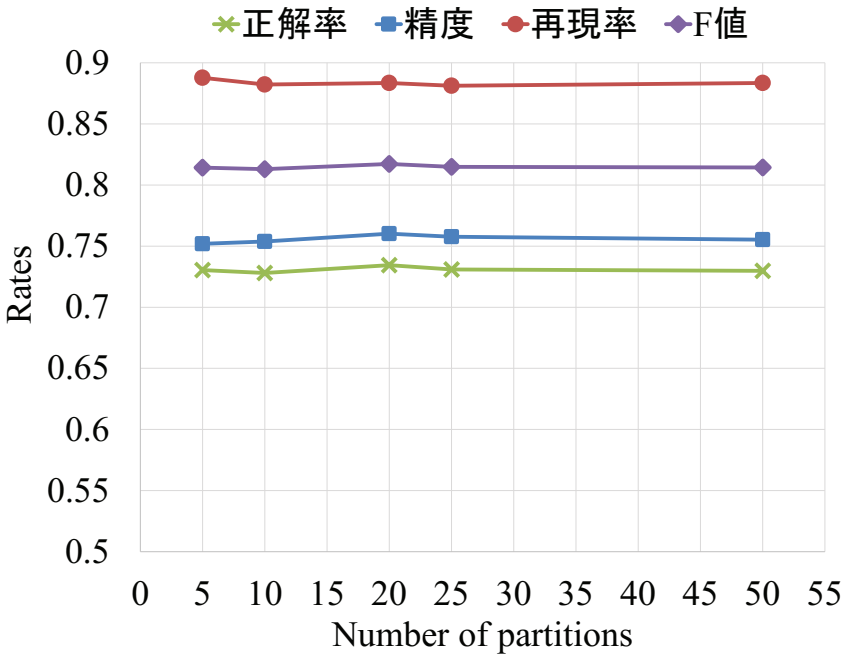


図 9: トピック「大雪」のツイート分類の交差検定

精度を示した．

2.5.3　時空間クラスタリングの評価実験

次に，時空間クラスタリングの評価を行う．本実験では，パラメータは $\epsilon = 5km$, $\tau = 3,600sec$, $MinRGT = 5$ を用いた．実験期間である 2 ヶ月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[*1]から，記事に記載されている各地域を抽出した．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 であった．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 2 と表 3 に 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，各日付において新聞記事で「大雨」,「大雪」と報道されていた地域数，検出数と検出率を表 2 と表 3 に示す．検出できたかどうかは報道された地域において時空間クラスタが抽出されているか，またツイートの内容がトピックを含んでいるかを手作業で確認した．

表 2 より，トピック「大雨」の検出率が 0.50 以上となったのは 16 日中 9 日である．表 3 より，トピック「大雪」の検出率が 0.50 以上となったのは 11 日中 6 日である．全体の検出

[*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 2: ツイート数とトピック「大雨」の検出率

日付	ツイート数	関連ツイート数	大雨の地域	検出数	検出率
2014/6/4	325,095	2,249	9	2	0.22
2014/6/6	312,145	6,401	8	4	0.50
2014/6/7	330,540	4,433	2	0	0.00
2014/6/13	346,507	2,589	2	0	0.00
2014/6/16	340,675	750	3	2	0.67
2014/6/22	411,863	4,172	1	0	0.00
2014/6/23	355,384	700	4	2	0.50
2014/6/25	393,441	2,331	12	11	0.92
2014/6/29	441,959	4,838	6	4	0.67
2014/7/3	341,770	4,738	13	6	0.46
2014/7/7	376,734	4,173	4	3	0.75
2014/7/8	366,887	1,405	3	2	0.67
2014/7/9	374,707	4,704	3	2	0.67
2014/7/10	395,061	4,803	1	0	0.00
2014/7/11	383,704	1,763	3	0	0.00
2014/7/19	412,403	5,369	3	2	0.67

率は，トピック「大雨」については 0.52，トピック「大雪」については 0.66 となった．しかしながら，検出できていない地域も存在している．検出できなかった理由としては，1 日中雨が降っているような日では，対象の地域に複数のジオタグ付きツイートが存在していたとしてもジオタグ付きツイート間の投稿間隔がパラメータ $\tau = 3,600sec$ 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．

2.5.4　抽出された時空間クラスタの確認

最後に，抽出された時空間クラスタを Web アプリケーション上で確認し評価を行う．図 10 に，トピックを「大雨」としたときの 2014 年 7 月 3 日の Web アプリケーションのスクリーンショットを示す．この日の午前中，西日本では大雨が観測されており，特に，九州地方では集中豪雨が観測されている．図 10 より，提案手法は大雨が観測された西日本において多くの時空間クラスタを抽出できた．

図 11 に，トピックを「大雪」としたときの 2013 年 12 月 20 日の Web アプリケーションのスクリーンショットを示す．この日，西日本では大雪が観測された．図 11 より，提案手

表 3: ツイート数とトピック「大雪」の検出率

日付	ツイート数	関連ツイート数	大雪の地域	検出数	検出率
2014/1/10	282,370	2,665	4	1	0.25
2014/1/14	284,215	981	1	0	0.00
2014/1/17	283,809	995	3	0	0.00
2014/2/6	284,065	2,821	2	0	0.00
2014/2/8	350,867	27,823	33	27	0.82
2014/2/11	289,628	3,564	1	1	1.00
2014/2/13	306,106	3,953	1	1	1.00
2014/2/14	378,368	21,834	18	15	0.83
2014/2/15	256,378	10,060	20	12	0.60
2014/2/16	307,708	5,121	5	1	0.20
2014/2/18	262,145	2,325	1	1	1.00

法は大雪が観測された西日本において多くの時空間クラスタを抽出できた．また，時空間クラスタのツイートを確認したところ，大雪の状況を伝えている内容のツイートが多く含まれていた．

図 12 に，トピックを「大雨」としたときの 2014 年 7 月 17 日 16 時の Web アプリケーションのスクリーンショットを示す．図 12 には，名古屋で抽出された時空間クラスタのツイートと雨雲レーダーを示している．この時間帯，名古屋ではゲリラ豪雨が観測されている．図 12 より，ゲリラ豪雨によって実際に雨が観測されている地域で多くのツイートが時空間クラスタとして抽出できているのが分かる．また，雨雲レーダーに沿って多くのツイートが抽出されており，雨が観測されていない地域ではツイートが少ない．そして，ツイートの内容を見ることによって，ゲリラ豪雨の発生とその状況を確認できる．以上の結果より，提案手法によって抽出された時空間クラスタを Web アプリケーション上で確認することによってトピックの発生，その変化と消滅を捉えることができた．

2.6　まとめ

本章では，Twitter 上に投稿されるジオタグ付きツイートを用いてトピックを時空間分析するための手法，密度に基づく時空間分析手法を提案した．提案手法は次の 3 つの処理によって行われる．

- ナイーブベイズ分類器を用いてジオタグ付きツイートを分類し，モニタリングの対象

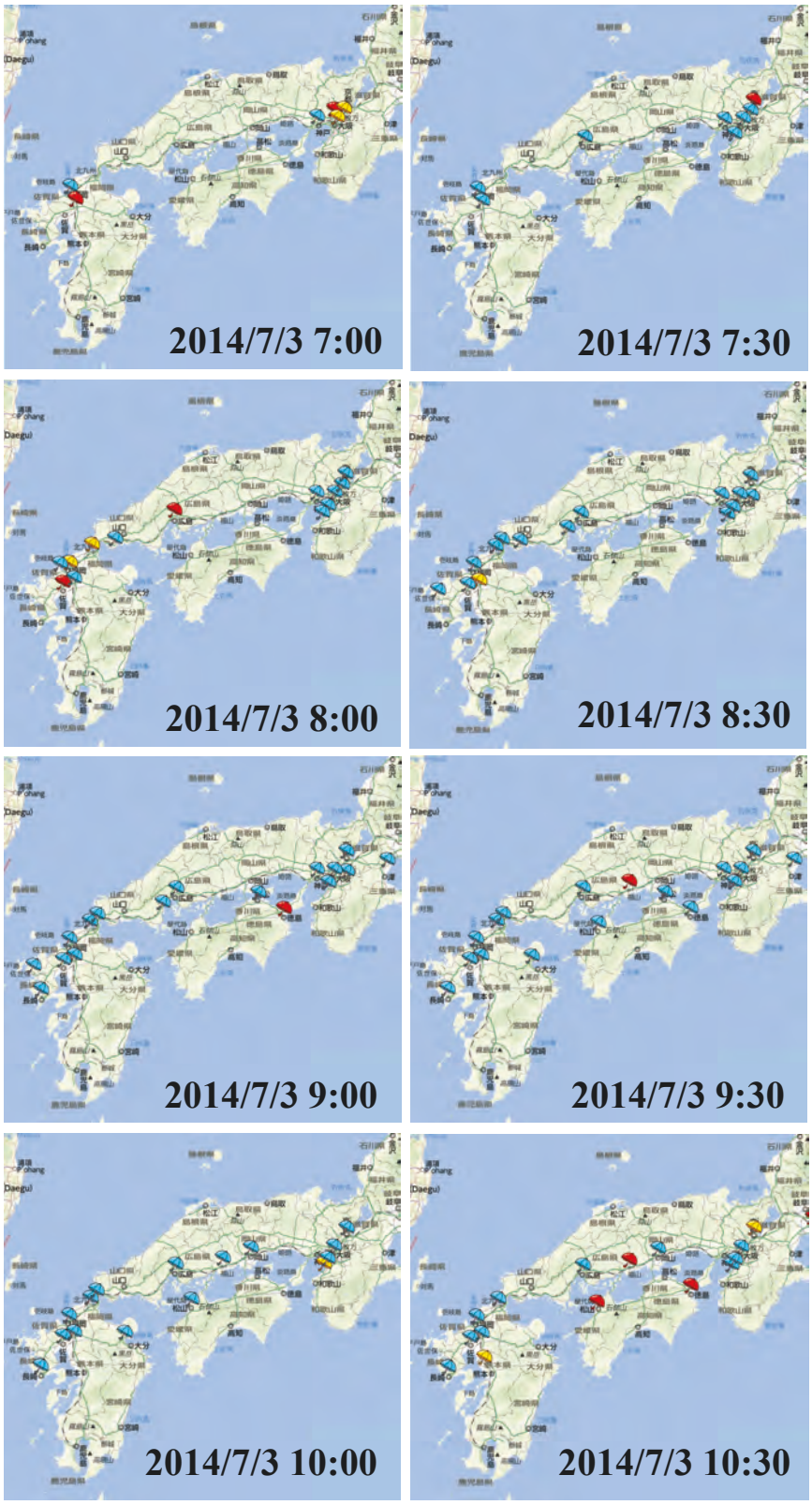


図 10: トピック「大雨」の 2014 年 7 月 3 日の Web アプリケーション

図 11: トピック「大雪」の 2013 年 12 月 20 日の Web アプリケーション

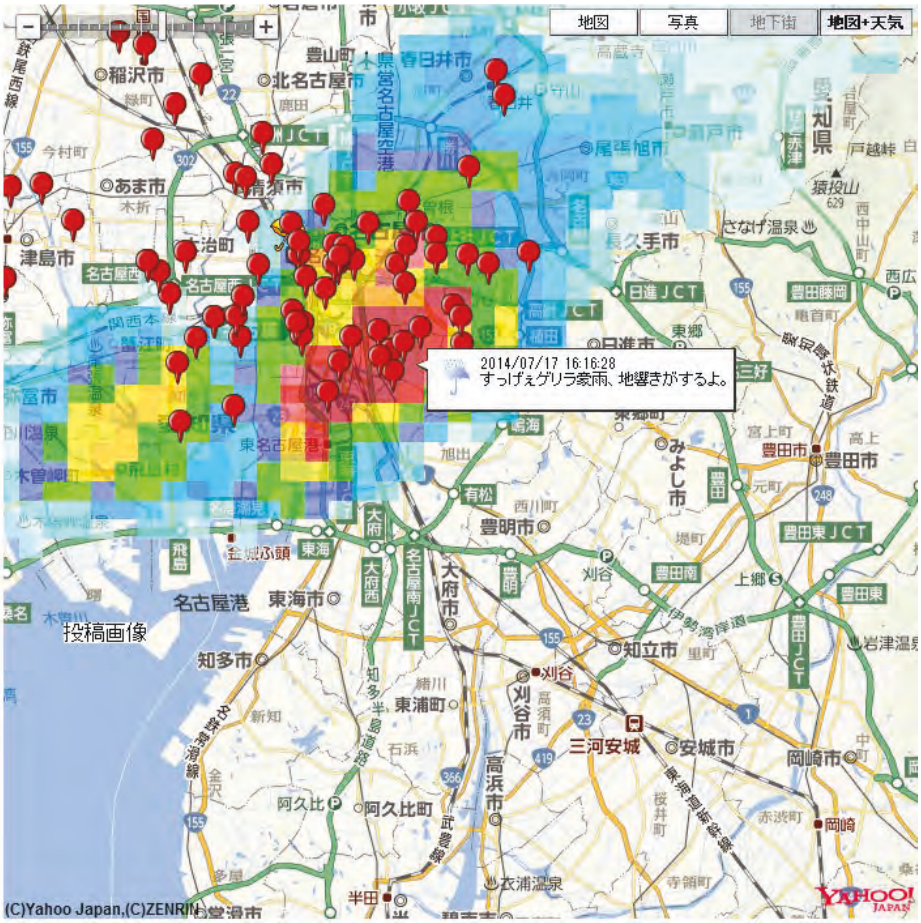


図 12: トピック「大雨」の 2014 年 7 月 17 日 16 時の Web アプリケーション

となっているトピックに関連するジオタグ付きツイートを抽出する．

- (ϵ, τ)-密度に基づく時空間クラスタリングを用いて，トピックが注目されている地域を時空間クラスタとして取り出す．また，インクリメンタルなアルゴリズムにより，注目されている地域をリアルタイムに抽出する．
- 抽出された時空間クラスタについてツイート内容と画像データを Web アプリケーション上に提示する．

提案手法を用いることによって，ジオタグ付きツイートからモニタリングの対象としているトピックが注目されている地域を抽出し，Web アプリケーション上で内容を確認することができる．

実際に Twitter 上からジオタグ付きツイートを収集し，トピックを「大雨」と「大雪」に設定し，評価実験を行った．評価実験の結果，ツイート分類の交差検定における F 値として，トピック「大雨」では 0.78，トピック「大雪」では 0.81 を示した．よって，一定以上の精度で，対象のトピックに関連するジオタグ付きツイートを抽出できることを確認できた．ま

た，(ϵ, τ)-密度に基づく時空間クラスタリングを用いてトピックが注目されている地域を検出できたか評価を行った結果，検出率としてトピック「大雨」では 0.52，トピック「大雪」では 0.66 を示した．さらに，抽出された時空間クラスタを Web アプリケーション上で確認することで，トピックの発生，その変化と消滅を捉えることができた．

本研究の今後の課題としては，以下の 3 点が挙げられる

- ツイート分類において深層学習を用いた新しい分類手法を開発し，分類性能を向上させる．
- ジオタグ付きツイートと気温や降雨量などの気象観測データを組み合わせた手法を開発し，より高性能な時空間分析を行えるようにする．
- トピックが注目されている地域の発生，その変化と消滅を捉えることができたか定量的な評価を行う．

第 3 章　密度に基づく適応的な時空間クラスタリング

本章では，密度に基づく適応的な時空間クラスタリングについて説明する．

3.1　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 [60, 61, 62]．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．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の分析が可能となる．

第 2 章にて，Twitter 上に投稿されるジオタグ付きツイートを用いて対象となっているトピックをリアルタイムに時空間分析するための手法，密度に基づく時空間分析手法を提案した．密度に基づく時空間分析手法は，最初に，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，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを用いて，当該トピックが注目されている地域を時空間クラスタとしてリアルタイムに抽出することで，当該トピックの発生，その変化と消滅を捉えることができる．

2.5 節にて行った評価実験の結果，密度に基づく時空間分析手法は (ϵ, τ)-密度に基づく時空間クラスタリングを用いることによって，トピック「大雨」と「大雪」の発生，その変化と消滅を捉えられることを確認した．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，適切に閾値 $MinRGT$ を設定しなければならないためである．投稿数の多い場合と少ない場合に分けて閾値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．

図 13 に，この問題が起こる例を示す．図 13 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．高密度な地域，つまりツイート集合 A または B に合わせて閾値を設定すると，図 13 のツイート集合 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，ツイート集合 C または D に合わせて閾値を設定

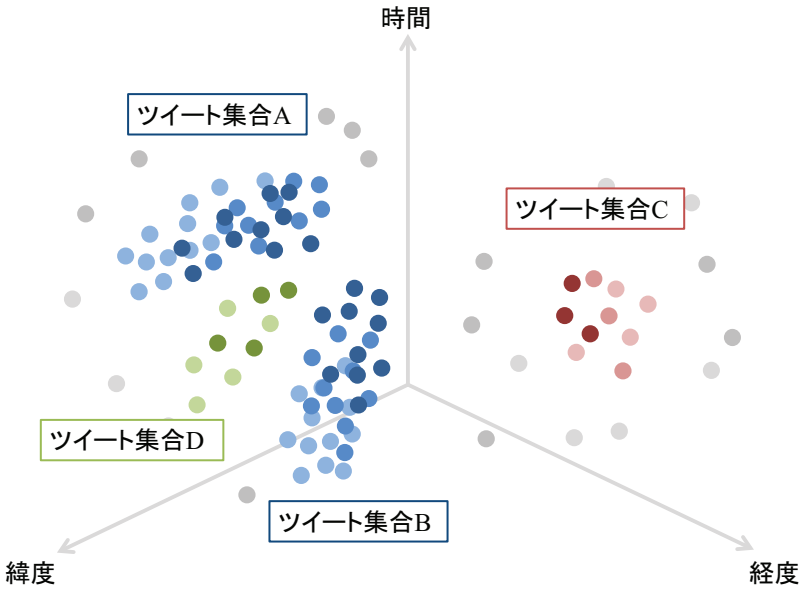


図 13: 密度に基づく時空間分析手法の問題点

すると，投稿数の多い地域では，図 13 のツイート集合 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．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な密度には違いがあり，ツイート集合 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．

本章では，地域や時間帯による投稿数の差異によって生じる問題を解決するために，(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する．(ϵ, τ)-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．よって，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない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．

提案手法を評価するために，提案手法を密度に基づく時空間分析手法に導入し，評価実験を行った．評価実験の結果，提案手法はトピック「大雨」と「大雪」について，既存手法よりも高性能にトピックの発生を捉えることができた．

本章の構成は以下の通りである．3.2 節では，関連研究について説明する．3.3 節では，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説明する．3.4 節では，評価実験の実験結果を示し，3.5 節で本章をまとめる．

3.2　関連研究

密度に基づく空間クラスタリングにおいて，データの密度の基準は空間上で異なることを問題として挙げ，パラメータを適切に設定するための研究が行われている [63, 64]．代表的な手法として，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，Ordering points to identify the clustering structure（OPTICS）が提案されている [65]．OPTICS は，密度に基づく空間クラスタリングの距離の基準となるパラメータを様々な値に設定した場合に，各パラメータ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を決めてクラスタリングを行う．投稿数の差異による問題を解決するためには，パラメータを各地域，各時間帯によって適応的に変化させる必要がある．本研究は，ソーシャルメディア上のデータに対する密度に基づく空間クラスタリングのパラメータを適応的に変化させて設定するという点について，初めての試みとなる．

3.3　(ϵ, τ)-密度に基づく適応的な時空間クラスタリング

本節では，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説明する．

3.3.1　概要

3.1 節にて述べた問題を解決するためには，各地域，各時間帯によって適切な閾値を設定する必要がある．(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，各時間帯によって変化する時空間上における適応的な閾値を定義する．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．

3.3.2　時空間上における適応的な閾値

本項では，時空間上における適応的な閾値について説明する．

定義 7（時空間上における適応的な閾値） 関連ジオタグ付きツイート $rgt_p^{(mtp)}$ の時空間上における適応的な閾値を $STAT(rgt_p^{(mtp)}, MaxMinRGT)$ と表記し，以下のように定義

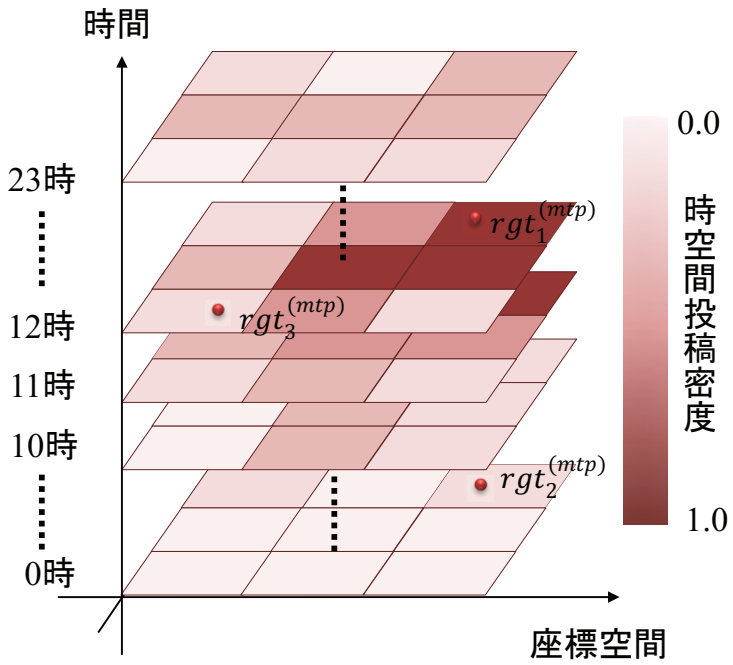


図 14: 時空間投稿密度の例

する．

$$STAT(rgt_p^{(mtp)}, MaxMinRGT) = (MaxMinRGT - 1) \times lstd(rgt_p^{(mtp)}) + 1 \quad (7)$$

関数 $lstd(rgt_p^{(mtp)})$ は $rgt_p^{(mtp)}$ の時空間投稿密度を返す関数であり $(0 \leq lstd(rgt_p^{(mtp)}) \leq 1)$，$MaxMinRGT$ 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．

$$lstd(rgt_p^{(mtp)}) = \frac{numstg(geo\_gid(rgt_p^{(mtp)})) - numstg_{min}}{numstg_{max} - numstg_{min}} \quad (8)$$

ただし，関数 $numstg(i)$ 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．$geo\_gid(rgt_p^{(mtp)})$ は $rgt_p^{(mtp)}$ が属する時空間グリッドの ID を返す関数である．$numstg_{max}$ と $numstg_{min}$ は最大数と最小数である．

図 14 に時空間投稿密度の例を示す．この例では，全時空間が 3 × 3 × 24 の時空間グリッドに分割され，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 $rgt_1^{(mtp)}$ と $rgt_2^{(mtp)}$ 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 $rgt_1^{(mtp)}$ は日中であり，$rgt_1^{(mtp)}$ の時空間投稿密度は $rgt_2^{(mtp)}$ の時空間投稿密度よりも高い．関連ジオタグ付きツイート $rgt_3^{(mtp)}$ は $rgt_1^{(mtp)}$ 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．

3.3.3　諸定義

本項では，(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 8（核関連ジオタグ付きツイート，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 $rgt_p^{(mtp)}$ について，$|N_{(\epsilon,\tau)}(rgt_p^{(mtp)})| \geq STAT(rgt_p^{(mtp)}, MaxMinRGT)$ を満たすとき，$rgt_p^{(mtp)}$ を核関連ジオタグ付きツイートと呼ぶ．反対に，$|N_{(\epsilon,\tau)}(rgt_p^{(mtp)})| < STAT(rgt_p^{(mtp)}, MaxMinRGT)$ であるとき，$rgt_p^{(mtp)}$ を周辺関連ジオタグ付きツイートと呼ぶ．

定義 9（(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 $rgt_q^{(mtp)}$ が $rgt_p^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍に存在し，$|N_{(\epsilon,\tau)}(rgt_p^{(mtp)})| \geq STAT(rgt_p^{(mtp)}, MaxMinRGT)$ を満たす時，関連ジオタグ付きツイート $rgt_q^{(mtp)}$ は $rgt_p^{(mtp)}$ 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 10（(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 $rgt_{p+1}^{(mtp)}$ が $rgt_p^{(mtp)}$ 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 $(rgt_p^{(mtp)}, rgt_{(p+1)}^{(mtp)}, \cdots, rgt_{(p+n)}^{(mtp)})$ を考える．この時，関連ジオタグ付きツイート $rgt_{(p+n)}^{(mtp)}$ は $rgt_p^{(mtp)}$ 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 11（(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 $rgt_p^{(mtp)}$ と $rgt_q^{(mtp)}$ とが $rgt_o^{(mtp)}$ から (ϵ, τ)-密度に基づいて適応的に到達可能であり，$rgt_o^{(mtp)}$ が $|N_{(\epsilon,\tau)}(rgt_o^{(mtp)})| \geq STAT(rgt_o^{(mtp)}, MaxMinRGT)$ を満たす時，$rgt_p^{(mtp)}$ と $rgt_q^{(mtp)}$ 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．

3.3.4　(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．特に，

局所的に密集している関連ジオタグ付きツイート集合も (ϵ, τ)-密度に基づく適応的な時空間クラスタとして定義される．

定義 12（(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 RGT 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 $rgt_p^{(mtp)} \in RGT$ と $rgt_q^{(mtp)} \in RGT$ 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 $rgt_p^{(mtp)}$ が所属（$rgt_p^{(mtp)} \in astc$）し，関連ジオタグ付きツイート $rgt_q^{(mtp)}$ が $rgt_p^{(mtp)}$ から (ϵ, τ)-密度に基づいて適応的に到達可能であれば，$rgt_q^{(mtp)}$ は $astc$ に所属（$rgt_q^{(mtp)} \in astc$）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 $rgt_p^{(mtp)} \in astc$ と $rgt_q^{(mtp)} \in astc$ は，(ϵ, τ)-密度に基づいて適応的に接続している．

3.3.5　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 Algorithm3 に示す．新しく取得した関連ジオタグ付きツイート $nrgt^{(mtp)}$，これまでに取得した関連ジオタグ付きツイート集合 $RGT^{(mtp)}$，前回抽出された時空間クラスタ集合 $CASTC^{(mtp)}$，パラメータ ϵ，τ と $MaxMinRGT$ を入力し，更新された時空間クラスタ集合 $NASTC^{(mtp)}$ を出力する．

Algorithm3 を詳しく説明する．

（1）前回抽出された時空間クラスタ集合 $CASTC^{(mtp)}$ を $NASTC^{(mtp)}$ に挿入する．

（2）関数 GetNeighborhood を用いて，新たに取得した関連ジオタグ付きツイート $nrgt^{(mtp)}$ 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 $RRGT^{(mtp)}$ に挿入する．

（3）$RRGT^{(mtp)}$ から関連ジオタグ付きツイート $rgt_p^{(mtp)}$ を 1 つ取り出す．ただし，$RRGT^{(mtp)}$ が空ならば（11）へ進む．

（4）$rgt_p^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍を取得する．また，関数 GetSTDens を用いて $rgt_p^{(mtp)}$ の時空間投稿密度を取得する．

（5）$rgt_p^{(mtp)}$ が核関連ジオタグ付きツイートであれば，（6）へ進む．$rgt_p^{(mtp)}$ が核関連ジオタグ付きツイートでなければ（3）へ戻る．

（6）関数 IsClustered を用いて，$rgt_p^{(mtp)}$ が時空間クラスタに所属しているかチェック

Algorithm 3 (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

Input: $nrgt^{(mtp)}$, $RGT^{(mtp)}$, $CASTC^{(mtp)}$, ϵ, τ, $MaxMinRGT$
Output: $NASTC^{(mtp)}$

$NASTC^{(mtp)} \leftarrow CASTC^{(mtp)}$
$RRGT^{(mtp)} \leftarrow$ GetNeighborhood($nrgt^{(mtp)}, \epsilon, \tau$)
for $i \leftarrow 1$ to $|RRGT^{(mtp)}|$ **do**
  $rgt_p^{(mtp)} \leftarrow rrgt_i^{(mtp)} \in RRGT^{(mtp)}$
  $N_{(\epsilon,\tau)} \leftarrow$ GetNeighborhood($rgt_p^{(mtp)}$, ϵ, τ)
  $lstd \leftarrow$ GetSTDens($rgt_p^{(mtp)}$)
  if $|N_{(\epsilon,\tau)}| \geq STAT(rgt_p^{(mtp)}, MaxMinRGT)$ **then**
    if IsClustered($rgt_p^{(mtp)}$) $== false$ **then**
      $astc^{(mtp)} \leftarrow$ MakeNewCluster($rgt_p^{(mtp)}$)
    else
      $astc^{(mtp)} \leftarrow$ GetCluster($rgt_p^{(mtp)}$, $NASTC^{(mtp)}$)
    end if
    $Q \leftarrow \phi$
    EnQueue(Q, $N_{(\epsilon,\tau)}$)
    while Q is not empty **do**
      $rgt_q^{(mtp)} \leftarrow$ DeQueue(Q)
      if IsClustered($rgt_q^{(mtp)}$) $== true$ **then**
        $N_{(\epsilon,\tau)} \leftarrow$ GetNeighborhood($rgt_q^{(mtp)}$, ϵ, τ)
        $lstd \leftarrow$ GetSTDens($rgt_q^{(mtp)}$)
        if $|N_{(\epsilon,\tau)}| \geq STAT(rgt_q^{(mtp)}, MaxMinRGT)$ **then**
          $astc^{(mtp)'} \leftarrow$ GetCluster($rgt_q^{(mtp)}$, $NASTC^{(mtp)}$)
          $astc^{(mtp)} \leftarrow$ AppendClusters($astc^{(mtp)}$, $astc^{(mtp)'}$)
        end if
      else
        $N_{(\epsilon,\tau)} \leftarrow$ GetNeighborhood($rgt_q^{(mtp)}$, ϵ, τ)
        $lstd \leftarrow$ GetSTDens($rgt_q^{(mtp)}$)
        if $|N_{(\epsilon,\tau)}| \geq STAT(rgt_q^{(mtp)}, MaxMinRGT)$ **then**
          EnUniqueQueue(Q, $N_{(\epsilon,\tau)}$)
        end if
        $astc^{(mtp)} \leftarrow astc^{(mtp)} \cup rgt_q^{(mtp)}$
      end if
    end while
    $NASTC^{(mtp)} \leftarrow NASTC^{(mtp)} \cup astc^{(mtp)}$
  end if
end for
RemoveOldCluster($NASTC^{(mtp)}$, τ)
return($NASTC^{(mtp)}$)

