2012年度 修士論文

マイクロブログのジオタグを用いたユーザの行動分析

Analysis of Human Behavior of a User using Geotag of Microblog

酒巻　智宏

Tomohiro, Sakamaki

東京大学大学院 新領域創成科学研究科

社会文化環境学専攻

Abstract

リアル・ワールドと Web との距離が近くなる中で，Web からリアル・ワールドの情報を抽出する技術に注目が集まっている．本研究では，現在世界中で広く使われるようになった Twitter に注目し，Twitter とジオタグを用いることでユーザのリアル・ワールドでの日常的な行動と，行動の意図を抽出する．ただし，Tweet Data は膨大になるため，クラスタリングすることでデータを処理する．ジオタグ付き Tweet は，投稿位置と投稿時間，投稿内容を属性として持っており，これらの情報を利用し時空間＋テキストでクラスタリングを行い，ユーザの日常的に活動している地点を推定する．また，投稿内容を基にした位置へのラベリングにより，ユーザの行動について推定する．評価実験では，位置情報だけでなく，時間やキストを利用することで，より精度の良いクラスタリングができることを確認した．ただし，ユーザによって重視すべき項目が異なり，より精度を上げてクラスタリングを行うためにはユーザの分類が必要であると考えられる．また，700 以上のジオタグ付き Tweet があれば，ユーザの行動推定をある程度の誤差で行うことができることを示した．

より精度よくユーザの行動を抽出する手法の検討，そしてユーザの行動情報を利活用するアプリケーションの提案が課題として考えられる．

Recently, with the advent of portable devices such as smartphones, “context-aware services” that support the work and everyday life are greatly expected. Context-aware services are essential to understand the user’s every 24 hours behavior. I focus on Twitter now widely used around the world for analysis of user behavior context. “Geotag”, a feature of Twitter, stores the user behavior history so we can analyze user behavior context. In this study, geotag from Twitter is used to estimate the scope and patterns of user mobility. I aim to analyze user behavior context with clustering based on time, location and text of geotagged Tweets. Clusters labeling is based on the contents of Tweets as well. In the experiment, it was confirmed that clustering can be more accurate using time and text content than criteria. When using time and text, average error improve 100 meter comparing with using only location. As a result, using 700 or more geotagged Tweets, it is possible to estimate user behavior with some degree of error. In futre work, I try to improve the precision of clustering and labeling, and make some application which use human behavior data from my method.

目 次

第1章　序論

1.1　背景

Web は本来サイバースペース上で完結するものであり，リアル・ワールドとは別の世界であると考えられていた．しかし 2000 年前後より，インターネット接続環境の進歩，SNS(ソーシャルネットワーク) の登場などにより，Web はリアル・ワールドと密接につながり，融合し，補完しあうものとして存在している [1] [2] ．例えば，ブログは現実世界で起きた出来事を Web 上に書き込むものであり，SNS 上ではリアル・ワールド上の人間関係をそのまま移植し，交流が行われている．

さらに，新しい概念として，近年，小型で高性能の処理能力を持った携帯電話であるスマートフォンが登場した．スマートフォンの登場により，人々は場所や時を選ばず Web に接続し，情報を収集・投稿できるようになった．現在では，Web はよりリアルタイムに，また密にリアル・ワールドとリンクしている．

このような状況下で，Twitter や tumblr，plurk，emote，squeelr，identi.ca といったマイクロブログサービスが興隆している．これらのサービスは，利用者の状況・行動・思索を短文で投稿することで，利用者同士がコミュニケーションを図るサービスである．スマートフォンを用いることで，利用者は室内・室外を問わず，任意にマイクロブログに自身の状況を投稿できる．マイクロブログを閲覧することで利用者の一日の行動を知ることができるため，マイクロブログサービスはコミュニケーションツールとして人気を博している．

これらマイクロブログサービスは研究対象としても注目されている．例えば，マイクロブログから祭りや花火大会といったイベントを抽出する研究 [3] や災害を地震速報より早く検出する研究 [4] など，様々な研究が行われている．また，リアル・ワールドでの近接関係と Web 上での交友関係を調査した研究 [5] では，これら両者がある程度一致していることが示されており，これらの研究の優位性を示している．

このような研究は，多くの場合マイクロブログ全体からリアル・ワールドの情報を抽出する場合が多い．そこで本論文では，マイクロブログの個々の利用者に注目し，利用者ひとりひとりのリアル・ワールドでの行動をマイクロブログから抽出することに着目する．具体的には，ある利用者がどこに住んでおり，どこへ通勤通学し，どのような場所で遊び，食事をし，どのような交通機関を利用しているのかについて，情報を抽出する．

このような人々の行動情報は様々な分野において重要である．一つは，都市における人の流れの把握という観点である．交通計画，都市計画，政策立案などの分野において，そもそも人々が日々どのように移動し，経済活動を行なっているのか把握することは重要である．このよう

な情報を取得するために，現在はパーソントリップ調査 [1] が行われているが，調査間隔が長いこと，費用がかかることなど，問題点も指摘されており [6]，パーソントリップ調査の手法を改善・発展させる研究も存在している [7] [8]．また，地球規模で人がどのように行動しているかを調査する必要性も指摘されており [9]，人の行動調査は国ごとの対応では不十分となりつつある．

もう一つは，個人に対する行動アシストの観点である．スマートフォンや無線情報通信網の発達により，人々はいつでも現在の状況に応じて情報検索やナビゲーションなどの高度な情報サービスを受けることが可能になっている．この中で，生活や活動を支援する「行動文脈依存型」の情報サービス (Context-Aware Service) への期待が高まっており，実現するための試みも行われている [10]．Context-Aware Service の実現には，そもそもサービス利用者がどのような活動を行なっているかを把握する必要がある．行動情報を用いることで，サービス利用者の興味・思考に基づいた目的地予測や，おかれた状況や行動状態に応じた活動の支援を行うことができる．

1.2 目的と手法の概要

上記の背景を踏まえ，本研究の目的を「マイクロブログのデータより，個人の行動情報の抽出を行う」と定義する．具体的には,「ユーザが定期的にどこで活動しているか」「ユーザはその場所でどのような活動をしているか」の２点についてマイクロブログのデータを用いて解析を行う．例えば，平日は三鷹から新宿に通勤していて，休日は渋谷で買い物をしている，といった情報を取得する．

本研究では，数あるマイクロブログサービスのうち，利用者が最大である Twitter を対象とする．Twitter で利用者は，いま起きていることを Tweet と呼ばれる 140 文字以内のメッセージで投稿することでコミュニケーションを行う．

Twitter 利用者が「いつ」「どこで」「なにをしているか」という情報を Twitter から取得し，利用する.「いつ」に対応する情報は Tweet の投稿時刻から,「どこで」は Twitter のジオタグと呼ばれる機能,「何をしているか」は Tweet の内容を用いる．

これらの 3 つの情報を用いたクラスタリングにより，ユーザが頻繁に活動を行っている場所を推定する．さらに，取得されたクラスタに含まれる Tweet のテキストを用いて，その場所に家や仕事場等のラベル付けを行うことで，ユーザにとってその場所がどのような意味を持つ場所であるかを推定する．これにより，上記の研究目的を達成することを目標とする．

ただし，本研究の仮定として Tweet とリアル・ワールドの行動がリンクしていることを前提としている．また，行動分析の対象は Twitter 利用者のみに限定される．

1.3 本論文の構成

本論文の構成と各章の概要は以下の通りである．

[1] http://www.tokyo-pt.jp/

- **第 1 章「序論」**では，本研究を行うにあたっての背景について述べ，続いて本研究の目的を示す.
- **第 2 章「**Twitter **マイニングと行動調査の現状」**では，関連研究分野についてまとめ，本研究の意義について明らかにする.
- **第 3 章「ジオタグ収集システム」**では，本研究で必要となるジオタグ付き Tweet の収集方法について示す.
- **第 4 章「クラスタリングによるユーザの活動地点の推定」**では，ジオタグ付き Tweet からユーザの活動地点を推定する手法について考察する.
- **第 5 章「ラベリングによるユーザの活動内容の推定」**では，ユーザがその場所でどのようなことをしているかを推定する手法について考察する.
- **第 6 章「評価実験」**では，4 章，5 章で提案した手法の有効性を検証する.
- **第 7 章「考察」**では，評価実験の結果から得られる知見について考察する.
- **第 8 章「結論」**では，本研究で得られた知見をまとめ，今後の課題について示す.

