

Tweet 分析による群衆行動を用いた地域特徴抽出

李 龍^{1,2,a)} 若宮 翔子¹ 角谷 和俊¹

受付日 2011年12月20日, 採録日 2012年4月7日

概要: 都市を特徴付けることは、人々が日々の生活や様々な活動の中で行っている情報処理過程の一部であり、都市での意思決定を行う際に重要となる。従来は、物理的な構造形態に基づく都市の機能や都市に対する人々の意識といった都市の外観に基づく特徴付けが行われていたが、実際に都市で生活する人々の活動を支援するためには、人々のライフスタイルを中心とした特徴付けが重要となる。都市における人々の活動は多様であり容易に把握することは困難であるが、近年のソーシャル・ネットワーク (SNS) の発達とスマートフォンの普及により、多くのユーザが実空間における活動や感情を自らの居場所の位置情報とともに自発的に発信するようになり、実世界の物理的空間と密接に関連した位置ベース SNS を通して、都市における群衆のライフスタイルを把握することが可能になっている。本研究では、位置ベース SNS に蓄積されている大量のユーザの時空間ライフログを用いて都市空間における群衆行動をモニタリングし、都市の地域特徴を抽出する手法を提案する。具体的には、Twitter に投稿されているジオタグ付き Tweets を用いてモニタリングした群衆行動をベクトル化し、地域と群衆行動特徴によって構成した行列を分析することで、特徴的な行動パターンとそれに対応する都市を抽出する手法を提案する。実験では、Twitter から取得した大量のジオタグ付き Tweets を用いて群衆行動を分析し抽出した地域特徴の意味付けを行うために、Yahoo! ロコが提供している店舗や施設のジャンルを調査した結果について示す。

キーワード: 位置ベース SNS, 地域特徴付け, Twitter, 群衆行動パターン

Urban Characteristics Extraction Based on Crowd Behavior by Tweets Analysis

RYONG LEE^{1,2,a)} SHOKO WAKAMIYA¹ KAZUTOSHI SUMIYA¹

Received: December 20, 2011, Accepted: April 7, 2012

Abstract: Characterizing urban space is critical to understand the space and conduct decision-makings in our daily urban lives and activities. Conventional methods have attempted to characterize urban space using urban functionalities based on physical configuration and people's conscious mind to the space. However, in order to support residents' activities in urban space, it is essential to extract characteristics focusing on crowd's urban lifestyles. However, it is a non-trivial work to monitor crowd activities and lifestyles in large-scale regions. In order to solve this problem, we can exploit current crowd's power on behalf of the proliferation of smartphones as well as the recent development of location-based social networks, where massive users voluntarily share their lifelogs and thoughts together with their whereabouts. Therefore, we can easily monitor crowd behavior through such location-based social networks. In this work, we propose a method to characterize urban space in terms of crowd behavior by utilizing enormous number of users' spatio-temporal lifelogs archived over social networks. Specifically, we derive latent classes of urban characteristics in terms of crowd behavioral patterns and relevant urban areas which are extracted using geo-tagged Tweets over Twitter.

Keywords: location-based social networks, urban area characterization, Twitter, crowd behavioral pattern

¹ 兵庫県立大学
University of Hyogo, Himeji, Hyogo 670-0092, Japan
² 独立行政法人情報通信研究機構

National Institute of Information and Communications
Technology, Soraku, Kyoto 619-0289, Japan
a) lee.ryong@gmail.com

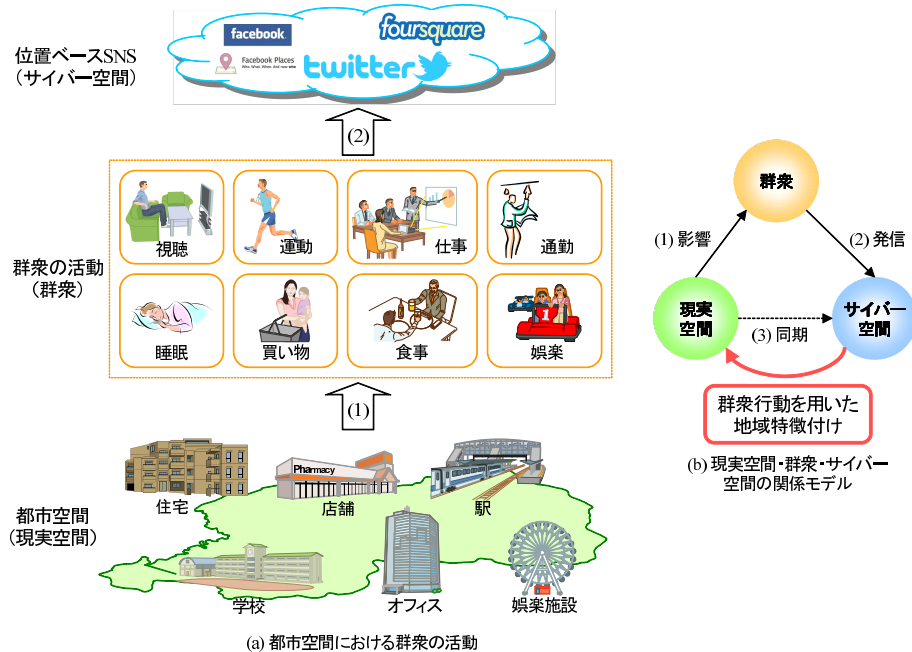


図 1 群衆に基づく都市特徴付けのためのアプローチ

Fig. 1 Research model of crowd-based urban characterization.

1. はじめに

都市は人間の生活のために重要な空間であり、人々の様々な活動を支えるためのインフラ設備や学校・病院・店舗のような施設によって構成されている。また、地形や気候による災害や自然現象が頻繁に発生したり、そこで生活している人々が関わる社会的イベントが開催されたりする複雑な空間である。そのため、人々の生活や様々な活動に関わる意思決定をするために必要となる都市の特徴を把握することが重要である。

都市の地域特徴を抽出するために、都市の物理的な構造形態に基づく機能性や都市に対する人々の意識調査による特徴分析など、様々なアプローチが行われてきた。しかし、これらのアプローチは、都市の外観や客観的な人々の活動に着目しているため、実際に都市で生活する人々の活動を支援するには不十分であり、より人々のライフスタイルに着目した特徴付けが必要とされている。

しかし、都市空間における人々の多様な活動やライフスタイルを簡単にかつ頻繁に把握することは困難である。たとえば、政府による人口推計^{*1}やパーソントリップ調査^{*2}のような社会統計的な手法では、調査員が調査対象地域を訪れ、現地の人々の活動の様子を直に観察したり、インタビューや質問票によるアンケート調査を行うことにより、人々のライフスタイルを分析している。このような方法は、調査者だけでなく、被験者にとっての負担も大きく、

コストがかかる。実際に、総務省が5年ごとに実施している社会生活基本調査^{*3}では、選定された約8万4千世帯に配布する調査票を作成したり、調査員を使ってそれらを配布・回収したりするために多大な金銭的および時間的コストがかかっている。また、調査世帯の人々にとっても、調査票に回答を記入して提出しなければならず負担がある。さらに、従来の手法では、調査範囲に時間的・空間的な制約があるため、広域的なエリアにおける膨大な数の人々の活動を逐一観察することは現実的に不可能である。

そこで、本研究では、人々のコミュニケーションを中心としたソーシャル・ネットワーク・サイト (SNS) が発展し、現実空間の居場所を中心とした人々の活動や感情を共有することによって形成されている位置ベース SNS に着目する。Twitter [20], Foursquare [7] や Facebook places [5] などに代表される位置ベース SNS は、位置情報を容易に取得・発信をすることができるスマートフォンの普及とともに著しく成長しており、地域関連性も高い。したがって、このような位置ベース SNS には、図 1 (a) に示すように、単に群衆の活動や感情が反映されているだけでなく、群衆の活動や感情を通して都市空間の様子が映し出されているといえる。たとえば、朝や夕方に人が集まるという行動から、ラッシュ時の通勤・通学客で混雑した駅の様子を簡単に想像することができる。このように、群衆の活動は都市空間を構成している住宅、学校、オフィスといった施設の機能や役割などに依存することが多い。そのため、位置ベース SNS に反映されている実空間における群衆の活

*1 総務省 人口推計: <http://www.stat.go.jp/data/jinsui/index.htm>

*2 国土交通省 パーソントリップ調査: <http://www.mlit.go.jp/crd/tosiko/pt.html>

*3 総務省 社会生活基本調査: <http://www.stat.go.jp/data/shakai/2011/2.htm#h05e1-1>

動から、地域特徴を把握することができると考えられる。

本研究では、現実空間・群衆・位置ベース SNS の 3 つの要素からなるモデルを作成し、群衆を通して位置ベース SNS 上に映し出されている現実空間の様子を観察して分析するための手法を検討する。本論文では、群衆の活動に基づく都市空間の地域特徴を抽出するために、1) 位置ベース SNS を通して地域ごとに群衆の行動を観察し、2) 特徴的な群衆行動パターンと対応する地域を抽出する。具体的には、ユーザ発信型の時空間ライフログとして Twitter から収集した大量のジオタグ付き Tweets を用いて地域ごとに群衆行動をモニタリングし、特定の時間帯ごとに集約した群衆行動特徴をベクトルによって表現する。次に、NMF (非負値行列因子分解: non-Negative Matrix Factorization) を用いて地域と群衆行動ベクトルの行列を因子分解し、潜在的特徴クラスを抽出する。そして、各特徴クラスに含まれる都市に存在する店舗や施設のジャンル情報を調査することにより、抽出した地域特徴の意味付けを行う。

本論文の構成は以下のとおりである。まず、2 章では本研究の概要を説明し、関連研究を紹介する。3 章では、位置ベース SNS を用いて群衆行動特徴をベクトル化し、特徴的な行動パターンとそれに対応する都市を抽出するための詳細なプロセスについて説明する。4 章では、Twitter から取得した大量のジオタグ付き Tweets を用いて抽出した潜在的特徴クラスを示し、施設のジャンル情報を用いて地域特徴の意味付けを行った結果について述べる。5 章で本研究に関する考察を行い、最後に 6 章でまとめを述べる。