する．時空間クラスタに所属していなければ，関数 MakeNewCluster を用いて新しく時空間クラスタ $astc^{(mtp)}$ を作成する．時空間クラスタに所属していれば，関数 GetCluster を用いて $rgt_p^{(mtp)}$ が所属している時空間クラスタを取得し，$astc^{(mtp)}$ とする．

（7）$rgt_p^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍をキュー Q に挿入する．
（8）キュー Q から関連ジオタグ付きツイート $rgt_q^{(mtp)}$ を取り出し，次の処理を行う．
　（a）$rgt_q^{(mtp)}$ が時空間クラスタに所属しているかチェックを行う．もし，時空間クラスタに所属していれば，（i）を実行する．時空間クラスタに所属していなければ（ii）を実行する．
　　i. $rgt_q^{(mtp)}$ 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，関数 GetCluster を用いて $rgt_q^{(mtp)}$ が所属する時空間クラスタ $astc^{(mtp)'}$ 取得する．そして，関数 AppendClusters を用いて $astc^{(mtp)}$ と $astc^{(mtp)'}$ を結合する．
　　ii. $rgt_q^{(mtp)}$ 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，$rgt_q^{(mtp)}$ の (ϵ, τ)-密度に基づく近傍を関数 EnUniqueQueue を用いてキュー Q に挿入する．挿入では，Q に存在せず，他の時空間クラスタに所属していない関連ジオタグ付きツイートのみを Q に挿入する．また，$rgt_q^{(mtp)}$ を $astc^{(mtp)}$ に挿入する．
（9）キュー Q が空であれば（10）へ進む．空でなければ（8）へ戻る．
（10）時空間クラスタ集合 $NASTC^{(mtp)}$ に，時空間クラスタ $astc^{(mtp)}$ を加え，（3）へ戻る．
（11）関数 RemoveOldCluster を用いて，$NASTC^{(mtp)}$ 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，$NASTC^{(mtp)}$ を更新された時空間クラスタ集合として出力する．

3.4　評価実験

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

3.4.1　実験内容

評価実験では，密度に基づく時空間分析手法の時空間クラスタリングの部分について，2.3 節で説明した (ϵ, τ)-密度に基づく時空間クラスタリング（DBSTC と表記する）を用いた場合と，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（DBASTC と表記する）を用いた場合の比較を行う．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．新聞報道に基づきそれぞれのトピックが注目された地域を検出できたかを比較する．特に，DBSTC の問題点は時空間クラスタを抽出できていない地域や時間帯があること

表 4: DBSTC の抽出クラスタ数

$MinRGT$	トピック「大雨」のクラスタ数	トピック「大雪」のクラスタ数
2	1,325	1,762
3	928	1,150
4	670	809
5	534	633
6	420	497
7	342	437
8	304	375
9	246	320
10	224	280

表 5: DBASTC の抽出クラスタ数

$MaxMinRGT$	トピック「大雨」のクラスタ数	トピック「大雪」のクラスタ数
2	2,163	2,879
3	2,125	2,813
4	1,880	2,424
5	1,714	2,291
6	1,598	2,039
7	1,485	1,897
8	1,372	1,750
9	1,306	1,720
10	1,202	1,652

から，本実験でこの点について改善されているか確認する．

パラメータとしては，$\epsilon = 5km$，$\tau = 3600sec$ を用いた．DBSTC のパラメータ $MinRGT$ と DBASTC 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 $div_{lng} = 1,000$，$div_{lat} = 1,000$，$div_{time} = 24$ 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．

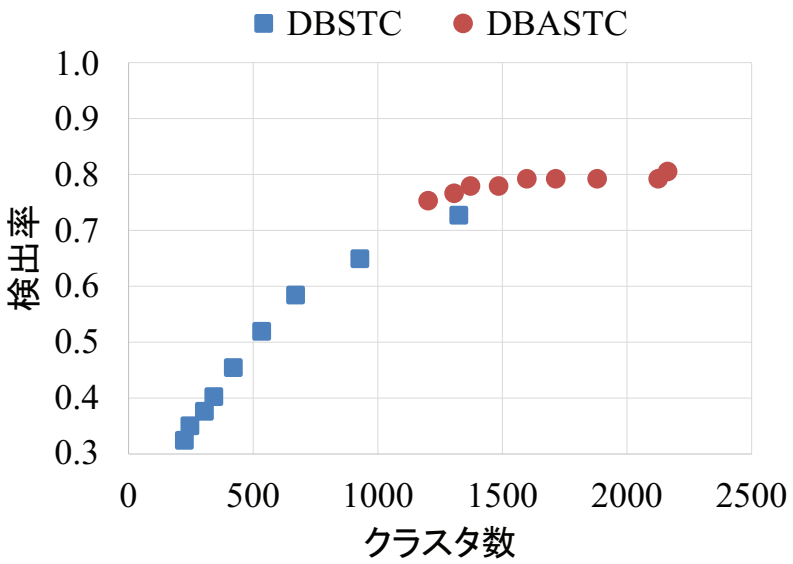


図 15: 閾値を変化させたときのトピック「大雨」の検出率

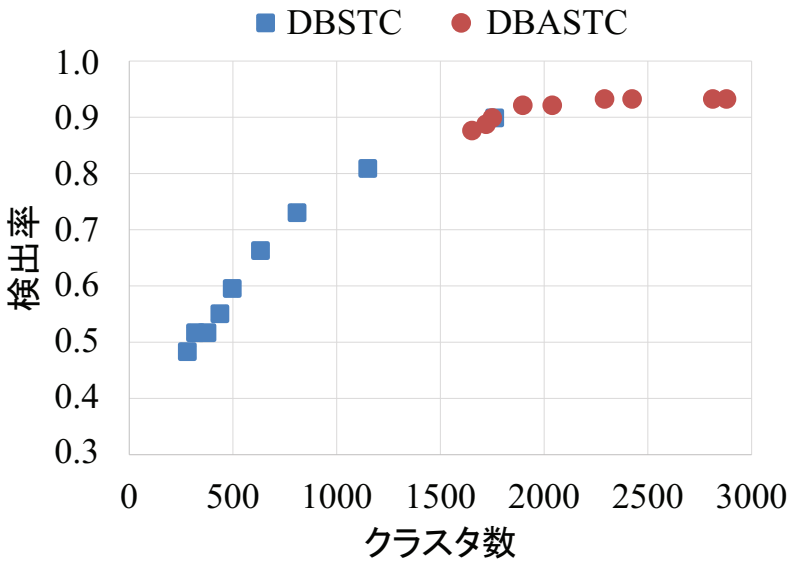


図 16: 閾値を変化させたときのトピック「大雪」の検出率

3.4.2　検出率の比較

最初に，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

表 6: トピック「大雨」の検出率（$MinRGT = 5$，$MaxMinRGT = 5$）

日付	関連ツイート数	大雨の地域（A）	DBSTC の検出数（B）	DBASTC の検出数（C）	DBSTC の検出率（B/A）	DBASTC の検出率（C/A）
2014/6/4	2,249	9	2	5	0.22	0.56
2014/6/6	6,401	8	4	6	0.50	0.75
2014/6/7	4,433	2	0	1	0.00	0.50
2014/6/13	2,589	2	0	1	0.00	0.50
2014/6/16	750	3	2	2	0.67	0.67
2014/6/22	4,172	1	0	1	0.00	1.00
2014/6/23	700	4	2	3	0.50	0.75
2014/6/25	2,331	12	11	12	0.92	1.00
2014/6/29	4,838	6	4	5	0.67	0.83
2014/7/3	4,738	13	6	11	0.46	0.85
2014/7/7	4,173	4	3	4	0.75	1.00
2014/7/8	1,405	3	2	3	0.67	1.00
2014/7/9	4,704	3	2	2	0.67	0.67
2014/7/10	4,803	1	0	0	0.00	0.00
2014/7/11	1,763	3	0	1	0.00	0.33
2014/7/19	5,369	3	2	3	0.67	1.00

事[*2]から，記事に記載されている地域（市町村）を抽出した．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 4 と表 5 に DBSTC と DBASTC について，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す．また，図 15 と図 16 に，$MinRGT$ と $MaxMinRGT$ を 2 から 10 まで変化させたときのトピック「大雨」と「大雪」の検出率をそれぞれ示す．図 15 と図 16 は，各 $MinRGT$，各 $MaxMinRGT$ における実験結果を，横軸を抽出クラスタ数，縦軸を検出率とした散布図で示している．検出できたかどうかは報道された地域において時空間クラスタが抽出されているか，またツイートの内容がトピックを含んでいるかを手作業で確認した．図 15 より，DBASTC を用いた場合のトピック「大雨」の検出率は $MaxMinRGT$ を 2 から 10 まで変化させても 0.80 から大きな変化がないのが分かる．一方，DBSTC を用いた場合，トピック「大雨」の検出率は 0.73 から 0.32 まで落ちている．また，DBSTC で抽出クラスタ数 1325 のときの検出率は 0.73，DBASTC で抽出クラスタ数 1306 のときの検出率は 0.77 となっており，提案手法が抽出クラスタ数が少ないにもかかわらず，検出率は高くなっている．図 16 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，DBASTC が DBSTC よりも高検出率であることが分かる．DBSTC で $MinRGT$ を増加させると高密度な地域のみが抽出されるが，DBASTC では閾値が地域と時間帯によって適応的に変化するため，検出率の低下を防ぐことができたと

[*2] 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 7: トピック「大雪」の検出率（$MinRGT = 5$，$MaxMinRGT = 5$）

日付	関連ツイート数	大雪の地域（A）	DBSTC の検出数（B）	DBASTC の検出数（C）	DBSTC の検出率（B/A）	DBASTC の検出率（C/A）
2014/1/10	2,665	4	1	3	0.25	0.75
2014/1/14	981	1	0	1	0.00	1.00
2014/1/17	995	3	0	2	0.00	0.67
2014/2/6	2,821	2	0	1	0.00	0.50
2014/2/8	27,823	33	27	33	0.82	1.00
2014/2/11	3,564	1	1	1	1.00	1.00
2014/2/13	3,953	1	1	1	1.00	1.00
2014/2/14	21,834	18	15	18	0.83	1.00
2014/2/15	10,060	20	12	18	0.60	0.90
2014/2/16	5,121	5	1	4	0.20	0.80
2014/2/18	2,325	1	1	1	1.00	1.00

いえる．

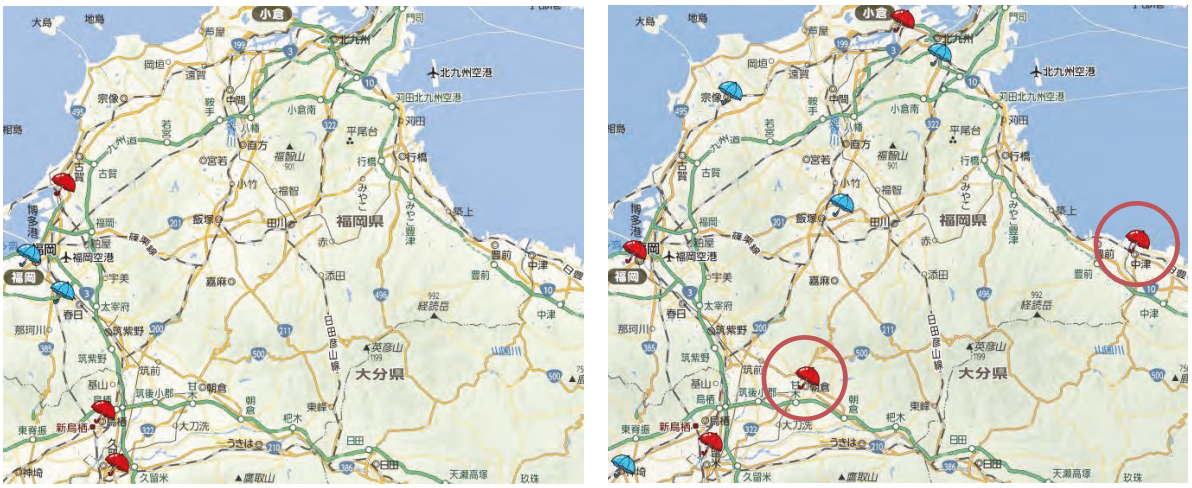
表 6 と表 7 に実験期間にツイート分類によって抽出された関連ジオタグ付きツイート数，新聞記事で「大雨」，「大雪」と報道されていた地域数（A），$MinRGT = 5$ としたときの DBSTC の検出数（B）と検出率（B/A），$MaxMinRGT = 5$ としたときの DBASTC の検出数（C）と検出率（C/A）を示す．DBASTC では，トピック「大雨」の検出率は 16 日中 14 日で 0.50 以上となっている．一方，DBSTC では，検出率が 0.50 以上となったのは 16 日中 9 日である．表 7 より，トピック「大雪」についても，DBASTC は DBSTC よりも高検出率であることが分かる．以上の新聞報道に基づいた検出率より，DBASTC は DBSTC よりも高い検出率でトピックの発生を捉えることができた．

3.4.3　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 17 に，トピックを「大雨」としたときの DBSTC と DBASTC によって，2014 年 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 DBASTC 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 8 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 8 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 18 に，トピックを「大雪」としたときの DBSTC と DBASTC によって，2014 年 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



(a) DBSTC　　(b) DBASTC

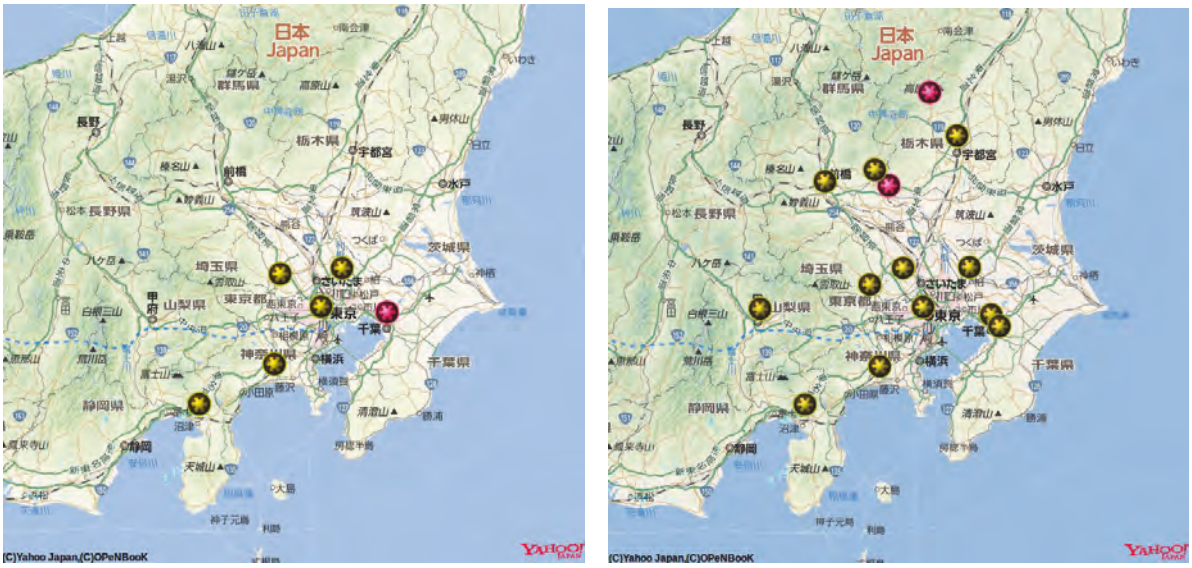
図 17: トピック「大雨」の 2014 年 7 月 3 日に九州地方で抽出された時空間クラスタ

表 8: 7 月 3 日に朝倉市と中津市で抽出された時空間クラスタのツイート

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

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 18 より，深夜の投稿が少ない時間帯において DBASTC は DBSTC 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．

図 19 に 7 月 3 日に DBASTC によって北九州にて抽出された時空間クラスタの遷移を示す．図 19 には，7 月 3 日の 7 時から 15 時までの間の時空間クラスタを地図上に示している．また，時空間クラスタの発生時間によってアイコンの色を変えており，赤，黄，青の順番で時空間クラスタの発生がその時から早いことを表している．7 月 3 日の北九州では 4 時ごろから大雨が観測され，時間が経つにつれて雨は弱まり，15 時ごろに雨は止んでいる．抽出された時空間クラスタを見てみると，午前中に多くの時空間クラスタが発生し，午後になるにつれて時空間クラスタは消滅しており，トピック「大雨」の時間的な変化を捉えることができた．



(a) DBSTC　　(b) DBASTC

図 18: トピック「大雪」の 2014 年 2 月 8 日に関東地方で抽出された時空間クラスタ

3.5　まとめ

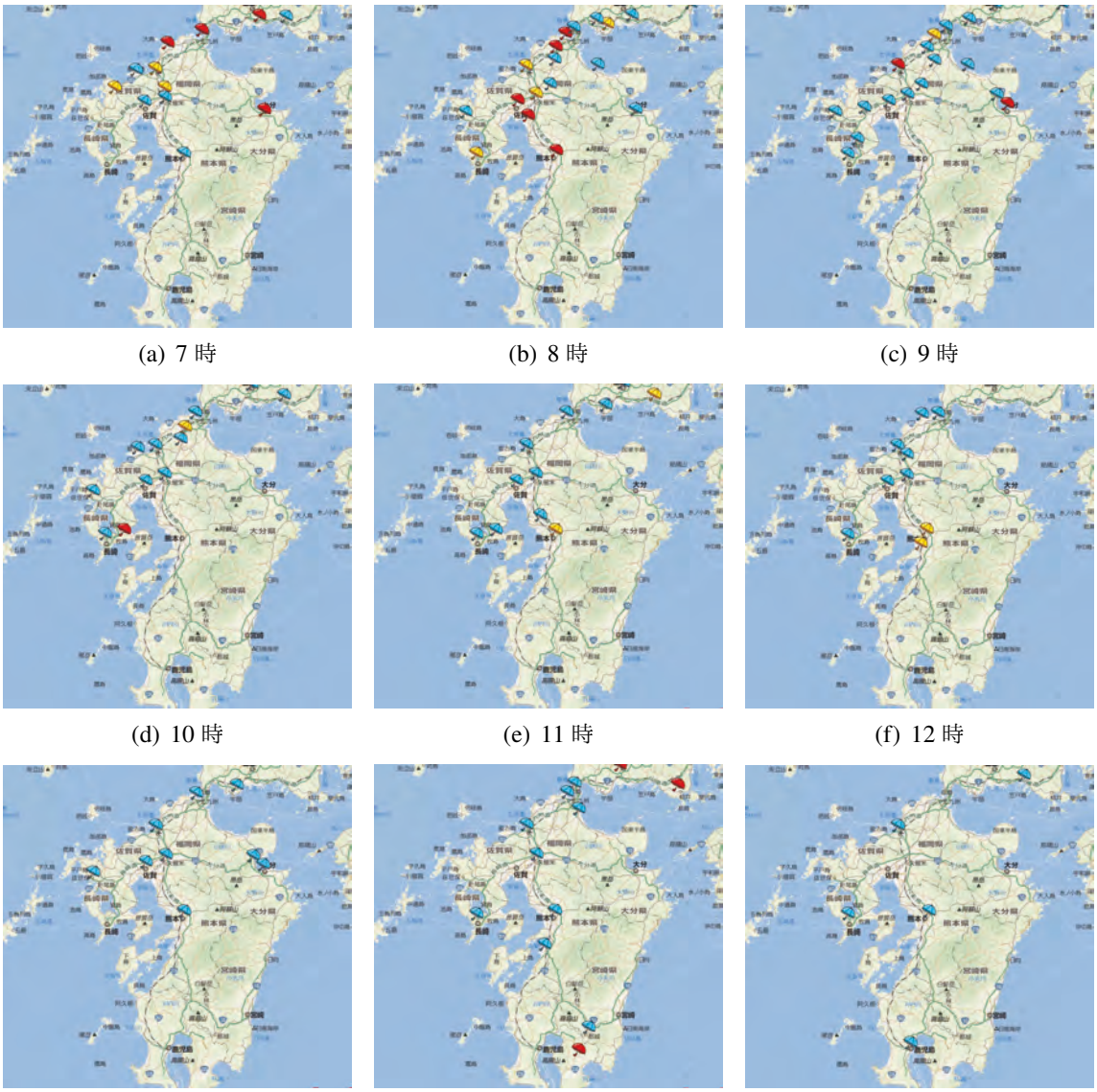
本章では，(ϵ, τ)-密度に基づく適応的な時空間クラスタリングを提案した．(ϵ, τ)-密度に基づく適応的な時空間クラスタリングでは各地域，また各時間帯の統計的な投稿密度を用いることで，時空間クラスタを抽出するときに基準となる閾値を適応的に変化させている．提案手法を用いることで，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない時間帯を区別することなく時空間クラスタを抽出することができる．

提案手法を第 2 章にて提案した密度に基づく時空間分析手法に導入し，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．提案手法と既存手法を比較した結果，トピック「大雨」について，既存手法ではパラメータ $MinRGT$ を 2 から 10 まで変化させた場合，検出率が 0.73 から 0.32 まで落ちるのに対して，提案手法はパラメータ $MaxMinRGT$ を 2 から 10 まで変化させたとしても，0.80 から大きく変化することなく，高い検出率を示すことができた．

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．

本研究の今後の課題としては，以下の 2 点が挙げられる．

- ユーザが設定する ϵ や τ などのパラメータを適応的または自動的に設定する手法を導



(a) 7 時　(b) 8 時　(c) 9 時

(d) 10 時　(e) 11 時　(f) 12 時

(g) 13 時　(h) 14 時　(i) 15 時

図 19: トピック「大雨」の 2014 年 7 月 3 日に北九州で抽出された時空間クラスタの遷移

入する．地域によって投稿数だけでなく時空間的なスケールが異なるため，例えば都会であれば，ϵ と τ を小さく，地方であれば，ϵ と τ を大きく自動的に設定する必要がある．

- 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挙げられる．評価実験において，検出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抽出できることが分かったが，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．

第 4 章　密度に基づく時空間分析手法における画像分類

本章では，密度に基づく時空間分析手法における画像分類について説明する．

4.1　はじめに

近年，インターネット上のユーザはソーシャルメディアサイトを通して気象状況や発生した自然災害に関する内容を投稿するようになってきている．例えば，代表的なソーシャルメディアサイトの Twitter では，大雨，大雪，地震や台風などを目の当たりにした人がテキストメッセージや写真によってそれらの状況を伝えている．通常，このような投稿には位置情報（ジオタグ）が付与されており，これらのジオタグ付きツイートから気象状況や自然災害などの緊急性のあるトピックを検出し，活用することが期待されている．

第 2 章にて，Twitter 上に投稿されるジオタグ付きツイートを用いて対象となっているトピックをリアルタイムに時空間分析するための手法，密度に基づく時空間分析手法を提案した．密度に基づく時空間分析手法は，最初に，モニタリングの対象とするトピックを定め，当該トピックの内容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，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している．時空間クラスタリングについては，第 3 章にて提案した (ϵ, τ)-密度に基づく適応的な時空間クラスタリングを用いることで，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない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．また，抽出された時空間クラスタのツイートと画像データを地図上に表示し，Web アプリケーションから閲覧することで当該トピックの発生，その変化と消滅のモニタリングを可能にしている．2.5 節にて行った評価実験では，密度に基づく時空間分析手法によってトピック「大雨」と「大雪」の発生，その変化と消滅を捉えられることを確認できた．

気象状況や自然災害などの緊急性のあるトピックを正確にユーザへ伝えるためには，テキストを提示するよりも，写真を提示する方がトピックの内容を一目で具体的に把握できるため有効である．密度に基づく時空間分析手法によって抽出されるジオタグ付きツイートには，画像データが付与されているものがある．しかしながら，全ての画像データがモニタリング対象のトピックを表しているわけではない．表 9 にモニタリング対象のトピックを「大雨」としたときの密度に基づく時空間分析手法において，2014 年 8 月 1 日から 10 日に抽出された時空間クラスタに含まれていた画像データ数と，その中でトピック「大雨」を表していない画像データ数を示す．表 9 より，抽出された多くの画像データがトピック「大雨」を

表 9: 密度に基づく時空間分析手法において抽出された画像データ数

日付	画像データ数	トピック「大雨」を表していない画像データ数
2014/8/1	130	75
2014/8/2	217	157
2014/8/3	203	132
2014/8/4	63	49
2014/8/5	63	47
2014/8/6	87	68
2014/8/7	53	39
2014/8/8	230	181
2014/8/9	412	329
2014/8/10	572	383

表していないことが分かる．つまり，ツイートの内容が当該トピックに関連していたとしても，付与されている画像データは当該トピックに関連していないことがある．

そこで本章では，モニタリングの対象となっているトピックに関連する画像データを抽出するための画像分類を提案し，密度に基づく時空間分析手法に導入する．提案手法は，Bag-of-Features（BoF）[19] または学習済み深層ネットワークを用いて画像データから特徴ベクトルを抽出する．次に，抽出した画像データの特徴ベクトルを使用して Support Vector Machine（SVM）を学習させ，当該トピックに関連する画像データかどうか分類する．そして，当該トピックに関連する画像データのみを Web アプリケーション上に提示することで，トピックの時空間分析の有効性を向上することができる．

提案手法を密度に基づく時空間分析手法に導入し，トピックを「大雨」と「大雪」と設定して評価実験を行った．評価実験の結果，提案手法は高性能にトピックに関連する画像データを分類することができた．

本章の構成は以下の通りである．4.2 節では，関連研究について説明する．4.3 節では，提案手法である画像分類について説明する．4.4 節では，評価実験の実験結果を示し，4.5 節で本章をまとめる．

4.2　関連研究

近年，インターネット上のユーザはソーシャルメディアサイト上に，携帯電話やスマートフォンで撮影した地域の話題やイベントの写真を画像データとして投稿し，情報発信を盛ん

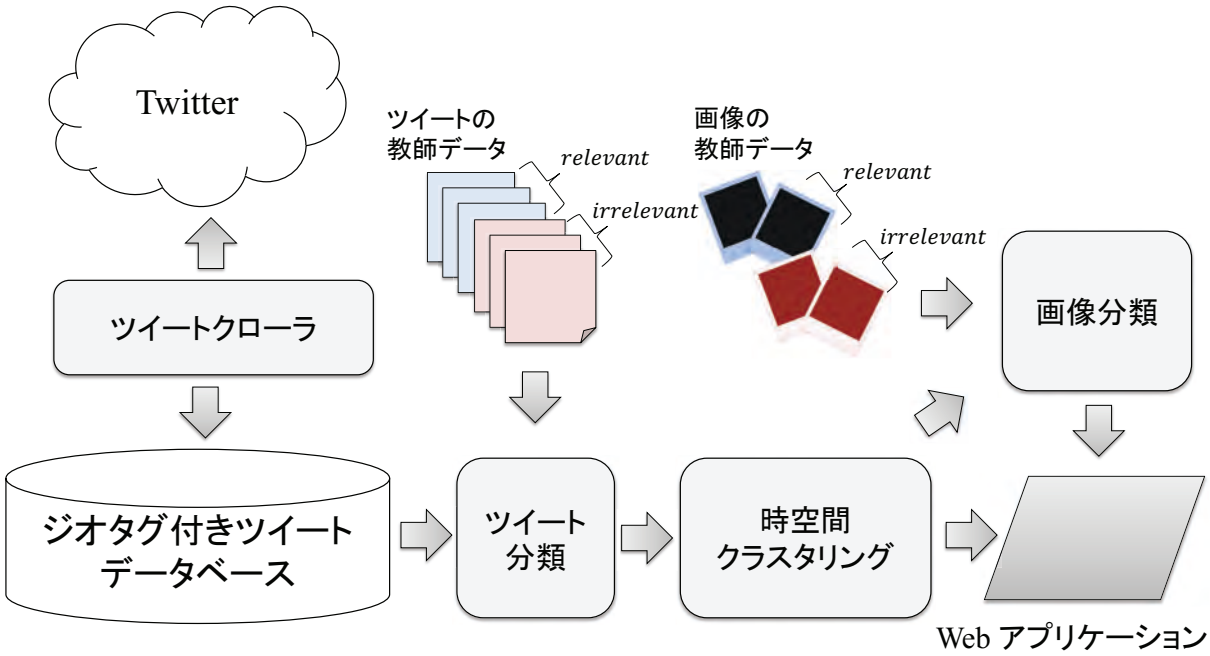


図 20: 画像分類を導入した密度に基づく時空間分析手法

に行うようになってきている．Flickr のような写真共有サイトや Twitter 上に投稿された写真は，個人的な趣味だけでなく社会的なトピックを含んでいるため，分析対象として注目を集めている．そこで，ソーシャルメディアサイト上に投稿された画像データを用いた研究が盛んに行われている [43]．Jaffe ら [41] は，Flickr 上に投稿されたジオタグ付き画像データをクラスタリングする手法を提案した．場所情報を用いて階層的に画像データをクラスタリングし，ホットスポットの検出を行っている．Yanai ら [42] は，k-means 法を使ってジオタグ付き画像データをクラスタリングする手法を提案している．また，Fruin ら [26] は，ジオタグ付きツイートに付与された画像データを用いてニュースを提示するための TweetPhoto と呼ばれるシステムを開発した．TweetPhoto は，ツイートを分類し，付与されている画像データにスコア付けを行っている．本研究は，ソーシャルメディアサイト上に投稿された位置情報付きのデータに対して分類やクラスタリングのみを行うのではなく，ツイート分類，時空間クラスタリングと画像分類を組み合わせて，トピックの時空間上における変化の分析を可能にする．

