氏名（本籍）	酒井（さかい） 達弘（たつひろ）（島根県）
学位の種類	博士（情報科学）
学位記番号	甲第133号
学位授与年月日	平成30年9月25日
学位授与の要件	広島市立大学大学院学則第36条第2項及び広島市立大学学位規程第3条第2項の規定による
学位論文題目	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 A Study on Spatiotemporal Data Mining for Social Media
論文審査委員	主査 教授 竹澤 寿幸 副査 准教授 田村 慶一 副査 教授 松原 行宏 副査 教授 北上 始（広島工業大学）

論文内容の要旨

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。

論文審査の結果の要旨

平成30年8月7日（火）10：40から11：55まで博士学位論文発表会（公聴会）を開催した。学位申請者が論文内容について説明を行い、その後、論文内容に関する質疑応答および議論を行った。発表会終了後の11：55から12：10まで審査委員会を開催し、論文の合否に関する審議を行った。

本論文では、ソーシャルメディア上の位置情報付きデータを利用して、実世界のトピックの時間変化のみならず空間上における変化を分析する手法を提案し、実験により、その効果を確認している。まず、「大雨」「大雪」といったトピックを単語で与え、ジオタグ付きツイートを用いて、自然災害が発生している地域と当該事象の時間変化を分析する手法を提案し、実際に動作するWebアプリケーションを構築し、実験により効果を確認している。都会と田舎のように、投稿数が多い地域と少ない地域が存在するという課題を指摘し、密度に基づく適応的な時空間クラスタリングを次に提案している。さらに、ジオタグ付きツイートには写真が含まれていることがあるため、テキスト情報のみならず、画像データを分類して活用する手法を提案している。これらの提案手法を役立てるためには、大規模データをリアルタイムでクラスタリングする課題があることを指摘し、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という高速化手法を提案し、人工データと実データを用いて、実験により効果を確認している。最後に、トピック語を与えずに、ジオタグ付きツイートのテキストと画像をもとに、トピックを推定するための密度に基づくマルチモーダル空間クラスタリングを提案している。

本論文の主な成果は、学位申請者が筆頭著者で執筆した査読付き論文誌4編で公表済である。

博士学位論文発表会（公聴会）では、学位申請者から研究の内容が的確に説明され、質疑応答も適切であった。聴講者や審査委員との間では、将来の発展的な課題も含めた議論が活発に行われた。関連する研究成果として、学位申請者が筆頭で論文執筆と発表を英語で行った査読付き国際会議が12編あることから、学位申請者は十分な外国語（英語）能力を有すると判断された。

以上より、学位申請者は博士（情報科学）を取得するのに十分な専門知識と資格を有するものと認め、審査委員会は試験（試問）を合格と判定した。