第2章　Twitterマイニングと行動調査の現状

本章では，本研究に関連する研究分野について調査した結果を示し，さらに本研究と比較し本研究の意義について考察する．

まず,「マイクロブログと位置情報系のサービス」の節では，マイクロブログサービス，特にTwitter の現状について触れる.「マイクロブログや地理情報付きデータからの知識抽出」の節では，Twitter をはじめとして，Web に関連するデータセットを用いたデータマイニングに関する研究について紹介する.「人々の行動パターンの抽出と分析」の節では，人々の行動パターンについて行われている調査・研究について紹介する．

そして,「本研究の方向性の定義」の節では，これらの関連分野と本研究の目的を比較しながら，本研究の意義・新規性・貢献について考察し，研究の方向性を示す．

2.1　マイクロブログと位置情報系サービス

本節では，マイクロブログと位置情報系サービスの現状について紹介する．

2.1.1　マイクロブログサービス

マイクロブログ（ミニブログとも）は，短いメッセージを投稿することで，コミュニケーションを行うサービスである．2000 年頃から，容易に Web 上に日記を投稿できるウェブログ（ブログ）サービスが注目され，ユーザの Web 上での情報発信活動に対する敷居を大きく下げた．しかし，ブログは記事を書くために大きな労力を必要とすることが欠点であった．そこで，記事を書くための労力を軽減し，ユーザが Web 上で情報発信をより簡単に行える，マイクロブログが登場する．マイクロブログは，ブログと比較して一つの投稿が短文であり，かつ投稿頻度が高いことが特徴である．登場当初は大きく注目を集めることはなかったマイクロブログであるが，2006 年に **Twitter** が登場すると徐々に利用者数が増え始め，2009 年頃に爆発的にヒットし，マイクロブログが一般に認知されることになる．

現在，マイクロブログサービスは，Twitter 以外にも情報収集をメインとした **tumblr**，時間軸を重視した **plurk**，オープンソースの **identi.ca** など様々なサービスが存在する．しかし，この中で 2012 年 1 月現在最も利用者数が多いのは Twitter である．

2.1.2　Twitter とジオタグ

Twitter はいま起きていることを Tweet と呼ばれる 140 文字以内のショートメッセージで投稿しコミュニケーションを行う，マイクロブログサービスである．Twitter は，2011 年 11 月現在で日本国内で約 1400 万人が利用しており，2011 年 9 月には全世界でのアクティブユーザ数

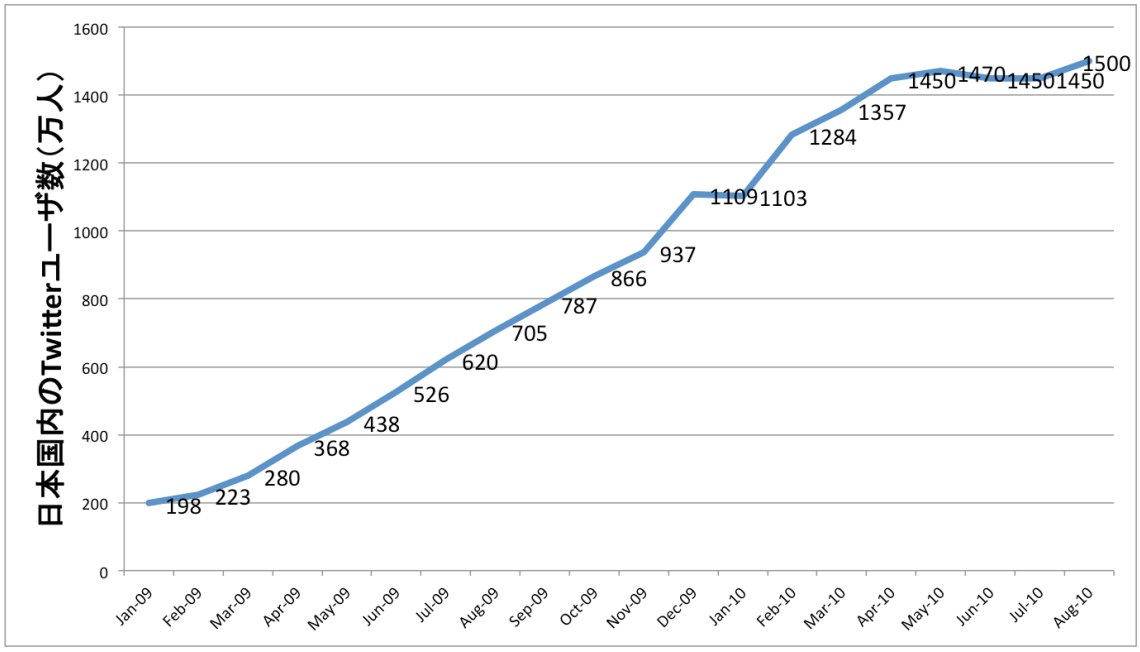

Fig. 2.1　国内の Twitter ユーザ数の動向 (ニールセン社の調査を参考に作成)

が 1 億人を突破した.

Twitter の利用方法について調査した研究 [11] [12] [13] では，Twitter の利用目的として他のユーザとのコミュニケーションだけでなく，情報提供・情報収集，ニュースに対する意見や考えの投稿など，様々な用途で利用していることをアンケート調査を通して示している．また，Twitter の利用方法に関する考察が田中らによって行われている [14] ．この研究で筆者は一つの Tweet は以下の 4 つの属性で分類できるとしている.

- 情報発信性 (自分の状況をただ発信するだけの自己完結型か，他のユーザーが見ることを意識した発信か)
- リアルタイム性 (速報性があるか)
- 社会性 (専門的な内容か，一般的な内容か)
- 有用性 (情報としての有用性もしくはネタとしての有用性があるか)

また，同様にユーザ自身の分類として，ユーザは以下の 5 つの属性で表されるとしている.

- 利用目的 (コミュニケーション目的もしくは bot，またはマーケティングなどの広告宣伝など)
- 有名度 (ユーザのリアル・ワールドでの有名度)
- 嗜好 (趣味や興味の差)