4.3　提案手法

本節では，提案手法である画像分類について説明する．

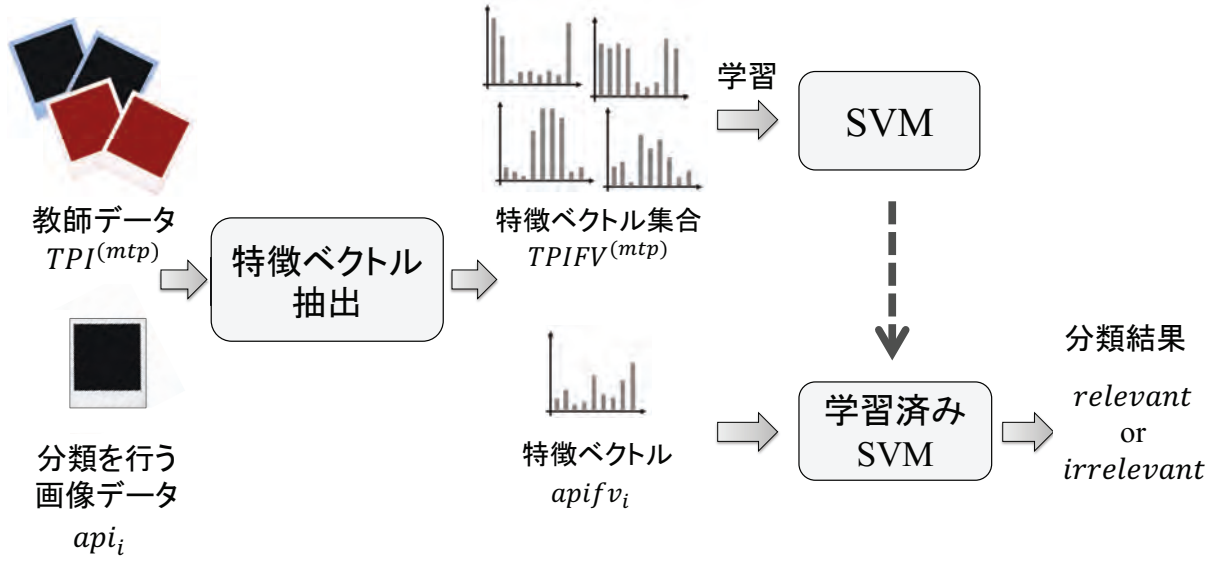


図 21: 提案する画像分類の処理手順

4.3.1　概要

図 20 に画像分類を導入した密度に基づく時空間分析手法の処理手順を示す．時空間クラスタに所属しているジオタグ付きツイートから付与されている画像データを抽出し，画像分類を行う．ここで，モニタリングの対象としているトピックを mtp とする．クラス集合を $CLASS^{(mtp)} = \{\text{“}relevant,\text{”}\ \text{“}irrelevant\text{”}\}$ とし，モニタリングの対象となっているトピックを含む画像データを $relevant$ クラス，それ以外の画像データを $irrelevant$ クラスとして二つに分類する．そして，$relevant$ クラスに分類された画像データを Web アプリケーション上に出力する．

画像分類では，各クラスの内容を表している画像データ集合を教師データとして予め用意しておく．教師データを $TPI^{(mtp)}$ とし，$TPI^{(mtp)} = \{(tpi_1^{(mtp)}, tc_1), (tpi_2^{(mtp)}, tc_2), \cdots, (tpi_n^{(mtp)}, tc_n)\}$ とする．ここで，$tpi_i^{(mtp)}$ は画像データ，tc_i は手動でラベル付けを行ったクラス（$tc_i = \{\text{“}relevant\text{”}\ or\ \text{“}irrelevant\text{”}\} \in CLASS^{(mtp)}$）とする．

4.3.2　処理手順

提案する画像分類の処理手順を図 21 に示す．最初に，教師データ $TPI^{(mtp)}$ の各画像データ $tpi_i^{(mtp)}$ から，BoF[19] または学習済み深層ネットワークを用いて特徴ベクトル

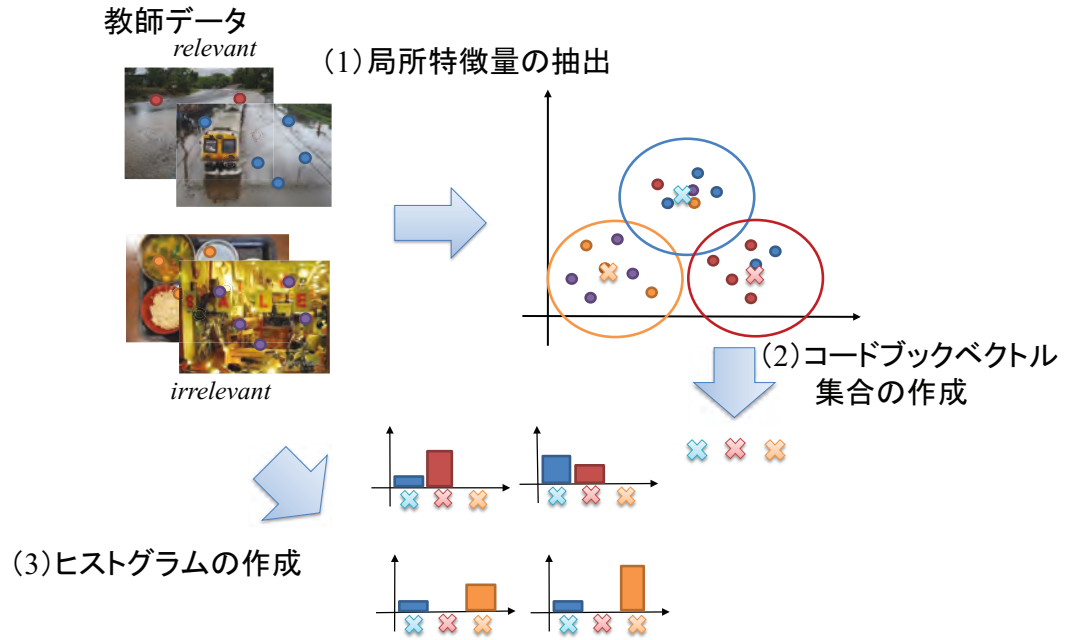


図 22: Bag-of-Features を用いた画像データの特徴ベクトル抽出

$tpifv_i^{(mtp)}$ を抽出する．特徴ベクトル抽出については，4.3.3 項と 4.3.4 項にて説明する．教師データから作成した特徴ベクトル集合を $TPIFV^{(mtp)}$ として次の式で表す．

$$TPIFV^{(mtp)} = \{(tpifv_1^{(mtp)}, tc_1), (tpifv_2^{(mtp)}, tc_2), \cdots, (tpifv_n^{(mtp)}, tc_n)\} \quad (9)$$

そして，作成した $TPIFV^{(mtp)}$ を用いて，SVM を学習させておく．

次に，密度に基づく時空間分析手法によって抽出された時空間クラスタから，ジオタグ付きツイートに付与されている画像データを抽出する．時空間クラスタに所属するジオタグ付きツイートに付与されている画像データ集合を次の式で表す．

$$API = \{(api_1, ac_1), (api_2, ac_2), \cdots, (api_m, ac_m)\} \quad (10)$$

なお，画像分類を行う前の各 ac_i は $ac_i = NULL$ となる．次に，API の各画像データ api_i から特徴ベクトル $apifv_i$ を抽出する．そして，$TPIFV^{(mtp)}$ を用いて学習させた SVM に API の各画像データの特徴ベクトル $apifv_i$ を入力し，$relevant$ または $irrelevant$ クラスに分類を行う．

4.3.3　Bag-of-Features を用いた特徴ベクトル抽出

図 22 に BoF を用いた画像データの特徴ベクトル抽出方法を示す．BoF を用いた画像データの特徴ベクトル抽出は（1）局所特徴量の抽出，（2）コードブックベクトル集合の作成，（3）ヒストグラムの作成の 3 ステップから構成される．

（1）最初に，各教師データから局所特徴量を抽出する．局所特徴量の抽出には，Speeded-up Robust Features（SURF）[66] を用いる．教師データ $tpi_i^{(mtp)}$ の局所特徴量を $TPILF_i^{(mtp)} = \{tpilf_{i,1}^{(mtp)}, tpilf_{i,2}^{(mtp)}, \cdots, tpilf_{i,numtlf(i)}^{(mtp)}\}$ とする．ここで，$tpilf_{i,j}^{(mtp)}$ は抽出された局所特徴量，$numtlf(i)$ は $tpi_i^{(mtp)}$ の局所特徴量数とする．

（2）次に，コードブックベクトル集合の作成を行う．まず，教師データから抽出された全ての局所特徴量を k-means 法を用いてクラスタリングを行う．クラスタリング後，各クラスタの中心ベクトル $cbv_i^{(mtp)}$ をコードブックベクトルとし，コードブックベクトル集合 $CBV^{(mtp)} = \{cbv_1^{(mtp)}, cbv_2^{(mtp)}, \cdots, cbv_k^{(mtp)}\}$ を作成する．

（3）最後に，画像データのヒストグラムを作成する．画像データの各局所特徴量がどのコードブックベクトルから近いか判定し，k 個の要素からなるヒストグラムを作成する．特徴ベクトル抽出を行う画像データを pi_i とし，$numplf(i)$ を $pi_i^{(mtp)}$ の局所特徴量数とする．pi_i の各局所特徴量 $PILF_i = \{pilf_{i,1}, pilf_{i,2}, \cdots, pilf_{i,numplf(i)}\}$ を次の式により分類する．

$$cpilf_{i,j} = \arg\min_c dist(pilf_{i,j}, cbv_c^{(mtp)}) \tag{11}$$

この式では，局所特徴量 $pilf_{i,j}$ が最も近いコードブックベクトルの番号を返している．ここで，pi_i のヒストグラムを $pihist_i$ とし，$pihist_i$ の各要素を $pihist_i[j]$（$1 \leq j \leq k$）としたとき，pi_i のヒストグラムは次の式で求められる．

$$pihist_i[j] = \sum_{l=1}^{numplf(i)} f(cpilf_{i,l}^{(mtp)}, j),$$
$$f(c,j) = \begin{cases} 1 & (if\ c = j) \\ 0 & (otherwise) \end{cases} \tag{12}$$

つまり，ヒストグラムは各コードブックベクトルに対して類似する局所特徴量を何個もっているかを示している．そして，抽出したヒストグラムを画像データ $pi_i^{(mtp)}$ の特徴ベクトルとして使用する．

4.3.4　学習済み深層ネットワークを用いた特徴ベクトル抽出

Twitter に投稿される画像データはバリエーションが多く，決まった種類の画像データが存在しないために，BoF では画像データの特徴を十分に捉えることができない可能性がある．そこで，大規模画像データによって学習させた学習済みの畳み込みニューラルネットワーク（CNN）を特徴ベクトル抽出に使用する．

CNN は数多く提案されている深層ネットワークの中でも，特に画像認識の分野に応用されているニューラルネットワークである．CNN は中間層に畳み込み層とプーリング層

入力層	150528
第1層（畳み込み層）	64
第2層（畳み込み層）	64
プーリング層	
第3層（畳み込み層）	128
第4層（畳み込み層）	128
プーリング層	
第5層（畳み込み層）	256
第6層（畳み込み層）	256
第7層（畳み込み層）	256
プーリング層	
第8層（畳み込み層）	512
第9層（畳み込み層）	512
第10層（畳み込み層）	512
プーリング層	
第11層（畳み込み層）	512
第12層（畳み込み層）	512
第13層（畳み込み層）	512
プーリング層	
第14層（全結合層）	4096
第15層（全結合層）	4096
第16層（出力層）	1000

ユニット数

図 23: VGG-16 の構造

が存在し，画像データの局所的な特徴を自動的に学習することができる．提案手法では，VGG-16，VGG-19[67] と GoogLeNet[68] の 3 つの CNN のネットワークモデルを用いて特徴ベクトル抽出を行う．これらのネットワークの学習は，大規模画像認識コンペティションの ILSVRC[69] にて提供された ImageNet と呼ばれる 1000 分類，120 万枚の画像を用いて行われている．ImageNet は一般的な内容の画像データを含んでおり，ImageNet を用いて学習したネットワークは画像データの様々な分析に応用可能な汎用知識を学習できているといわれている．

VGG-16 の具体的な構造を図 23 に示す．VGG-16 は 16 層から形成される．畳み込み層では，入力に対して重みフィルタの内積を計算する．各畳み込み層は前層の出力に対して畳み込み処理を行い，次の層の入力となる特徴マップを出力する．プーリング層では，畳み込み層から出力された特徴マップを縮小する．VGG-16 では，最大値プーリングを用いている．全結合層では，重み付き結合を計算し，活性化関数によりユニットの値を求める．VGG-16

表 10: 実験期間に抽出された画像データ数

日付	画像データ数
8/1	130
8/2	217
8/3	203
8/4	63
8/5	63
8/6	87
8/7	53
8/8	230
8/9	412
8/10	572

日付	画像データ数
2/10	210
2/11	206
2/12	23
2/13	87
2/14	2,196
2/15	940
2/16	306
2/17	152
2/18	58

では活性化関数として，ReLU を用いている．また，VGG-19 は VGG-16 よりも中間層が 3 層多い 19 層から形成されるネットワークモデルである．GoogLeNet は 22 層から形成される CNN のネットワークモデルである．GoogLeNet の特徴としては，サイズの異なる複数のフィルターで畳み込み処理を行い，その複数の出力を結合するインセプション構造を用いている．

提案手法では，ネットワークモデルの出力層手前の全結合層から特徴ベクトルを抽出する．つまり，VGG-16 を使用する場合は，4,096 次元の特徴ベクトルが抽出される．VGG-19 では 4,096 次元，GoogLeNet では 1,024 次元の特徴ベクトルが抽出される．3 つのネットワークモデルの学習に用いられている ImageNet には気象や自然災害に関する分類を含む画像データは無い．しかしながら，出力層手前の中間層には画像データの汎用的な特徴が表れるため，ソーシャルメディア上に投稿される画像データを区別する特徴ベクトルとして利用できると考える．

4.4　評価実験

提案した画像分類を評価するために，評価実験を行った．本節では，評価実験の結果を示す．

4.4.1　実験内容

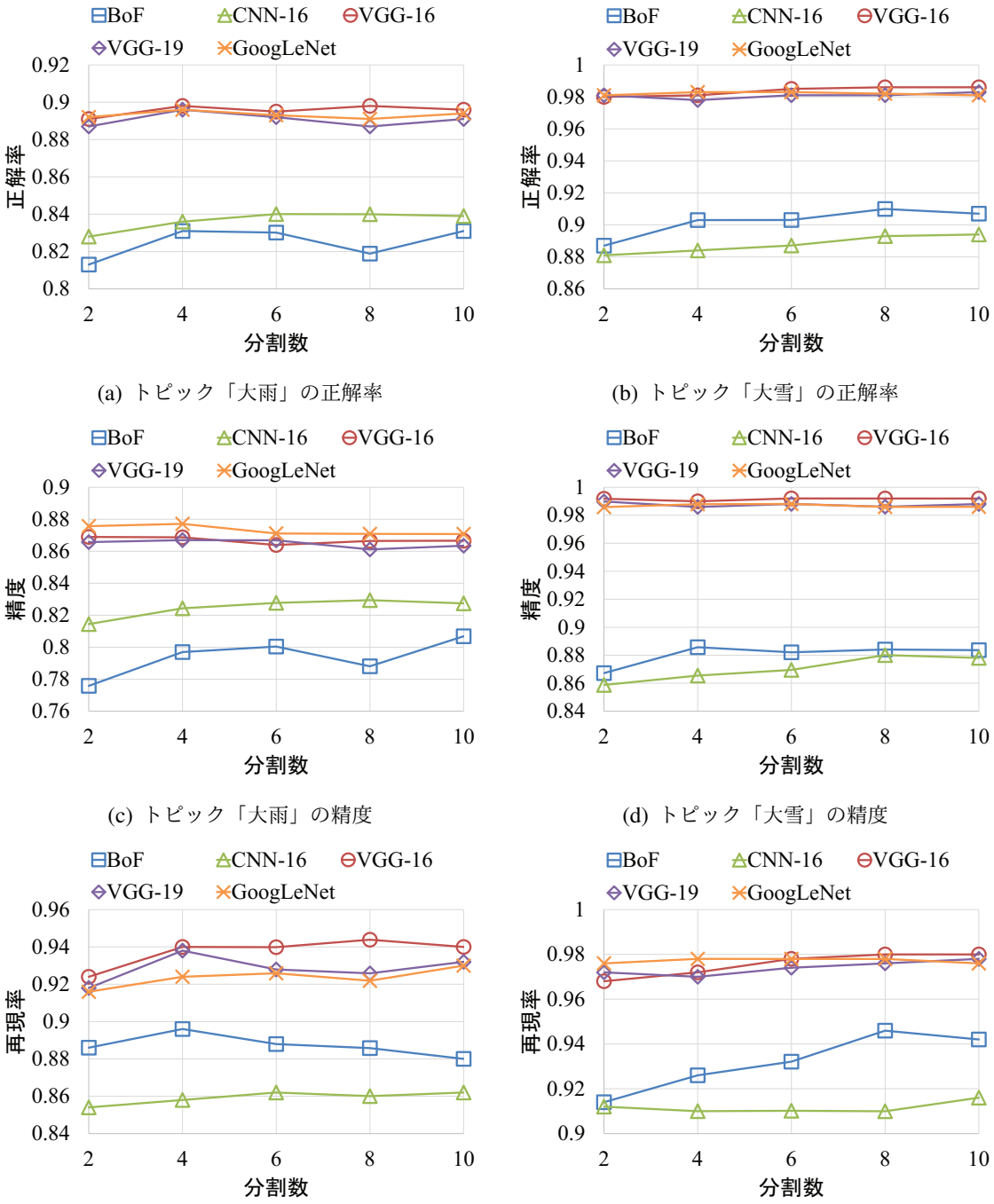
評価実験では，モニタリングをするトピックを「大雨」と「大雪」としてそれぞれ評価を行う．トピック「大雨」についての教師データとして，2014 年 7 月 3 日から 11 日に投稿された関連ジオタグ付きツイートに付与されている画像データから，「大雨」に関するトピックを含む $relevant$ クラス 500 枚，「大雨」に関するトピックを含まない $irrelevant$ クラス 500 枚を用いた．また，トピック「大雪」についての教師データとして，2014 年 1 月 10 日から 2 月 8 日に投稿された関連ジオタグ付きツイートに付与されている画像データから，「大雪」関するトピックを含む $relevant$ クラス 500 枚，「大雪」に関するトピックを含まない $irrelevant$ クラス 500 枚を用いた．実験では，特徴ベクトル抽出手法として BoF を用いた手法，教師データを用いて学習させた VGG-16 と同じ構造をした 16 層の CNN を用いた手法，VGG-16 を用いた手法，VGG-19 を用いた手法と GoogLeNet を用いた手法（それぞれ BoF，CNN-16，VGG-16，VGG-19，GoogLeNet と表記する）を比較する．BoF のコードブックベクトル数は 4,096 とした．また，VGG-16 については，特徴ベクトルを抽出する層による結果の違いを調べるために，14 層から特徴ベクトルを抽出した場合（VGG-16_{14} と表記する）と 15 層から特徴ベクトルを抽出した場合（VGG-16_{15} と表記する）を比較する．

評価方法としては，最初に交差検定による評価を行う．次に，実際に密度に基づく時空間分析手法によって抽出された時空間クラスタから画像データを取り出し，画像分類を行った結果を評価する．実験期間としては，トピック「大雨」については，日本で台風が観測され全国各地で大雨となった 2014 年 8 月 1 日から 10 日としている．トピック「大雪」については，日本全国各地で降雪のあった 2014 年 2 月 10 日から 2 月 18 日とする．表 10 に実験期間に抽出された時空間クラスタに含まれていた画像データ数を示す．

4.4.2　交差検定

最初に交差検定による評価を行う．交差検定の分割数は 2，4，6，8 と 10 分割を用いた．図 24 に，BoF，CNN-16，VGG-16，VGG-19 と GoogLeNet の交差検定の結果を示す．図 24 には，トピックを「大雨」と「大雪」としたときの正解率，精度と再現率をそれぞれ示している．図 24 より，VGG-16，VGG-19 と GoogLeNet は，BoF と CNN-16 よりも高性能であることが分かる．トピック「大雨」では，正解率と再現率は VGG-16，精度は GoogLeNet が最も良い結果となり，トピック「大雪」では，正解率と精度は VGG-16，再現率は GoogLeNet が最も良い結果となった．VGG-16 の正解率は，トピック「大雨」では 0.89，トピック「大雪」では 0.98 となった．以上の結果より，交差検定において提案手法は高性能に画像分類を行うことができ，特に，VGG-16 を用いた手法が最も良い結果となった．

図 25 に VGG-16_{14} と VGG-16_{15} の正解率，精度と再現率をそれぞれ示す．図 25 より，正



(a) トピック「大雨」の正解率

(b) トピック「大雪」の正解率

(c) トピック「大雨」の精度

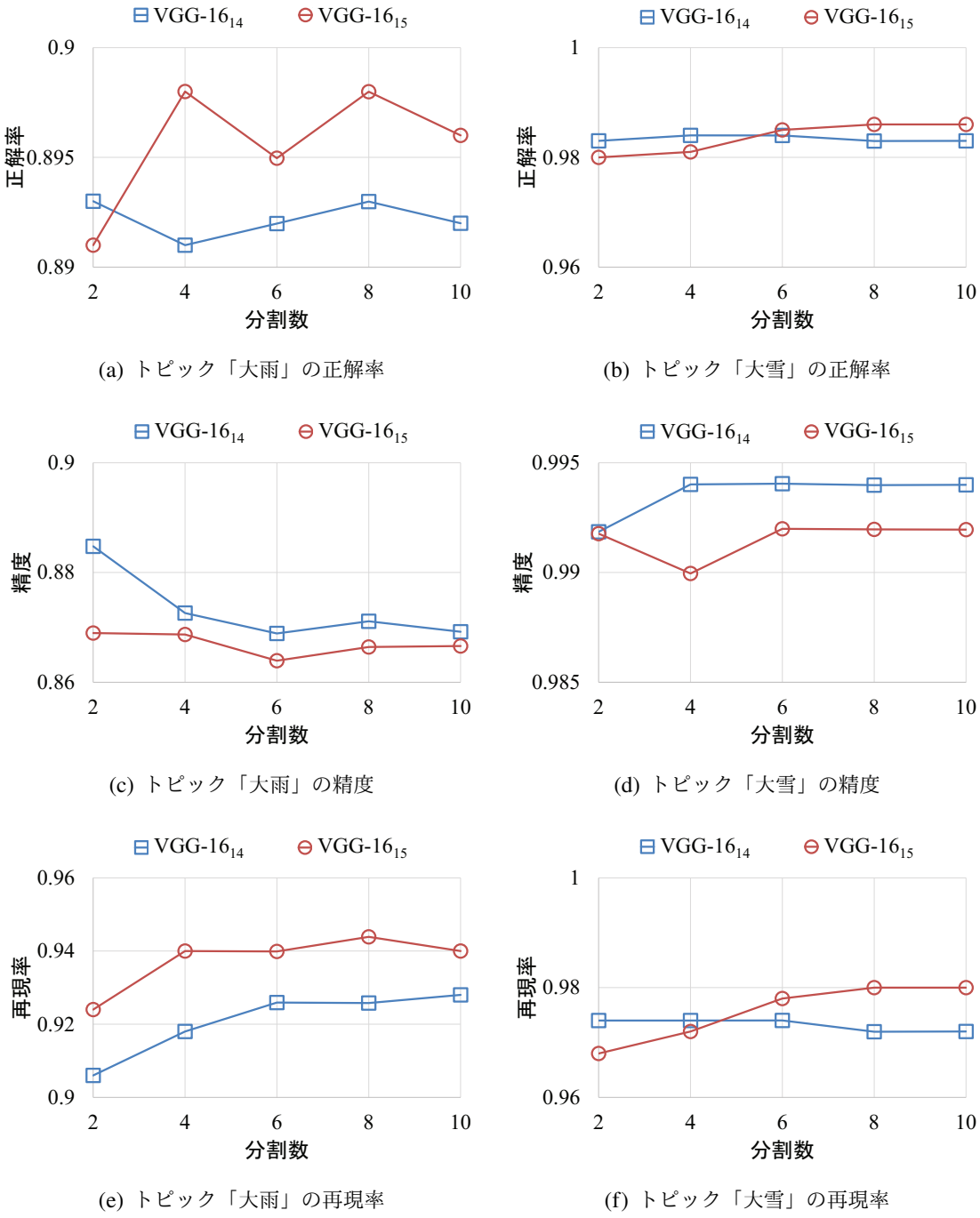
(d) トピック「大雪」の精度

(e) トピック「大雨」の再現率

(f) トピック「大雪」の再現率

図 24: 画像分類の交差検定の結果

解率と再現率は VGG-16$_{15}$ が良く，精度は VGG-16$_{14}$ が良い結果となった．



(a) トピック「大雨」の正解率　(b) トピック「大雪」の正解率

(c) トピック「大雨」の精度　(d) トピック「大雪」の精度

(e) トピック「大雨」の再現率　(f) トピック「大雪」の再現率

図 25: VGG-16 の特徴ベクトルを変更して行った交差検定の結果

4.4.3　密度に基づく時空間分析手法における評価

次に密度に基づく時空間分析手法における評価を行う．実験期間の各日付の画像分類の正解率，精度と再現率を表 11，12 と 13 にそれぞれ示す．表 11，12 と 13 には，その日付

表 11: 密度に基づく時空間分析手法における画像分類の正解率

日付	BoF	CNN-16	VGG-16	VGG-19	GoogLeNet
8/1	0.68	0.73	**0.85**	0.80	0.82
8/2	0.76	0.71	0.83	0.84	**0.85**
8/3	0.75	0.78	0.82	**0.83**	0.80
8/4	**0.83**	0.67	0.81	**0.83**	0.81
8/5	0.67	0.70	**0.78**	0.73	0.75
8/6	0.61	0.63	0.64	0.64	**0.66**
8/7	0.58	0.58	0.70	**0.74**	0.68
8/8	0.62	0.71	**0.74**	**0.74**	0.72
8/9	0.70	0.70	0.77	**0.78**	0.77
8/10	0.73	0.74	0.81	0.81	**0.83**
2/10	0.68	0.68	**0.80**	0.79	0.76
2/11	0.70	0.71	**0.81**	0.79	0.77
2/12	0.65	0.70	0.70	**0.74**	**0.74**
2/13	0.67	0.68	**0.77**	0.76	0.71
2/14	0.78	0.77	**0.87**	**0.87**	0.84
2/15	0.75	0.81	**0.90**	**0.90**	0.88
2/16	0.69	0.71	**0.85**	0.83	0.82
2/17	0.68	0.69	0.80	**0.82**	**0.82**
2/18	0.57	**0.66**	**0.66**	0.64	**0.66**

で最も性能が良い値に下線を引いている．正解率と再現率は VGG-16 がそれぞれ 19 日間中 10 日と 12 日，精度は VGG-19 が 19 日間中 10 日で最も良い結果となった．BoF では $relevant$ クラスへ正しく分類できなかったが，VGG-16 では $relevant$ クラスへ正しく分類することができた画像データを確認したところ，トピック「大雨」と「大雪」以外に人物や物体が写っているものが多いことが分かった．VGG-16 を用いて抽出された特徴ベクトルは，モニタリングをしているトピック以外の人物や物体が写っていたとしても，画像データ中の当該トピックの部分の特徴を捉えることができたといえる．以上の結果より，密度に基づく時空間分析手法における評価では，VGG-16 を特徴ベクトル抽出手法として用いた場合が最も良い結果となった．

表 14 に VGG-16_{14} と VGG-16_{15} の密度に基づく時空間分析手法における画像分類の正解率，精度と再現率を示す．正解率は VGG-16_{14} が 19 日間中 15 日，精度は VGG-16_{14} が 19

表 12: 密度に基づく時空間分析手法における画像分類の精度

日付	BoF	CNN-16	VGG-16	VGG-19	GoogLeNet
8/1	0.59	0.66	**0.75**	0.70	0.72
8/2	0.54	0.49	0.63	0.65	**0.66**
8/3	0.60	0.67	0.69	**0.70**	0.66
8/4	**0.57**	0.37	0.54	**0.57**	0.55
8/5	0.43	0.45	**0.53**	0.48	0.50
8/6	0.34	0.36	0.38	0.38	**0.39**
8/7	0.39	0.37	0.47	**0.50**	0.45
8/8	0.34	0.41	0.44	**0.45**	0.43
8/9	0.39	0.38	0.46	**0.47**	**0.47**
8/10	0.56	0.58	0.66	0.66	**0.68**
2/10	0.74	0.75	**0.83**	0.82	0.82
2/11	0.74	0.76	**0.81**	**0.81**	0.79
2/12	0.29	0.38	0.38	**0.43**	**0.43**
2/13	0.61	0.63	**0.71**	**0.71**	0.68
2/14	0.82	0.84	**0.89**	**0.89**	**0.89**
2/15	0.86	0.90	0.93	**0.94**	**0.94**
2/16	0.72	0.76	**0.82**	0.80	0.81
2/17	0.66	0.71	0.73	0.76	**0.77**
2/18	0.44	**0.52**	0.51	0.50	0.51