2. 位置ベース SNS 上の群衆の行動に基づく都市特徴付け

2.1 研究モデル

本章では、位置ベース SNS を用いて都市空間における群衆の活動を推定することで、都市における人々のライフスタイルに関わる地域特徴を抽出するための手法を提案する。まず、図 2 に本研究の基盤となるモデルを示す。このモデルは、実世界に存在する物理的な空間である「現実空間」と、現実空間における地域を基盤とした人々の集まりとしての「群衆」、そしてウェブや SNS に代表される「サイバー空間」の 3 つの要素から構成されている。各要素について、以下で具体的に述べる。

都市空間: 現実存在する物理的な空間として、人間が生活を維持するために必要な施設やインフラ設備が整っており、実際に人々が生活している空間を都市空間と定義する。従来、都市空間は農村空間と対比することによってとらえられてきたが、近年の情報通信技術に代表される産業の発展により、農村空間における人々のライフスタイルと都市空間における人々のライフスタイルに大きな変化が見られなくなった。さらに、都市から周辺地域への人口流動が活発になっていることなどから、都市空間と農村空間を明確に区別することが難しくなっている [26]。そこで本論文では、人々が生活する空間として、都市をより一般的に次のように定義する。都市空間は、住宅、学校、オフィスビルのよ

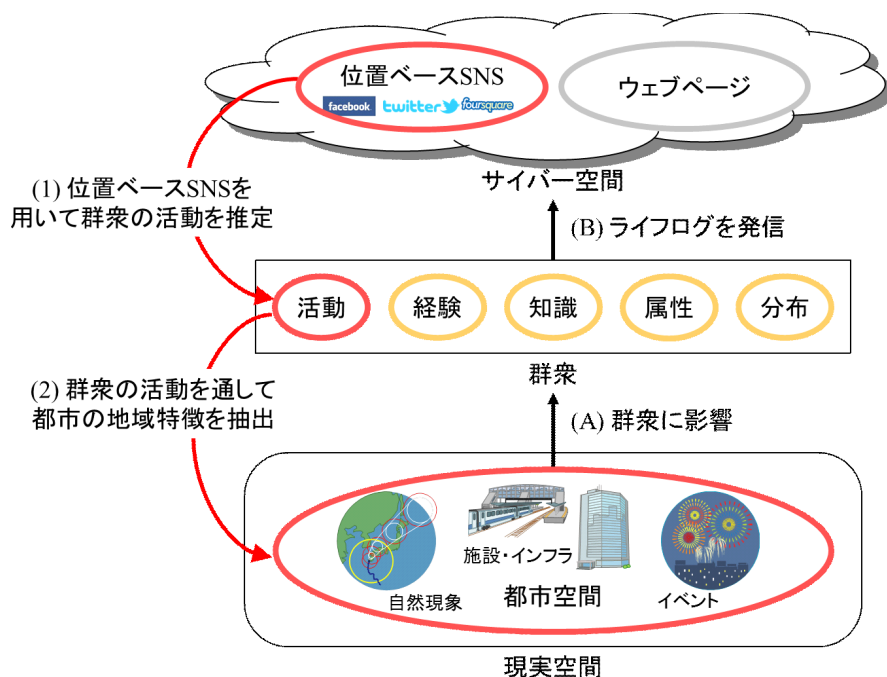


図 2 都市の地域特徴抽出のための研究モデル

Fig. 2 Research model for urban characterization.

うな施設や駅などのインフラ設備だけでなく、その地域の地形や気候に影響して生じる地震や台風などの自然現象や、地域社会において開催される花火大会などのイベントを含む複雑な空間であり、図 2(A) に示しているように、都市空間で生活している群衆のライフスタイルに影響を与える。

群衆の活動： 現実空間とサイバー空間に存在する不特定多数の人々を群衆と定義する。従来、群衆は現実空間で生活したり様々な活動をしたりすることによって地域社会を形成する存在としてとらえられてきた。しかし、近年のウェブや SNS に代表されるサイバー空間の発展により、情報を検索したり、現実空間での経験あるいは様々な現象やイベントなどに関する情報を発信したりする群衆が増加している。その結果、現実空間とサイバー空間に存在し、両者をつなぐ存在として群衆をとらえることができるようになっていく (図 2(B))。また、このような群衆は、現実空間やサイバー空間における活動、経験や知識、そして性別や年齢のような属性や分布、人とのつながりといった多様な観点からとらえられる。本研究では、これらの中でも都市空間における日常生活を最も直接的に表している群衆の活動に着目する。

位置ベース SNS： 位置ベース SNS は、人々のコミュニケーションを中心としたソーシャル・ネットワーク・サイト (SNS) が発展し、現実空間の居場所を中心とした人々の活動や感情を共有できるようになったことで形成された SNS の一種である。特に、位置情報を容易に取得・発信をすることができるスマートフォンの普及にともない、その成長が目覚ましい。位置ベース SNS には、群衆が自発的に発信している都市空間での活動や感情、あるいは起こった現象やイベントなどに関する情報が大量に蓄積されている。たとえば、位置ベース SNS を代表する Twitter には、Tweets と呼ばれる群衆の時空間ライフログが蓄積されている。この Tweets にはそれぞれテキストメッセージ、ジオタグ、ユーザ ID、発信時間のような基本的なメタデータが含まれているだけでなく、ユーザアイコンに用いられている写真、Tweets を投稿する際に用いた端末、フォロー (ユーザの Tweets を受信する人々) の数、フォロー (ユーザが Tweets を受信する人々) の数、ハッシュタグ、外部メディアへの URL といった様々なメタデータが付与されている。したがって、位置ベース SNS を用いることで、群衆の活動はもちろん、群衆を通して社会全体を容易に把握することができると考えられる。

本研究では、これらの 3 つの要素間の関係を用いることで、現実空間の中でも特に複雑で、人々の生活や活動の基盤として重要な都市空間を特徴付けるために、都市空間に

おける群衆の活動に着目する。そして、都市空間における群衆の活動をモニタリングするために、群衆の活動に関する情報が大量に蓄積されている位置ベース SNS を利用し、群衆を介した都市空間の分析を行う。

2.2 関連研究

2.2.1 地域特徴抽出に関する研究

都市の地域特徴を抽出するために、これまで様々なアプローチが行われてきた。都市プランナであった Lynch は「都市のイメージ」[14] という著書の中で、都市空間の外観的な構成要素に焦点をあて、イメージアビリティが高い物理的な形態をパス、エッジ、ディストリクト、ノード、ランドマークの 5 つに分類することで、都市の機能性の観点からパブリックイメージを創造するための道筋を示した。また、Tezuka ら [19] は、人間の主観的な地理空間認知を反映したイメージマップを作成するために、ウェブ上のテキストにおける位置説明文を解析し、人々が思い浮かべる都市の特徴の抽出について研究を行った。

しかし、これらの研究は、都市の外観や現場での人々のライフスタイルなどの観察に着目しているため、実際に都市で生活する人々の活動を支援するには不十分であった。そこで、本研究では、都市空間における群衆の活動に着目し、地域の特徴を抽出する手法を提案する。

2.2.2 現実世界における群衆マイニングに関する研究

現実世界における群衆の行動をモニタリングしたり、分析したりする研究は重要である。日常生活における人間行動を観察するために用いられている一般的な手法として、環境に埋め込まれた様々な物理的なセンサを用いた研究がある。西田ら [29] は、天井埋め込み型の超音波レーダを用いて物理的に人間の行動を観察するシステムを提案した。しかし、このようなシステムは、モニタリングすることができる環境や規模が限られているため、現実世界における群衆の活動をモニタリングするには不十分である。

一方で、近年は、位置ベース SNS に対して人々が発信する情報のリアルタイム性に着目し、人間を 1 つの社会的なセンサと見なして現実世界をモニタリングする研究がさかんに行われている。このような研究は、図 1(b)(1) の現実空間から群衆への影響関係と図 1(b)(2) の群衆からサイバー空間への情報発信という行動を通して、図 1(b)(3) のサイバー空間に反映されている現実空間の状況を推定して分析する研究と位置付けられる。Sakaki ら [17] は、Tweets のテキストメッセージをモニタリングすることで、地震のような突発的に起こる自然災害を検知するシステムを開発した。Lee ら [11], [12] は、Twitter を用いて群衆の行動に基づく地域の通常性をそれぞれ推定し、異常を判定することによって地域イベントを検出する手法を提案した。

図 1(b)(2) において群衆からサイバー空間へ発信されるライフログを用いた群衆マイニングに位置付けられる研究

として、Wakamiya ら [23], [24], [27] は、多くの人々が視聴中のテレビ番組に対する意見や感想、あるいは番組に関する情報を Twitter へ発信していることに着目し、テキスト・時間・空間の観点から Tweets とテレビ番組との類似度を算出して視聴者を発見し、それぞれの視聴行動を分析する手法を提案している。

本研究では、図 1(b)(1) の現実空間から群衆への影響関係と図 1(b)(2) の群衆からサイバー空間への情報発信という行動に着目して、図 1(b)(3) に示すようにサイバー空間に同期している都市空間の地域特徴を抽出する。そこで、位置ベース SNS におけるユーザ発信型の時空間ライフログに共通して付与されているジオタグ、ユーザ ID および発信時間を用いて基本的な群衆行動をモデリングし分析する。本論文では、定量的な群衆行動を分析することに焦点をあてているためテキストメッセージの分析を行っていないが、本研究の次のステップとして、群衆行動から推定される群衆の活動を意味的にとらえる際には、テキストメッセージの分析が必要になると考えられる。

3. 位置ベース SNS を用いた群衆行動分析による地域特徴付け

本章では、位置ベース SNS に投稿されている膨大な数の群衆の時空間ライフログを用いて、群衆行動に基づく地域特徴を抽出するための詳細な手法について述べる。具体的には、1) 位置ベース SNS を代表する Twitter からの大量のジオタグ付き Tweets の収集 (3.1 節)、2) 群衆行動をモニタリングする地域境界の設定 (3.2 節)、3) 収集した Tweets を用いた群衆行動のモデリングおよびベクトル化 (3.3 節)、そして 4) 群衆行動のベクトル分析による特徴的な行動パターンと対応する地域からなる潜在的な特徴クラスの抽出 (3.4 節) について述べる。