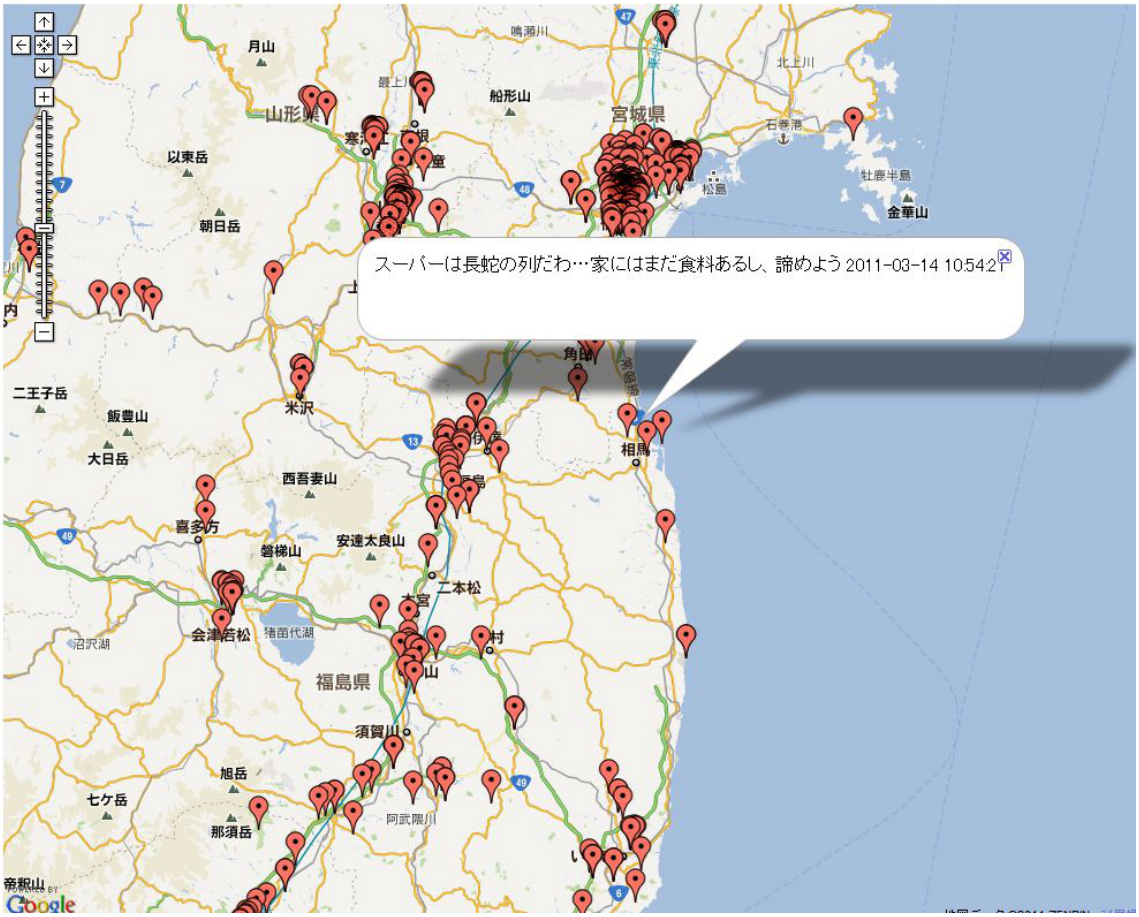

Fig. 2.2　東日本大震災時の Tweet 例

- 情報収集意識 (情報収集のために Twitter を使っているか)
- 情報発信意識 (情報を発信する意図があるか)

このように，Twitter の使われ方は多岐に及ぶ．近年では，中東で革命のための情報共有手段として Twitter や Facebook が利用されたことが記憶に新しい．

単なるコミュニケーションツールとしてでなく，情報共有・議論の場として活用されている Twitter は，2011 年 3 月の東日本大震災の際にも重要な情報共有インフラとして注目された．生活必需品の店頭での在庫状況や行方不明者の生存情報，また阪神大震災経験者からのアドバイスなど有用な情報が共有された．これらの情報は，時にはジオタグと呼ばれるユーザの現在位置情報を付加し，その情報がどの場所に関連するものであるかを示して投稿された．図 2.2 は東日本大震災の際に情報共有された情報を地図上にマッピングしたものである．

ジオタグは，2009 年 11 月にリリースされた位置情報付き投稿を可能にする Twitter の付加機能である．Tweet がどの地点で行われているかを付加することで，今，”どこで ”，何をしているか？という情報を他のユーザに伝えることができる．ジオタグにより，リアルタイムなコミュニケーションが可能な Twitter において，より実世界とリンクしたコミュニケーションが

可能となった．

ただし，ジオタグの利用者数は少ない．フランスの Semiocast の調査では，ジオタグ付き Tweet の全 Tweet に対する割合は全世界で 0.5 %であるとしている．全ジオタグ付き Tweet のうち，最もジオタグ付き Tweet 数が多かったのは米国で，全体の 43 %を占める．日本は 7 %で，インドネシアより少ないという結果になった [15]．自分の位置を他人に公開することに対してプライバシー上の懸念を抱く人は多く，ジオタグの普及を妨げていると考えられる．

2.1.3 位置情報系サービス

Twitter のジオタグのように，位置情報を用いることでユーザの利便性を高めるサービスは近年増加している．位置情報を投稿することでユーザ同士のコミュニケーションを図ることを狙いとしたサービスとして，**Foursquare**，**Gowalla** などがあげられる．これらのサービスは，スマートフォンの発展に伴いユーザ数が伸びており，研究対象としても注目されている．

既存のサービスに位置情報を追加することで，利便性を高めた例としては，**yelp**，**食べログ**，**ぐるなび**，**30min** などがあげられる．これらのサービスは，従来は自宅など屋内での利用が主であった．しかし，位置情報を利用することで，現在位置から近い店を探すことが出来るなど，外出先での利用も可能になりユーザの利便性が飛躍的に高まった．

コロニーな生活や**国盗り合戦**など，位置情報を利用したゲームも流行している．このようなサービスは，単純に娯楽としての側面だけでなく，特定の地域でしか入手できないアイテムを用意することにより地域に人を呼び込むという，いわゆる町おこしの面からも注目されている．

さらに，自分の位置情報をロギングすることによるライフログアプリケーションも注目されている．例えば，ランニングの履歴を残すことができるアプリケーションなどが利用されている．ライフログと位置情報は相性がよく，多くのライフログ系アプリケーションで位置情報は活用されている．

位置情報サービスはユーザに対して新しく利便性の高いサービスを提供するが，一方で問題もある．それは，プライバシーの問題である．位置情報を利用したアプリケーション・サービスは，ユーザのプライバシーの管理が問題となる場合が多い．実験用アプリケーションを用いて被験者の位置情報公開に対するプライバシー意識について調査を行った研究 [16] では，駅などの様々な人が訪問する場所はプライバシー意識が低く，自分の位置情報を公開することに対する真理障壁は低いが，自宅や職場など，その場所を訪問する場所が少ない場所に自分がいることを公開することに対する真理障壁は高いことが示された．また，ユーザは自身の位置情報を公開する範囲や対象などを詳細に設定できることを好むことも実験結果から分かった．位置情報をを利用したサービスやアプリケーションを作成する，または，このような情報を利用して研究を行う際には，ユーザのプライバシーに対して十分に考慮することが求められる．

2.2 マイクロブログや地理情報付きデータからの知識抽出

本項では，Twitter をはじめとして，Web に関連するデータセットを用いたデータマイニングに関する研究について紹介する．

2.2.1 地理情報付きデータセットからの知識抽出

地理情報付きデータセットから，有益な情報を抽出する研究例は多く存在する．

渡邊らは，地理情報に基づく画像管理方法を提案している [17] ．この研究では，画像のグルーピングにおいて，写真が撮影された地点の位置情報を用いてクラスタリングすることで，写真管理の労力を削減できると主張している．この研究と同様に，位置情報を用いてグルーピングを行うことで有益な知識を抽出する研究例は多い．

また，Lovett らは，カレンダーの情報とリアル・ワールドでの活動の不一致に注目し，位置情報や SNS を用いて現実世界の活動をセンシングし，カレンダーの補正を行った [18] ．この研究では，室内に設置したセンサーによってセンシングした情報を解析することで，被験者のカレンダー上での行動を補正することに成功している．この研究のように，センシングした位置情報データを用いて新たな知識を抽出する研究も近年多く行われている．特に，スマートフォンの普及により，センシングのためのデバイスを容易に利用できるようになり，大規模なセンシングを行うことで新たな知識獲得を目指す研究も可能である．

近年では，集合知的アプローチによるユーザ参加型の地図，OpenStreetMap も注目されている．OpenStreetMap は，自由に作成・改変・閲覧する事のできる地図であり，企業に依存しないために利用に制限が少ないことが特徴である．また，OpenStreetMap の地図情報を Linked Data 化し，再利用を容易にした研究もある [19] ．Linked Data とは，情報同士がリンクで接続され，コンピュータでの処理を容易に行えるようなデータのことを指し，Web 上に存在する莫大な情報を適切に扱うための重要な要素技術の一つである．このようなデータセットを用いることで，コストを低く抑えた知識抽出が可能になる．

2.2.2 Twitter からの知識抽出

前述の通り，Twitter は今までにない新しさを持ったボリュームのあるデータセットとして注目されており，データマイニング，自然言語処理，その他多くの分野で研究が行われている．この項では，Twitter に関する研究のうち，特にデータマイニングに関する研究を紹介する．