日間中 15 日，再現率は VGG-16_{15} が 19 日間中 15 日で良い結果となった．密度に基づく時空間分析手法における評価では，VGG-16 から画像データの特徴ベクトルを抽出する層を変化させたとしても分類性能に大きな変化は無かった．

4.5　まとめ

本章では，密度に基づく時空間分析手法における画像分類を提案した．提案した画像分類では，BoF または学習済み深層ネットワークを用いて画像データから特徴ベクトルを抽出し，SVM を用いて画像分類を行う．そして，モニタリングの対象となっているトピックに関連している画像データのみを Web アプリケーション上に提示することで，密度に基づく時空間分析手法の有効性を向上することができる．評価実験の結果，特徴ベクトル抽出手法

表 13: 密度に基づく時空間分析手法における画像分類の再現率

日付	BoF	CNN-16	VGG-16	VGG-19	GoogLeNet
8/1	0.76	0.76	**0.95**	0.91	0.91
8/2	**0.97**	0.82	0.92	0.95	0.93
8/3	0.86	0.73	0.86	**0.90**	**0.90**
8/4	**0.93**	0.71	**0.93**	**0.93**	0.86
8/5	0.94	0.88	**1.00**	0.94	**1.00**
8/6	0.84	0.89	0.95	**1.00**	**1.00**
8/7	**1.00**	0.79	**1.00**	**1.00**	**1.00**
8/8	0.82	0.80	0.90	**0.92**	**0.92**
8/9	**0.90**	0.82	0.87	**0.90**	**0.90**
8/10	0.88	0.81	0.90	0.89	**0.91**
2/10	0.73	0.71	**0.84**	0.83	0.78
2/11	0.82	0.80	**0.92**	0.87	0.86
2/12	0.40	**0.60**	**0.60**	**0.60**	**0.60**
2/13	0.75	0.73	**0.85**	0.80	0.70
2/14	0.86	0.80	**0.93**	**0.93**	0.87
2/15	0.79	0.84	**0.93**	**0.93**	0.90
2/16	0.80	0.76	**0.96**	0.94	0.91
2/17	0.77	0.68	0.95	**0.96**	0.92
2/18	0.71	0.62	**0.90**	0.86	**0.90**

として VGG-16 を用いた場合，交差検定における正解率として，トピック「大雨」では 0.89，トピック「大雪」では 0.98 を示し，提案手法は高性能にトピック「大雨」と「大雪」に関連する画像データを分類することができた．また，特徴ベクトル抽出手法としては学習済み深層ネットワークである VGG-16 を用いた手法が最も高性能であることを示した．本研究の今後の課題としては，分類性能の向上のために学習済み深層ネットワークを再学習させて新しいモデルを作成し，特徴ベクトル抽出に用いることが挙げられる．学習済みの深層ネットワークを初期値とし，特定のトピックに関する画像データを用いて再学習することで，汎用性があり，また特定のトピックに適した深層ネットワークモデルの作成が期待できる．

表 14: VGG-16 の特徴ベクトルを変更して行った密度に基づく時空間分析手法における画像分類の評価

	正解率			精度			再現率	
日付	VGG-16_{14}	VGG-16_{15}	日付	VGG-16_{14}	VGG-16_{15}	日付	VGG-16_{14}	VGG-16_{15}
8/1	0.83	**0.85**	8/1	0.73	**0.75**	8/1	**0.95**	**0.95**
8/2	**0.84**	0.83	8/2	**0.65**	0.63	8/2	**0.95**	0.92
8/3	0.81	**0.82**	8/3	0.67	**0.69**	8/3	**0.87**	0.86
8/4	**0.81**	**0.81**	8/4	**0.55**	0.54	8/4	0.86	**0.93**
8/5	**0.78**	**0.78**	8/5	**0.53**	**0.53**	8/5	**1.00**	**1.00**
8/6	**0.66**	0.64	8/6	**0.38**	**0.38**	8/6	**0.95**	**0.95**
8/7	**0.70**	**0.70**	8/7	**0.47**	**0.47**	8/7	**1.00**	**1.00**
8/8	**0.74**	**0.74**	8/8	**0.44**	**0.44**	8/8	0.88	**0.90**
8/9	0.76	**0.77**	8/9	0.45	**0.46**	8/9	0.83	**0.87**
8/10	**0.82**	0.81	8/10	**0.67**	0.66	8/10	0.89	**0.90**
2/10	**0.81**	0.80	2/10	**0.85**	0.83	2/10	**0.84**	**0.84**
2/11	**0.81**	**0.81**	2/11	**0.81**	**0.81**	2/11	0.91	**0.92**
2/12	**0.74**	0.70	2/12	**0.43**	0.38	2/12	**0.60**	**0.60**
2/13	**0.78**	0.77	2/13	**0.72**	0.71	2/13	**0.85**	**0.85**
2/14	**0.87**	**0.87**	2/14	**0.89**	**0.89**	2/14	0.92	**0.93**
2/15	**0.91**	0.90	2/15	**0.94**	0.93	2/15	**0.94**	0.93
2/16	0.83	**0.85**	2/16	0.81	**0.82**	2/16	0.92	**0.96**
2/17	**0.81**	0.80	2/17	**0.75**	0.73	2/17	**0.95**	**0.95**
2/18	**0.69**	0.66	2/18	**0.54**	0.51	2/18	**0.95**	0.90

第 5 章　最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN

本章では，最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN について説明する．

5.1　はじめに

ビッグデータへの関心の高まりとともに，データマイニング分野においてデータクラスタリングの高速化に関する研究が注目を集めている．密度に基づくクラスタリングはデータの密度をクラスタリングの基準とした手法で，Ester ら [13, 14] によって提案された．k-means 法はユーザが予め指定した数の円状のクラスタを抽出するのに対し，密度に基づくクラスタリングは予めクラスタ数を指定することなく任意形状のクラスタを自動的に抽出可能な手法として，地理空間情報，情報推薦，異常検知などの様々な分野で利用されている．

密度に基づくクラスタリングの代表的な手法として Ester らは Density-based spatial clustering of applications with noise（DBSCAN）を提案している．DBSCAN では密度の基準として，距離 ϵ 以内の近傍に含まれるデータ数を用いる．そして，データの距離 ϵ 以内に $MinPts$ 個以上のデータが存在するとき当該データをコアデータと呼び，コアデータを結合していくことでクラスタを抽出することができる．DBSCAN はシンプルなアルゴリズムで動作するため，高速化やアプリケーションへの利用など数多くの研究が行われている．

DBSCAN には，各データの近傍を求めるための範囲検索とクラスタを形成するための到達可能データ探索という計算コストの大きい二つの処理がある．空間インデックスを使用しない場合，データ数を n とすると DBSCAN の計算量は $O(n^2)$ となることが知られている．そこで，多くの研究者によって DBSCAN の高速化が行われてきた．高速化の研究では，クラスタリング結果が厳密解となる手法と近似解を許す手法に分かれて研究が行われており，本研究では厳密解の得られる手法を研究対象とする．厳密解の得られる DBSCAN の高速化としてはグリッドベースの手法 [70] が提案されているが，最悪計算量は $O(n^2)$ となる．そこで，範囲検索と到達可能データの探索に依存しない手法として，セルベースの DBSCAN[15] が提案された．セルベースの DBSCAN は範囲検索の回数を大幅に削減し，小さく分割されたセルを基準にクラスタを形成することで到達可能データ探索の処理を必要としないため，処理の高速化が実現できている．

セルベースの DBSCAN は 2 次元の場合において，データセット全体を一辺が長さ $\epsilon/\sqrt{2}$ の小さいセルに分割する．ここで，セル内のデータ数が $MinPts$ 以上である場合，セル内の全てのデータは自動的にコアデータであると判定できる．次にコアデータを含むセルを結

合していく．二つのセル間に距離 ϵ 以内の任意のコアデータのペアがある場合にのみ，その二つのセルを結合する．このセルの結合判定は，単純な方法を用いると条件を満たすペアが見つからない場合に，二つのセル間の全てのコアデータ間の距離計算を行わなければならないため多くの処理時間を要する．

そこで本章では，最小外接矩形（MBR）とセルの再帰分割を用いてセルの結合判定を高速に行う新しいセルベースの DBSCAN を提案する．提案手法の特長は以下の通りである．

- セル中のコアデータを囲む MBR を作成し，MBR 間の距離をセルの結合判定に用いることで，条件を満たす場合にコアデータ間の距離を計算することなく高速にセルの結合判定ができる．
- セルを再帰的に分割し，対象となるコアデータの候補を減らしていくことで，高速にセルの結合判定ができる．

本章の構成は以下の通りである．5.2 節では，関連研究を述べる．5.3 節では，DBSCAN とセルベースの DBSCAN を簡単に説明する．5.4 節では，提案手法について説明する．5.5 節では，評価実験の結果を示し，5.6 節で本章をまとめる．

5.2　関連研究

近年，ビッグデータへの注目の高まりとともに，データベースの大規模化と多次元化が進んでいる．密度に基づくクラスタリング [13, 14] はデータの密度をクラスタリングの基準とした多次元データに有効なクラスタリング手法であり，多くの研究者によってその代表的な手法である DBSCAN の高速化が行われてきた．しかしながら，その多くはクラスタリング結果が近似解となることを許す手法 [30, 71, 72, 73, 74, 75] である．これらの手法はパラメータやデータ分布によっては得られるクラスタリング結果が厳密解とは異なる場合があり，精度の求められるアプリケーションでの利用には有効ではない．

クラスタリング結果が厳密解となることを保ちつつ高速化が可能な手法として，グリッドベースの DBSCAN が提案されている [70, 76]．グリッドベースの DBSCAN の主な特徴として，データセット全体をいくつかのグリッドに分割し，グリッドごとに処理を行うことで範囲検索の処理を高速化できる．そして，グリッドごとに求めたクラスタリング結果を統合することによって，厳密解のクラスタリング結果が得られる．

DBSCAN の高速化手法の一つとして，セルベースの DBSCAN が提案されている [15]．セルベースの DBSCAN は範囲検索の回数を大幅に削減し，小さく分割されたセルを基準にクラスタを形成するため到達可能データ探索の処理を必要としない．Gan ら [16] は，セルベースの DBSCAN の近似解を許す手法と厳密解の得られる手法を提案している．近似解を許す ρ-approximate DBSCAN は既存手法と比較して大幅な高速化ができている．

しかしながら，厳密解の得られる手法は十分な高速化を行うことができていない．厳密解の得られる手法では，セルの結合判定に Bichromatic closest pair（BCP）[77] の考えを用いることで高速化している．本研究では，厳密解の得られるセルベースの DBSCAN に焦点を当て，セルの結合判定に MBR とセルの再帰分割を用いることでさらなる高速化を行う．

Mai ら [78] は乱択アルゴリズムに基づく手法を提案している．Mai らの手法では，ランダムに選択したデータに対して範囲検索を行うことで随時クラスタを形成し，クラスタの状態に変化を与えないデータに対しては範囲検索を行わない．その結果，範囲検索の回数を削減でき，高速に厳密解を得られることを示している．しかしながら，一回の範囲検索で多くのデータをカバーできないようなデータ分布の場合，範囲検索の回数が多くなるため処理時間が大きくなる可能性がある．また，範囲検索に依存する手法であるため，kd-tree といった索引構造の構築が必要となり，セルベースの DBSCAN に比べて多くのメモリを要する．Mai らは，評価実験において，範囲検索の高速化のために kd-tree を索引構造として用いて DBSCAN やセルベースの DBSCAN よりも高速であることを実験結果として示している．ただし，評価実験において，Gan らのパラメータ設定とは異なる実験のみでしか評価を行っていない．

5.3　事前準備

本節では，DBSCAN とセルベースの DBSCAN について簡単に説明する．

5.3.1　DBSCAN

本項では，DBSCAN の諸定義を説明する．d 次元のデータ集合を DT とし，ϵ と $MinPts$ をユーザが与えるパラメータとする．データ $dt_p \in DT$ について，dt_p から距離 ϵ 以内に存在するデータ集合を dt_p の ϵ-近傍と呼び，$N_\epsilon(dt_p)$ と表記する．本研究では，セルベースの DBSCAN と同じくデータ間の距離はユークリッド距離と定める．

定義 13（コアデータ） データ dt_p の ϵ-近傍について，$|N_\epsilon(dt_p)| \geq MinPts$ を満たすとき，dt_p をコアデータと呼ぶ．

定義 14（ϵ-密度的に到達可能） データ列 $(dt_1, dt_2, \cdots, dt_n)$ について，以下の条件を満たすとき，dt_1 から dt_n へ ϵ-密度的に到達可能であると表現する．

(1)　$dt_1, dt_2, \cdots, dt_{n-1}$ がコアデータである．
(2)　dt_{i+1} が dt_i の ϵ-近傍に存在（$dt_{i+1} \in N_\epsilon(dt_i)$）する．

図 26 に定義 13 と定義 14 の例を示す．$MinPts = 5$ とする．データ dt_1 の ϵ-近傍は dt_1

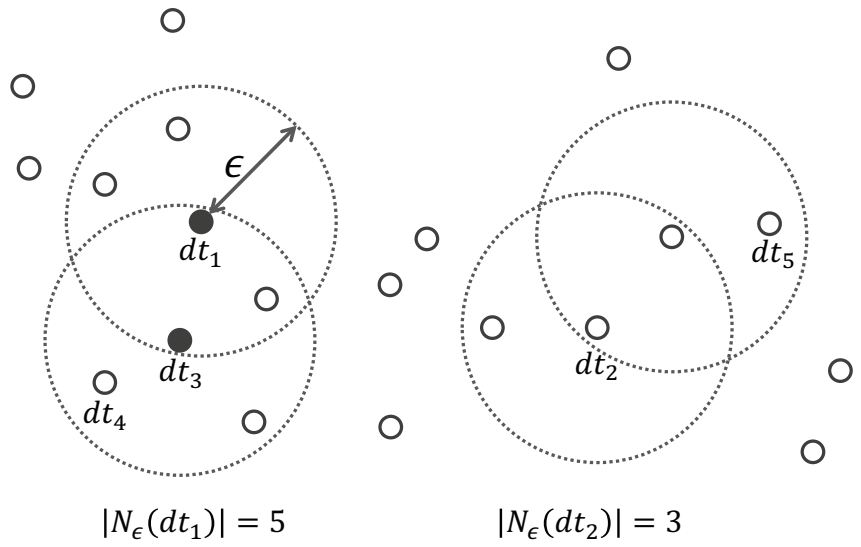


図 26: 定義 13 と定義 14 の例

を含めると $|N_\epsilon(dt_1)| = 5$ となるので，dt_1 はコアデータとなる．反対に，データ dt_2 の ϵ-近傍は $|N_\epsilon(dt_2)| = 3$ であるため，dt_2 はコアデータではない．また，dt_1 から dt_3 と dt_4 へは ϵ-密度的に到達可能であるが，dt_2 から dt_5 へは ϵ-密度的に到達可能ではない．

DBSCAN において，密度に基づくクラスタはコアデータから ϵ-密度的に到達可能なデータを再帰的に接続していくことで形成されていく．コアデータではないがコアデータから ϵ-密度的に到達可能なデータをボーダデータと呼ぶ．また，どのコアデータからも ϵ-密度的に到達可能ではないデータをノイズと呼ぶ．密度に基づくクラスタの定義を以下に示す．

定義 15（密度に基づくクラスタ） データ集合 DT において，密度に基づくクラスタ DBC は以下の条件を満たす部分データ集合である．

(1)　任意のコアデータ $dt_p \in DBC$ について，DBC に属する全てのデータは dt_p から ϵ-密度的に到達可能である．

(2)　任意のデータ dt_p，$dt_q \in DBC$ について，dt_p と dt_q へ ϵ-密度的に到達可能な $dt_o \in DBC$ が存在する．

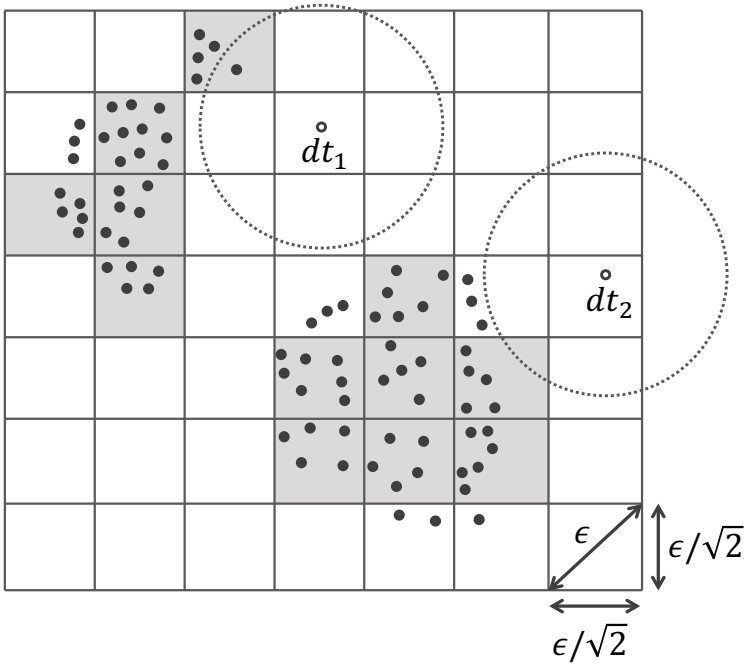


図 27: セルベースの DBSCAN

5.3.2　セルベースの DBSCAN

本項では，セルベースの DBSCAN とその問題点について説明する．セルベースの DBSCAN はセル分割，コアデータの判定，セル結合とボーダデータとノイズの判定の四つの処理から構成される．四つの処理が完了すると，各データはどのクラスタに属するか，またクラスタに属さないかが決定する．

5.3.2.1　セル分割

最初に，データセット全体を小さいセルに分割し，各データを一つのセルに割り当てる．セルは一辺の長さが同じ超立方体とする．d をデータの次元数とすると，セル一辺の長さは $\epsilon/\sqrt{d}$ となる．セル一辺の長さを $\epsilon/\sqrt{d}$ とすることでセルの対角線の長さは ϵ となる．図 27 にセルベースの DBSCAN の例を示す．図 27 は $d = 2$ であり，$\epsilon/\sqrt{2} \times \epsilon/\sqrt{2}$ のセルに分割されている．

5.3.2.2　コアデータ判定

次に，各データがコアデータであるかそうでないか判定する．セル内のデータ間の距離は最大でも高々 ϵ であり，セル内にデータが $MinPts$ 以上存在する場合，そのセル内のデータは全てコアデータと判定できる．セル内のデータ数が $MinPts$ 未満の場合，近隣するセルに存在するデータとの距離計算を行い，コアデータであるか判定する．図 27 の例を見ると，$MinPts = 5$ とすると，グレーで示しているセルにはデータが 5 以上あり，これらのセルに属するデータは全てコアデータとなる．

5.3.2.3　セル結合

次に，コアデータを含むセルを結合しクラスタを形成していく．二つのセル間に距離 ϵ 以内の任意のコアデータのペアがある場合，二つのセルに含まれる全データは ϵ-密度的に到達可能となり，二つのセルを結合し一つのクラスタとする．図 27 の例を見ると，左側に密集しているセルに属するデータ集合と，右側に密集しているセルに属するデータ集合とで二つのクラスタが形成される．

5.3.2.4　ボーダデータとノイズの判定

最後に，コアデータではないデータに対してクラスタに属するボーダデータ，もしくはノイズとなるか判定する．判定するデータについて，もし ϵ 以内にコアデータが存在する場合，そのデータはボーダデータとなり，最も近いコアデータが属するクラスタに属する．反対に，ϵ 以内にコアデータが存在しない場合，そのデータはクラスタに属さないノイズとなる．図 27 の例を見ると，データ dt_1 から距離 ϵ 以内にコアデータが存在しており，dt_1 はボーダデータとなる．反対に，データ dt_2 から距離 ϵ 以内にはコアデータが存在していないので，dt_2 はノイズとなる．

5.3.2.5　問題点

セルベースの DBSCAN で最も処理時間を要するのはセル結合の処理である．セルの結合判定方法として，単純な方法としては二つのセルのコアデータ間の距離計算を総当りで行い，距離 ϵ 以内のコアデータのペアが見つかり次第，判定を終える方法がある．しかしながら，条件を満たすペアが見つからない場合に二つのセル間の全てのコアデータ間の距離計算を行わなければならないため多くの処理時間を要する．

そこで，既存手法 [16] では BCP[77] の考えを導入し高速化を行っている．既存手法にて用いられている判定方法は以下の通りである．

(1) 二つのセルのコアデータ数をかけた値が m 未満の場合は総当りによる判定を行う．

(2) 全てのコアデータについて相手のセルとの距離を計算し昇順ソートを行う．
(3) 距離が近い順にコアデータ間の距離計算を行う．

BCP は二つのデータ集合間の最も距離の近いペアを見つける問題であり，距離 ϵ 以内のペアがある場合には高速にそのペアを見つけることができる．しかしながら，BCP を用いた手法では，距離 ϵ 以内のコアデータのペアが無い場合には全てのコアデータ間の距離を計算し，結合できないと判定する必要がある．

5.4　提案手法

本節では，提案手法である MBR とセルの再帰分割を用いたセルベースの DBSCAN について説明する．

5.4.1　概要

提案手法では，セルの結合判定に MBR[79] とセルの再帰分割を用いる．提案手法の判定方法は以下の通りである．

(1) 二つのセルのコアデータ数をかけた値が m 未満の場合は総当りによる判定を行う．
(2) 二つのセルに含まれるコアデータを囲む MBR をそれぞれ作成し，MBR 間の MINMAXDIST と MINMINDIST を用い，二つのセルが結合できるかできないか判定する．MINMAXDIST とは二つの MBR に含まれるデータ間の最大距離になりうる値の最小値であり，MINMINDIST とは二つの MBR に含まれるデータ間の最小距離になりうる値の最小値である．
(3) MBR を用いたセルの結合判定が行えない場合は二つのセルを再度分割し，お互いに近い領域に存在するコアデータのみを対象に再度 MBR を作成し，セルの結合判定を行う．
(4) MBR を用いた判定とセルの分割を k 回繰り返し行い，判定できない場合には，BCP を用いた手法で判定する．

5.4.2　MBR を用いたセルの結合判定方法

あるセル C_p 内のコアデータを囲む MBR を $CM_p(MAX_p, MIN_p)$ と表記する．ここで，MAX_p はセルに含まれるコアデータの各次元の最大値のベクトルであり，$MAX_p = (max_{p,1}, max_{p,2}, \cdots, max_{p,d})$ とする．また，MIN_p はセルに含まれるコアデータの各次元の最小値のベクトルであり，$MIN_p = (min_{p,1}, min_{p,2}, \cdots, min_{p,d})$

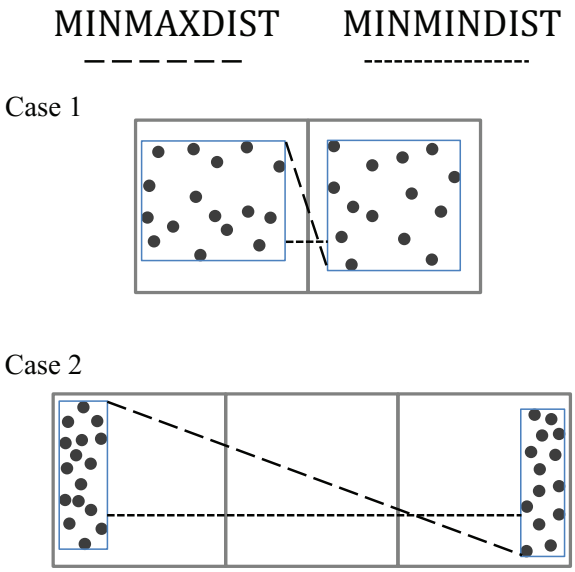


図 28: MBR を用いたセルの結合判定の例

とする．提案手法では，処理時間の削減のため MINMAXDIST については文献 [79] で定義された上界を用いる．ある二つのセル C_p と C_q 内のコアデータを囲む MBR を $CM_p(MAX_p, MIN_p)$ と $CM_q(MAX_q, MIN_q)$ とすると，この二つのセル間の MINMAXDIST の上界，$MMAD(CM_p, CM_q)$ は次の式で表される．

$$MMAD(CM_p, CM_q) = \sqrt{\min_{1 \le j \le d} \left\{ x_j^2 + \sum_{i=1, i \neq j}^{d} y_i^2 \right\}} \tag{13}$$

ここで，$x_j = \min\{|max_{p,j} - max_{q,j}|, |max_{p,j} - min_{q,j}|, |min_{p,j} - max_{q,j}|, |min_{p,j} - min_{q,j}|\}$，$y_i = \max\{|max_{p,i} - min_{q,i}|, |min_{p,i} - max_{q,i}|\}$ である．また，この二つのセル間の MINMINDIST，$MMID(CM_p, CM_q)$ は次の式で表される．

$$MMID(CM_p, CM_q) = \sqrt{\sum_{i=1}^{d} y_i^2}$$

$$y_i = \begin{cases} min_{q,i} - max_{p,i}, & if\ min_{q,i} > max_{p,i} \\ min_{p,i} - max_{q,i}, & if\ min_{p,i} > max_{q,i} \\ 0, & otherwise \end{cases} \tag{14}$$

補題 1（MBR を用いたセルの結合判定） C_p と C_q 間の最も距離の近いコアデータのペア間の距離を $MID(C_p, C_q)$ とすると，MINMAXDIST と MINMINDIST の関係から以下の式が成り立つ．

$$MMID(CM_p, CM_q) \le MID(C_p, C_q) \le MMAD(CM_p, CM_q) \tag{15}$$

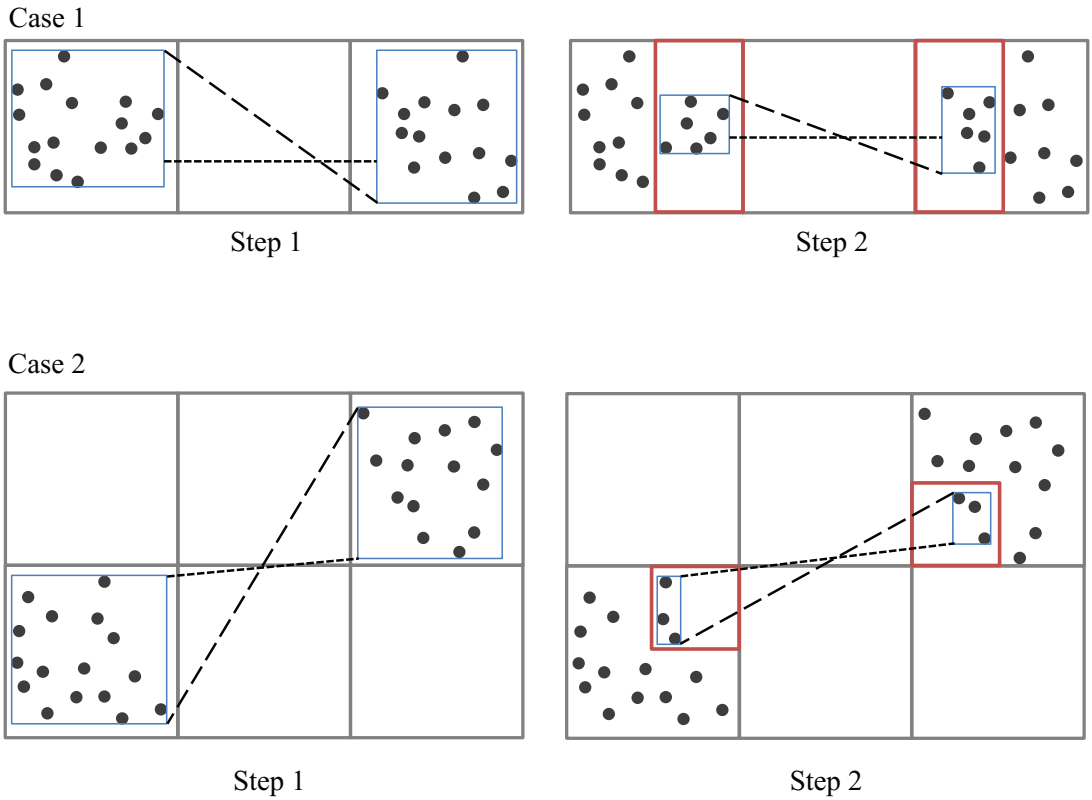


図 29: セルの再帰分割と MBR を用いたセルの結合判定の例

また，式（15）より，C_p と C_q の結合について以下のように判定できる．

- $MMAD(CM_p, CM_q) \leq \epsilon$ であれば，二つのセル C_p と C_q は結合できる．
- $MMID(CM_p, CM_q) > \epsilon$ であれば，二つのセル C_p と C_q は結合できない．

証明　$MMAD(CM_p, CM_q) \leq \epsilon$ のとき，式（15）より，$MID(C_p, C_q) \leq MMAD(CM_p, CM_q) \leq \epsilon$ となり，$MID(C_p, C_q) \leq \epsilon$ が成り立つ．このとき，C_p と C_q に距離 ϵ 以内の任意のコアデータのペアが存在するため，C_p と C_q は結合できる．$MMID(CM_p, CM_q) > \epsilon$ のとき，式(15)より，$\epsilon < MMID(CM_p, CM_q) \leq MID(C_p, C_q)$ となり，$MID(C_p, C_q) > \epsilon$ が成り立つ．このとき，C_p と C_q の全てのコアデータのペア間の距離が ϵ よりも大きくなるため，C_p と C_q は結合できない．　□