3.1 Twitter からの群衆のライフログの取得

Twitter を用いて都市空間における群衆の活動を推定するために、まず、図 3(1) に示すように、Twitter からジオタグ付き Tweets を収集する。このとき、広域的なエリアを対象として大量の Tweets を網羅的に取得することは困難である。実際に、Twitter がデータを共有するために公開している API [21] では、クエリで指定した中心の地理座標と半径に基づくエリア内で発信された Tweets を取得することが可能であるが、1 回の問合せで取得可能な Tweets 数は最新の 1,500 件までに制限されている。そのため、API を用いた単純な問合せで広域エリアを指定した場合、取得される 1,500 件の Tweets の大半をユーザが多い大都市（たとえば、東京、名古屋や大阪など）から発信された Tweets が占めることになると考えられる。そのため、地域をより小さな地域に分割してデータの出現をモニタリングすることが必要となる。たとえば、日本全域をグリッドで一定の

大きさ（1km 四方など）に分割して簡単にモニタリングすることも可能である。しかしながら、日本全域を対象としてグリッドを設定すると、セルの数だけ Twitter への問合せが行われる。同時に複数の問合せを行うことも可能ではあるが、その数が何万件にもなるとシステムの運用そのものに悪影響を及ぼす恐れがあるためほとんど許可されていない。しかしながら、実際には、データの出現密度は地域ごとにばらつきがあるため、対象地域全体を同様に扱って同時に詳細に調べる必要はなく、データの出現密度が高い地域には小さいセルを、密度が低い地域には大きいセルを割り当てて調べるというように、セルの大きさを調節することによって全地域を調べるコストを極端に減らすことが可能である。我々の以前の研究 [11], [12] では、この方法を実装した Twitter 向けの地理的マイクロブログ収集システムを開発している。このシステムは、Quad-tree の生成アルゴリズム [6] を用いてデータの出現密度を調査しているため、対象領域をより少ない問合せでより早く調査することが可能である。また、発生頻度に応じてセルを調べる順番を動的に決定することにより、対象領域内で発信されているデータをバランス良く取得している。

Tweets のジオタグは、テキストによる地名もしくは緯度・経度からなる地理座標の形式で表現されている*4。各 Tweet がいつどこで発信されたものであるかを正確に把握するために、ジオタグのデータ形式がテキストである場合には、地名に対応する地理座標に変換するジオコーディングを行う必要がある。地理的マイクロブログ収集システムでは、Google Map のジオコーディング・サービス [8] をマッシュ・アップ・サービスとして用いている。具体的には、Twitter に投稿されている Tweets のジオタグは、ユーザが利用している携帯端末やブラウザから送信される正確な位置データ（経度と緯度）に基づき位置データを取得し、Twitter が正確な位置データあるいはテキストによる地名（駅名、ランドマーク名や市区町村名）、あるいはその両方を各 Tweet に追加する。このとき、テキストによる地名表現は、「姫路市、兵庫県」のように市区町村レベルの粒度で付与されることが多い。本論文では、Twitter が提供しているローカルサーチによって得られた結果に対してジオコーディングを行い、正確な位置データへと変換している。これにより、図 3(1) に示すように、取得したすべての Tweets を地図上に配置することが可能になる。

3.2 群衆分布に基づく地域境界の設定

本研究は、群衆の活動に焦点をあてた都市の地域特徴分析を目的としている。そのため、群衆の行動をモデリングし、地域特徴を抽出する単位として都市を定義する必要がある。一般的に、都市は都道府県や市区町村のような行政

*4 Twitter 位置情報追加機能について：
<https://support.twitter.com/articles/250973>

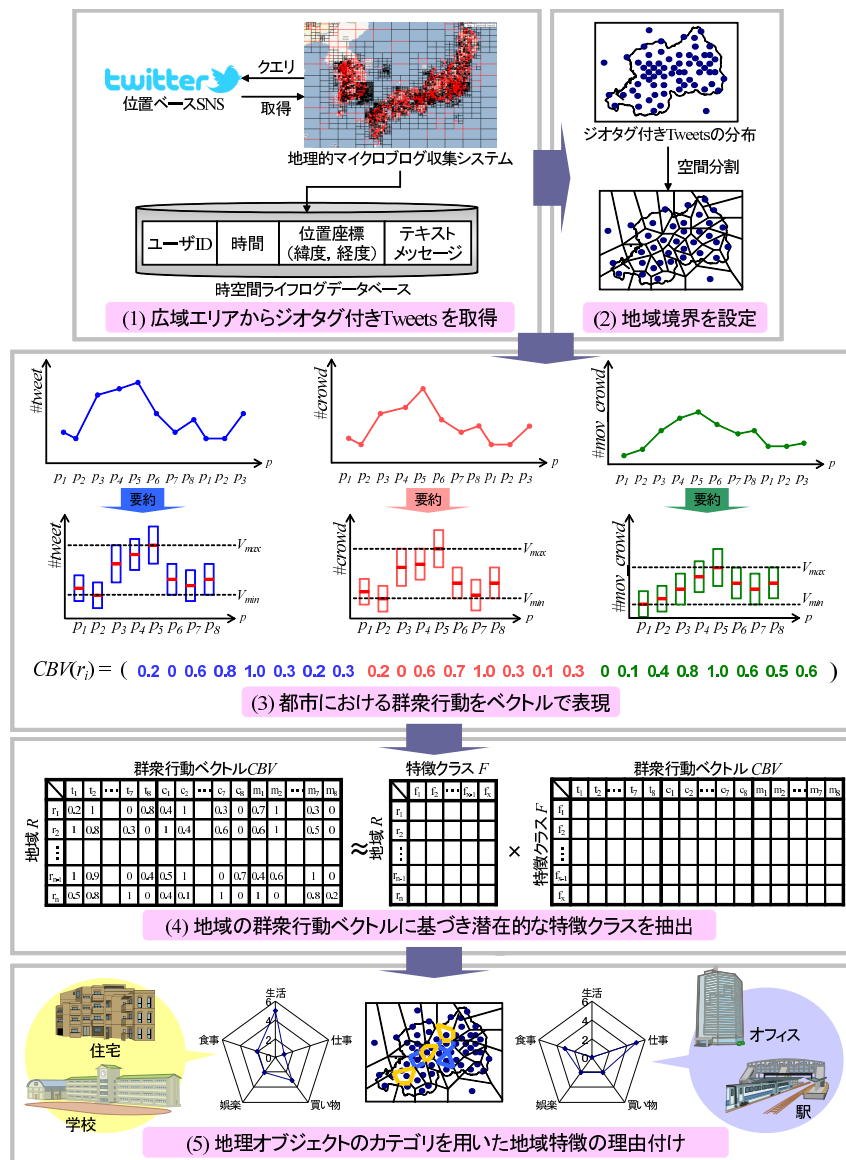


図 3 Twitter を用いた都市空間の地域特徴抽出のプロセス

Fig. 3 Process of Twitter-based urban area characterization.

区域でとらえられることが多い。しかし、車や公共交通機関などの移動手段の発達により、行政区域を越えた人々の活動が頻繁に起こり、複数の行政区域にまたがる生活圏が存在している。したがって、群衆に着目した地域境界を設定することが有用であると思われる。このとき、群衆の密度が高い地域は、密度が低い地域に比べてより多様な活動が起こると考えられる。そのため、群衆の地理的な分布を考慮し、群衆密度が高い部分をより細かく分割することができる地域境界の設定が有用であると考えられる。

そこで本研究では、Tweets の地理座標に基づき群衆をクラスタリングし、生成されるクラスタに基づく空間分割を行う (図 3(2))。以前の研究 [22], [25] では、Tweets の地理座標データに k-means アルゴリズム [13] を適用することで群衆をクラスタリングした。その結果、少ない計算量で指定した数のクラスタにデータをクラスタリングする

ことができた。しかし、適当なクラスタ数を決めることが困難であるという問題点があった。そこで本論文では、自動的にクラスタ数を決定することができる EM アルゴリズム [1] を適用して群衆をクラスタリングした。しかし、EM アルゴリズムは計算量が大きいため、大量の Tweets の地理座標データを処理することは現実的でない。そこで、データサイズを減らすために、データセットを特徴クラスとノイズクラスに分類する最近傍クリーニングアルゴリズム (Nearest-Neighbor Cleaning, NNClean) [3], [4] を適用した。具体的には、図 4(a) のように地図上にマッピングした大量の Tweets に最近傍クリーニングアルゴリズムを適用することによって、高頻度の地点集合 (図 4(b)) と低頻度の地点集合 (図 4(c)) の 2 つのグループに分類する (図 4(1))。このとき、最近傍クリーニングアルゴリズムでは、高頻度の地点集合を特徴クラス、低頻度の地点集

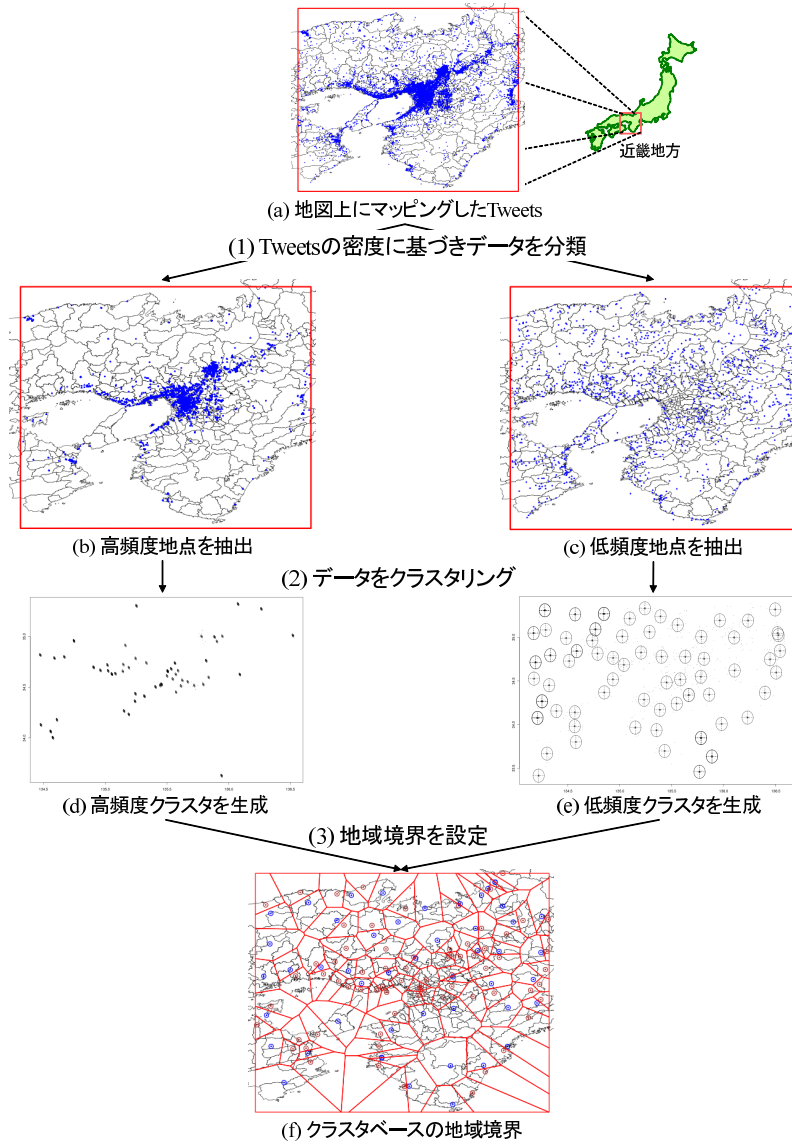


図 4 クラスタに基づく地域境界の設定プロセス

Fig. 4 Process of constructing socio-geographic boundaries based on clustering.