1 章でも述べたように，ジオタグを用いた Tweet 解析は，Twitter 全体から情報を取得するものが多い．例えば，岡崎らの研究 [4] が挙げられる．この研究は，Twitter から知識を抽出することで災害を地震速報より早く検出することに成功している．この研究では，まず Tweet を分類器を用いて地震に関するものどうか，分類する．そして，地震に関連する Tweet の投稿者のプロフィールの位置情報をカルマンフィルタとパーティクルフィルタを通すことでフィルタリングし，地震の震源を求める．この研究は Twitter からリアル・ワールドの事象を抽出できることを示している．さらに，虹や台風も同様の手法で検出できることを示唆している．この研究は，**Toretter** という名前でサービス化されている (図 2.3)．

Fig. 2.3　Toretter:Tweet から地震を検知するサービス

また，Twitter とジオタグから祭りや花火大会といったイベントを抽出する研究 [3] も行われている．この研究は，ジオタグを用いて地域のイベントを検出する．通常時とイベント時の Tweet 数の差分を考慮することで，ある地域の異常な Tweet 数を検出し，その地域に何らかのイベントが起きていることを推定する．また，アイルランドでも同様の研究が行われている [20]．

また，Singh らの研究 [21] では，災害や地域イベントだけでなく，インフルエンザの流行状況も Twitter から抽出することができることを示している．

データソースとして注目されているジオタグだが，上記で指摘したように，ジオタグを利用しているユーザ数が少なくデータのボリュームが不足している．そこで，ジオタグが付加されていない Tweet から，ユーザの位置情報を推定しようという研究例もある [22]．この研究では，Tweet の内容を自然言語処理で解析することで，地域と Tweet 内容の相関関係を見出し，それを元にユーザの住居地域を推定する．

また，Tweet から都市の特性を抽出する試みも行われている [23]．この研究では，昼もしくは夜に Tweet 数が増える地域を発見することで，その地域がビジネスタウンなのかベッドタウンなのかを推定できることを示した．

このように，Twitter からの知識抽出は，人間の行動に関するものだけでなく，災害やイベント，病気の流行など，多岐に渡る分野で行われている．研究事例の豊富さが Twitter のデータソースとしての価値の高さを示していると言える．

2.3　人々の行動パターンの抽出と分析

本節では，人々の行動パターンの抽出と分析に関する分野として，パーソントリップ調査，携帯電話基地局を用いた行動調査，そして行動の目的抽出について紹介する．

2.3.1 パーソントリップ調査

本研究は人の行動を推定することを目的としているが，人々の行動に関する調査として日本では東京や大阪，広島，札幌などの各都市圏にて 1 章でも紹介したパーソントリップ調査が行われている．パーソントリップ調査は，災害や経済活動のために都市空間における日々の人々の流れを把握し，都市を可視化することを目的としている．パーソントリップ調査の結果は様々な分野・領域で利用される．パーソントリップ調査の結果を用いた研究例として，インフルエンザの伝搬の数式モデリングを行った研究 [24] や 都内の分娩施設のアクセス評価を行った研究 [25] などが挙げられる．

パーソントリップ調査では，調査票に個人の 1 日における移動状況を記入してもらい,「いつ，誰が，どのような方法で移動しているか」を把握する．パーソントリップ調査は全国で行われており，首都圏では東京都市圏交通計画協議会が昭和 43 年以降，10 年に一度のスパンで調査を行っている．前回調査は 2008 年 10 月頃に約 500 万人を対象として行われた．

都市における経済活動を把握するという目的において，パーソントリップ調査は有効な調査である．しかし，調査上の問題点も多く指摘されている [6] ．主に指摘されている点として，以下の項目が挙げられている．

- 多くの被験者に対して調査を行うために，多大な調査費用がかかる．
- 人員，予算の都合上，頻繁に調査を行うことができない．
- 個人情報保護意識の高まりや振り込め詐欺の影響により，被験者の同意を得にくい．
- 調査員の負担への不満，賃金の不満（調査内容が複雑で，対象者への説明も難しい）

これらの問題を解決する手法として，GPS 付き携帯電話端末によるパーソントリップ調査 [7] が提案されている．この研究では，アンケート調査の記入漏れ，時刻情報の不正確さ，コストなど，指摘されているパーソントリップ調査の問題を解決する手法として，パーソントリップ調査に GPS を利用することを提案している．これにより，作業負担の軽減と費用の削減が可能になる．

国や自治体で行われる大規模なパーソントリップ調査以外にも，あるイベントや観光スポットなどで試験的に行われたパーソントリップ調査も存在する．四国で観光客の観光スポット立ち寄りについて調査した研究 [8] では，携帯電話の GPS を利用して，被験者の交通モード，立ち寄り場所を高精度で抽出した．このような研究は，GPS ロガーや携帯電話などの端末に加えて，Web 上や紙ベースのアンケート調査を行う場合が多い．これらの研究の多くは GPS 端末を用いたものであるため，被験者数が限られてしまうという問題点があるが，少数の被験者でイベントや特定地域のような局所的な空間における人々の行動調査においては有効であるといえる．

これら，GPS 端末を用いてパーソントリップ調査を行う研究は，被験者の負担軽減や記憶に頼らない調査が可能である点など，様々な利点がある．ただし，欠点もある．携帯電話の GPS

特性についての研究 [26] で示されているように，GPS は位置情報を正確に取得できない．この調査では，GPS の精度は端末によるが屋外で 10m から 20m 程度，屋内だと 200m 程度，最大で数キロずれるという結果となった．特に，市街地ではビル陰や道の密集性などから，誤差が生じてしまう．また，現在のスマートフォンを始めとした携帯端末では，バッテリーの問題から GPS 機能を長時間利用することは難しいといった欠点もある．また，実験用端末の確保も難しい．一版に GPS をロガーは高価であり，被験者数を多数確保することは難しい．また，プライバシーの観点からも，調査に個人の携帯電話を調査に使用することに対して合意を取ることは難しいといった問題点も残る．

2.3.2 携帯電話基地局を用いた行動調査

前項ではパーソントリップ調査についての概略と問題点について示した．この中で，近年携帯電話基地局を利用した行動調査の手法が注目を浴びている．

NTT ドコモは，モバイル空間統計と題して携帯電話基地局情報を用いた人々の行動調査を行なっている [27]．携帯電話は，各ユーザの携帯電話と電話やメールを行う際に基地局と通信を行う．この情報を元に，ユーザがある時間にどの携帯電話基地局の周辺に滞在しているかを知ることができる．よって，基地局の粒度でユーザの滞在地点と時刻を把握できる．また，日本人の多くは携帯電話を利用するため，調査の母数も大きく有用な調査が可能になる．近年では，東日本大震災の際の人々の動きを可視化した調査が注目を浴びた (図 2.4)．

また，米国の通信会社 AT & T と共同で行った Sibren らの研究 [28] では，通信会社の協力の下，同様の分析を行なっている．実験はニューヨークとロサンゼルスで行われ，対象地域のユーザをランダムでサンプリングし各ユーザの 78 日分の通信記録を収集した．また，アンケート調査を行い，37 人の行動データを収集し，正解データとして評価実験を行なった．まず，行動地域を推定するために，閾値を 1mile とし，基地局データをクラスタリングする．その後，各クラスタに対してロジスティック回帰を用いて重要度を設定し，重要度の高いクラスタをその人の主要な行動地域として抽出する．ロジスティクス回帰のための特徴量としては，総通信時間や通信間隔，総通信回数などを利用している．また，論文内では，このような行動調査の結果から，ユーザの通勤時間を推定したり，カーボンフットプリントを調査したりしており，携帯電話基地局情報を用いた行動調査の利用方法について提案している．

このような，携帯電話基地局情報を利用した研究は，近年世界中で行われており，ひとつのトレンドとして注目されている．

2.3.3 行動目的の抽出

近年では，単純にユーザの行動した地点を推定するだけでなく，なぜユーザがその場所へ移動したかというユーザの行動目的を推定する研究が行われている．ここでは，行動目的の抽出を行なっている研究を挙げ，行動目的抽出に関する研究の現状について考察する．

その位置が個々人によってどのような意味を持つかを，被験者実験のデータを用いてモデル化した研究が行われている [29]．この研究は，人々がある位置をどのように名付けるかを分析