図 28 に MBR を用いたセルの結合判定の例を示す．図 28 の青い四角は MBR を示している．Case 1 では，二つのセルの MINMAXDIST が ϵ 以下であるため，結合できると判定できる．反対に，Case 2 では，二つのセルの MINMINDIST が ϵ よりも大きく，結合できないと判定できる．

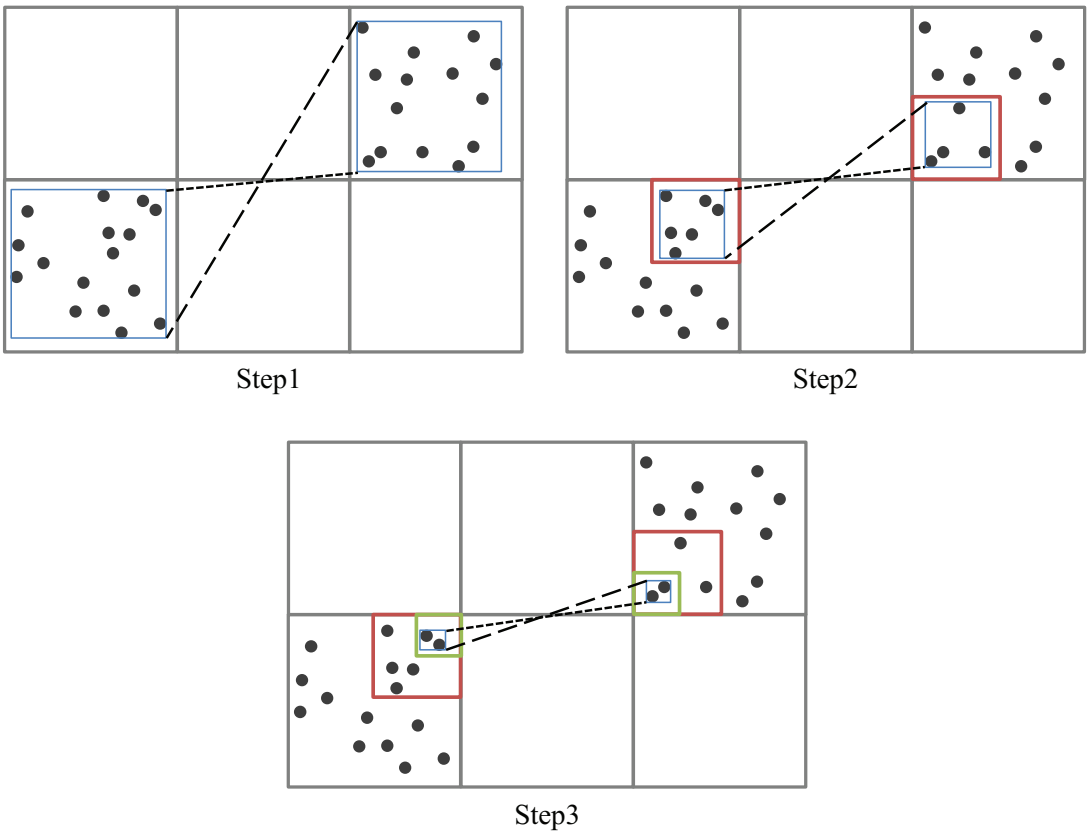


図 30: セルが再帰的に分割されていく例

5.4.3　セルの再帰分割方法

$MMID(CM_p, CM_q) \leq \epsilon < MMAD(CM_p, CM_q)$ のとき，MBR を用いたセルの結合判定を行うことはできない．このとき，二つのセルを再度分割し，お互いに近い領域に存在するコアデータのみを対象に再度 MBR を作成し，セルの結合判定を行う．具体的には，二つのセルが各次元について何番目に位置しているか，各次元の位置番号を求める．そして，二つのセルの位置番号が異なる次元について，セルを半分に分割し，お互いに近い領域のコアデータを対象とする．

図 29 にセルの再帰分割を行い，MBR を用いたセルの結合判定の例を示す．図 29 の Case 1 と Case 2 の Step 1 では，二つのセルからなる MBR によって判定を行ったとしても，条件に当てはまらず判定できない．Case 1 では，横軸の次元について位置が異なるので，二つのセルを横軸について半分に分割し，お互いに近い領域に存在するコアデータのみを対象に再度 MBR を作成している．その結果，二つのセルの MINMAXDIST が ϵ 以下となり，結合できると判定できる．Case 2 では，縦軸と横軸の次元について位置が異なるので，二つの

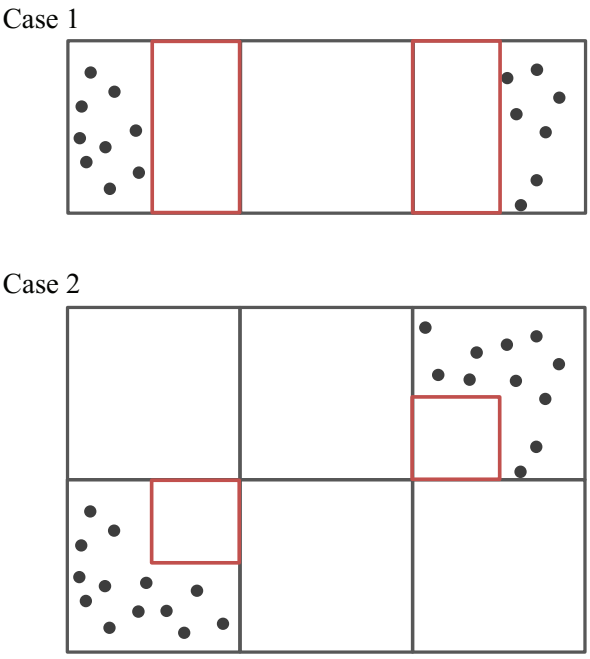


図 31: セルの再帰分割を用いたセルの結合判定の例

セルを両方の軸について半分に分割し，お互いに近い領域に存在するコアデータのみを対象に再度 MBR を作成している．その結果，二つのセルの MINMINDIST が ϵ よりも大きくなり，結合できないと判定できる．

セルの分割と MBR を用いた判定を再帰的に行い，二つのセルの対象のコアデータ数をかけた値が m 未満となれば総当りに切り替えて判定する．また，k 回分割を行っても判定できない場合は BCP を用いた手法で判定する．図 30 にセルが再帰的に分割されていく例を示す．図 30 では，Step 1 と Step 2 で作成した MBR ではセルの結合判定ができない．Step 2 で分割されたセルをさらに分割することで，結合できると判定できる．

また，セルの再帰分割によって自動的にセルの結合判定ができる場合がある．図 31 の二つのセルの結合判定の例では，分割後に結合判定の対象となる二つのセル中のコアデータ数が 0 となっている．このように隣接していないセルの結合判定において，再帰分割を行い対象のコアデータが無くなった場合，距離 ϵ 以内のペアが存在する可能性は無いため，結合できないと判定できる．

5.4.4　アルゴリズム

本項では，提案手法のアルゴリズムについて説明する．提案手法の流れを Algorithm 4 に示す．Algorithm 4 は入力として，データセット DT，次元数 d，セルの結合判定において総当たりに切り替える計算数 m，MBR を用いた判定とセルの再帰分割の繰り返し回数 k，ϵ と

Algorithm 4 提案手法のセルベースの DBSCAN のアルゴリズム

Input: DT, d, m, k, ϵ, $MinPts$
Output: $ClusterList$, $NoiseList$
1: /*セル分割*/
2: データセット全体を一辺の長さが $\epsilon/\sqrt{d}$ の超立方体に分割し，$CellList$ を作成
3: **for** $i = 0$ to $|DT|$ **do**
4:　　データ dt_i を一つのセルに割り当てる
5: **end for**
6: データ数が 0 個のセルを $CellList$ から削除
7: /*コアデータ判定*/
8: **for** $i = 0$ to $|CellList|$ **do**
9:　　**if** $|DN(C_i)| \geq MinPts$ **then**
10:　　C_i 内のデータをコアデータとする
11:　　**else**
12:　　**for** $j = 0$ to $|DN(C_i)|$ **do**
13:　　C_i 内の j 番目のデータに対して，近隣のセル内のデータと距離計算を行いコアデータ判定を行う
14:　　**end for**
15:　　**end if**
16: **end for**
17: /*セル結合*/
18: **for** $i = 0$ to $|CellList|$ **do**
19:　　**for** $j = (i + 1)$ to $|CellList|$ **do**
20:　　**if** C_i と C_j の距離が ϵ 以下 **then**
21:　　**if** CheckConnectCells(C_i, C_j, d, m, k, ϵ) == true **then**
22:　　C_i と C_j を結合
23:　　**end if**
24:　　**end if**
25:　　**end for**
26: **end for**
27: 結合されたセル内のデータ集合をクラスタとし，$ClusterList$ に追加
28: /*ボーダデータとノイズの判定*/
29: **for** $i = 0$ to $|DT|$ **do**
30:　　**if** dt_i がコアデータではなく，dt_i の ϵ 以内にコアデータが存在 **then**
31:　　dt_i から最も近いコアデータが属するクラスタに dt_i を追加
32:　　**else**
33:　　dt_i を $NoiseList$ に追加
34:　　**end if**
35: **end for**
36: return $ClusterList$, $NoiseList$

$MinPts$ を入力として受け取り，クラスタリスト $ClusterList$ とノイズリスト $NoiseList$ を返す．

最初に，データセット全体を一辺の長さが $\epsilon/\sqrt{d}$ の超立方体に分割してセルリスト $CellList$ を作成し，各データを一つのセルに割り当て，データ数が 0 個のセルは削除する（2〜6 行）．次に，各データがコアデータであるかそうでないか判定する（8〜16 行）．Algorithm 4 中で $DN(C_i)$ はセル C_i に含まれるデータ集合とする．次に，セルを結合しク

Algorithm 5 CheckConnectCells(C_p, C_q, d, m, k, ϵ)

Input: C_p, C_q, d, m, k, ϵ
Output: true または false
1: **if** $|CN(C_p)| \times |CN(C_q)| < m$ **then**
2:　return $CN(C_p)$ と $CN(C_q)$ の総当りによる判定結果
3: **else**
4:　$CN(C_p)$ と $CN(C_q)$ を囲む MBR，CM_p と CM_q をそれぞれ作成
5:　**if** $MMAD(CM_p, CM_q) \leq \epsilon$ **then**
6:　return true
7:　**else if** $MMID(CM_p, CM_q) > \epsilon$ **then**
8:　return false
9:　**else**
10:　**if** $k == 0$ **then**
11:　return $CN(C_p)$ と $CN(C_q)$ の BCP を用いた判定結果
12:　**end if**
13:　**for** $i = 0$ to d **do**
14:　**if** i 次元について C_p と C_q の位置番号が異なる **then**
15:　$CN(C'_p) = \emptyset$
16:　$CN(C'_q) = \emptyset$
17:　**for** $j = 0$ to $|CN(C_p)|$ **do**
18:　**if** i 次元について $CN(C_p)$ の j 番目のデータが，C_q へ C_p の中央よりも近い **then**
19:　$CN(C'_p)$ へ $CN(C_p)$ の j 番目のデータを追加
20:　**end if**
21:　**end for**
22:　**for** $j = 0$ to $|CN(C_q)|$ **do**
23:　**if** i 次元について $CN(C_q)$ の j 番目のデータが，C_p へ C_q の中央よりも近い **then**
24:　$CN(C'_q)$ へ $CN(C_q)$ の j 番目のデータを追加
25:　**end if**
26:　**end for**
27:　$CN(C_p) = \emptyset$
28:　$CN(C_q) = \emptyset$
29:　$CN(C'_p)$ と $CN(C'_q)$ を $CN(C_p)$ と $CN(C_q)$ へそれぞれ取得
30:　**end if**
31:　**end for**
32:　**if** C_p と C_q が隣接していない && $CN(C_p) == \emptyset$ && $CN(C_q) == \emptyset$ **then**
33:　return false
34:　**else**
35:　$k = k - 1$
36:　return CheckConnectCells(C_p, C_q, d, m, k, ϵ)
37:　**end if**
38:　**end if**
39: **end if**

ラスタを形成する（18〜26 行）．セルの結合判定には，関数 CheckConnectCells を用いる．結合されたセル内のデータ集合をクラスタとし，$ClusterList$ に追加する（27 行）．最後に，コアデータではないデータに対して，ボーダデータかノイズか判定を行う（29〜35 行）．ノイズと判定されたデータは $NoiseList$ に追加する．

セルの結合判定を行う関数 CheckConnectCells の流れを Algorithm 5 に示す．Algorithm 5 は入力として，判定対象の二つのセル C_p および C_q と，d，m，k と ϵ を入力として受け取り，二つのセル C_p と C_q が結合できる（true）か結合できない（false）かを返す．Algorithm 5 中で $CN(C_p)$ と $CN(C_q)$ はそれぞれ C_p と C_q のコアデータ集合とする．

最初に，C_p と C_q のコアデータ数をかけた値が m 未満であれば，総当りによる判定を行う（1〜2 行）．そうでなければ，C_p と C_q のコアデータを囲む MBR をそれぞれ作成し，MINMAXDIST と MINMINDIST を用いて判定する（4〜8 行）．判定できなかった場合，再帰呼び出しのパラメータ k が 0 であれば，BCP を用いた判定を行う（10〜11 行）．そうでなければ，位置番号が異なる次元について C_p と C_q を分割し，お互いに近い領域のコアデータのみを取得する（13〜31 行）．セルの分割によって自動的に判定できる場合は，判定する（32 行〜33 行）．k をデクリメントし，C_p，C_q，d，m，k，ϵ を入力とし CheckConnectCells を再帰的に呼び出す（35 行〜36 行）．

5.4.5　計算量と厳密性の考察

提案手法の時間計算量について考察する．提案手法は既存のセルベースの DBSCAN とセルの結合判定以外の処理は同じである．ここでは，二つのセル C_p と C_q に対するセルの結合判定の最悪時間計算量について提案手法と BCP を用いた手法とを比較する．

C_p と C_q のコアデータ数の合計を N とする．MBR 作成の時間計算量は $O(dN)$ である．MINMAXDIST と MINMINDIST の時間計算量はそれぞれ $O(d)$ である [79]．セルの再帰分割の時間計算量は，各次元についてセルを二分割し，C_p と C_q の各コアデータがどちらに属するか判定するので，$O(dN)$ となる．MBR を用いた判定とセルの再帰分割の繰り返し回数を k とすると，$O(k(2dN + d))$ となり定数を除くと，提案手法のセルの結合判定の時間計算量は $O(kdN)$ となる．

一方，既存手法の BCP を用いたセルの結合判定において，各コアデータと相手のセルとの距離計算を行い，距離が近い順にソートを行うと，時間計算量は $O(dN + N \log N)$ となる．例外として，両手法ともに C_p と C_q のコアデータ数をかけた値が m 未満であれば，時間計算量は $O(dm)$ となる．k の値が提案手法の時間計算量に影響するが，k は一桁程度であるため，提案手法は既存手法の BCP を用いた手法よりも高速であるといえる．また，提案手法でも BCP を用いてセルの結合判定を行うことがあるが，BCP を用いた判定は他の方法を用いた判定と比べて回数が少ないことを 5.5.2 項に示す．

提案手法の空間計算量について考察する．提案手法は既存のセルベースの DBSCAN と同様に，データとセル格納のためのスペースに比例した空間計算量を要する．データ数を n とすると，全データの空間計算量は $O(dn)$ である．セル数を l とすると，各セルは自身を構成する二つの座標を持つので，空間計算量は $O(2dl)$ となる．また，各データをセルに割り

表 15: 使用したデータセットの詳細

データセット名	次元数	データ数
*SSD*2	2	1,000,000 から 10,000,000
*SSD*3	3	1,000,000 から 10,000,000
*SSD*5	5	1,000,000 から 10,000,000
*SSD*7	7	1,000,000 から 10,000,000
*PAMAP*2	4	3,850,505
FARM	5	3,627,086
HOUSEHOLD	7	2,049,280

当て，各セルはデータ ID を保持するため，$O(n)$ が必要となる．提案手法の空間計算量は $O(dn + dl + n)$ となり，d と l は n に比べると非常に小さいので，n について線形となることが分かる．

既存のセルベースの DBSCAN と異なる点は，セルの結合判定の際に，セルの再帰分割のために判定対象の二つのセルに属するコアデータを複製し，メモリを使用する．しかしながら，対象の二つのセルの結合判定が終了するとそのメモリは解放される．全データとセルの空間計算量に比べると二つのセルのコアデータ数は非常に少ないため，空間計算量についての影響はないといえる．

次に，提案手法が出力するクラスタリング結果の厳密性について考察する．セルベースの DBSCAN は，セルの結合判定において，二つのセル間に距離 ϵ 以内の任意のコアデータのペアがある場合にセルを結合できる，ペアがない場合にセルを結合できないと判定することで厳密解であることを保つことができる [15]．提案手法は MBR を用いた判定とセルの再帰分割を行い，判定できない場合でも，BCP を用いて判定を行うことで厳密解であることを保つことができている．MBR を用いた判定は，補題 1 より，正しく判定できることを示している．セルの再帰分割では，距離 ϵ 以内のペアになる可能性があるコアデータを対象から外さないように，分割したお互いに近い領域のデータを残していくことで厳密解であることを保つことができている．

5.5　評価実験

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

5.5.1　実験内容とデータセット

評価実験では，提案手法と既存手法の厳密解の得られる DBSCAN の処理時間の比較を行う．比較手法としては，セルの結合判定に総当りを用いる手法（CDBSCAN と表記する），セルの結合判定に BCP を用いる手法（$\mathrm{CDBSCAN_{BCP}}$ と表記する）[16]，乱択アルゴリズムに基づく手法（AnyDBC と表記する）[78] と提案手法（$\mathrm{CDBSCAN_{MBR}}$ と表記する）を比較する．実験に用いる計算機としては，CPU が Intel Xeon E5-1270 V2 @3.5 GHz，メモリが 32GB の計算機を使用した．プログラミング言語は C++ を用いて実装し，コンパイラとして g++ 4.9.4，オプションとして-O3 を使用した．

データセットは文献 [16] にて開発された Seed Spreader を使用して作成した人工データ（*SSD*2，*SSD*3，*SSD*5，*SSD*7）と文献 [16] にて使用された実データ（*PAMAP*2，*FARM*，*HOUSEHOLD*）[80, 81] を使用した．Seed Spreader は空間内に密集したデータ集合をいくつかランダムに配置し，少量のノイズをランダムに配置することで，クラスタリングのベンチマークに適した分布のデータセットを作成できる．各データセットの詳細を表 15 に示す．クラスタリングのパラメータも文献 [16] と同様に，$m = 10,000$，$MinPts = 100$，ϵ は 5,000 から結果が一つのクラスタになるまで変化させて実験を行った．また，セルの再帰分割回数は実験的にいくつかの値を用いて検証を行い，その中から良い結果を示した $k = 5$ とした．

5.5.2　人工データを使用した実験結果

最初に，各データ数の人工データを使用した実験結果を図 32 に示す．図 32 の実験結果には，全て $\epsilon = 5,000$ を用いた場合の結果を示している．図 32 より，セルベースの DBSCAN を比較すると，*SSD*2 では $\mathrm{CDBSCAN_{MBR}}$ と $\mathrm{CDBSCAN_{BCP}}$ の結果に大きな差は無いが，全ての人工データにおいて $\mathrm{CDBSCAN_{MBR}}$ が最も高速であることが分かる．特に，データの次元数が大きくなるほど $\mathrm{CDBSCAN_{MBR}}$ は $\mathrm{CDBSCAN_{BCP}}$ と比較して高速になっている．AnyDBC と比較しても，*SSD*2 の 1,000 万件のデータセット以外では，$\mathrm{CDBSCAN_{MBR}}$ の方が高速である．*SSD*2 では $\mathrm{CDBSCAN_{MBR}}$ は高速化できなかった理由については後述する．

図 33 にデータ数 2,000,000 件のデータを使用して，パラメータ ϵ を変化させて行った実験結果を示す．なお，AnyDBC の実験結果について，処理時間が 5,000 秒を超えた結果は記載していない．図 33 より，セルベースの DBSCAN を比較すると，ϵ を変えて設定した場合でも，$\mathrm{CDBSCAN_{MBR}}$ が最も高速であることが分かる．*SSD*2 や ϵ を大きい値に設定した結果では，AnyDBC が最も高速な場合がある．一般に AnyDBC は範囲検索によってカバーするデータが多くなると高速に動作するので，ϵ が十分に大きくなると処理が高速にな

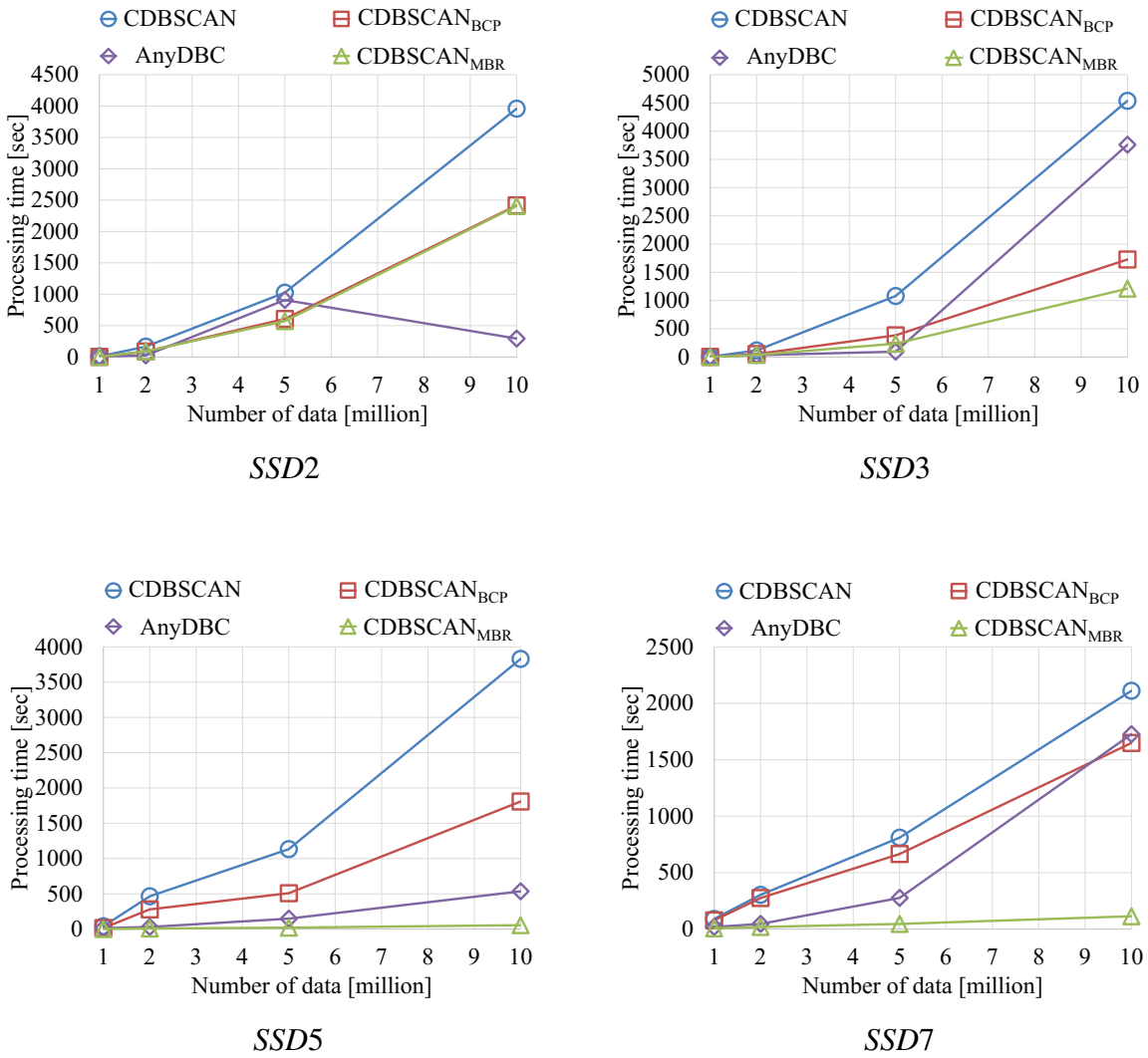


図 32: 人工データを使用した実験結果

る．しかしながら，乱択であることから処理時間が安定していないために，データセットやパラメータによってはセルベースの DBSCAN よりも処理時間が大きくなっている．以上の結果より，人工データを使用した場合，CDBSCAN$_{\mathrm{MBR}}$ は CDBSCAN$_{\mathrm{BCP}}$ よりも高速であり，AnyDBC よりも安定して高速化できることを示した．

表 16 に，データ数 2,000,000 件と $\epsilon = 5,000$ を使用した場合の CDBSCAN$_{\mathrm{MBR}}$ のセルの結合判定に用いられた各方法の回数と割合を示す．表 16 より，各データセットについて約 5 割から 9 割が MINMAXDIST，MINMINDIST またはセルの再帰分割によってセルの結合判定ができており，BCP を用いた判定回数が少ないのが分かる．また，*SSD5* では MINMINDIST とセルの再帰分割を用いた判定が，*SSD7* ではセルの再帰分割を用いた判定が *SSD2* および *SSD3* と比較して多いのが分かる．処理時間の比較では高次元のデータになるほど CDBSCAN$_{\mathrm{MBR}}$ はより高速化ができていることから，この二つの方法を用いた判定

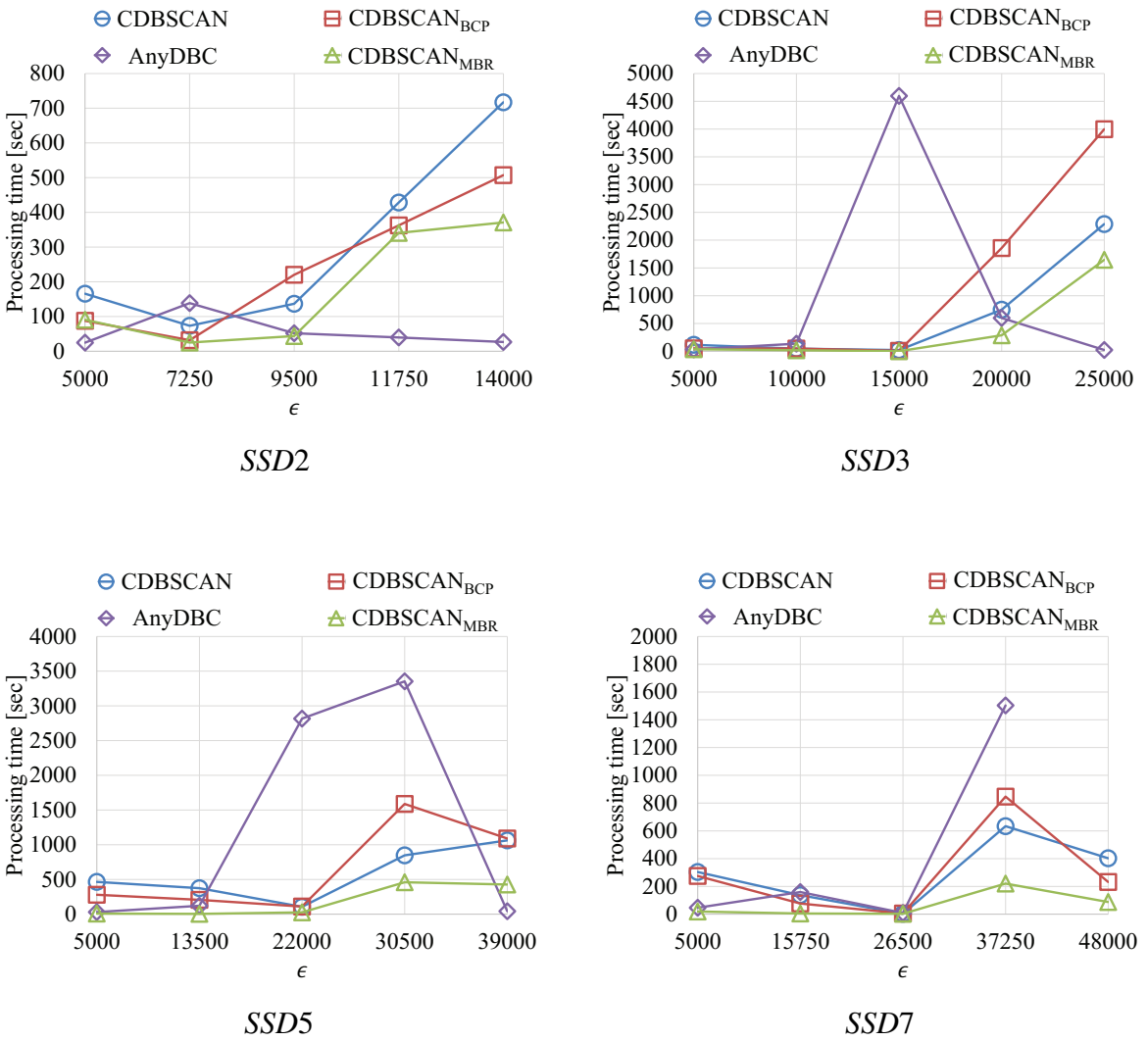


図 33: 人工データを使用し，ϵ を変化させた実験結果

表 16: CDBSCAN$_{MBR}$ におけるセルの結合判定の各方法の回数と割合

データセット	総当り	MINMAXDIST	MINMINDIST	セルの再帰分割	BCP
SSD2	60(0.30)	115(0.58)	14(0.07)	3(0.02)	7(0.03)
SSD3	81(0.08)	766(0.75)	100(0.10)	46(0.05)	26(0.03)
SSD5	6,606(0.16)	17,151(0.42)	7,487(0.18)	8,249(0.20)	1,565(0.04)
SSD7	178,176(0.44)	107,320(0.26)	46,621(0.11)	66,902(0.16)	9,670(0.02)