合をノイズクラスと見なす。しかし、特徴クラスに分類されたデータだけでは、低頻度で Tweets が発信されている地域を考慮することができないという問題点がある。そのため、本手法では2つのグループに分割したデータに対してEM アルゴリズムをそれぞれ適用し、2種類のクラスタ集合を生成する（図 4(d) と (e)）。そして、生成されたクラスタの中心地理座標（緯度、経度）を集約し、各点に領域を割り当てるためにボロノイ図 [2] を描画して地域境界（図 4(f)）を設定することで、対象領域 R を多角形のクラスタ r_i に分割する。

$$r_i \in R \ (1 \leq i \leq \#cluster) \quad (1)$$

ここで R は地域境界を設定することによって分割されたすべてのクラスタ r_i の集合を意味する。この手法により、2種類のクラスタ集合に含まれるクラスタの総数 $\#cluster$ と同数の地域が形成される。図 7 に、近畿地方で発信され

た Tweets に基づく地域境界によって分割された 147 の地域を示す。

3.3 群衆活動ベクトルの生成

Twitter から取得したジオタグ付き Tweets を用いて都市における群衆の行動を表現するために、Tweet の位置情報、ユーザ ID および発信時間から算出される基本的なパラメータを用いる。本論文では、3つのパラメータ（ $\#tweet$, $\#crowd$, $\#mov_crowd$ ）を用いて群衆行動を定義する。3つのパラメータの詳細は次のとおりである。

$\#tweet(r_i, p_k)$: 群衆の活発さを示すパラメータである。任意の地域 r_i において特定の時間帯 p_k に発信される Tweet 数とする。

$\#crowd(r_i, p_k)$: 群衆の存在とその規模を表すパラメータであり、任意の地域 r_i において特定の時間帯 p_k に観測される重複のない Twitter ユーザ数とする。

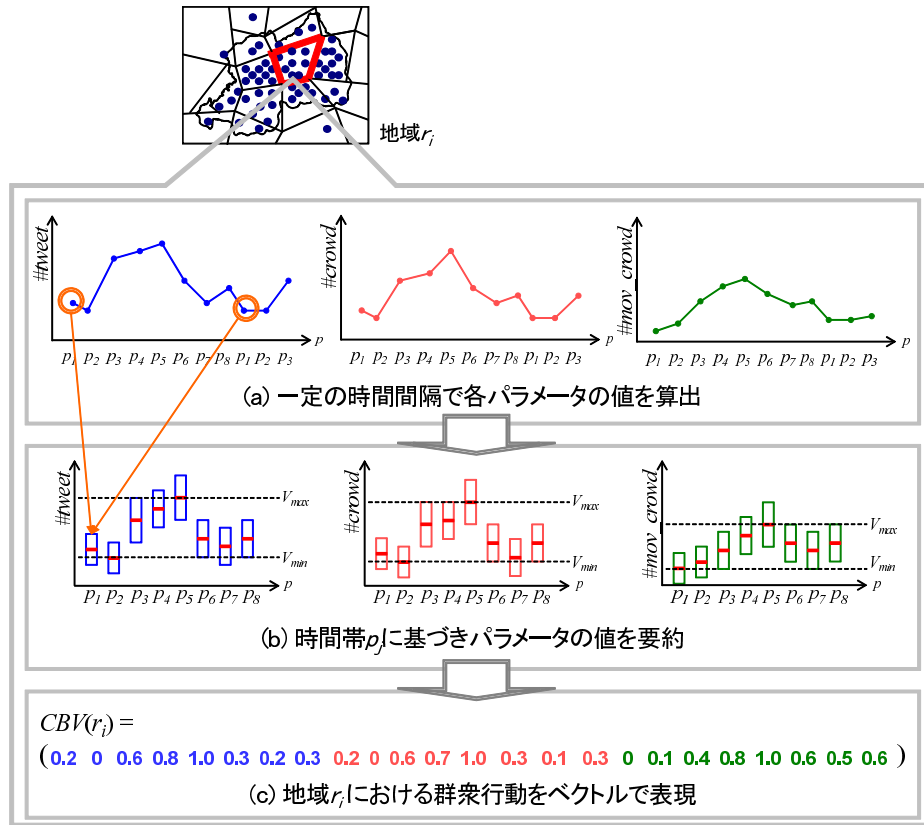


図 5 群衆行動ベクトルの生成

Fig. 5 Generating crowd behavioral vectors.

$\#mov_crowd(r_i, p_k)$: 実空間での群衆の移動を反映したパラメータであり、任意の地域 r_i において特定の時間帯 p_k に観測される移動ユーザの数とする。移動タイプとして次の3つを扱い、いずれかに該当する場合、移動ユーザと見なし加算する。a) 内部での移動 (inner): 同一時間帯に同一地域の異なる位置で Tweet を発信したユーザの数, b) 外部から内部への移動 (incoming): 地域 r_i 以外の地域 r_j で Tweet を発信した後、同一時間帯に地域 r_i で Tweet を発信したユーザの数, c) 内部から外部への移動 (outgoing): 地域 r_i で Tweet を発信した後、同一時間帯に地域 r_i 以外の地域 r_j において Tweet を発信したユーザの数とする。

これらのパラメータの値は、3.2 節で述べた手法によって分割した地域 r_i ($r_i \in R$) ごとに算出する。本論文では、群衆の日常的な活動に着目するために、日単位で群衆行動をモデリングする。ここで、1 日の群衆の活動は時間帯によって異なると考えられるため、1 日を任意の時間間隔で分割し、特定の時間帯 p_k におけるパラメータの値をそれぞれ算出する (図 5 (a))。4 章の実験では、1 日を 3 時間間隔に分割することによって、1 つの地域につき p_1 から p_8 までの 8 つの時間帯における群衆行動のパラメータの値を算出した。

$$p_k = [3(k-1) : 00, 3k : 00) \quad (1 \leq k \leq 8) \quad (2)$$

ここで、 $x : 00$ は時刻の表記であり、時間帯 p_1 は 00 : 00 から 03 : 00 までを意味する。

次に、これらのパラメータの値に基づき群衆行動をベクトルを用いて表現する。ある地域 r_i における群衆行動ベクトル $CBV(r_i)$ を次のように定義する。

$$CBV(r_i) = (CB_{tweet}(r_i), CB_{crowd}(r_i), CB_{mov_crowd}(r_i)) \quad (3)$$

$$CB_{tweet}(r_i) = \text{normalize}(tw(r_i, p_1), \dots, tw(r_i, p_8)) \quad (4)$$

$$CB_{crowd}(r_i) = \text{normalize}(cr(r_i, p_1), \dots, cr(r_i, p_8)) \quad (5)$$

$$CB_{mov_crowd}(r_i) = \text{normalize}(mv(r_i, p_1), \dots, mv(r_i, p_8)) \quad (6)$$

$$tw(r_i, p_k) = \text{median}\left(\bigcup \#tweet(r_i, p_k)\right) \quad (7)$$

$$cr(r_i, p_k) = \text{median}\left(\bigcup \#crowd(r_i, p_k)\right) \quad (8)$$

$$mv(r_i, p_k) = \text{median}\left(\bigcup \#mov_crowd(r_i, p_k)\right) \quad (9)$$

ここで、式 (4), (5), (6) の $CB_{tweet}(r_i)$, $CB_{crowd}(r_i)$, $CB_{mov_crowd}(r_i)$ は、任意の地域 r_i におけるパラメータの p_1 から p_8 の各時間帯の中央値の列を、関数 normalize を用いて一般化した値の列によって構成されている。この一般化は、地域ごとにばらつきがある各パラメータの値を共通の基準で分析するために行っており、この結果、群衆

行動ベクトルの特徴量は図 5(c) のように 0 から 1.0 の値で表現される。また、式 (7), (8), (9) で用いられている関数 *median* は、入力されたデータセットの中央値を算出する関数である。たとえば、1 週間分のデータを用いる場合、同一時間帯の値が 7 つ存在することになる。このようにばらつきのあるデータセットを扱うために、地域ごとに特定の時間帯における平均的な群衆行動を抽出する。そのために、関数 *median* を用いて、データのばらつきを統計的に要約し、視覚的に表現する箱ヒゲ図 (boxplot) [16] を描画するための要約統計量を算出し、中央値を取得する。図 5(b) は、ある地域における各時間帯のパラメータの値を箱ヒゲ図によって表現したものであり、箱内部の線が中央値を意味している。

$$\text{normalize}(V) = (nv_1, nv_2, \dots, nv_7, nv_8) \quad (10)$$

$$nv_l = \begin{cases} \frac{v_l - V_{\min}}{V_{\max} - V_{\min}} & (V_{\max} \neq V_{\min}) \\ 0 & (V_{\max} = V_{\min}) \end{cases} \quad (11)$$

$$v_l \in V \quad (12)$$

$$V_{\max} = \max(V) \quad (13)$$

$$V_{\min} = \min(V) \quad (14)$$

ここで、式 (10) の関数 *normalize* は、式 (4), (5), (6) の $CB_{\text{tweet}}(r_i)$, $CB_{\text{crowd}}(r_i)$, $CB_{\text{mov_crowd}}(r_i)$ を算出するときに用いた関数であり、引数としてデータセット V が与えられたとき、 nv_l の列を返す。変数 nv_l は、データセット V の最大値 $V_{\max}$ と最小値 $V_{\min}$ が異なる場合、変数 v_l と最小値 $V_{\min}$ の差を最大値 $V_{\max}$ と最小値 $V_{\min}$ の差で除算することによって算出される 0 以上 1.0 以下の値である。そして、最大値 $V_{\max}$ と最小値 $V_{\min}$ が同一である場合には、ゼロ除算となるため 0 を代入する (式 (11))。この最大値 $V_{\max}$ と最小値 $V_{\min}$ は、それぞれ式 (13) の関数 *max* と式 (14) の関数 *min* にデータセット V を引数として与えると算出される値である。