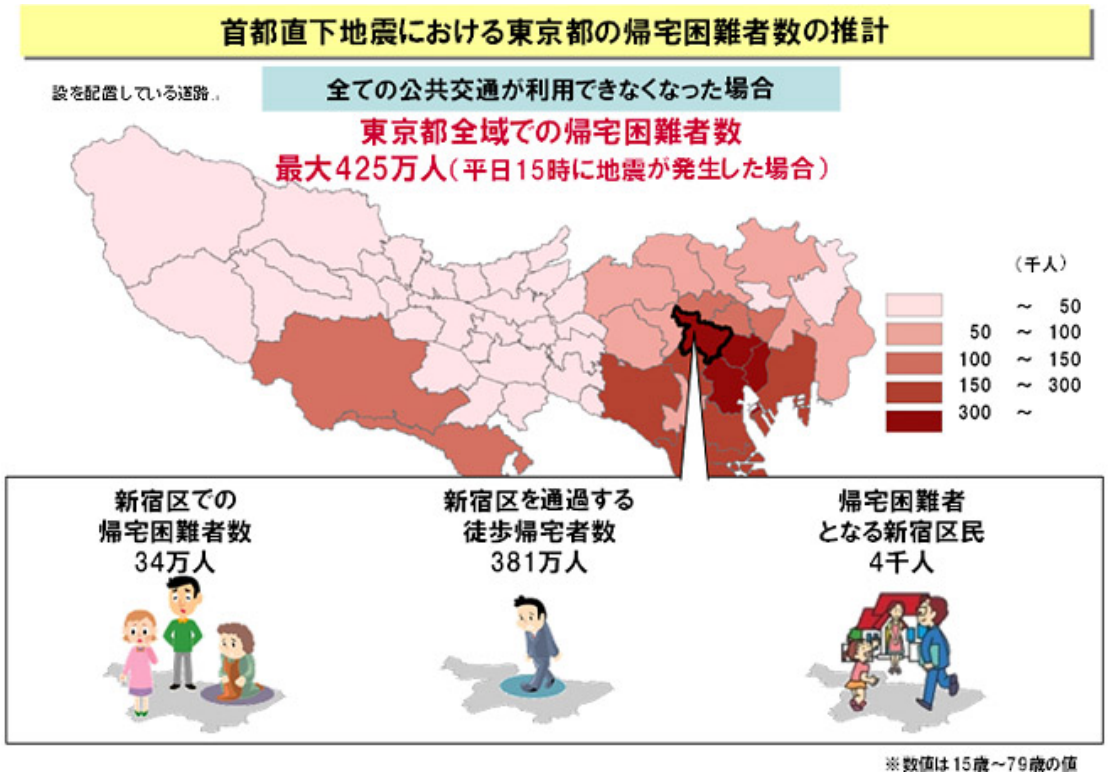

Fig. 2.4　東日本大震災の際の携帯電話基地局を利用した人々の行動分析（NTT ドコモプレスリリースより引用)

したものであり，ある物理的な位置に対して複数の名前を付けられる可能性が高いことを示したものである．例えば,「A 大学」は，学生にとっては「学校」であるが，教員にとっては「職場」である．この事実は，個人によって，位置は様々な意味を持つことを意味する．この研究では，まず GPS を被験者に持たせ，被験者が活動している地点を収集する．その後，被験者の活動している地点に対して被験者が「学校」「職場」といったようにユーザが任意に名前をつける．そして，それらの地点を

- Semantic(「自宅」「自分の職場」といったように文脈を踏まえた意味的な命名規則による名前)
- Geographic(地名や住所のように，地理的な命名規則による名前)
- Hybrid(両者の混合)

の 3 つに分類する．そして被験者がどのように自分の活動している位置に対して名前を設定するかを分析する．

実験結果では，ある位置に対して複数のユーザは様々な命名を行い，ひとつの場所に対して平均で 2.78 個の名前が付けられた．また，被験者は自身のプライバシーに関する位置（自宅など）については，正確な情報を記述したがらない傾向があることも示された．

筆者らは，実験結果を元に，位置に対して自動で名前付けをすること目標にしている．しかし，目標を実現するためには幾つかの問題があると論文内で指摘している．まず，名前のカテゴリを自動で推定するモデル精度が悪い．これは，GPS の精度が悪いという点や，位置情報だけでは位置に対して名前を自動でつけるための情報量が少ないという点などが問題となっている．このため，精度良く位置に対して名前付けを行うためには，位置情報だけでなく他の情報を利用することが必要であると指摘している．一例として，SNS 等を用いた草の根的手法はコストもかからず有効であるとしている．

また，Web 上に存在するテキスト解析することにより，ユーザの行動目的を抽出する研究も行われている [30] ．この研究では，Web 上に存在する人や車両などの移動情報を収集することで，交通の流れを把握することを目的としている．質問回答サイトの情報を利用し，形態素と品詞を素性とした SVM により，ある質問に現れる移動目的を交通センサス [1]にて定義されている１６種類の移動目的に分類する

Web から人間行動を抽出する研究 [31] では，条件付き確率場と自己教師あり学習で行動主，動作，対象，時間，場所を自動で抽出している．また，Twitter やブログのようなライフストリームから同様に行動を抽出する研究例もある [32] ．この研究では，行動調査のために動詞を行動動詞，非行動動詞に分け，さらに行動動詞は外面的行動動詞と内面的行動動詞に分け，テキストから行動に関する動詞のみを抽出する，．さらに，助詞や助動詞を元にテキストを what,when,where,whom,how の 5 つのパターンに当てはめ，知識の抽出を行う．この研究で，動詞の分類に利用されているのが行動辞書 [33] である．この辞書は，テキストから行動を抽出することを目的として作成されている．上記の研究例のように，この辞書を用いたパターンマッチングも行動目的抽出の手法として利用できる．

2.4 本研究の方向性の定義

本節では，ここまで概覧してきた本研究に関連する分野の現状をまとめ，本研究の目指す方向性を定義する．

「マイクロブログと位置情報系サービス」では，マイクロブログサービスの現状と，マイクロブログサービスの代表としての Twitter，そして Twitter の機能であるジオタグについて紹介した．この節で示したように，現在マイクロブログサービス，特に Twitter のユーザ数は増加しており，研究分野にとどまらず，社会的，文化的な面からも注目されている．特に研究分野においては，社会状況を示すデータセットとして注目されている．

本研究の目指す Twitter を用いた行動調査においても,「今，何をしているか」を投稿する Twittert は有効であると考えられる．特に，ユーザの具体的な現在位置を取得できるジオタグを用いることで，高精度の行動調査ができると予想される．

ユーザの位置情報が付加されるジオタグ付き Tweet は，ユーザの日々の行動履歴の性質を持っていると考えられる．また，ジオタグ付き Tweet は，通常の Tweet と同じくユーザが「今

[1]国土交通省が実施している交通状況に関する調査

何をしているか？」という文字列情報も付加されている．これらの情報から，その位置がそのユーザにとってどのような意味を持つかを推定することが可能であると考えられる．例えば「会議始まった」というジオタグ付き Tweet があれば，その地点はそのユーザにとって仕事に関連する場所であることがわかる．ただし，ジオタグを利用しているユーザは少ないことを考慮しなければならない．

「マイクロブログや地理情報付きデータからの知識抽出」では，マイクロブログ，特に Twitter に関連する研究について紹介した．Twitter に関する研究は豊富に存在する．また Twitter から有用な知識を抽出する研究では，Twitter とリアル・ワールドとの関連性について示している．これらの研究は Twitter 全体から情報を取得する例が多く，個人の情報を利用する例は少ない．また，ジオタグを利用している研究例は少ない．このため，本研究では，個人をターゲットとして Twitter マイニングを行い，個人の行動調査を行うことを目標とする．

ここで，Twitter を用いてどのような行動調査が可能であるかを考察する．行動調査には，日々の周期的な行動の調査と，ある地点から地点への特定の移動とその手段の調査がある．前者の研究例としては，上記でも紹介した [28] 等が挙げられ，後者としては [34] [35] [36] [37] といった研究が挙げられる．