が高速であったため，高次元になるほどより全体の処理時間の高速化ができたといえる．

表 17 に，CDBSCAN$_{BCP}$ と CDBSCAN$_{MBR}$ のセルの結合判定に用いられた各方法の処理時間を示す．表 17 より，低次元のデータになるほど総当たりによる判定に処理時間を要しているのが分かる．低次元の方が高次元と比較してセル数が少なく，一つのセル中

表 17: $\mathrm{CDBSCAN}_{\mathrm{BCP}}$ と $\mathrm{CDBSCAN}_{\mathrm{MBR}}$ におけるセルの結合判定の各方法の処理時間

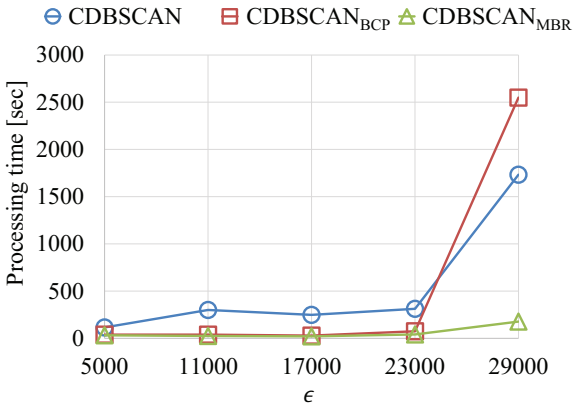
手法	データセット	総当り	MINMAXDIST	MINMINDIST	セルの再帰分割	BCP
$\mathrm{CDBSCAN}_{\mathrm{BCP}}$	*SSD*2	107.55	-	-	-	2.29
	*SSD*3	44.17	-	-	-	14.82
	*SSD*5	0.25	-	-	-	275.07
	*SSD*7	2.77	-	-	-	261.22
$\mathrm{CDBSCAN}_{\mathrm{MBR}}$	*SSD*2	91.32	0.05	0.01	0.01	0.10
	*SSD*3	36.05	0.09	0.01	0.03	0.04
	*SSD*5	0.16	0.53	0.13	1.34	3.40
	*SSD*7	1.99	0.58	0.12	3.07	3.88

のデータが多くなり，この違いが出てきている．表 16 より，*SSD*7 において総当たりによる判定回数の割合は多いが，一つのセル中のデータが少ないため，処理時間は少なくなっている．$\mathrm{CDBSCAN}_{\mathrm{MBR}}$ はより高次元のデータほど高速化できるが，低次元のデータになるほど総当たりによる判定が全体の処理時間を占める割合が多くなり，$\mathrm{CDBSCAN}_{\mathrm{BCP}}$ と $\mathrm{CDBSCAN}_{\mathrm{MBR}}$ で大きな差が出ないことが分かった．以上の理由により，*SSD*2 では，$\mathrm{CDBSCAN}_{\mathrm{MBR}}$ は高速化できていないことから相対的に AnyDBC が最も高速となった．

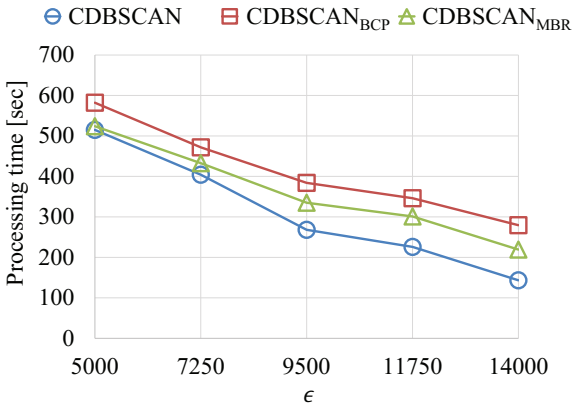
5.5.3　実データを使用した実験結果

次に，実データを使用し ϵ を変化させて行った実験結果を図 34 に示す．本実験では，CDBSCAN，$\mathrm{CDBSCAN}_{\mathrm{BCP}}$ と $\mathrm{CDBSCAN}_{\mathrm{MBR}}$ の実験結果を比較する．図 34 より，*PAMAP*2 では $\mathrm{CDBSCAN}_{\mathrm{MBR}}$ が最も高速であるが，*FARM* と *HOUSEHOLD* では CDBSCAN が最も高速な場合がある．本実験で使用した実データは人工データと比べてデータが空間全体にまばらに広がっているという違いがある．各実データで $\epsilon = 5,000$ を使用した結果において，9 割以上のセルの結合判定が二つのコアデータ数をかけた値が m 未満になったことによる総当たりでの判定であった．つまり，データが空間全体にまばらに広がっているようなデータセットでは，低次元のデータと同様に総当たりによる判定が全体の処理時間を占める割合が多くなり，セルベースの DBSCAN の手法の結果に大きな差が出ないことが分かった．

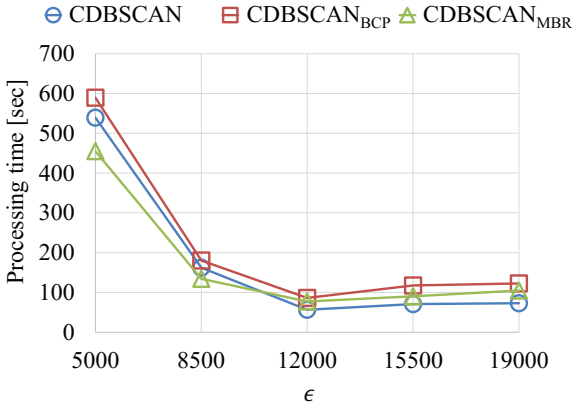
セルの結合判定は ϵ 以内のコアデータのペアを見つけることができれば，その時点で判定を終えることができる．実際に *FARM* と *HOUSEHOLD* では，総当りの早い段階で終わるようなセルの結合判定が多いことを確認した．そこで，$\mathrm{CDBSCAN}_{\mathrm{MBR}}$ において，最初の 100 件のコアデータは総当りによって投機的にセルの結合判定を行い，判定できなかった場合に MBR とセルの再帰分割を行うように処理を改良した．データセット *FARM* と *HOUSEHOLD* を使用し，CDBSCAN と改良した $\mathrm{CDBSCAN}_{\mathrm{MBR}}$ の実験結果を図 35 に示



PAMAP2



FARM



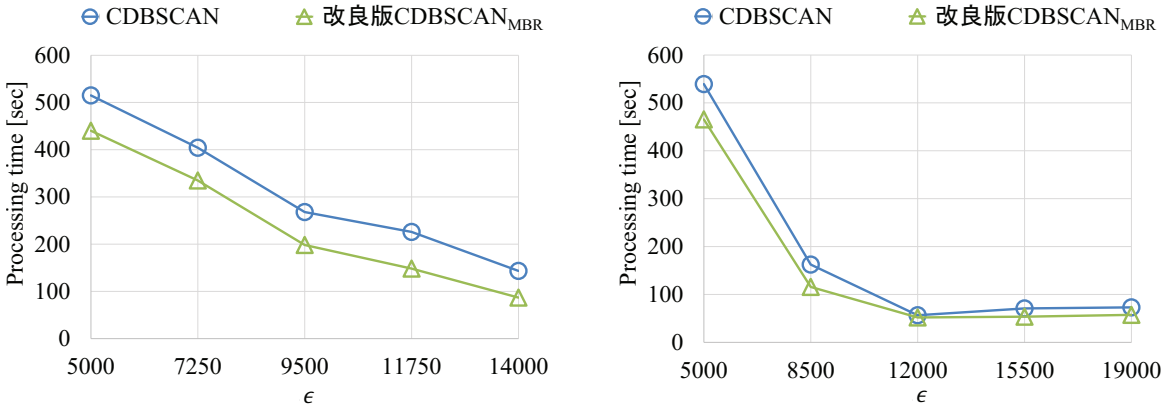
HOUSEHOLD

図 34: 実データを使用した実験結果

す．図 35 より，改良した CDBSCAN$_{MBR}$ は CDBSCAN よりも高速であることが分かる．しかしながら，通常，データクラスタリングを行う際には分布が分からない状態で行われる．最初にいくつかのデータで投機的に総当りで判定し，判定できなかった場合に高度な処理を行っていく方法は有効であるが，総当りを行うデータ数を自動決定する方法を考える必要がある．

5.6　まとめ

本章では，密度に基づくクラスタリングの代表的な手法である DBSCAN の高速化のために，MBR とセルの再帰分割を用いた新しいセルベースの DBSCAN を提案した．提案手法はセルベースの DBSCAN のセルの結合判定について，MBR を用いた結合判定とセルの分



FARM　　*HOUSEHOLD*

図 35: 改良した CDBSCAN$_{\mathrm{MBR}}$ を使用した実験結果

割を再帰的に行うことによって高速に処理することができる．

評価実験の結果より，人工データを用いた場合，低次元のデータでは既存手法と比較して大幅な高速化はできなかったが，高次元のデータになるほど大幅な高速化ができた．実データを用いた場合，データが空間全体にまばらに広がっている分布であることから，高速化できなかったデータセットもあった．

しかしながら，投機的に総当りを行う方法を用いることで全てのデータセットで高速化ができた．ただし，データの分布を何らかの方法で高速に判定し，投機的に総当りを行う回数を自動的に決定する必要がある．本研究の今後の課題としては，DBSCAN の更なる高速化のために，投機的な総当り数の自動決定方法と本章で提案した手法の並列化を行うことが挙げられる．また，提案手法はデータ間の距離にユークリッド距離を用いた DBSCAN のみに対応している．そのため，ユークリッド距離以外の距離計算方法を用いる密度に基づくクラスタリングに対応することが課題として挙げられる．

第 6 章　密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出

本章では，密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出について説明する．

6.1　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から有益な知識を抽出する研究が盛んに行われている [82, 83, 84, 85, 86, 87, 88, 89]．例えば，写真共有サイトである Flickr に投稿される観光関連の写真（画像データ）について，写真に付けられたコメント（テキストデータ）を用いてユーザの関心が高い観光スポットを取り出す研究が行われている [90]．また，近年，GPS 付きスマートフォンの普及と GPS に連動したアプリケーションが利用できる環境が整うとともに，位置情報付きの画像データがソーシャルメディア上に投稿されるようになっている．つまり，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 [91]，新しい空間情報としてその利活用が期待されている [92, 93, 94]．

本章では，ソーシャルメディア上に投稿される，画像，テキストと位置情報を持つデータのことをジオソーシャル画像データと呼ぶ．ジオソーシャル画像データとして，Twitter に投稿されるジオタグ付きのツイートで画像データが付与されているツイートに着目する．Twitter に投稿されるジオソーシャル画像データは，個人的な趣味や話題だけでなく，ユーザが各地域で日々目にした事象や話題を含んでおり [95, 96]，各地域のトピックを抽出することができれば，観光情報，マーケティングや動向分析に活用することができる．

そこで，本章では，ジオソーシャル画像データから各地域のトピックを抽出するための手法を提案する．内容が類似したジオソーシャル画像データが多数投稿される地域はユーザの関心の高いトピックが存在すると考えられる．提案手法では，最初に，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングを用いて，空間的また内容的にも類似したジオソーシャル画像データが密に投稿されている注目領域をマルチモーダル空間クラスタとして抽出する．次に，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出し分かりやすくするために，マルチモーダル空間クラスタ中に含まれるジオソーシャル画像データ集合の類似度グラフから，ネットワークベースの重要度算出手法を用いて，代表画像データを抽出する．

(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングは，密度に基づく空間クラスタリング [13, 14] の拡張であり，各地域において投稿数の差異があること，つまり，都市部と郊

外とでは投稿数が異なることを考慮し，空間クラスタの抽出基準を変化させている．また，ジオソーシャル画像データ間の類似度は，テキストと画像データ間の二つの類似度を考慮する必要がある．そこで，提案手法では，テキストと画像データの異なる種類のデータ間の類似度を計算するために，マルチモーダル類似度を定めている．提案するマルチモーダル類似度はテキストデータ間の類似度と画像データ間の類似度のトレードオフとなっている．テキストデータ間の類似度は，テキストデータに含まれる語句集合のコサイン類似度によって求める．画像データ間の類似度は，Bag-of-Features（BoF）[19] または学習済み深層ネットワークを用いて抽出した画像データの特徴ベクトル間のコサイン類似度によって求める．

提案手法を評価するために，Twitter 上に投稿される画像データを含むジオタグ付きツイートを用いて評価実験を行った．評価実験の結果，各地域のトピックをマルチモーダル空間クラスタとして取り出すことできた．また，マルチモーダル空間クラスタから代表画像データを抽出することで，空間クラスタが持つトピックが分かりやすくなることも確認できた．

本章の構成は以下の通りである．6.2 節では，関連研究を述べる．6.3 節では，データ構造と提案手法の全体の処理手順を説明する．6.4 節では，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングについて説明する．6.5 節では，ネットワークベースの重要度算出手法について説明する．6.6 節では，評価実験の実験結果を示し，6.7 節で本章をまとめる．

6.2　関連研究

インターネット上のデータを分析し，実世界の情報を抽出することは，データマイニング分野で盛んに行われている研究テーマとなっている [97]．近年，インターネット上のユーザはソーシャルメディアサイトに，携帯電話やスマートフォンで撮影した地域の話題やイベントの写真を画像データとして投稿し，情報発信を盛んに行うようになってきている．Flickr のような写真共有サイトや Twitter 上に投稿された写真は，個人的な趣味だけでなく社会的なトピックを含んでいるため，分析対象として注目を集めている．インターネット上のユーザは，投稿される画像データを通して各地域のトピックを収集し，また，画像データを情報発信のツールとして活用している．このように，インターネット上に投稿される写真はジオソーシャル画像データとして大切な情報源として扱われるようになってきているため，ジオソーシャル画像データに対するデータマイニング技術の確立は重要な研究課題の一つとなっている．

ソーシャルメディア上のデータからトピックやイベントを抽出する研究が盛んに行われている．Canneyt ら [22] は，サポートベクターマシン（SVM）を用いて，ソーシャルメディアサイト上のデータから新しい名所を検出する手法を提案した．Lee ら [32, 33] は，ボロノイ図を用いて領域を分割し，投稿数などが急増した領域を特定することで地域的なイベントを検出する手法を提案した．奥ら [39] は，旅行先などの現地ならではの語句を自動的に抽出

する手法を提案した．飲食店や観光施設，娯楽施設など実空間上の位置情報と関連付けられたオブジェクトをスポットと定義し，旅行先などの対象地域の中心から半径 r 以内に存在するスポットデータを用いて，地域限定性スコアという地域限定語句を抽出するための尺度を定義している．杉谷ら [34] は，ジオタグ付きツイートから時空間クラスタリングを用いて，地域的なイベントを抽出する手法を提案している．山本ら [23] は，食事，交通，災害や気象などの様々な生活の局面を設定しておき，それぞれの局面にツイートを分類することで，実世界トピックの抽出を行った．新田ら [25] はユーザが与えたクエリ（キーワード）に対して，そのクエリと短期的かつ長期的に共起して投稿されているキーワードを，関連度が高いキーワードとして定め，そのキーワードを含むツイートをクエリに関する実世界観測情報として抽出した．

Flickr 上に投稿された画像データを分析する研究がこれまでに盛んに行われている．Jaffe ら [41] は，Flickr 上に投稿されたジオソーシャル画像データをクラスタリングする手法を提案している．場所情報を用いて階層的に画像データをクラスタリングし，ホットスポットの検出を行っている．また，Rattenbury ら [90] は，Flickr 上に投稿される画像データからイベントサイトを検出する研究を行っている．タグデータを使って画像データがどのイベントであるかを推定する手法も提案している．Kennedy ら [98] は，地理的なランドマークに関連する画像データを検索するためにタグデータを用いた検索システムを提案している．Yanai ら [42] は，k-means 法を使ってジオソーシャル画像データをクラスタリングする手法を提案している．このように Flickr 上における研究は多数存在するが，クラスタリング部分については，位置情報やタグのみで行われている．

提案手法は，密度に基づくクラスタリングを用いている．Kisilevich ら [11] は，ジオタグ付き画像データに対して DBSCAN を用いて空間クラスタを求める手法を提案している．ジオタグ付き画像データを投稿したユーザ数を密度と考え空間クラスタを抽出している．Chen ら [17] は，頻度の高いタグを持つ画像データを密度に基づくクラスタリングである DBSCAN を用いてクラスタリングし，時空間上でイベント検出する手法を提案している．Shirai ら [31] は，DBSCAN を用いてジオタグ付き画像データからホットスポットを見つける手法を提案している．これらの研究は空間クラスタリングを用いているが，クラスタリングの過程では位置情報のみが使用され，データ間の類似度は考慮されていない．

Ji ら [99] は，Flickr 上のジオソーシャル画像データからスペクトルクラスタリングを用いて代表画像を抽出する手法を提案している．また，Schinas ら [28] は，類似したテキストによるトピックモデルを用いて，各トピックにおいて，グラフベースのアルゴリズムで類似した画像データを取り出す手法を提案している．また，McParlane ら [27] は，トピックに関連した画像を取り出し，ランキングする手法を提案している．

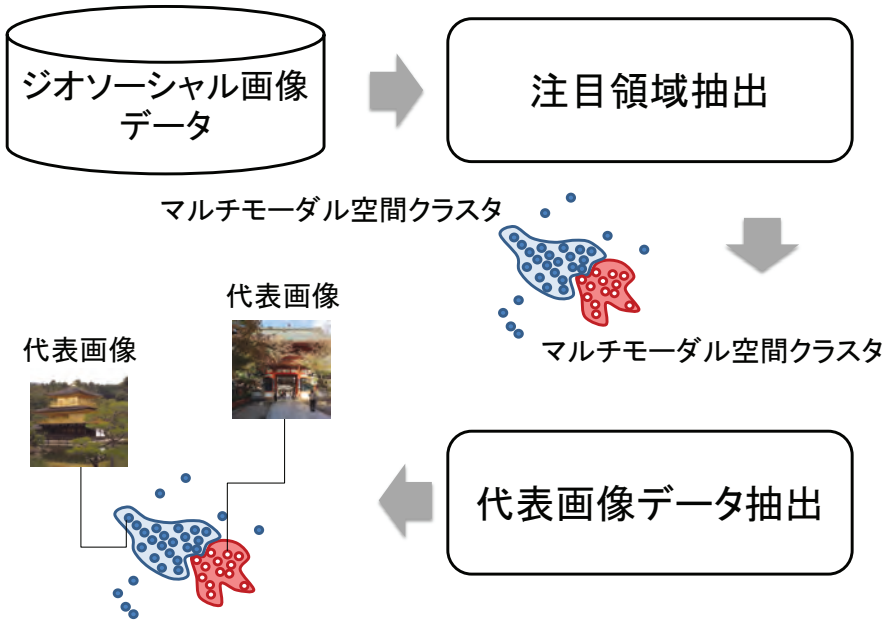


図 36: 密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出の概要

6.3　提案手法

本節では，ジオソーシャル画像データのデータモデルと提案手法の処理手順について述べる．

6.3.1　データモデル

ジオソーシャル画像データ集合を $GSI = \{gsi_1, gsi_2, \cdots, gsi_n\}$ とする．ジオソーシャル画像データ gsi_i は，次の 3 つの要素から構成される画像データである．

$$gsi_i = <pl_i, text_i, photo_i> \tag{16}$$

ここで，pl_i は投稿位置（例えば，施設名や経度・緯度などのジオタグなど），$text_i$ は画像と同時に投稿されたテキストデータ（例えば，タイトル，コメントやタグなど），$photo_i$ は画像データ，もしくは画像データが格納されているオンライン上の URL である．

6.3.2　概要

類似したジオソーシャル画像データが密に投稿されている空間上の領域はユーザの関心が高い注目領域であり，注目領域とその内容を特定することでジオソーシャル画像データ集合からトピックを抽出する．図 36 に提案手法の概要を示す．提案手法では，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングを用いて，注目領域をマルチモーダル空間クラスタとして抽出する．そして，抽出した各マルチモーダル空間クラスタについて，クラスタに含まれるジオソーシャル画像データから代表画像データをネットワークベースの重要度算出手法を用いて選出する．

6.3.3　注目領域抽出

ジオソーシャル画像データ集合中で類似したジオソーシャル画像データはユーザの関心が高いトピックを含んでいる可能性が高い．また，類似したジオソーシャル画像データの投稿が集中している領域は，各地域においてユーザの関心が高いトピックである可能性が高い．そこで，密度に基づくマルチモーダル空間クラスタリングを用いて，空間的に近く，画像データまたはテキストデータの内容が類似したジオソーシャル画像データ集合をマルチモーダル空間クラスタとして抽出する．ジオソーシャル画像データが持つ情報は，画像データだけでなく，位置情報，テキストデータを含むためマルチモーダルであり，マルチモーダルなデータの空間クラスタリングとして (ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングを用いる．(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングについては 6.4 節で詳しく説明する．

6.3.4　代表画像データ抽出

密度に基づくマルチモーダル空間クラスタリングを用いて抽出したマルチモーダル空間クラスタは注目領域であるが，各マルチモーダル空間クラスタが持つトピックは，ジオソーシャル画像データを一つ一つ閲覧しないと分からない．そこで，各マルチモーダル空間クラスタに含まれるジオソーシャル画像データの類似度グラフを作成し，ネットワークベースの重要度算出手法を用いて，重要度を算出する．重要度の高いジオソーシャル画像データを取り出し，代表画像データとしてユーザに提示することで，各注目領域が持つトピックを分かりやすくする．ネットワークベースの重要度算出手法については 6.5 節で詳しく説明する．

6.4　(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリング

本節では，類似したジオソーシャル画像データが空間的に密に投稿されている領域を取り出すための (ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングについて説明する．

6.4.1　空間密度尺度

(ϵ, σ)-密度に基づくマルチモーダル空間クラスタは，密度に基づく空間クラスタリング [13, 14] の諸定義を拡張して定義される．密度に基づく空間クラスタリングでは，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．密集しているかどうかの判断は，各データの近傍 ϵ 以内に，$MinPts$ 以上のデータが存在するかどうかで判定する．

本研究では，あるジオソーシャル画像データについて，類似したジオソーシャル画像データが $MinGSI$ 以上，近傍に存在する場合，当該ジオソーシャル画像データ周辺の密度は高いと判断する．また，マルチモーダル空間クラスタを形成する閾値である $MinGSI$ を各地域の過去の投稿率で逓減することで，地域毎の密度に応じた空間クラスタを抽出可能にしている．

6.4.2　マルチモーダル類似度

ジオソーシャル画像データはマルチモーダルであり，本研究では，テキストと画像データの二つを考慮したマルチモーダル類似度を定める必要がある．例えば，ある二つのジオソーシャル画像データについて，テキストの内容は類似していなくとも画像データが類似した場合，反対に，テキストの内容は類似しているが画像データは類似していない場合において，2 つのソーシャル画像データは類似していると判断する尺度が必要となる．

本研究では，テキストデータ間の類似度と画像データ間の類似度のトレードオフであるマルチモーダル類似度関数 $msim$ を次のように定める．

$$msim(gsi_i, gsi_j) = w \times tsim_{i,j} + (1 - w) \times psim_{i,j}, \tag{17}$$

ここで，w は重み $(0 \leq w \leq 1)$，$tsim_{i,j}$ はテキストデータ間の類似度，$psim_{i,j}$ は画像データ間の類似度である．

テキストデータ間の類似度 $tsim_{i,j}$ は，語句ベースのコサイン類似度を使用する．ここで，$TW_i = \{tw_{i,1}, tw_{i,2}, \cdots, tw_{i,|TW_i|}\}$ を $text_i$ に含まれる語句集合とし，$|TW_i|$ は $text_i$ に含まれる語句の数とする．本研究では，テキストデータ中に現れる名詞，動詞，形容詞を語句

として扱う．テキストデータ間の類似度 $tsim_{i,j}$ は次の式で表される．

$$tsim_{i,j} = \frac{|TW_i \cap TW_j|}{\sqrt{|TW_i||TW_j|}}. \tag{18}$$

画像データ間の類似度 $psim_{i,j}$ については，BoF[19] または学習済み深層ネットワークを用いて画像データ $photo_i$ から特徴ベクトル pfv_i を抽出し，コサイン類似度によって求める．画像データ $photo_i$ の特徴ベクトルについて，l 次元目の値を $pfv_i[l]$ と表記すると，画像データのコサイン類似度は，

$$psim_{i,j} = \frac{\sum_{l=1}^{dim} pfv_i[l] \times pfv_j[l]}{\sqrt{\sum_{l=1}^{dim} pfv_i[l]^2}\sqrt{\sum_{l=1}^{dim} pfv_j[l]^2}}, \tag{19}$$

となる．

6.4.2.1　Bag-of-Features を用いた特徴ベクトル抽出

BoF を用いた画像データの特徴ベクトル抽出は，（1）局所特徴量の抽出，（2）コードブックベクトル集合の作成，（3）ヒストグラムの作成の 3 ステップから構成される．

（1）最初に，各画像データから局所特徴量を抽出する．局所特徴量の抽出には，Speeded-up Robust Features（SURF）[66] を用いる．画像データ $photo_i$ の局所特徴量を $PLF_i = \{plf_{i,1}, plf_{i,2}, \cdots, plf_{i,numplf(i)}\}$ とする．ここで，$plf_{i,j}$ は抽出された局所特徴量，$numplf(i)$ は $photo_i$ の局所特徴量数とする．

（2）次に，コードブックベクトル集合の作成を行う．まず，画像データから抽出された全ての局所特徴量を k-means 法を用いてクラスタリングを行う．クラスタリング後，各クラスタの中心ベクトル cbv_i をコードブックベクトルとし，コードブックベクトル集合 $CBV = \{cbv_1, cbv_2, \cdots, cbv_k\}$ を作成する．

（3）最後に，画像データのヒストグラムを作成する．画像データの各局所特徴量がどのコードブックベクトルから近いか判定し，k 個の要素からなるヒストグラムを作成する．$photo_i$ の各局所特徴量 $PLF_i = \{plf_{i,1}, plf_{i,2}, \cdots, plf_{i,numplf(i)}\}$ を次の式により分類する．

$$cplf_{i,j} = \arg\min_c dist(plf_{i,j}, cbv_c) \tag{20}$$

この式では，局所特徴量 $plf_{i,j}$ が最も近いコードブックベクトルの番号を返している．ここで，$photo_i$ のヒストグラムを $phist_i$ とし，$phist_i$ の各要素を $phist_i[j]$

（$1 \leq j \leq k$）としたとき，$photo_i$ のヒストグラムは次の式で求められる．

$$phist_i[j] = \sum_{l=1}^{numplf(i)} f(cplf_{i,l}, j),$$
$$f(c,j) = \begin{cases} 1 & (if\ c = j) \\ 0 & (otherwise) \end{cases} \tag{21}$$

つまり，ヒストグラムは各コードブックベクトルに対して類似する局所特徴量を何個もっているかを示している．そして，抽出したヒストグラムを画像データ $photo_i$ の特徴ベクトル pfv_i として使用する．

6.4.2.2　学習済み深層ネットワークを用いた特徴ベクトル抽出

Twitter に投稿される画像データはバリエーションが多く，決まった種類の画像データが存在しないために，BoF では画像データの特徴を十分に捉えることができない可能性がある．そこで，大規模画像データによって学習させた学習済みの畳み込みニューラルネットワーク（CNN）を特徴ベクトル抽出に使用する．

CNN は数多く提案されている深層ネットワークの中でも，特に画像認識の分野に応用されているニューラルネットワークである．CNN は中間層に畳み込み層とプーリング層が存在し，画像データの局所的な特徴を自動的に学習することができる．提案手法では，CNN モデルである VGG-16[67] を用いて特徴ベクトル抽出を行う．VGG-16 のネットワークの学習は，大規模画像認識コンペティションの ILSVRC[69] にて提供された ImageNet と呼ばれる 1000 分類，120 万枚の画像を用いて行われている．ImageNet は一般的な内容の画像データを含んでおり，ImageNet を用いて学習したネットワークは画像データの様々な分析に応用可能な汎用知識を学習できているといわれている．

VGG-16 は 16 層で形成されており，畳み込み層，プーリング層と全結合層から成る．畳み込み層では，入力に対して重みフィルタの内積を計算する．各畳み込み層は前層の出力に対して畳み込み処理を行い，次の層の入力となる特徴マップを出力する．プーリング層では，畳み込み層から出力された特徴マップを縮小する．VGG-16 では，最大値プーリングを用いている．全結合層では，重み付き結合を計算し，活性化関数によりユニットの値を求める．VGG-16 では活性化関数として，ReLU を用いている．

提案手法では，ネットワークモデルの出力層手前の全結合層から特徴ベクトルを抽出する．VGG-16 の出力層手前の全結合層のユニット数は 4,096 であるため，4,096 次元の特徴ベクトルが抽出される．VGG-16 の学習に用いられている ImageNet には気象や自然災害に関する分類を含む画像データは無い．しかしながら，出力層手前の中間層には画像データの汎用的な特徴が表れるため，ジオソーシャル画像データを区別する特徴ベクトルとして利用できると考える．