3.4 群衆行動ベクトルの分析による特徴的な行動パターンの抽出と地域の分類

本節では、都市空間での生活や様々な活動に関わる意思決定をするために必要となる都市の特徴を抽出するために、地域と平均的な 1 日の群衆行動の関係をマイニングし、特徴的な群衆の行動パターンとそれに対応する地域からなる潜在的な特徴クラスを抽出する手法について述べる。本論文では、行列を因子分解することによって潜在的な特徴クラスを発見することができる NMF (non-Negative Matrix Factorization: 非負値行列因子分解) を適用し、特徴的な行動パターンを抽出して地域を分類する。具体的に、このアルゴリズムは観測データとして任意の二次元行列 X を与えたとき、負値を含まない基底行列 W と係数行列 H の 2 つの行列の積に分解する。NMF は、すべての要素を非

負に制限することで、観測データを行列の加算で表現することができ、元の行列 X が持つ潜在的要素を明確に示すことができる。一般的に、NMF は次のように定義される。

$$X = W \times H + \alpha \quad (15)$$

実際に、NMF を用いた行列の分解は、 X の基底行列 W と係数行列 H におけるすべての非負の要素を更新しながら反復計算を行う。このとき、要素の値を更新するたびに、元の行列 X と分解された行列 WH との距離 α を計算し、この距離を最小化していく。距離が 0 になった場合、あるいはあらかじめ決めておいた反復回数だけ更新を行った場合に行列の分解は完了する。一般的に、距離関数として二乗誤差の最小化や KL ダイバージェンス (Kullback-Leibler divergence) [10] などが用いられている。

4 章の実験では、147 の地域と 3.3 節で述べた 24 次元 (3 つのパラメータ $\times$ 8 つの時間帯) の群衆行動ベクトルによって構成される地域-群衆行動ベクトル行列 X ($R \times CBV$) に NMF を適用し、地域-特徴クラス行列 W ($R \times F$) と特徴クラス-群衆行動ベクトル行列 H ($F \times CBV$) に分解した。このとき、KL ダイバージェンスを適用することで元の行列と分割された行列との距離を計算した。この基底行列 W と係数行列 H を要素の値の大きさに従って解析するとき、基底行列 W の各要素は、各特徴クラス f_x ($f_x \in F$) の地域 r_i ($r_i \in R$) に対する重要度を示しており、係数行列 H の各要素は、各群衆行動ベクトル CBV (r_i, p_j) の各特徴クラス f_x ($f_x \in F$) に対する重要度を表している。このとき、クエリとして適当な群衆行動ベクトルの集合があったとして、まず H から、その群衆行動ベクトル CBV の集合と最も適合する特徴クラス F を得ることができる。そしてその特徴クラスに最も適合すると思われる行を W から選ぶと、地域を得ることができる。ここで得られる地域は、最初のクエリとして与えた群衆行動ベクトルの集合に関連する地域である。このように、NMF によって地域 R と群衆行動ベクトル CBV の関係をマイニングし、潜在的な特徴クラスとして局所的な群衆行動パターンを抽出し、パターンごとに対応する地域を分類することができる。

4. 実験

本章では、Twitter から取得した大量の Tweets を用いて行った実験について述べる。まず、4.1 節では実験に用いたデータセットについて説明し、4.2 節では Tweets のパラメータを用いて表現した群衆行動ベクトルを分析し、地域の潜在的特徴として抽出した群衆行動パターンを示し、そのパターンによってクラスタリングされた地域について考察する。最後に、4.3 節では、クラスタリングされた地域に実在する施設に付与されているジャンル情報を調べることで、地域特徴の意味付けを行う。

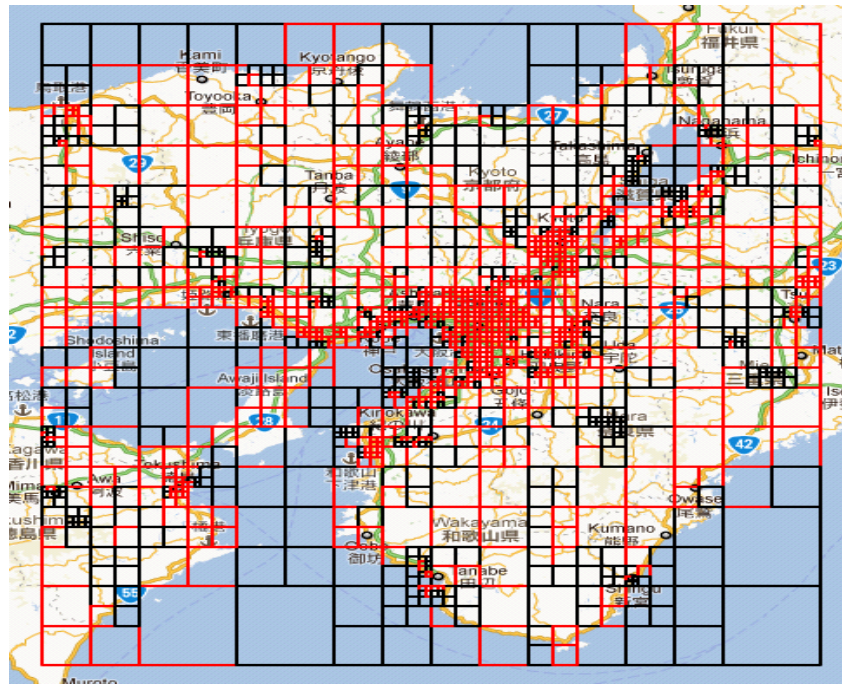


図 6 地理的マイクロブログ収集システムのインターフェース

Fig. 6 An interface of geographic microblog gathering system.

4.1 データセット

本実験では、地域特徴抽出の対象地域として近畿地方を設定した。そして、近畿地方で発信された Tweets を取得するために、地理的マイクロブログ収集システムによるモニタリング領域として近畿地方を含む緯度 [33.384555:35.839419]、経度 [134.126551:136.58890] の領域を設定し、2011 年 10 月 9 日から 2011 年 10 月 22 日までの間に発信されたジオタグ付き Tweets を取得した (図 6)。その結果、重複のない 8,272 人のユーザが発信した 75,279 件の Tweets を取得することができた。

次に、取得した Tweets を用いて、3.2 節で説明したクラスタに基づく地域境界の設定を行い、近畿地方を 147 の地域に分割した。その結果を図 7 に示す。ここで、ボロノイ図を用いているため、147 のクラスタのうちデータが存在しない 2 つのクラスタが生成された。この 2 つのクラスタでは群衆の活動を観測することができないため、本研究で定義する都市として適切でないと判断し、残りの 145 のクラスタを本実験における特徴付けの対象地域とした。

4.2 群衆行動ベクトル分析に基づく潜在的特徴クラスの抽出

4.1 節で述べたデータセットを用いて、145 の地域について 3 つのパラメータ (#tweet, #crowd, #mov_crowd) の値を 3 時間ごとにそれぞれ算出し、それらの値に基づき群衆行動ベクトルを生成した。実際に生成した群衆行動ベクトルの例を図 8 に示す。

次に、145 の地域と 24 次元の群衆行動ベクトルによ

て地域-群衆行動ベクトル行列 $X: R \times CBV$ を生成する。この行列に NMF を適用することで、地域-特徴クラス行列 $W: R \times F$ と特徴クラス-群衆行動ベクトル行列 $H: F \times CBV$ に分解した。今回の実験では、元の行列と分解した行列の距離が最小となった時点で分解を終了した。その結果、群衆行動パターンと対応する地域をそれぞれまとめた 13 の潜在的特徴クラスを抽出することができた。図 9 に、ヒートマップを用いて表現した行列分解の結果を示す。図 9(a) は地域と特徴クラスからなる基底行列 W 、図 9(b) は特徴クラスと群衆行動ベクトルからなる係数行列 H である。このヒートマップにおいて、色の濃淡は特徴の強さを表しており、色が濃いほど特徴が強い。図 9(a) の横軸と図 9(b) の縦軸は f_1 から f_{13} までの特徴クラスを示しており、図 9(a) では左から順に、図 9(b) では上から順に並んでいる。一方、図 9(a) の縦軸は地域を表しており、各地域に割り振った番号が示されている。図 9(b) の横軸は 24 次元の群衆行動ベクトルであり、左から $mv_{p_2}, mv_{p_8}, mv_{p_7}, mv_{p_6}, mv_{p_3}, mv_{p_1}, mv_{p_4}, mv_{p_5}, tw_{p_3}, cr_{p_3}, tw_{p_2}, cr_{p_2}, tw_{p_1}, cr_{p_1}, tw_{p_8}, cr_{p_8}, tw_{p_7}, cr_{p_7}, tw_{p_6}, cr_{p_6}, tw_{p_4}, cr_{p_4}, tw_{p_5}, cr_{p_5}$ である。この分解結果から得られた特徴クラス $f_1, f_3, f_5, f_6, f_8, f_9, f_{12}, f_{13}$ には、それぞれの時間帯 p_k における tw_{p_k} と cr_{p_k} の組から構成されるパターンが含まれており、 CB_{tweet} と CB_{crowd} の間に強い相関関係があることが分かった。

図 10(A)(a), (B)(a), (C)(a) は、特徴クラス f_3, f_8, f_9 に分類された地域を地図上で強調して示したものである。この結果から、群衆行動パターンによってグルーピングさ