Twitter は，ユーザの投稿は非連続かつ任意のタイミングで行われることが特徴として挙げられる．そのため，Twitter から連続的なデータを得ることは難しい．行動調査においても，ある地点からある地点への移動，またその手段（移動モード）を判定するような研究は難しい．一方で，ユーザの Tweet を長期間観測することで，日々の周期的な行動を取得することは可能である．本研究では，行動パターン抽出のような，ユーザの周期的な行動の抽出を目標とする．

日々の習慣的な行動の抽出には，データマイニングの手法が有効であると考えられる．そこで本研究では，クラスタリングを利用することで雑多なジオタグ付き Tweet から有効な知識を抽出する．ジオタグ付き Tweet は，主に位置情報，時間情報，投稿内容の 3 つの情報を持っている．クラスタリングには，これら 3 つの属性を用いる．

位置情報と時間情報を用いたクラスタリングは時空間クラスタリングと呼ばれ，病気の伝搬 [38] といった空間疫学や，犯罪学などの研究分野で主に研究が行われている．ただし，ジオタグ付き Tweet のような，位置情報，時間情報，投稿内容という属性を持つデータのクラスタリングに関する研究例はまだ少ない．

特にテキスト情報を扱ってクラスタリングを行う研究例として，ユーザの投稿によりリアルタイムにイベントを検知する研究 [39] を挙げる．この研究では，位置情報とテキスト情報を利用してクラスタリングを行なっている．具体的には，テキストを分類器を用いてカテゴリに分け，両者のコサイン距離によって結合するクラスタを決定する手法を用いている．しかし，この研究では位置情報とテキストの重み付けのバランスについては考慮していない．そこで，本研究ではジオタグ付き Tweet のようなデータセットを扱う際に位置情報，時間情報，テキスト情報をどのような重みで利用するべきかについて考察する．

「人々の行動パターンの抽出と分析」では，行動調査に関する既存研究について紹介した．

近年では，パーソントリップ調査の代用として，携帯電話基地局を利用した行動調査が注目されている．携帯電話基地局を利用した手法は，多数の人々に対して行動調査を行えることが大きな利点である．

そのような研究に対して，本研究でも同様の調査を行うことを目的としつつも，近年研究が進んでいる行動文脈の調査も行うことも目的とする．具体的には，ユーザが頻繁に活動する地点に対して，その位置が家なのか，大学なのかといった「行動の意図」を推定する．本研究は，Twitter のデータを利用するため，Tweet の投稿内容を用いて詳細に行動の意図を推定することを目標にする．

行動目的の抽出のために，ここでもデータマイニングの手法を用いる．本研究では，分類器を用いたクラス分類によって，ある地点のユーザの活動を，家・職場・学校・娯楽・移動などのラベルに分類することで推定する．

ここまで，既存研究を紹介し，本研究の方向性について考察を行った．ここまでの考察をまとめると，本研究の目指す方向性は

- 個人のジオタグ付き Tweet を元に，ユーザの日々の行動を，位置情報，時間情報，テキスト情報を持ったジオタグ付き Tweet をクラスタリングすることで抽出する
- Tweet の投稿内容を用いて，その地点でのユーザの活動を分類し，ラベリングする

となる．また，既存研究を踏まえた上での，研究分野に対する本研究の貢献としては

- Twitter から個人の行動を推定する手法の提案．具体的には，ユーザがどこで活動しているか，その場所でどのような活動をしているかの把握をジオタグ付き Tweet を用いて行う．
- ジオタグ付き Tweet のように，位置情報と時間情報，テキスト情報を持つ対象に関して，どのようにクラスタリングを行うべきかの考察

といった点が挙げられる．

また，行動調査の対象としては，まずは日本国内のユーザを対象とする．具体的には，関東圏で主に活動を行なっているユーザを対象として，行動調査を行なっていく．

次章以降で，この目的を達成するためのデータセットの収集，手法の提案，評価実験について述べていく．

第3章　ジオタグ収集システム

ジオタグ付きTweetを用いた行動調査のために，まずTwitterからジオタグ付きTweetを収集する．本章では，ジオタグ収集のためのシステムについて説明する．

3.1　Twitter APIの利用

Twitterに関するデータは，Twitter社が提供するAPI [1]を通して自由に利用することができる．ただし，一時間あたりのAPIコール回数や取得できる過去Tweetに制限があるなど，Twitterが保持しているログのすべてを利用することはできない．Twitter APIから取得されるデータは，図3.1のようにjsonもしくはxml形式で得られる．

Twitterには，Streaming API，Search API，REST APIなどが存在するが，本研究では，このうちREST APIとSearch APIを利用する．また，APIをコールするためのプログラム言語としてPythonを，データベースとしてMySQLを利用している．

REST APIは，通常の投稿，閲覧などを行う機能を持つAPIである．本研究では，REST APIを用いて各ユーザのTweetを収集している．

また，Search APIは検索機能を持ったAPIである．Search APIを用いて，ある地点の周辺に分布しているTweetを収集するシステムを作成し，ジオタグ付きTweetを収集している．具体的には，本研究で対象としている東京都市圏のジオタグ付きTweetを収集している．

3.2　ジオタグ付きTweet収集・閲覧システム

上記のAPIを用いて，ジオタグ付きTweetを収集した．また，これらのデータを可視化するためのアプリケーションを作成した．

図3.2は，本研究で利用しているジオタグ付きTweet収集・閲覧システムの概念図である．

本研究で用いたシステムは，大きく

- ジオタグ収集システム：Twitter APIをコールしてジオタグ付きTweetを取得するシステム
- ジオタグ閲覧システム：クライアントからのリクエストに応答し，収集したジオタグを可視化するシステム

の2つに分かれる．

[1]Application Program Interfaceの略．あるシステムやプラットフォームを操作するための手続きを定めた規約

Fig. 3.1　Twitter API の取得結果の例

Table 3.1　ジオタグ収集システムにて集めたジオタグ付き Tweet のデータ

項目	値
収集期間	2010/05/21 から 2012/01/30 まで (588 日間)
収集した全ジオタグ付き Tweet 数	8,224,805
一日あたりの平均 Tweet 数	約 13,988 Tweets

まず，Twitter API をコールし，ジオタグ付き Tweet を収集するジオタグ収集システムについて説明する．このシステムは大きく 3 つのモジュールに分かれる．はじめに，モジュール A が Twitter Search API を用い，東京の控除を中心として半径 30 ｋｍ以内で投稿された Tweet を検索し，該当する Tweet をデータベースに保管する．次に，モジュール B は収集した Tweet をチェックし，頻繁にジオタグ付き Tweet を投稿しているユーザを発見する．具体的には，100Tweet 以上の投稿を行なっているユーザを抽出し，ユーザ情報をデータベースに保管する．さらに，モジュール C はモジュール B が抽出したユーザの全 Tweet を収集し，データベースに保管する．この 3 ステップで，ジオタグ付き Tweet を投稿するユーザを発見し，各ユーザの Tweet を収集する．