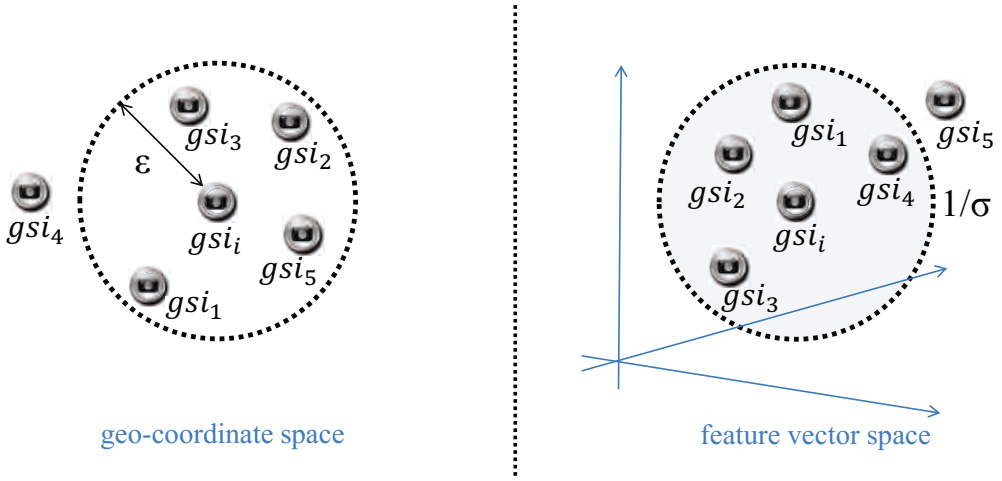


図 37: (ϵ, σ)-密度に基づくマルチモーダル近傍の例

6.4.3　諸定義

本項では，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングの諸定義を説明する．

定義 16　((ϵ, σ)-密度に基づくマルチモーダル近傍) ジオソーシャル画像データ gsi_i の (ϵ, σ)-密度に基づくマルチモーダル近傍を $N_{(\epsilon,\sigma)}(gsi_i)$ と表記し，次のように定義する．

$$N_{(\epsilon,\sigma)}(gsi_i) = \{gsi_j \in GSI | dist(gsi_i, gsi_j) \leq \epsilon \ and \ msim(gsi_i, gsi_j) \geq \sigma\}, \quad (22)$$

関数 $dist$ は緯度・経度など座標値を使って，ジオソーシャル画像データ gsi_i と gsi_j 間の空間上の距離を求める関数であり，本研究では Lambert-Andoyer の公式を用いて計算した距離を返す．

図 37 に (ϵ, σ)-密度に基づくマルチモーダル近傍の例を示す．この例では，gsi_i のマルチモーダル近傍は，$N_{(\epsilon,\sigma)}(gsi_i) = \{gsi_1, gsi_2, gsi_3\}$ である．gsi_4 は，gsi_i との類似度が σ 以上であるが，距離が ϵ よりも離れているため，$N_{(\epsilon,\sigma)}(gsi_i)$ には含まれていない．gsi_5 は，gsi_i から ϵ 以内に存在しているが，gsi_i との類似度が σ よりも小さいため，$N_{(\epsilon,\sigma)}(gsi_i)$ には含まれていない．

定義 17（空間上における適応的な閾値） ジオソーシャル画像データ gsi_i が存在する地域の空間投稿密度を $lsd(gsi_i)$ と表記する．ジオソーシャル画像データ gsi_i に対する空間上における適応的な閾値 SAT を次のように定義する．

$$SAT(gsi_i, MinGSI) = (MinGSI - 1) \times lsd(gsi_i) + 1. \quad (23)$$

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，対象とする空間領域を 2 次元の空間グリッドに分割（$div_{lng} \times div_{lat}$）する．次に，各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．グリッド k 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 $numsg(k)$ とし，関数 $geo\_gid(gsi_i)$ をジオソーシャル画像データ gsi_i が位置するグリッドの ID を求める関数とすると，gsi_i の空間投稿密度 $lsd(gsi_i)$ は次の式で定義される．

$$lsd(gsi_i) = \frac{numsg(geo\_gid(gsi_i)) - numsg_{min}}{numsg_{max} - numsg_{min}}. \tag{24}$$

ただし，$numsg_{min}$ は最も投稿数が少ないグリッドに含まれるジオソーシャル画像データ数であり，$numsg_{max}$ は最も投稿数の多いグリッドに含まれるジオソーシャル画像データ数である．

空間投稿密度 $lsd(gsi_i)$ を用いることによって，過去の投稿数が少ない，つまり低密度な地域では，適応的な閾値 SAT が小さくなる．過去の投稿数が多い，つまり高密度な地域では，SAT が大きくなり，各地域の投稿数に応じたマルチモーダル空間クラスタの抽出が可能となる．

定義 18　(核ジオソーシャル画像データ，周辺ジオソーシャル画像データ) もし，ジオソーシャル画像データ gsi_i が，$|N_{(\epsilon,\sigma)}(gsi_i)| \geq SAT(gsi_i, MinGSI)$ を満たすなら，gsi_i を核ジオソーシャル画像データと呼ぶ．そうでなければ，周辺ジオソーシャル画像データと呼ぶ．

図 37 において，$MinGSI = 4$，$lsd(gsi_i) = 0.8$ とする．このとき，$|N_{(\epsilon,\sigma)}(gsi_i)| < SAT(gsi_i, MinGSI)(SAT(MinGSI, lsd(gsi_i)) = 3.4)$ となるため，gsi_i は周辺ジオソーシャル画像データとなる．$MinGSI = 4$，$lsd(gsi_i) = 0.2$ とすると，$|N_{(\epsilon,\sigma)}(gsi_i)| \geq SAT(gsi_i, MinGSI)(SAT(MinGSI, lsd(gsi_i)) = 1.6)$ となり，gsi_i は核ジオソーシャル画像データとなる．

定義 19　((ϵ,σ)-密度に基づいて直接到達可能) ジオソーシャル画像データ gsi_j がジオソーシャル画像データ gsi_i の (ϵ,σ)-密度に基づくマルチモーダル近傍に存在し，$|N_{(\epsilon,\sigma)}(gsi_i)| \geq SAT(gsi_i, MinGSI)$ を満たす時，gsi_j は gsi_i から (ϵ,σ)-密度に基づいて直接到達可能であると表現する．

定義 20　((ϵ,σ)-密度に基づいて到達可能) ジオソーシャル画像データ gsi_{i+1} がジオソーシャル画像データ gsi_i から (ϵ,σ)-密度に基づいて直接到達可能である，ジオソーシャル画像データ列 $(gsi_i, gsi_{(i+1)}, \cdots, gsi_{i+n})$ を考える．このとき，$gsi_{(i+n)}$ は gsi_i から，(ϵ,σ)-密度に基づいて到達可能であると表現する．

定義 21　((ϵ, σ)-密度に基づいて接続) ジオソーシャル画像データ gsi_i とジオソーシャル画像データ gsi_j とが，ある任意のジオソーシャル画像データ gsi_o と (ϵ, σ)-密度に基づいて到達可能であり，gsi_o が $|N_{(\epsilon,\sigma)}(gsi_o)| \geq SAT(gsi_o, MinGSI)$ を満たす時，gsi_i と gsi_j は (ϵ, σ)-密度に基づいて接続していると表現する．

6.4.4　(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ

(ϵ, σ)-密度に基づくマルチモーダル空間クラスタを次のように定義する．

定義 22　((ϵ, σ)-密度に基づくマルチモーダル空間クラスタ) ジオソーシャル画像データ集合 GSI において，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc は以下の 2 つの条件を満たす部分ジオソーシャル画像データ集合である．

（1）　任意のジオソーシャル画像データ $gsi_i \in GSI$ と $gsi_j \in GSI$ について，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc に gsi_i が所属（$gsi_i \in msc$）し，gsi_j が gsi_i から (ϵ, σ)-密度に基づいて到達可能であれば，gsi_j は (ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc に所属（$gsi_j \in msc$）する．

（2）　(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc に所属する任意のジオソーシャル画像データ $gsi_i \in msc$ と $gsi_j \in msc$ は，(ϵ, σ)-密度に基づいて接続している．

(ϵ, σ)-密度に基づくマルチモーダル空間クラスタは，距離だけでなく画像とテキストの内容も類似したジオソーシャル画像データから構成される．つまり，複数の類似したジオソーシャル画像データが近くに集まっていることを示し，抽出された場所で注目されているトピックを含んでいる．

6.4.5　アルゴリズム

Algorithm6 に (ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングのアルゴリズムを示す．ジオソーシャル画像データ集合 GSI において，核ジオソーシャル画像データを見つけ，マルチモーダル空間クラスタ msc の核とする．そして，(ϵ, σ)-密度に基づいて直接到達可能なデータを再帰的に msc に加えていくことで，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタを抽出する．

Algorithm6 は，ジオソーシャル画像データ集合 GSI，パラメータ ϵ，σ と $MinGSI$ を入力として，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ集合 MSC を出力する．Algorithm6 の内容を詳しく説明する．

（1）ジオソーシャル画像データ集合 GSI からジオソーシャル画像データ gsi_p を 1 つ取

Algorithm 6 (ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングアルゴリズム

Input: GSI, ϵ, σ, $MinGSI$

Output: MSC

1: $MSC \leftarrow \phi$
2: **for** $i \leftarrow 1$ to $|GSI|$ **do**
3: $gsi_p \leftarrow gsi_i \in GSI$
4: **if** IsClustered(gsi_p) $== false$ **then**
5: $N_{(\epsilon,\sigma)} \leftarrow$ GetNeighborhood(gsi_p, ϵ, σ)
6: $lst \leftarrow$ GetSDens(gsi_p)
7: **if** $|N_{(\epsilon,\sigma)}| \geq SAT(gsi_p, MinGSI)$ **then**
8: $msc \leftarrow$MakeNewCluster(gsi_p)
9: $Q \leftarrow \phi$
10: EnQueue(Q, $N_{(\epsilon,\sigma)}$)
11: **while** Q is not empty **do**
12: $gsi_q \leftarrow$ DeQueue(Q)
13: $N_{(\epsilon,\sigma)} \leftarrow$ GetNeighborhood(gsi_q, ϵ, σ)
14: $lst \leftarrow$ GetSDens(gsi_q)
15: **if** $|N_{(\epsilon,\sigma)}| \geq SAT(gsi_p, MinGSI)$ **then**
16: EnUniqueQueue(Q, $N_{(\epsilon,\sigma)}$)
17: **end if**
18: $msc \leftarrow msc \cup gsi_q$
19: **end while**
20: $MSC \leftarrow MSC \cup msc$
21: **end if**
22: **end if**
23: **end for**
24: return MSC

り出す．ただし，GSI が空ならば，（8）へ進む．

（2）関数 IsClustered を用いて，gsi_p が (ϵ, σ)-密度に基づくマルチモーダル空間クラスタに所属しているかチェックする．(ϵ, σ)-密度に基づくマルチモーダル空間クラスタに所属していなければ，関数 GetNeighborhood を用いて gsi_p の (ϵ, σ)-密度に基づくマルチモーダル近傍を取得する．

（3）関数 GetSDens を用いて gsi_p の空間投稿密度を取得し，gsi_p が核ジオソーシャル画像データであるか確認する．gsi_p が核ジオソーシャル画像データであれば，（4）へ進

む．gsi_p が核ジオソーシャルデータでなければ，（1）へ戻る．

（4）関数 MakeNewCluster を用いて新たに (ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc を作成する．

（5）ここで，gsi_p の (ϵ, σ)-密度に基づくマルチモーダル近傍に存在するジオソーシャル画像データ集合を関数 EnQueue を用いてキュー Q に挿入する．

（6）キュー Q からジオソーシャル画像データ gsi_q を取り出し，次の処理を行う．

（a）gsi_q の (ϵ, σ)-密度に基づくマルチモーダル近傍と空間投稿密度を取得する．もし，gsi_q が核ジオソーシャル画像データであれば，gsi_q の (ϵ, σ)-密度に基づくマルチモーダル近傍を関数 EnUniqueQueue を用いてキュー Q に挿入する．挿入では，Q に存在せず，他の (ϵ, σ)-密度に基づくマルチモーダル空間クラスタに所属していないジオソーシャル画像データのみを Q に挿入する．

（b）gsi_q を (ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc に挿入する．キュー Q が空であれば，（7）へ進む．空でなければ（6）へ戻る．

（7）(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ集合 MSC に msc を加え，（1）に戻る．

（8）MSC を出力する．

6.5　ネットワークベースの重要度算出手法

本節では，ネットワークベースの重要度算出手法について説明する．図 38 にネットワークベースの重要度算出手法の概要を示す．ネットワークベースの重要度算出手法は類似度グラフ SG の作成と，ノードの重要度の算出の二段階によって行う．

（1）(ϵ, σ)-密度に基づくマルチモーダル空間クラスタに対して類似度グラフ SG を作成する．(ϵ, σ)-密度に基づくマルチモーダル空間クラスタ msc に所属するジオソーシャル画像データ集合を $msc = \{gsi_1, gsi_2, \cdots, gsi_{|msc|})\}$ とする．類似度グラフ $SG = (V, E)$ は，ノード集合 V と辺集合 E から構成される．ノード $v_i \in V$ はジオソーシャル画像データ gsi_i に対応し，辺 $e = (j, k) \in E$ はノード v_j とノード v_k 間に辺が存在すること示す．ここで，

$$E = \{(j, k) \mid v_j \in V, v_k \in V, psim_{j,k} \geq \alpha\} \tag{25}$$

と定義する．α はパラメータであり，類似度が高いノード間の辺のみに制限することができる．

（2）媒介中心性 [100] を用いて各ノードの重要度を算出する．媒介中心性とは，ノードがどれくらいネットワーク上で重要な媒介を行っているかを示し，通常，ノード間の最

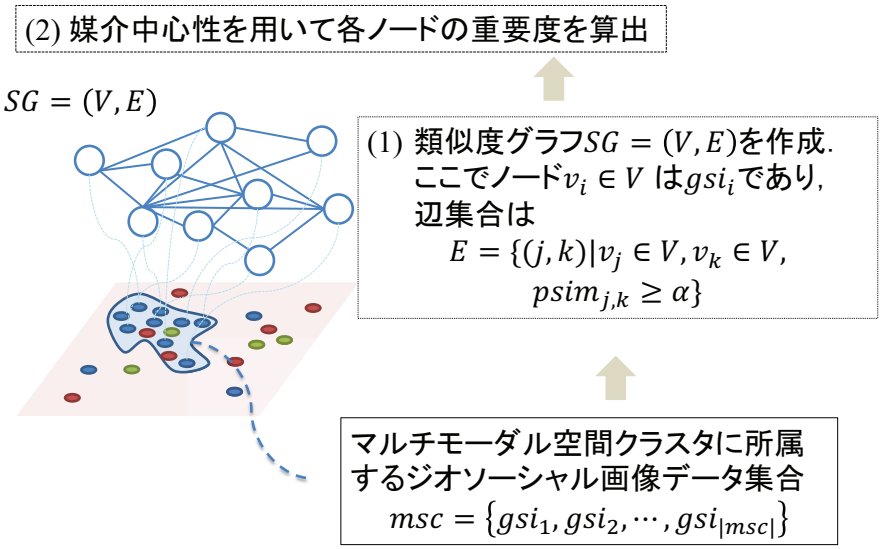


図 38: ネットワークベースの重要度算出手法

短路が何本通っているかで算出され，v_i の重要度 $BC(v_i)$ は以下の式で表される．

$$BC(v_i) = \sum_{s \neq t \neq v_i} \frac{path_{s,t}(v_i)}{path_{s,t}} \tag{26}$$

ここで，$path_{s,t}(v_i)$ はノード s とノード t 間の経路でノード v_i を通る経路の数，$path_{s,t}$ はノード s とノード t 間の経路の総数である．各ノードの $BC(v_i)$ の値を計算し，$BC(v_i)$ の値を，そのノードが示すジオソーシャル画像データの重要度とする．

6.6　評価実験

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

6.6.1　実験内容

評価実験では，Twitter 上に投稿されたジオタグ付きで画像 URL を持つツイートをジオソーシャル画像データとして扱い実験を行う．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用いて，ジオソーシャル画像データの集合 GSI を構築した．実験では，京都府庁舎から半径 30km 以内のデータ 11,189 件を用いて，京都周辺のトピックを取り出した．

(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングのパラメータは，$\epsilon = 500m$，$\sigma = 0.6$，$MinGSI = 5$ を用いた．ネットワークベースの重要度算出手法のパラメータは，$\alpha = 0.5$ を用いた．マルチモーダル類似度 $msim$ のパラメータとしては，$w = 1.0$（WDBMSC と表記する）と $w = 0.5$ を用いた場合の比較を行う．$w = 0.5$ を用いた場合については，画像データの特徴ベクトル抽出手法として BoF[19] を用いた場合（$\mathrm{MDBMSC_{BoF}}$ と表記する）と学習済み深層ネットワークである VGG-16[67] を用いた場合（$\mathrm{MDBMSC_{VGG}}$ と表記する）の比較を行う．WDBMSC は画像データ間の類似度を考慮していない語句に基づく手法となり，$\mathrm{MDBMSC_{BoF}}$ と $\mathrm{MDBMSC_{VGG}}$ は語句と画像データに基づく手法となる．空間投稿密度の算出に用いる統計データは 3,301,605 件のジオタグ付きツイートを用いた．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) からなる矩形を対象に，空間投稿密度を求めた．パラメータは $div_{lng} = 1,000$，$div_{lat} = 1,000$ を用いた．また，形態素解析には MeCab を用いた．

6.6.2　実験結果

表 18，19 と 20 に，WDBMSC，$\mathrm{MDBMSC_{BoF}}$ と $\mathrm{MDBMSC_{VGG}}$ のトピック抽出結果をそれぞれ示す．表 18，19 と 20 には，ツイート数の上位 5 件のクラスタを示しており，各クラスタのツイート数，緯度，経度と頻出語句上位 5 件を示している．WDBMSC では 98 クラスタ，$\mathrm{MDBMSC_{BoF}}$ では 89 クラスタ，$\mathrm{MDBMSC_{VGG}}$ では 83 クラスタが抽出された．各手法において，1 位のクラスタは「京都駅」と「京都タワー」，2 位のクラスタは「清水寺」に関するトピックを含むクラスタとなった．1 位のクラスタのツイート数を見ると，WDBMSC は 217 ツイート，$\mathrm{MDBMSC_{BoF}}$ は 220 ツイート，$\mathrm{MDBMSC_{VGG}}$ は 252 ツイートとなっており，$\mathrm{MDBMSC_{VGG}}$ の場合はクラスタのツイート数が多くなっている．$\mathrm{MDBMSC_{VGG}}$ では，「京都駅」と「京都タワー」に関するトピックだけでなく京都駅周辺の飲食店や観光名所が同じクラスタとして抽出されている．2 位の「清水寺」のクラスタは三手法に大きな違いはなかった．

WDBMSC の 3 位のクラスタは京都駅付近で投稿されたツイートから構成されている．このクラスタのツイートの内容を確認したところ，注目されているようなトピックは含まれていなかった．また，ツイートの本文に投稿した場所の住所が書かれており，全てのツイートに「東塩小路高倉町」が含まれていた．WDBMSC では，付与されている画像データの内容を考慮していないために，このような同じ住所が書かれたテキスト同士が類似して，トピックが含まれないクラスタが抽出された．また，WDBMSC の 4 位は「祇園」，「三条大橋」と「恵美須神社」，5 位は「南禅寺」と「平安神宮」に関するクラスタとなっており，それぞれ内容の異なるトピックが同じクラスタとして抽出された．異なるトピックが同じクラスタとして抽出された理由としては，クラスタ中の多くのツイートに「なう」が含まれてお

表 18: WDBMSC のクラスタリング結果

ID	ツイート数	緯度	経度	頻出語句上位 5 件
1	217	34.9819–34.9937	135.7500–135.7694	京都，駅，なう，タワー，新幹線
2	100	34.9936–34.9984	135.7758–135.7857	清水寺，清水，東山，京都，舞台
3	48	34.9829–34.9917	135.7555–135.7624	京都，下京，駅，東塩小路高倉，東塩小路
4	45	34.9997–35.0130	135.7657–135.7772	なう，京都，祇園，三条大橋，恵美須神社
5	37	34.0091–35.0161	135.7773–135.7944	南禅寺，なう，平安神宮，三門，紅葉

表 19: $\mathrm{MDBMSC_{BoF}}$ のクラスタリング結果

ID	ツイート数	緯度	経度	頻出語句上位 5 件
1	220	34.9809–34.9908	135.7492–135.7643	京都，駅，なう，タワー，新幹線
2	108	34.9936–34.9984	135.7758–135.7857	清水寺，清水，東山，京都，舞台
3	84	35.0091–35.0161	135.7773–135.7955	南禅寺，永観堂，京都，左京，なう
4	55	35.0103–35.0174	135.6725–135.6815	寺，天龍，京都，右京，嵯峨
5	51	35.0112–35.0155	135.6725–135.6862	渡月，橋，嵐山，なう，紅葉

表 20: $\mathrm{MDBMSC_{VGG}}$ のクラスタリング結果

ID	ツイート数	緯度	経度	頻出語句上位 5 件
1	252	34.9819–34.9897	135.7500–135.7643	京都，駅，ない，下京，新幹線
2	100	34.9936–34.9984	135.7766–135.7857	清水寺，清水，東山，京都，舞台
3	48	35.0112–35.0142	135.6773–135.6813	月橋，渡，京都，嵯峨，右京
4	42	35.0363–35.0417	135.7270–135.7333	金閣寺，鹿苑寺，京都，なう，綺麗
5	28	34.9998–35.0044	135.7724–135.7829	京都，東山，八坂神社，祇園北側，祇園南側

り，異なるトピックを表すツイートでも，テキスト同士が類似してしまったためである．以上の結果より，WDBMSC は $\mathrm{MDBMSC_{BoF}}$ と $\mathrm{MDBMSC_{VGG}}$ に比べて，画像データ間の類似度を考慮していないため，正確にジオソーシャル画像データ間の類似度を計算することができなかった．

$\mathrm{MDBMSC_{BoF}}$ のクラスタ 3 は，「南禅寺」や「平安神宮」など様々なトピックが含まれる広い範囲で投稿されたツイートから構成されるクラスタとなった．$\mathrm{MDBMSC_{VGG}}$ では，「南禅寺」と「平安神宮」に関するトピックはそれぞれ独立したクラスタとして抽出されていた．VGG-16 を用いて抽出した特徴ベクトルは BoF を用いて抽出した特徴ベクトルよりも，より正確にジオソーシャル画像データ間の類似度を算出できたといえる．

その他のクラスタは，$\mathrm{MDBMSC_{BoF}}$ の 4 位は「天龍寺」，5 位は「渡月橋」，$\mathrm{MDBMSC_{VGG}}$ の 3 位は「渡月橋」，4 位は「金閣寺」，5 位は「八坂神社」に関するクラスタとなっており，それぞれ京都で注目されているトピックを抽出することができた．

$\mathrm{MDBMSC_{VGG}}$ で抽出された表 20 の各クラスタから，代表画像データを抽出した．また，

重要度の上位 5 件をクラスタの代表画像データとした．表 20 の 1 位のクラスタでは，「京都駅」の構内や看板が代表画像データとして抽出された．2 位のクラスタでは，「清水寺」の境内や「清水寺」周辺の紅葉などが抽出され，クラスタの内容に適している画像データを抽出することができた．3 位のクラスタでは「渡月橋」，4 位のクラスタでは「金閣寺」，5 位のクラスタでは「八坂神社」で撮影された写真が代表画像データとして抽出されており，クラスタの内容を自動的に抽出することができた．

6.7　まとめ

本章では，ジオソーシャル画像データから各地域で注目されているトピックを抽出するために，密度に基づくマルチモーダル空間クラスタリングを用いたトピックの抽出手法を提案した．提案手法は，最初に，(ϵ, σ)-密度に基づくマルチモーダル空間クラスタリングを用いて，注目領域をマルチモーダル空間クラスタとして抽出する．次に，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出し分かりやすくするために，代表画像データを抽出する．Twitter 上に投稿されたジオタグ付きで画像 URL を持つツイートを用いて評価実験を行った．評価実験の結果，京都の「清水寺」，「渡月橋」や「金閣寺」などのトピックをマルチモーダル空間クラスタとして抽出することができた．また，各マルチモーダル空間クラスタから代表画像データを抽出することで，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出することができた．

本研究の今後の課題としては，マルチモーダル類似度の算出方法の改善，ユーザ情報や時間情報も加えてより詳細なトピック抽出を行えるようにすること，抽出したトピックを閲覧するための可視化システムの開発が挙げられる．具体的には，マルチモーダル類似度を求める際に，テキストデータ間の類似度と画像データ間の類似度を別々に算出しているため，それらを統一的に特徴ベクトル化し，類似度を算出する必要がある．

第 7 章　結論

7.1　本論文のまとめ

ソーシャルメディア上の位置情報付きのデータには個人的な話題だけでなく，ユーザが目にした事象や話題を含んでおり，位置情報付きのデータから実世界のトピックの分析や抽出を行うことは重要な研究課題の一つである．そこで，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対象にして，実世界で注目されているトピックの分析を行う研究が盛んに行われている．その多くはデータに付与された時間情報と位置情報に着目し，データが盛んに投稿されている時間帯または領域には何かしらの有益な知識があるという考えに基づいている．しかしながら，これらの情報とともに投稿されるテキストと画像データの内容を考慮した時空間データマイニング手法は確立しているとはいえない．

本論文では，ソーシャルメディア上に投稿される時間情報と位置情報が付与されたテキストと画像データに対する時空間データマイニング手法の確立を目指し，以下の五つの目的を達成した．

（1）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

Twitter 上に投稿されるジオタグ付きツイートを用いてトピックを時空間分析するための手法，密度に基づく時空間分析手法を提案した．提案手法は，対象となっているトピックの内容を含むジオタグ付きツイートを抽出するために，ナイーブベイズ分類器を用いてジオタグ付きツイートを分類する．次に，(ϵ, τ)-密度に基づく時空間クラスタリングのインクリメンタルなアルゴリズムを用いて，トピックが注目されている地域を時空間クラスタとしてリアルタイムに抽出する．そして，抽出された時空間クラスタについてその領域，ツイート内容と画像データを Web アプリケーション上に提示する．

実際に Twitter 上からジオタグ付きツイートを収集し，トピックを「大雨」と「大雪」と設定して評価実験を行った．評価実験の結果，ツイート分類の交差検定における F 値として，トピック「大雨」では 0.78，トピック「大雪」では 0.81 を示した．また，トピックが注目されている地域を検出できたか評価を行った結果，検出率としてトピック「大雨」では 0.52，トピック「大雪」では 0.66 を示した．そして，抽出された時空間クラスタを Web アプリケーション上で確認することによって，本研究が目的とするトピックの時空間分析が可能であることを確認できた．

（2）密度に基づく適応的な時空間クラスタリング

密度に基づく時空間分析手法において，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない時間帯を区別することなく時空間クラスタを抽出するため

に，(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．提案手法は，各地域，各時間帯における統計的な投稿数を用いて，時空間クラスタを抽出する基準となる閾値を適応的に変化させている．よって，投稿数が多い地域と少ない地域，また投稿数が多い時間帯と少ない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．

提案手法を密度に基づく時空間分析手法に導入し，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，トピック「大雨」について，既存手法は閾値を変化させた場合，検出率が 0.73 から 0.32 まで落ちるのに対して，提案手法は閾値を変化させたとしても 0.80 から大きく変化することなく高い検出率を示すことができた．また，トピック「大雪」についても，提案手法は既存手法と比較して，高い検出率を示すことができた．

（3）密度に基づく時空間分析手法における画像分類

密度に基づく時空間分析手法において，対象となっているトピックに関連している画像データのみを抽出するための画像分類を提案した．提案手法は，BoF または学習済み深層ネットワークを用いて画像データから特徴ベクトルを抽出する．次に，抽出した画像データの特徴ベクトルを使用して SVM を学習させ，当該トピックに関連する画像データかどうか分類する．そして，当該トピックに関連する画像データのみを Web アプリケーション上に提示することで，密度に基づく時空間分析手法の有効性を向上することができる．

提案手法を実際に密度に基づく時空間分析手法に導入し，評価実験を行った．トピックを「大雨」と「大雪」と設定して行った評価実験の結果，提案手法は特徴ベクトル抽出手法として学習済み深層ネットワークである VGG-16 を用いた場合，交差検定における正解率として，トピック「大雨」では 0.89，トピック「大雪」では 0.98 を示し，高性能にトピック「大雨」と「大雪」に関連する画像データを分類することができた．

（4）最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN

DBSCAN の高速化のために，最小外接矩形（MBR）とセルの再帰分割を用いたセルベースの DBSCAN を提案した．提案手法では，セルベースの DBSCAN のセルの結合判定について，セル中のデータを囲む MBR を作成し，MBR 間の距離を用いることで，条件を満たす場合に高速に判定することができる．また，セルを再帰的に分割し，計算の対象となるデータを減らしていくことで，高速にセルの結合判定ができる．

人工データと実データを用いて評価実験を行った．評価実験の結果より，人工データを用いた場合，低次元のデータでは既存手法と比較して大幅な高速化はできなかったが，高次元のデータになるほど大幅な高速化ができた．実データを用いた場合，

データが空間全体にまばらに広がっている分布であることから，高速化できなかったデータセットもあった．しかしながら，投機的に総当りを行う方法を用いることで全てのデータセットで高速化ができた．

（5）密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出

ジオソーシャル画像データから各地域で注目されているトピックを抽出するために，密度に基づくマルチモーダル空間クラスタリングを用いたトピックの抽出手法を提案した．提案手法は，(ϵ, σ)-密度に基づく空間マルチモーダルクラスタリングを用いて，空間的また内容的にも類似したジオソーシャル画像データが密に投稿されている注目領域をマルチモーダル空間クラスタとして抽出する．また，ジオソーシャル画像データ間の類似度を正確に算出するために，マルチモーダル類似度を定めている．そして，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出し分かりやすくするために，ネットワークベースの重要度算出手法を用いて，代表画像データを抽出する．

Twitter 上に投稿されたジオタグ付きで画像 URL を持つツイートを用いて評価実験を行った．評価実験の結果，京都の「清水寺」,「渡月橋」や「金閣寺」などのトピックをマルチモーダル空間クラスタとして抽出することができた．また，各マルチモーダル空間クラスタから代表画像データを抽出することで，マルチモーダル空間クラスタに含まれるトピックを自動的に抽出することができた．

これらの研究成果は，ソーシャルメディアに対する時空間データマイニングの基盤となる重要な技術であり，本研究の目的とする時空間データマイニング手法の確立を達成できたといえる．

7.2　今後の課題

今後の課題は以下の通りである．

（1）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の課題

ツイート分類において深層学習を用いた新しい分類手法を開発し，分類精度を向上させること，ソーシャルメディア上のデータと気温や降雨量などの気象観測データを組み合わせた手法を開発すること，トピックが注目されている地域の発生，その変化と消滅を捉えることができたか定量的な評価を行うことが挙げられる．

（2）密度に基づく適応的な時空間クラスタリングの課題

ユーザが設定する ϵ や τ などのパラメータを適応的または自動的に設定する手法を導入すること，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挙げられる．

（3）密度に基づく時空間分析手法における画像分類の課題

分類性能の向上のために，学習済み深層ネットワークを再学習させて新しいモデルを作成し，特徴ベクトル抽出に用いることが挙げられる．学習済みの深層ネットワークを初期値とし，特定のトピックに関する画像データを用いて再学習することで，汎用性があり，また特定のトピックに適した深層ネットワークモデルの作成が期待できる．

（4）最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN の課題

DBSCAN の更なる高速化のために，投機的な総当り数の自動決定方法と提案手法の並列化を行うことが挙げられる．また，ユークリッド距離以外の距離計算方法を用いる密度に基づくクラスタリングに対応することが挙げられる．

（5）密度に基づくマルチモーダル空間クラスタリングを用いたトピック抽出の課題

マルチモーダル類似度の算出方法について，テキストと画像データを統一的に特徴ベクトル化し類似度を算出すること，抽出したトピックを閲覧するための可視化システムを開発すること，ユーザ情報や時間情報も加えて，より詳細なトピック抽出を行えるようにすることが挙げられる．

謝辞

本研究を行うに当たり，格別なる御指導ならびに御鞭撻を賜りました北上始名誉教授，竹澤寿幸教授，田村慶一准教授に深甚なる感謝の意を表します．

松原行宏教授には，お忙しい中，博士論文の審査をお引き受けいただき，研究に関する御助言をいただきました．厚く御礼申し上げます．

黒木進准教授，森康真助教には，日頃から御助言と御指導をいただきました．厚く御礼申し上げます．

本論文をまとめるに当たって御協力いただいたデータ工学研究室の皆様に厚く御礼申し上げます．

参考文献

[1] John Chon and Hojung Cha. LifeMap: A smartphone-based context provider for location-based services. *IEEE Pervasive Computing*, Vol. 10, No. 2, pp. 58–67, 2011.