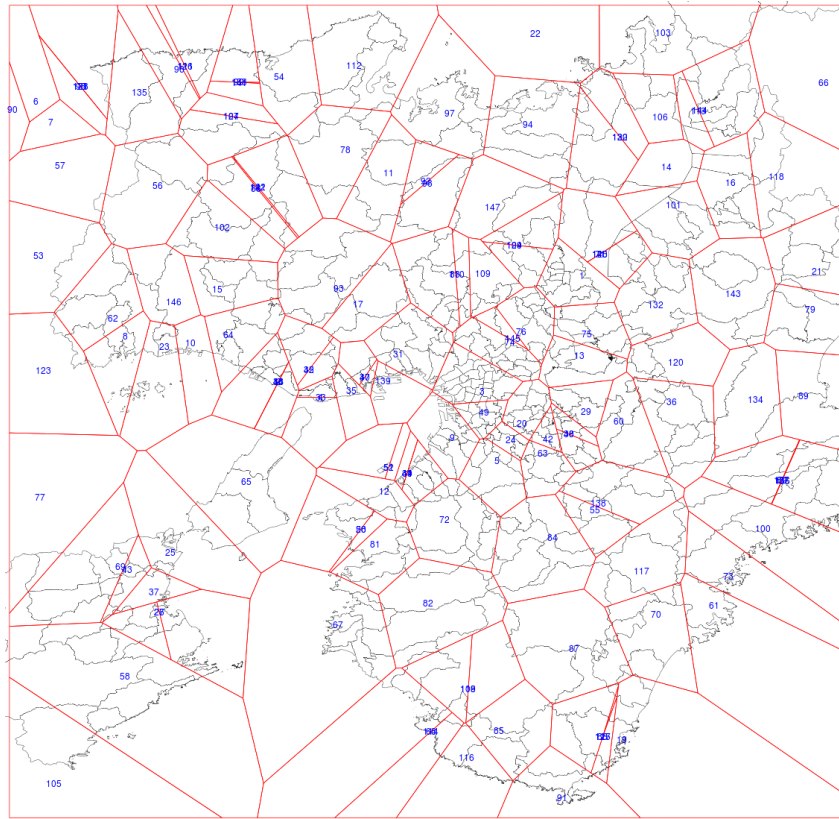
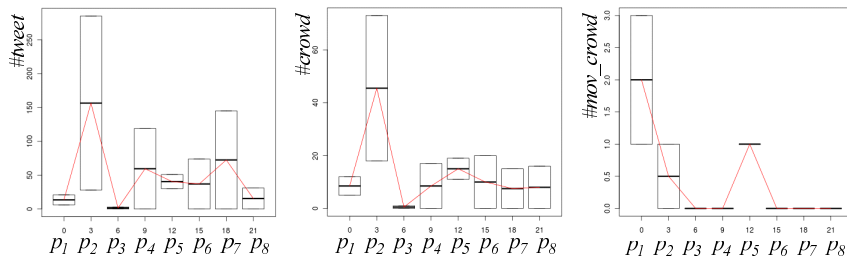


図 7 群衆の分布に基づく地域境界

Fig. 7 Geo-social boundary based on crowd distribution.



$$CBV(r_i) =$$

$$\begin{matrix} tw_{p1} & tw_{p2} & tw_{p3} & tw_{p4} & tw_{p5} & tw_{p6} & tw_{p7} & tw_{p8} & cr_{p1} & cr_{p2} & cr_{p3} & cr_{p4} & cr_{p5} & cr_{p6} & cr_{p7} & cr_{p8} & mv_{p1} & mv_{p2} & mv_{p3} & mv_{p4} & mv_{p5} & mv_{p6} & mv_{p7} & mv_{p8} \\ (0.1 & 1.0 & 0 & 0.4 & 0.3 & 0.2 & 0.5 & 0.2 & 0.2 & 1.0 & 0 & 0.2 & 0.4 & 0.3 & 0.2 & 0.2 & 1.0 & 0.4 & 0 & 0 & 0.6 & 0 & 0 & 0) \end{matrix}$$

図 8 生成された群衆活動ベクトルの例

Fig. 8 An example of crowd behavioral vector.

れた地域には次の 2 種類の傾向があることが分かる。まず、同一の特徴クラスに分類される地域は、隣接した地域に見られる場合が多いということである。これにより、群衆行動パターンに基づく隣接地域とのつながりを理解することが可能である。その一方で、地域間の境界線が明確に存在していることも分かる。つまり、同一の群衆行動パターンが見られる地域はある程度地理的にまとまっており、その隣接した地域集合を単位としてモニタリング領域に散らばっているといえる。この結果は、地域境界を設定する際に用いたクラスタリングの粒度に関わっている可能

性も考えられる。今回は、群衆が発信した時空間ライフログを活用するために、計算量と実験により経験的に最適な地域境界を設定している。クラスタリングの粒度は自由に調整することが可能であるため、細かくセルを分割することやさらに大きなセルを用いることも可能である。しかしながら、前者の場合には、セル内に含まれるデータ数が少なくなってしまう、地域特徴をうまくとらえることが困難になる可能性がある。逆に、セルが大きくなると、複数の地域の様々な要素が混ざり合うことになるため、地域特徴を適切にとらえることが困難になると考えられる。このよ

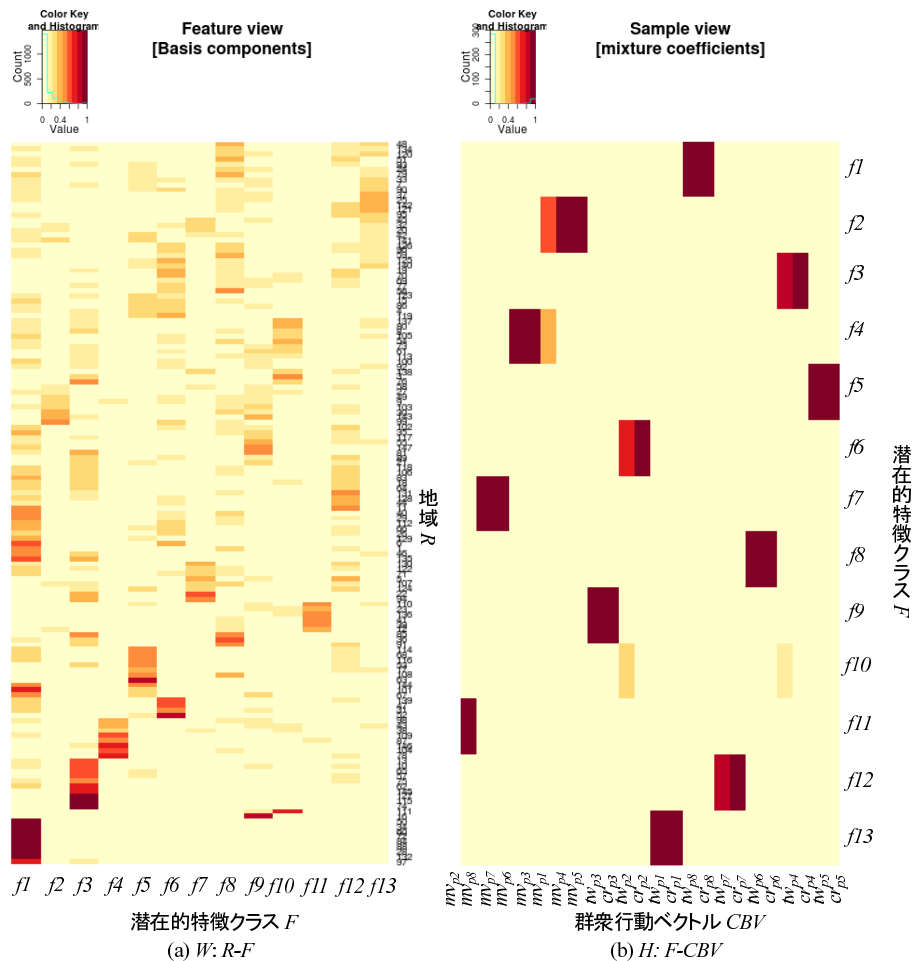


図 9 NMF を用いた地域における群衆活動パターンの抽出
Fig. 9 Extracting crowd behavioral patterns in urban areas using NMF.

うな観点も考慮すると、同一行動パターンによって分類された地域の地理的な分布の傾向は、群衆の行動に着目しているからこそ得られるものであるといえる。

4.3 群衆に基づく都市特徴付けの意味付け

群衆行動パターンに基づきクラスタリングされた都市の特徴を直感的に把握するために、都市に存在する施設のジャンル情報を調べた。今回は、Yahoo! Japan ロコ^{*5}においてそれぞれの施設に付与されているジャンルのデータを取得して活用した。Yahoo! Japan ロコの場合、各施設は階層的な構造を持つカテゴリによって分類されている。大カテゴリには「食べる」、「買う」、「遊ぶ」、「暮らす」の4つがあり、その中でさらに中カテゴリ、小カテゴリがある。それぞれのカテゴリの例として、大カテゴリ「食べる」の下には「居酒屋」、「ビアホール」、「中華」、「スイーツ」などの中カテゴリが存在し、中カテゴリ「居酒屋」、「ビアホール」の下には、「和風居酒屋」、「洋風居酒屋」、「アジア居酒屋」、「無国籍居酒屋」などの小カテゴリが存在する。また、ジャンルのデータが付与されていない施設もいくつか存在

する。それらの施設を手で確認したところ、主に産業に関わる施設であることが分かったため、大カテゴリとして「働く」を付与し、中カテゴリ、小カテゴリは総務省統計局の日本標準産業分類^{*6}を参照し、適当なジャンルデータを人手で付与した。施設のジャンルデータを用いた理由付けの目的は、都市に対する直感的なイメージを把握することであり、中カテゴリや小カテゴリのような施設を詳細に表現するカテゴリ名よりも、大カテゴリのように人々の目的を表現するようなものが適切であると考えられる。そのため、施設に付与された5つの大カテゴリに基づくジャンルデータを用いて、各特徴クラスの意味付けを行った。

今回は、地域ごとに「食べる (food)」、「買う (shopping)」、「遊ぶ (entertainment)」、「暮らす (life)」、「働く (work)」のそれぞれのジャンルデータが付与された施設の割合を調べた。しかし、地域によって面積がそれぞれ異なるため、施設数の絶対的な割合を比較することは適切でない。そこで、地域ごとに5つのジャンルデータに含まれる施設数の

^{*5} Yahoo! Japan ロコ : <http://maps.loco.yahoo.co.jp/>

^{*6} 日本標準産業分類 : <http://www.e-stat.go.jp/SG1/htoukeib/TopDisp.do;jsessionid=KJJwTqwNqdjTQhk0LXGpGDkcLFTHGMJ2TxP4XhKP7DTZ1RyLqRQn!-838196174!10291480?bKind=10>