本ジオタグ収集システムを用いて収集したジオタグのデータを表 3.1 にまとめた．

次に，クライアントからのリクエストに応答し，収集したジオタグを可視化するジオタグ閲

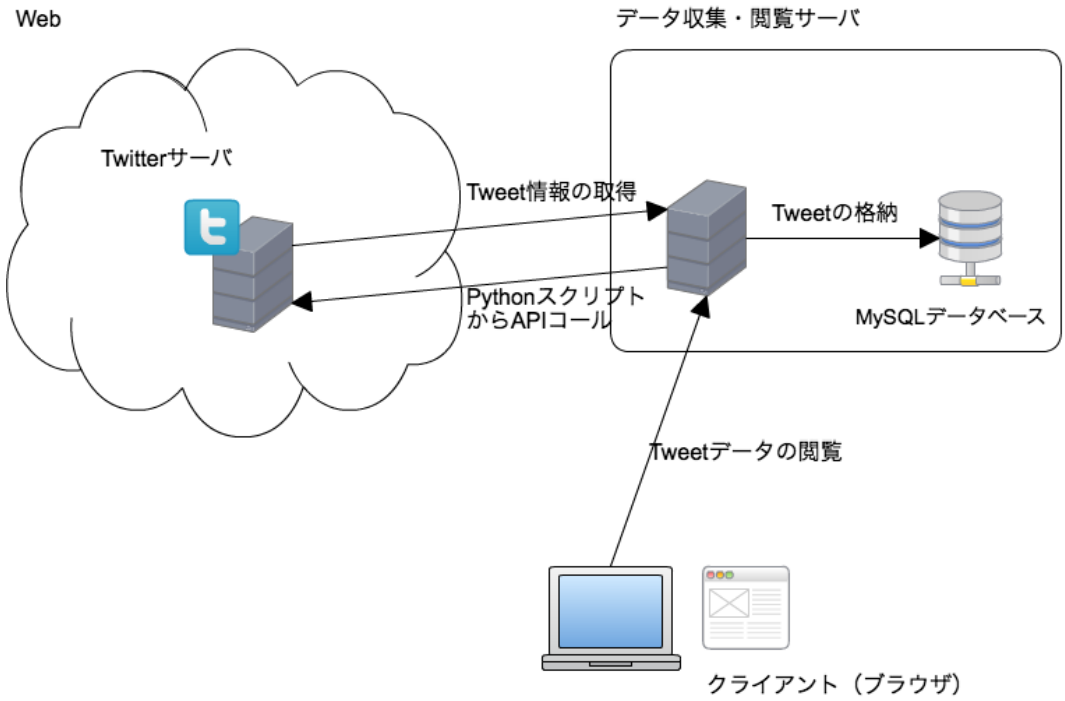

Fig. 3.2　ジオタグ付き Tweet 収集・閲覧システムの概念図

覧システムについて説明する．このシステムは，大きく 2 つのモジュールに分かれる．1 つ目のモジュールは，図 3.3 のように Twitter Search API で収集したジオタグ付き Tweet をマップ上にマッピングし，表示するモジュールである．このモジュールは，日時を指定することで，時間帯別・日にち別の Tweet を閲覧することができる．このモジュールを用いることで，頻繁にジオタグ付き Tweet が投稿されている地点を知ることができる．2 つ目のモジュールは，図 3.4 のように各ユーザごとのジオタグ付き Tweet をマッピングし，閲覧するモジュールである．このモジュールは，各ユーザごとにユーザのジオタグ付き Tweet を表示することができる．このモジュールを用いることで，ユーザの頻繁に活動する地点を可視化することができる．

図 3.5 は，本ジオタグ閲覧システムを用いて，2010 年 8 月 15 日に投稿されたジオタグ付き Tweet を時系列にまとめたものである．この日は，東京ビックサイトにて 3 日で 50 万人が訪れる同人即売イベントが開催されており，その開催時間帯に多数のジオタグ付き Tweet が東京ビックサイト周辺で投稿されている．この例のように，実際に社会で起きているイベントもジオタグ付き Tweet から知ることができる．