[2] Mor Naaman. Geographic information from georeferenced social media data. *SIGSPATIAL Special*, Vol. 3, No. 2, pp. 54–61, 2011.

[3] Andrea Kavanaugh, Edward Fox, Steven Sheetz, Seungwon Yang, Lin Li, Travis Whalen, Donald Shoemaker, Paul Natsev, and Lexing Xie. Social media use by government: From the routine to the critical. *Government Information Quarterly*, Vol. 29, No. 4, pp. 480 – 491, 2012.

[4] Mashaal Musleh. Spatio-temporal visual analysis for event-specific tweets. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1611–1612, 2014.

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

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

[7] Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the Second Workshop on Language in Social Media*, pp. 27–36, 2012.

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

[9] Yan-Tao Zheng, Zheng-Jun Zha, and Tat-Seng Chua. Research and applications on georeferenced multimedia: a survey. *Multimedia Tools and Applications*, Vol. 51, No. 1, pp. 77–98, 2011.

[10] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, Vol. 27, No. 6, pp. 52–59, 2012.

[11] Slava Kisilevich, Florian Mansmann, and Daniel Keim. P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In *Proceedings of the 1st International Conference and Exhibition on*

Computing for Geospatial Research & Application, pp. 38:1–38:4, 2010.

[12] Keiichi Tamura and Takumi 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*, pp. 2079–2084, 2013.

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

[14] 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.

[15] Ade Gunawan. A faster algorithm for DBSCAN. Master's thesis, Technische University of Eindhoven, 40 pages, 2013.

[16] Junhao Gan and Yufei Tao. DBSCAN revisited: Mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 519–530, 2015.

[17] Ling Chen and Abhishek Roy. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 523–532, 2009.

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

[19] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray. Visual categorization with bags of keypoints. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, pp. 1–22, 2004.

[20] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter Power: Tweets As Electronic Word of Mouth. *J. Am. Soc. Inf. Sci. Technol.*, Vol. 60, No. 11, pp. 2169–2188, 2009.

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

[22] Steven Van Canneyt, Olivier Van Laere, Steven Schockaert, and Bart Dhoedt. Using Social Media to Find Places of Interest: A Case Study. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pp. 2–8, 2012.

[23] 山本修平, 佐藤哲司. トピックと局面の対応関係に基づく実生活ツイートのマルチラベ

ル分類. 情報処理学会論文誌データベース（TOD）, Vol. 7, No. 2, pp. 24–36, 2014.

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

[25] 新田直子, 角谷直人, 馬場口登. 単語間の関係性の経時変化を考慮したマイクロブログからの実世界観測情報の抽出. 日本データベース学会和文論文誌, Vol. 13, No. 1, pp. 13–18, 2014.

[26] Brendan C. Fruin, Hanan Samet, and Jagan Sankaranarayanan. TweetPhoto: Photos from news tweets. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 582–585, 2012.

[27] Philip J. McParlane, Andrew James McMinn, and Joemon M. Jose. "Picture the scene...";: Visually summarising social MediaEvents. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pp. 1459–1468, 2014.

[28] Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris, and Pericles A. Mitkas. Visual event summarization on social media using topicmodelling and graph-based ranking algorithms. In *Proceedings of the 5th ACM on International Conference onMultimedia Retrieval*, pp. 203–210, 2015.

[29] Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, pp. 47–56, 2008.

[30] Cheng-Fa Tsai and Chien-Tsung Wu. GF-DBSCAN: A new efficient and effective data clustering technique for large databases. In *Proceedings of the 9th WSEAS International Conference on Multimedia Systems & Signal Processing*, pp. 231–236, 2009.

[31] Motohiro Shirai, Masaharu Hirota, Shohei Yokoyama, Naoki Fukuta, and Hiroshi Ishikawa. Discovering multiple hotspots using geo-tagged photographs. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pp. 490–493, 2012.

[32] Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 1–10, 2010.

[33] Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs. *World Wide Web*, Vol. 14, No. 4, pp. 321–349, 2011.

[34] Takuya Sugitani, Masumi Shirakawa, Takahiro Hara, and Shojiro Nishio. Detecting

local events by analyzing spatiotemporal locality of tweets. In *Proceedings of the Sixth International Symposium on Mining and Web*, pp. 191–196, 3 2013.

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

[36] Shinya Hiruta, Takuro Yonezawa, Marko Jurmu, and Hideyuki Tokuda. Detection, classification and visualization of place-triggered geotagged tweets. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 956–963, 2012.

[37] Ke Xie, Chaolun Xia, Nir Grinberg, Raz Schwartz, and Mor Naaman. Robust detection of hyper-local events from geotagged social media data. In *Proceedings of the Thirteenth International Workshop on Multimedia Data Mining*, pp. 2:1–2:9, 2013.

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

[39] 奥健太, 西崎剛司, 服部文夫. 地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出. 情報処理学会論文誌データベース（TOD）, Vol. 5, No. 3, pp. 97–116, 2012.

[40] Takamu Kaneko and Keiji Yanai. Visual event mining from geo-tweet photos. In *Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops*, pp. 1–6, 2013.

[41] Alexandar Jaffe, Mor Naaman, Tamir Tassa, and Marc Davis. Generating summaries and visualization for large collections of geo-referenced photographs. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 89–98, 2006.

[42] Keiji Yanai, Keita Yaegashi, and Bingyu Qiu. Detecting Cultural Differences using Consumer-generated Geotagged Photos. In *Proceedings of the 2nd International Workshop on Location and the Web*, pp. 12:1–12:4, 2009.

[43] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*, pp. 761–770, 2009.

[44] Ozer Ozdikis, Halit Oguztuzun, and Pinar Karagoz. Evidential location estimation for events detected in twitter. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, pp. 9–16, 2013.

[45] Junjie Yao, Bin Cui, Yuxin Huang, and Xin Jin. Temporal and social context based burst detection from folksonomies. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pp. 1474–1479, 2010.

[46] Tomoki Matsui, Keiichi Tamura, and Hajime Kitakami. Location-based burst detection algorithm for georeferenced document streams based on user's moving direction. In *Proceedings of the 2013 IEEE 6th International Workshop on Computational Intelligence and Applications*, pp. 57–62, 2013.

[47] Keiichi Tamura and Hajime Kitakami. Detecting Location-Based Enumerating Bursts in Georeferenced Micro-Posts. In *Proceedings of the 2013 Second IIAI International Conference on Advanced Applied Informatics*, pp. 389–394, 2013.

[48] Shota Kotozaki, Keiichi Tamura, and Hajime Kitakami. Identifying Burstiness of Local Topic using Location-based Burst Detection with a Classifier Technique. In *Proceedings of the 2014 IEEE 7th International Workshop on Computational Intelligence and Applications*, 2014.

[49] Shota Kotozaki, Keiichi Tamura, and Hajime Kitakami. A New Method for Identifying Location-Based Bursts in a Sequence of Batched Georeferenced Documents. In *Proceedings of the International Conference on INTERNET STUDIES*, 15 pages, 2014.

[50] Shota Kotozaki, Keiichi Tamura, and Hajime Kitakami. Identifying Local Burstiness in a Sequence of Batched Georeferenced Documents. *International Journal of Electronic Commerce Studies*, Vol. 6, No. 2, pp. 269–288, 2015.

[51] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we RT? In *Proceedings of the First Workshop on Social Media Analytics*, pp. 71–79, 2010.

[52] 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 *Proceedings of the 21st International Conference Companion on WWW*, pp. 653–656, 2012.

[53] Karl Kreiner, Aapo Immonen, and Hanna Suominen. Crisis management knowledge from social media. In *Proceedings of the 18th Australasian Document Computing Symposium*, pp. 105–108, 2013.

[54] Sarvnaz Karimi, Jie Yin, and Cecile Paris. Classifying microblogs for disasters. In *Proceedings of the 18th Australasian Document Computing Symposium*, pp. 26–33, 2013.

[55] Myung-Hwa Hwang, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Zhenhua Zhang. Spatiotemporal transformation of social media geostreams: A case study of twitter for flu risk analysis. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, pp. 12–21, 2013.

[56] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and visualizing microblogs for event ex-

ploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227–236, 2011.

[57] Dennis Thom, Harald Bosch, Steffen Koch, Michael Worner, and Thomas Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Proceedings of the 2012 IEEE Pacific Visualization Symposium*, pp. 41–48, 2012.

[58] 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 *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1749–1758, 2014.

[59] Kyoung-Sook Kim, Ryong Lee, and Koji Zettsu. *m*Trend: Discovery of topic movements on geo-microblogging messages. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 529–532, 2011.

[60] Hongzhi Yin, Zhiting Hu, Xiaofang Zhou, Hao Wang, Kai Zheng, Quoc Viet Hung Nguyen, and Shazia Sadiq. Discovering interpretable geo-social communities for user behavior prediction. In *Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering*, pp. 942–953, 2016.

[61] Hongzhi Yin, Xiafang Zhou, Bin Cui, Hao Wang, Kai Zheng, and Quoc Viet Hung Nguyen. Adapting to user interest drift for poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 10, pp. 2566–2581, 2016.

[62] Hongzhi Yin, Bin Cui, Xiaofang Zhou, Weiqing Wang, Zi Huang, and Shazia Sadiq. Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. *ACM Trans. Inf. Syst.*, Vol. 35, No. 2, pp. 11:1–11:44, 2016.

[63] Daoying Ma and Aidong Zhang. An Adaptive Density-based Clustering Algorithm for Spatial Database with Noise. In *Proceedings ofthe Fourth IEEE International Conference on Data Mining*, pp. 467–470, 2004.

[64] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10, 2010.

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

[66] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, Vol. 110, No. 3, pp. 346–359, 2008.

[67] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations 2015*, pp. 1–14, 2015.

[68] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[69] IMAGENET large scale visual recognition challenge (ILSVRC). http://image-net.org/challenges/LSVRC/.

[70] Shaaban Mahran and Khaled Mahar. Using grid for accelerating density-based clustering. In *Proceedings of the 2008 8th IEEE International Conference on Computer and Information Technology*, pp. 35–40, 2008.

[71] Shuigeng Zhou, Aoying Zhou, Jing Cao, Jin Wen, Ye Fan, and Yunfa Hu. Combining sampling technique with DBSCAN algorithm for clustering large spatial databases. In *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pp. 169–172, 2000.

[72] Manoranjan Dash, Huan Liu, and Xiaowei Xu. '1+1>2': Merging distance and density based clustering. In *Proceedings of the Seventh International Conference on Database Systems for Advanced Applications*, pp. 32–39, 2001.

[73] Xin Wang and Howard J. Hamilton. DBRS: A density-based spatial clustering method with random sampling. In *Proceedings of the 7th Pacific-Asia Conference, Advances in Knowledge Discovery and Data Mining*, pp. 563–575, 2003.

[74] B. Borah and D. K. Bhattacharyya. An improved sampling-based DBSCAN for large spatial databases. In *Proceedings of the International Conference on Intelligent Sensing and Information Processing, 2004*, pp. 92–96, 2004.

[75] Bing Liu. A fast density-based clustering algorithm for large databases. In *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, pp. 996–1000, 2006.

[76] Yasser El-Sonbaty, M. A. Ismail, and Mohamed Farouk. An efficient density based clustering algorithm for large databases. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pp. 673–677, 2004.

[77] Pankaj K. Agarwal, Herbert Edelsbrunner, Otfried Schwarzkopf, and Emo Welzl. *Discrete & Computational Geometry*, Vol. 6, No. 3, pp. 407–422, 1991.

[78] Son T. Mai, Ira Assent, and Martin Storgaard. AnyDBC: An efficient anytime density-based clustering algorithm for very large complex datasets. In *Proceedings of the 22nd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1025–1034, 2016.

[79] Antonio Leopoldo Corral Liria. *Algorithms for Processing of Spatial Queries using R-trees. The Closest Pairs Query and its Application on Spatial Databases*. PhD thesis, Department of Languages and Computation, University of Almeria, 67 pages, 2002.

[80] Manik Varma and Andrew Zisserman. Texture classification: Are filter banks necessary? In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 691–698, 2003.

[81] Kvin Bache and Moshe Lichman. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2013.

[82] Haiqin Yang, Shouyuan Chen, Michael R. Lyu, and Irwin King. Location-based topic evolution. In *Proceedings of the 1st international workshop on Mobile location-based service*, pp. 89–98, 2011.

[83] Andrea Kavanaugh, Edward A. Fox, Steven Sheetz, Seungwon Yang, Lin Tzy Li, Travis Whalen, Donald Shoemaker, Paul Natsev, and Lexing Xie. Social media use by government: From the routine to the critical. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pp. 121–130, 2011.

[84] Vanessa Murdock. Your mileage may vary: On the limits of social media. *SIGSPATIAL Special*, Vol. 3, pp. 62–66, 2011.

[85] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 2541–2544, 2011.

[86] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 155–164, 2012.

[87] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 769–778, 2012.

[88] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. EvenTweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, Vol. 6, No. 12, pp. 1326–1329, 2013.

[89] Sayan Unankard, Xue Li, and Mohamed A. Sharaf. Emerging event detection in social networks with location sensitivity. *World Wide Web*, Vol. 18, No. 5, pp. 1393–1417, 2015.

[90] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 103–110, 2007.

[91] Michael 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.

[92] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pp. 56–65, 2007.

[93] Brent J. Hecht and Darren Gergle. On the "Localness" of User-generated Content. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pp. 229–232, 2010.

[94] Sitaram Asur and Bernardo A. Huberman. Predicting the Future with Social Media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, pp. 492–499, 2010.

[95] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pp. 591–600, 2010.

[96] Yuki Arase, Xing Xie, Takahiro Hara, and Shojiro Nishio. Mining people's trips from large scale geo-tagged photos. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 133–142, 2010.

[97] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 37–45, 1998.

[98] Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of the 15th International Conference on Multimedia*, pp. 631–640, 2007.

[99] Rongrong Ji, Yue Gao, Bineng Zhong, Hongxun Yao, and Qi Tian. Mining flickr landmarks by modeling reconstruction sparsity. *ACM Trans. Multimedia Comput. Commun. Appl.*, Vol. 7S, No. 1, pp. 31:1–31:22, 2011.

[100] Linton. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, Vol. 40, No. 1, pp. 35–41, 1977.

発表論文一覧

論文誌

[i] 酒井　達弘，田村　慶一，北上　始，竹澤　寿幸．最小外接矩形とセルの再帰分割を用いたセルベースの DBSCAN の高速化．電子情報通信学会論文誌 D，Vol.J101-D，No.4，pp.690–701，2018 年.

[ii] Keiichi Tamura, Hajime Kitakami, and Tatsuhiro Sakai. Adaptive Distributed Modified Extremal Optimization for Maximizing Contact Map Overlap and Its Performance Evaluation. International Journal Computational Intelligence Studies, Vol.6, No.4, pp.288–310, 2017.

[iii] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Density-based Spatiotemporal Analysis System with Photo Image Classifier using the BoF Model. Information Engineering Express (IEE), Vol.1, No.4, pp.85–94, 2015.

[iv] Tatsuhiro Sakai and Keiichi Tamura. Real-time Analysis Application for Identifying Bursty Local Areas Related to Emergency Topics. SpringerPlus, 4:162, 17 pages, 2015.

[v] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Extracting Attractive Local-Area Topics in Georeferenced Documents using a New Density-based Spatial Clustering Algorithm. IAENG International Journal of Computer Science, Volume 41 Issue 3, pp.185–192，2014.

国際会議

査読有り

[i] Shuichi Hashida, Keiichi Tamura, and Tatsuhiro Sakai. Classifying Sightseeing Tweets using Convolutional Neural Networks with Multi-Channel Distributed Representation. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC2018), pp.178–183, 2018.

[ii] Tatsuhiro Sakai, Keiichi Tamura, Hajime Kitakami, and Toshiyuki Takezawa. Density-based Multimodal Spatial Clustering using Pre-trained Deep Network for Extracting Local Topics. Fifth International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data (GeoRich2018), pp.7–12, 2018.

[iii] Shuichi Hashida, Keiichi Tamura, and Tatsuhiro Sakai. Multi-Channel Distributed Representation for Classifying Tweets by using Convolutional Neural Networks. International

MultiConference of Engineers and Computer Scientists 2018 (IMECS2018), pp.279–284, 2018.

[iv] Tatsuhiro Sakai, Keiichi Tamura, Hajime Kitakami, and Toshiyuki Takezawa. Photo Image Classification using Pre-trained Deep Network for Density-based Spatiotemporal Analysis System. 2017 IEEE 10th International Workshop on Computational Intelligence and Applications (IWCIA2017), pp.207–212, 2017.

[v] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Cell-Based DBSCAN Algorithm Using Minimum Bounding Rectangle Criteria. Database Systems for Advanced Applications: DASFAA 2017 International Workshops: BDMS, BDQM, SeCoP, and DMMOOC, Suzhou, China, March 27–30, 2017, Proceedings, Lecture Notes in Computer Science (LNCS), Springer-Verlag, Vol.10179, pp.133–144, 2017.

[vi] Keiichi Tamura, Hajime Kitakami, and Tatsuhiro Sakai. Contact Map Overlap Maximization using Adaptive Distributed Modified Extremal Optimization. 2016 IEEE 9th International Workshop on Computational Intelligence and Applications (IWCIA2016), pp.87–92, 2016.

[vii] Keiichi Tamura, Tatsuhiro Sakai, and Takumi Ichimura. Time Series Classification using MACD-Histogram-based SAX and Its Performance Evaluation. 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC2016), pp.2419–2424, 2016.

[viii] Tatsuhiro Sakai, Keiichi Tamura, Kohei Misaki, and Hajime Kitakami. Parallel Processing for Density-based Spatial Clustering Algorithm using Complex Grid Partitioning and Its Performance Evaluation. The 22nd International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2016), pp.337–343, 2016.

[ix] Keiichi Tamura, Tatsuhiro Sakai, and Hajime Kitakami. Location-based Temporal Burst Detection using Outlier Factors in Geo-tagged Tweets. 5th International Congress on Advanced Applied Informatics (AAI2016), pp.191–196, 2016.

[x] Tatsuhiro Sakai, Keiichi Tamura, Shota Kotozaki, Tsubasa Hayashida, and Hajime Kitakami. Real-time Local Topic Extraction using Density-based Adaptive Spatiotemporal Clustering for Enhancing Local Situation Awareness. 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp.203–210, 2015.

[xi] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Identifying Main Topics in Density-based Spatial Clusters using Network-based Representative Document Extraction. 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA2015), pp.77–82, 2015.

[xii] Keiichi Tamura, Hajime Kitakami, Tatsuhiro Sakai, and Yoshifumi Takahashi. A New

Distributed Modified Extremal Optimization for Optimizing Protein Structure Alignment. 2015 IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA2015), pp.109–114, 2015.

[xiii] Keiichi Tamura, Tomoki Matsui, Hajime Kitakami, and Tatsuhiro Sakai. Identifying Local Temporal Burstiness using MACD Histogram. 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC2015), pp.2666–2671, 2015.

[xiv] Tatsuhiro Sakai and Keiichi Tamura. Summarizing Results of Keyword Search on Social Photos using Clustering-based Burst Detection. 4th International Congress on Advanced Applied Informatics (AAI2015), pp.715–716, 2015.

[xv] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Extracting Topic-related Photos in Density-based Spatiotemporal Analysis System for Enhancing Situation Awareness. 4th International Congress on Advanced Applied Informatics (AAI2015), pp.201–206, 2015.

[xvi] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Emergency Situation Awareness during Natural Disasters using Density-based Adaptive Spatiotemporal Clustering. Database Systems for Advanced Applications: DASFAA 2015 International Workshops, SeCoP, BDMS, and Posters, Hanoi, Vietnam, April 20–23, 2015, Revised Selected Papers, Lecture Notes in Computer Science (LNCS), Springer-Verlag, Vol.9052, pp.155–169, 2015.

[xvii] Keiichi Tamura and Tatsuhiro Sakai. Density-based Semantic Spatial Clustering for Extracting Areas of Interesting in Geo-tagged Photo Images. 30th International Conference on Computers and Their Applications (CATA2015), pp.201–206, 2015.

[xviii] Tatsuhiro Sakai and Keiichi Tamura. Identifying Bursty Areas of Emergency Topics in Geotagged Tweets using Density-based Spatiotemporal Clustering Algorithm. 2014 IEEE 7th International Workshop on Computational Intelligence and Applications (IWCIA2014), pp.95–100, 2014.

[xix] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. Density-based Adaptive Spatial Clustering Algorithm for Identifying Local High-Density Areas in Georeferenced Documents. 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC2014), pp.513–518, 2014.

[xx] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. A New Density-based Spatial Clustering Algorithm for Extracting Attractive Local Regions in Georeferenced Documents. International MultiConference of Engineers and Computer Scientists 2014 (IMECS2014), pp.360–365, 2014.

査読無し

[i] Tatsuhiro Sakai, Keiichi Tamura, Hajime Kitakami, and Toshiyuki Takezawa. Density-based Adaptive Spatial Clustering Algorithm for Identifying Local High-density Areas. The 8th International Workshop with Mentors on Databases, Web and Information Management for Young Researchers (iDB Workshop 2017), 6 pages, 2017.

国内研究会

[i] 橋田　修一，田村　慶一，酒井　達弘．部分系列のクラスタリングに基づく符号化を用いた CNN による時系列データの分類手法．2018 IEEE SMC Hiroshima Chapter 若手研究会，pp.89–96, 2018.

[ii] 橋田　修一，田村　慶一，酒井　達弘．畳み込みニューラルネットワークを用いた観光ツイートの分類手法．2018 年度人工知能学会全国大会，4 pages，2018.

[iii] 橋田　修一，酒井　達弘，田村　慶一．深層学習を用いたツイートからの観光情報抽出手法．2018 年電子情報通信学会総合大会，p.148，2018.

[iv] 犬塚　幹浩，酒井　達弘，田村　慶一．外れ値による補正に基づく位置を考慮したバースト検出手法．2018 年電子情報通信学会総合大会，p.48，2018.

[v] 酒井　達弘，田村　慶一，北上　始，竹澤　寿幸．密度に基づくマルチモーダル空間クラスタリングによるジオソーシャル画像からのトピック抽出．2018 年電子情報通信学会総合大会，p.23，2018.

[vi] 酒井　達弘，田村　慶一，北上　始，竹澤　寿幸．セルベースの DBSCAN のマルチコア CPU 上における並列化．第 10 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2018），7 pages，2018.

[vii] 橋田　修一，田村　慶一，酒井　達弘．分散表現の多チャンネル化による深層学習を用いたマイクロブログの分類手法．平成 29 年度（第 68 回）電気・情報関連学会中国支部連合大会，2 pages，2017.

[viii] 酒井　達弘，田村　慶一，北上　始，竹澤　寿幸．密度に基づく時空間分析システムにおける画像分類とその性能評価．2017 IEEE SMC Hiroshima Chapter 若手研究会，pp.91–97，2017.

[ix] 酒井　達弘，田村　慶一，北上　始．密度に基づく時空間分析システムにおける学習済み深層ネットワークを用いた画像分類．人工知能学会・第 15 回インタラクティブ情報アクセスと可視化マイニング研究会，pp.26–32，2017.

[x] Tatsuhiro Sakai, Keiichi Tamura, Kohei Misaki, and Hajime Kitakami. A New Paral-

lelization Model using Complex Grid Partitioning for Density-based Spatial Clustering Algorithm on a Multi-Core CPU. 第 109 回数理モデル化と問題解決研究発表会，2016-MPS-109(4)，4 pages，2016.

[xi] 酒井　達弘，田村　慶一，三崎　浩平，北上　始．複合グリッドを用いた密度に基づく空間クラスタリングの高速化とその性能評価．2016 IEEE SMC Hiroshima Chapter 若手研究会，pp.120–126，2016.

[xii] 犬塚　幹浩，田村　慶一，事崎　翔太，酒井　達弘，北上　始．ジオタグ付きツイートを用いた注目キーワードの抽出．2016 IEEE SMC Hiroshima Chapter 若手研究会，pp.103–107，2016.

[xiii] 酒井　達弘，田村　慶一，北上　始．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．第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2016），8 pages，2016.

[xiv] 酒井　達弘，田村　慶一．密度に基づく空間クラスタリングを用いたジオソーシャル画像からのトピック抽出．第 8 回コンピューテーショナル・インテリジェンス研究会，pp.31–38，2015.

[xv] 酒井　達弘，田村　慶一，事崎　翔太，林田　翼沙，北上　始．密度に基づく時空間分析手法を用いた実世界トピックの閲覧システム．第 17 回 IEEE 広島支部学生シンポジウム，pp.515–520，2015.

[xvi] 酒井　達弘，田村　慶一，事崎　翔太，林田　翼沙，北上　始．密度に基づく時空間分析手法を用いたローカルな話題のリアルタイム抽出．2015 IEEE SMC Hiroshima Chapter 若手研究会，pp.23–26，2015.

[xvii] 三崎　浩平，田村　慶一，酒井　達弘，北上　始．複合グリッドを用いた密度に基づく空間クラスタリングアルゴリズムの並列化．2015 IEEE SMC Hiroshima Chapter 若手研究会，pp.69–72，2015.

[xviii] 事崎　翔太，田村　慶一，酒井　達弘，北上　始．バースト度の地域的な変化の可視化に関する検討．2015 IEEE SMC Hiroshima Chapter 若手研究会，pp.93–94，2015.

[xix] 酒井　達弘，田村　慶一，北上　始．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．情報処理学会第 77 回全国大会，pp.1-649–1-650，2015.

[xx] 酒井　達弘，田村　慶一，北上　始，伊東　晴奈．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．第 101 回数理モデル化と問題解決研究発表会，2014-MPS-101(3)，6 pages，2014.

[xxi] 伊東　晴奈，酒井　達弘，田村　慶一，北上　始．ネットワークベースの要約手法を用いた空間クラスタからの代表文書抽出．第 16 回 IEEE 広島支部学生シンポジウム，B-63，8 pages, 2014.

[xxii] 酒井　達弘，田村　慶一，北上　始．時空間クラスタリングを用いた動向情報の時空間分析手法．第 16 回 IEEE 広島支部学生シンポジウム，B-61，6 pages，2014．

[xxiii] 酒井　達弘，田村　慶一．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．2014 IEEE SMC Hiroshima Chapter 若手研究会，pp.17–20，2014．

[xxiv] 酒井　達弘，田村　慶一，北上　始．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．2013 IEEE SMC Hiroshima Chapter 若手研究会，pp.17–20，2013．