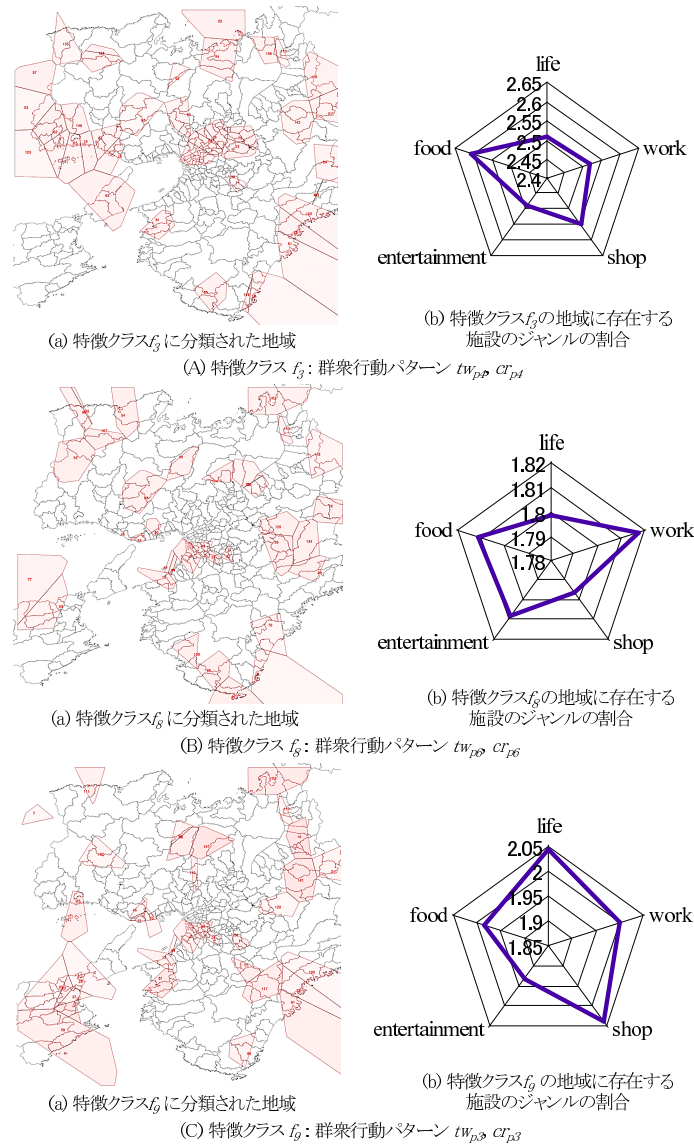


図 10 実験結果 (1): 群衆行動パターンによる地域分類と施設のジャンルを用いた意味付け

Fig. 10 Experimental result (1): Regions of latent urban characteristics and categories of local facilities.

合計を用いて、すべての地域を対象として各ジャンルに含まれる施設数の平均値を算出することにより、地域ごとに施設数の相対的な割合を求めている (式 (16)).

$$(C_{r_i} - \text{avg}(C)) / (\max(C) - \min(C)) + 2.0 \quad (16)$$

ここで、 C_{r_i} は地域 r_i に存在する施設の数を表しており、 avg 関数によって各カテゴリに含まれる施設数の平均を計算する。具体的には、各地域における施設数を他の地域と比較し、相対的にどの程度存在しているかの比率を算出し、レーダチャートを用いて表現した。

まず、特徴クラス f_3 に分類された地域に存在する施設のジャンルを調べたところ、図 10(A)(b) に示すように「食べる (food)」の割合が他の 4 つのジャンルに比べて高かった。この潜在的な特徴クラスは、朝 9 時から 12 時の時間帯に Tweets の発信数と群衆の数に特徴がある行動パター

ンとしてまとめられており、ランチタイムににぎわいを見せるような地域がまとめられていると推測することができる。特徴クラス f_8 に分類された地域における施設のジャンルは、図 10(B)(b) に示すように「働く (work)」の割合が他の 4 つのジャンルに比べて高かった。この潜在的な特徴クラスは、15 時から 18 時の時間帯に Tweets の発信数と群衆の数に特徴がある行動パターンとしてまとめられており、この時間帯はちょうど多くの会社の退社時刻や学校の講義の終了時間と一致しており、実際に、この特徴クラスにまとめられた地域には神戸や大阪などが含まれている。つまり、オフィス街などの地域が行動パターンによってまとめられていると考えられる。特徴クラス f_9 に分類された地域に存在する施設のジャンルに関しては、図 10(C)(b) に示すように「暮らす (life)」や「買う (shopping)」の割合が高く、次いで「働く (work)」や「食べる (food)」の割

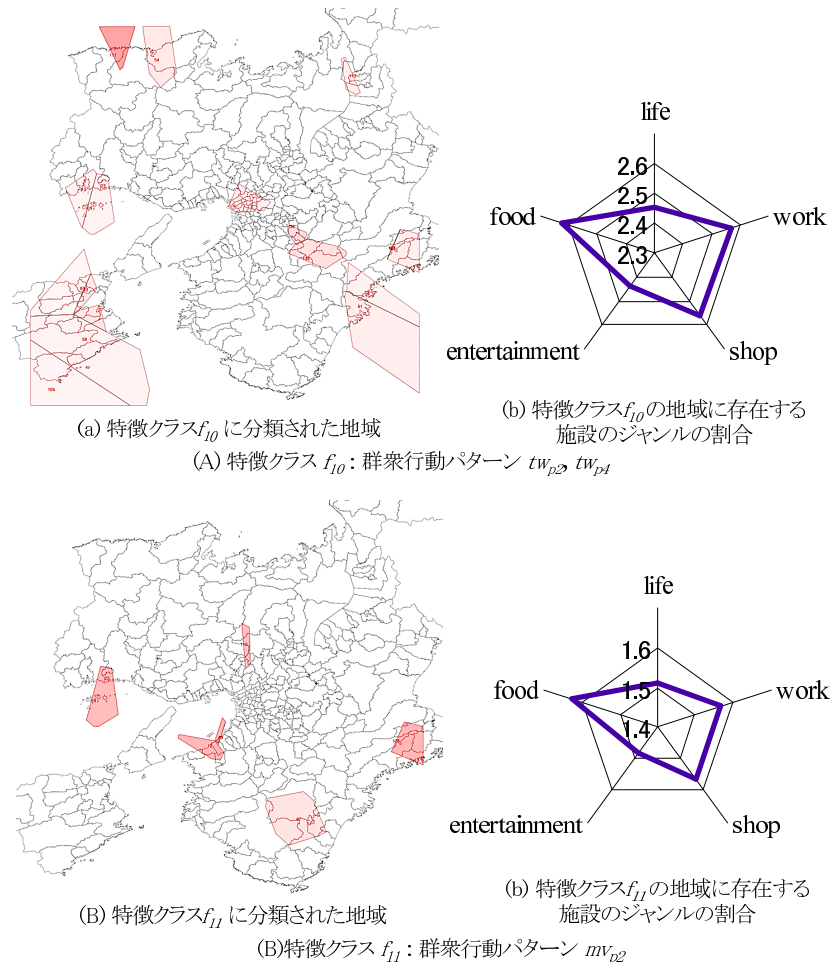


図 11 実験結果 (2): 群衆行動パターンによる地域分類と施設のジャンルを用いた意味付け

Fig. 11 Experimental result (2): Regions of latent urban characteristics and categories of local facilities.

合が高かった。この特徴クラスは、朝 6 時から 9 時の時間帯に Tweets の発信数と群衆の数に特徴がある行動パターンとしてまとめられているが、施設のジャンルの割合から住宅地を含んでいると予測することができる。

先に示した結果は、一般的に相関関係が高いと思われる tw_{pk} と cr_{pk} に基づくパターンであり、ある程度予測可能であった。そこで、次は、予測することが困難な活動パターンによってまとめられた潜在的特徴クラスについて、施設のジャンルを調べた結果を示し考察する。特徴クラス f_{10} には、深夜 3 時から 6 時の時間帯と朝 9 時から 12 時の時間帯に Tweets の発信数が増加するという特徴的な行動パターンが観測された地域が分類されている。図 11 (A)(a) に該当地域を強調して示す。このクラスの意味付けを行うために、該当地域に存在する施設のジャンルを調べたところ、図 11 (A)(b) に示すように「食べる (food)」の割合が最も高く、次いで、「買う (shopping)」と「働く (work)」の割合が高かった。この結果から、仕事や学校へ行く前に朝食をとったり、仕事や学校の休憩中に昼食をとったりする群衆活動を推定することが可能である。また、特徴クラス

f_{11} には、朝の 3 時から 6 時の時間帯に移動する群衆の数が増加するという行動パターンが特徴的な地域が分類されている。図 11 (B)(a) に強調して示した該当地域に存在する施設のジャンルは、特徴クラス f_{10} の場合と同様に、「食べる (food)」の割合が最も高く、次いで「買う (shopping)」と「働く (work)」の割合が高かった (図 11 (B)(b))。この結果から、この特徴クラスにまとめられた地域は短時間の移動が目立つ地域であり、仕事や学校へ行く前に立ち寄ることができる近場の商業地域などで観察されられる。このように、施設のジャンルのみを用いた場合には、施設のジャンルの割合が類似している地域として特徴クラス f_{10} と f_{11} に含まれる地域を扱うことになる。しかし、群衆行動を観察することで、都市空間・施設・群衆のライフスタイルの関係を考慮した地域特徴を抽出することができたといえる。

5. 議論：SNS を用いた社会調査

本研究では、位置ベース SNS に投稿されている時空間ライフログデータを用いて現実空間における群衆行動を

観察し、地域特徴を抽出する手法を提案した。社会生活調査、国勢調査やパーソントリップ調査のような従来の調査手法と比較すると、本提案手法で用いる Twitter のデータは母集団分布に対して当然偏りが存在すると考えられる。しかしながら、MarketingGum [15] の報告では、2011 年には 1 日平均 2 億ものメッセージが全世界から発信されていること、さらにその発信者の年齢層が若者中心から全世代へと広がっていることを示している。具体的には、18 歳から 25 歳では 13%、26 歳から 34 歳では 30%、35 歳から 44 歳では 27%、45 歳から 54 歳では 17%、55 歳以上では 9% ものユーザが Twitter を利用しているといわれている。このユーザ分布は全世界を対象に調査した結果であり、今回我々が対象とした日本におけるユーザ分布とは多少異なる可能性も考えられるが、都市社会を中心的に構成している世代が Twitter を積極的に活用しているユーザの世代と重複していることは明らかである。また、世代を区別して社会意見調査などを実施することも考えられるが、年齢などの個人データを大量の人々から直接的に得るには限界があるため、メッセージの内容を分析しユーザらの年齢や性別を推測する研究 [9] も行われている。

本研究では時空間ライフログデータを用いているが、取得可能なライフログデータ全体に占める時空間ライフログデータの割合がきわめて低いことは事実である。しかしながら、Sysomos [18] の調査では、2009 年には 44% のユーザらが、2010 年には 73% のユーザらが位置情報を明らかにして Twitter へメッセージを発信したという結果を得た。さらに、2012 年には GPS などの位置認識機能を標準搭載したスマートフォンの爆発的な普及にともない、より多くのユーザらからジオタグ付きのメッセージが発信されるものと予想される。

本研究で提案したアプローチは、従来の調査手法（社会生活調査や国勢調査、パーソントリップ調査など）とは大きく異なる。まず、データそのものに関して、従来手法は調査目的に応じたインタビュー、質問票やアンケートなどの手段を用いて、長い期間と膨大な費用をかけて収集したデータを用いている。一方、我々は SNS から取得したデータを用いており、これは日々の生活の中で大量の人々が様々な観点で発信しているライフログデータであるため、すべてのデータが意味を持つとは限らない。しかし、このようなユーザ参加型データは、データ規模、調査期間やコストなどに関わる従来手法の限界を克服し、実時間で大量の人々の経験や考えを集め、分析することができる理想的な調査環境を提供することを可能にしている。実際に、本研究では、日本全国の Twitter ユーザを対象にデータを収集して分析するシステムを構築し、地域特徴を抽出することで、その可能性を示すことができたと考えている。

6. おわりに

近年普及している位置ベース SNS を代表する Twitter を用いて都市空間における人々の活動を推定するために、群衆行動を分析し特徴的な行動パターンを抽出することで都市の特徴付けを行う手法を提案した。実験では、Twitter から大量のジオタグ付き Tweets を取得し、Tweets のユーザ ID、位置情報、時間情報を用いて地域と時間帯ごとに群衆行動をモデリングし、NMF アルゴリズムを適用して分析することで特徴的な行動パターンとそれに対応する地域からなる潜在的な特徴クラスを抽出した。そして、抽出した地域特徴を直感的に把握するために、各地域に存在する施設のジャンル情報を用いてその意味付けを行った。本論文では、Twitter に投稿されているジオタグ付き Tweets を用いているが、本手法は Foursquare や Facebook places のような他の位置ベース SNS において共有されているデータにも共通して含まれている基本的な時空間属性に着目し群衆行動をモデリングしている。それゆえに、Twitter に特化した手法ではなく、位置ベース SNS 全般に対して共通して適用することができる汎用性の高い手法であるといえる。

今後の課題として、都市空間で発生する様々な自然現象やイベントに関連する群衆行動のパターンを調査し、地域や時間を様々な粒度でとらえることにより、さらなる地域特徴や地域間の地理社会的な関係を分析する。また、今回は特定期間に発信されたすべての時空間データをクラスタリングすることで地域境界の設定を行ったが、休日と平日、特別なイベントが起こった期間や季節といった時間の粒度や空間的な粒度を考慮したデータのクラスタリングを行うことによる地域境界の変動やそれにとまなう地域特徴の変化、および目的に応じた最適化について分析・検討を行う計画である。

謝辞 本研究の一部は、第 8 回マイクロソフトリサーチ CORE 連携研究プログラムによるものです。ここに記して謝意を表します。

参考文献

- [1] Dempster, A.P., Laird, N. and Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistical Soc., Ser. B*, Vol.39, No.1, pp.1-38 (1977).
- [2] Aurenhammer, F.: Voronoi Diagrams – A Survey of a Fundamental Geometric Data Structure, *ACM Computing Surveys*, Vol.23, No.3, pp.345-405 (1991).
- [3] Baddeley, A.: *Analysing spatial point patterns in R*, CSIRO (2010).
- [4] Byers, S. and Raftery, A.: Nearest-neighbour Clutter Removal for Estimating Features in Spatial Point Processes, *Journal of the American Statistical Association*, Vol.93, pp.577-584 (1998).
- [5] Facebook Places, available from

- (<http://www.facebook.com/about/location>).
- [6] Finkel, R. and Bentley, J.: Quad Trees: A Data Structure for Retrieval on Composite Keys, *Acta Informatica*, Vol.4, No.1, pp.1-9 (1974).
 - [7] Foursquare, available from (<https://foursquare.com/>).
 - [8] Google Geocoding API, available from (<http://code.google.com/intl/ja/apis/maps/documentation/geocoding/>).
 - [9] KDDI 研究所: Twitter のつぶやきから年代や性別などが分かるプロフィール自動推定技術の開発に成功 (2011 年 2 月 2 日掲載), available from (<http://www.kddi.com/business/oyakudachi/square/labo/033/>).
 - [10] Kullback, S. and Leibler, R.: On Information and Sufficiency, *Annals of Mathematical Statistics*, Vol.22, No.1, pp.79-86 (1951).
 - [11] Lee, R. and Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection, *Proc. 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN '10*, pp.1-10 (2010).
 - [12] Lee, R., Wakamiya, S. and Sumiya, K.: Discovery of Unusual Regional Social Activities using Geo-tagged Microblogs, *World Wide Web (WWW) Special Issue on Mobile Services on the Web*, Vol.14, No.4, pp.321-349 (2011).
 - [13] Lloyd, S.P.: Least squares quantization in PCM, *IEEE Trans. Information Theory*, Vol.28, No.2, pp.129-137 (1982).
 - [14] Lynch, K.: *The Image of the City*, The MIT Press (1960).
 - [15] MarketingGUM, Twitter Statistics ? Updated stats for 2011, available from (<http://www.marketinggum.com/twitter-statistics-2011-updated-stats/>).
 - [16] McGill, R., Tukey, J.W. and Larsen, W.A.: Variations of Box Plots, *The American Statistician*, Vol.32, No.1, pp.12-16 (1978).
 - [17] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, *Proc. 19th International Conference on World Wide Web, WWW '10*, pp.851-860 (2010).
 - [18] Sysomos, Twitter Statistics for 2010: An in-depth report at Twitter's Growth 2010, compared with 2009, available from (<http://www.sysomos.com/insidetwitter/twitter-stats-2010/>).
 - [19] Tezuka, T., Lee, R., Takakura, H. and Kambayashi, Y.: Cognitive Characterization of Geographic Objects Based on Spatial Descriptions in Web Resources, *Proc. Workshop on Spatial Data and Geographic Information Systems (SpaDaGIS)* (2003).
 - [20] Twitter, available from (<http://twitter.com/>).
 - [21] Twitter Open API, available from (<http://apiwiki.twitter.com/Twitter-Search-API-Method%3A-search>).
 - [22] Wakamiya, S., Lee, R. and Sumiya, K.: Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from Twitter, *Proc. 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '11*, pp.10:1-10:9 (2011).
 - [23] Wakamiya, S., Lee, R. and Sumiya, K.: Crowd-powered TV viewing rates: measuring relevancy between tweets and TV programs, *Proc. 2nd International Workshop on Social Networks and Social Media Mining on the Web, SNSMW'11*, pp.390-401 (2011).
 - [24] Wakamiya, S., Lee, R. and Sumiya, K.: Towards better TV viewing rates: Exploiting crowd's media life logs

over Twitter for TV rating, *Proc. 5th International Conference on Ubiquitous Information Management and Communication, ICUIMC '11*, pp.39:1-39:10 (2011).

- [25] Wakamiya, S., Lee, R. and Sumiya, K.: Urban Area Characterization Based on Semantics of Crowd Activities in Twitter, *Proc. 4th International Conference on Geospatial Semantics, GeoS '11*, pp.108-123 (2011).
- [26] 経済産業省: 創業・起業促進型人材育成システム開発等事業大型閉鎖店舗再生等対策の総合プロデュース人材育成事業 I 都市の捉え方, available from (<http://www.meti.go.jp/report/data/jinzai.ikusei2004.10.html>).
- [27] 若宮翔子, 李 龍, 角谷和俊: Twitter-based TV Audience Behavior Estimation for Better TV Ratings, 電子情報通信学会第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011), Vol.1, pp.A1-4 (2011).
- [28] 手塚太郎, 李 龍, 高倉弘喜, 上林弥彦: ウェブ上の自然言語解析で描く都市のイメージ, 日本データベース学会 Letters, Vol.1, pp.22-30 (2002).
- [29] 西田佳史, 堀 俊夫, 本村陽一, 金出武雄: 日常空間における人間行動の観察技術とシミュレーション技術, デジタルヒューマンワーク基盤技術平成 15 年度成果報告書, pp.9-18 (2004).



李 龍 (正会員)

2001 年京都大学大学院情報学研究科社会情報学専攻修士課程修了。2003 年京都大学大学院情報学研究科社会情報学専攻博士課程修了。同年韓国 SAMSUNG 総合技術院入社。2008 年兵庫県立大学環境人間学部環境人間学科特任講師。2009 年兵庫県立大学環境人間学部環境人間学科特任准教授。2011 年情報通信研究機構専攻研究員。同年兵庫県立大学環境人間学部環境人間学科客員研究員。現在に至る。ソーシャルネットワーク, 群衆マイニング, ウェブ情報検索, 地理情報処理に関する研究に従事。博士 (情報学)。



若宮 翔子 (学生会員)

2011 年兵庫県立大学大学院環境人間学研究科博士前期課程修了。現在, 兵庫県立大学大学院環境人間学研究科博士後期課程在学中。日本学術振興会特別研究員 (DC2)。ソーシャルメディア, 群衆マイニングに関する研究に従事。日本データベース学会学生会員。



角谷 和俊 (正会員)

1988 年神戸大学大学院工学研究科修士課程修了。同年松下電器産業（株）入社。ソフトウェア開発環境，マルチメディアデータベース，データ放送の研究開発に従事。1998 年神戸大学大学院自然科学研究科博士後期課程情報メディア科学専攻修了。1999 年神戸大学都市安全研究センター都市情報システム研究分野講師，2000 年同助教授。2001 年京都大学大学院情報学研究科社会情報学専攻助教授。2004 年兵庫県立大学環境人間学部環境人間学科教授，現在に至る。博士（工学）。IEEE Computer Society, ACM, 電子情報通信学会，日本データベース学会等各会員。

(担当編集委員 牛尾 剛聡)