これらのシステムを用いることで，研究に必要なデータの収集，そして研究の方針を決めるためのジオタグ付き Tweet の分布の調査を行った．

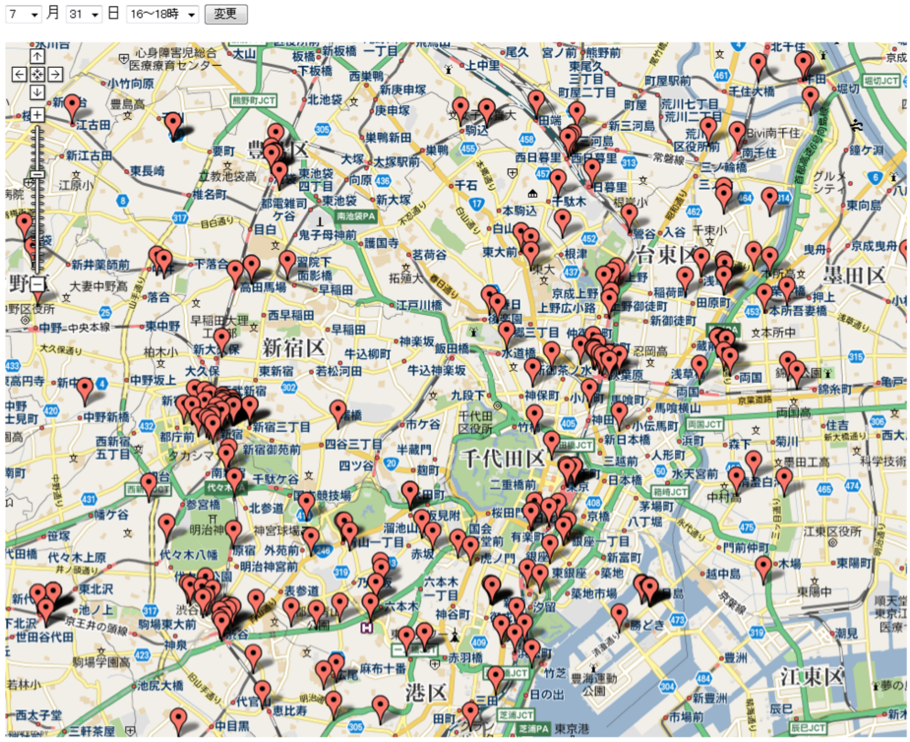

Fig. 3.3　全体のジオタグ閲覧システム

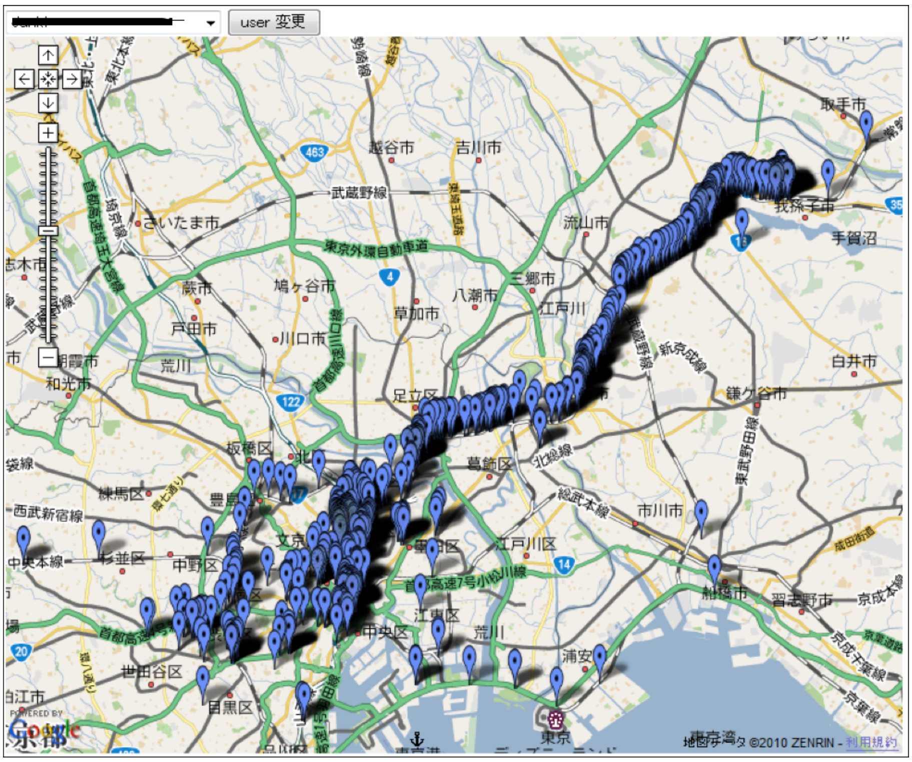

Fig. 3.4　各ユーザごとのジオタグ閲覧システム

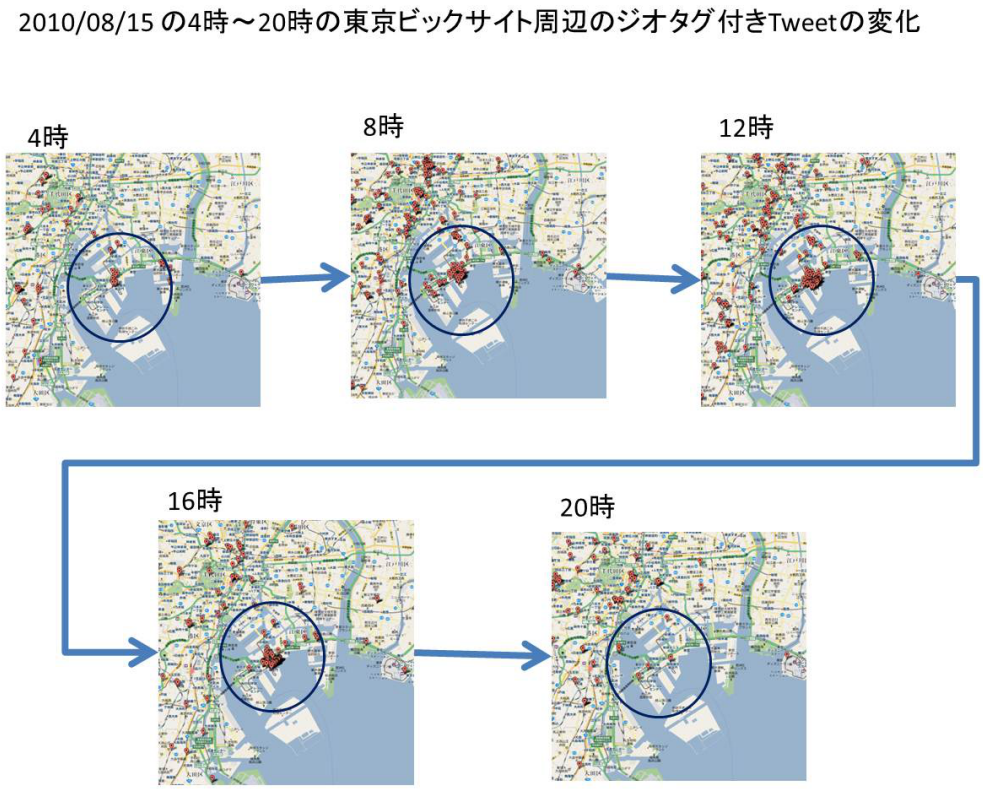

Fig. 3.5　ジオタグによるイベント検知

3.3　ジオタグの特徴分析

上記のシステムで収集したデータを用いて，ジオタグ付き Tweet についての基礎調査を行った [40] ．調査には 2010/05/21 から 2010/08/30 にかけて収集したデータを用いた．

まず，ジオタグを利用しているユーザがどの程度存在するのかについて調査した．14533 人のユーザに対して，直近の 200 件の投稿のうち，20 件以上ジオタグ付き投稿を行っているユーザを抜き出したところ，86 人のユーザ，つまり，約 0.6 %のユーザが該当した．2012 年 1 月現在の日本での Twitter のユニークユーザ数は約 1000 万人であり，単純計算では約 6 万人のユーザがジオタグを利用しているといえる．

また，ジオタグがどのように分布しているかを調べるために，地図上へのマッピングを行った．図 3.6 は，2010/8/15 に投稿されたジオタグ付き Tweet を Google Map 上にマッピングしたものである．

図 3.6 を見ると，東京駅周辺や新宿，渋谷などの都市圏に多くのジオタグが分布しており，そこから放射的に郊外に分布していることがわかる．これは，郊外から都心に通勤・通学する人々が電車乗車中に Tweet を投稿することが多いためだと考えられる．反対に，幹線道路沿いに Tweet が分布することは少ない．自動車を運転している間に Tweet を投稿することは難しいの

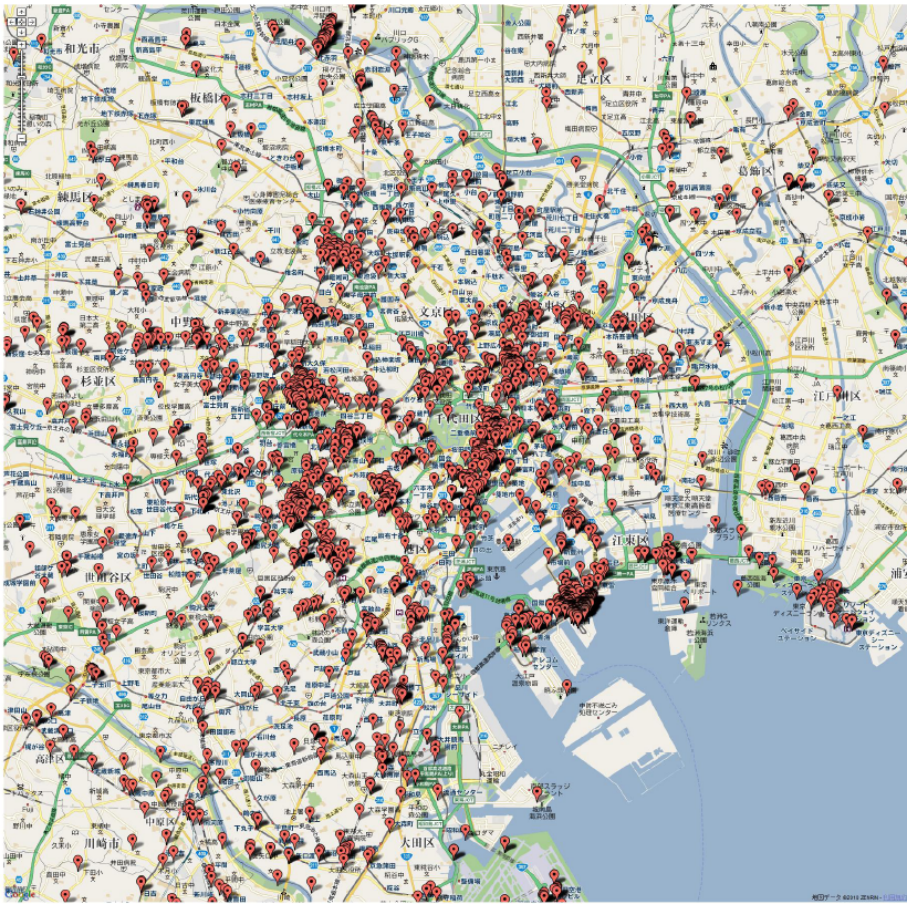

Fig. 3.6　2010/8/15 のジオタグの分布

で，このような結果になると考えられる．

Twitter からだけではなく，外部サービスとの連携によってジオタグ付き Tweet が投稿される場合がある．どのような外部サービスから，どの程度のジオタグ付き Tweet が行われているかを調査した結果が表 3.2 である．

Table 3.2　外部サービスによるジオタグ付き Tweet

サービス名	割合 (%)
Foursquare	14.7 %
ロケタッチ	3.9 %
Brightkite	1.0 %

表 3.2 は，2010/7/30〜2010/8/28 に投稿されたジオタグ付き Tweet，85353 件の解析から得られたものである．表 3.2 をみると，Foursquare やロケタッチといった，ロケーションベースの SNS からの投稿が多い．これらのサービスから投稿された Tweet は，施設名が記載されることが多く，ユーザの行動調査にも役立てることが可能であると考えられる．また，ジオタグ

付き Tweet 全体のうち外部サービスからの投稿は約 2 割弱であった.

第4章　クラスタリングによるユーザの活動地点の推定

はじめに，ユーザの活動地点を推定する手法の検討を行う．

ジオタグ付き Tweet は，図 4.1，図 4.2 のようにユーザの活動する場所に密に分布する傾向がある．このように分布したデータを集約するためには，距離に基づくクラスタリングが有効である．

また，ジオタグ付き Tweet は位置情報・時間情報・テキスト情報の 3 つの属性を持つ．これらの属性を元に，クラスタ間の距離を算出することで，クラスタリングを行い，ユーザの活動地点を適切に抽出することを目指す．

4.1　クラスタリングの概要

クラスタリング，またはクラスター分析とは，機械学習の一種で，データの構造が似ている個体を同じのグループにまとめるデータの処理方法である．また，神嶌はクラスタリングを「内的結合と外的分離が達成されるようなクラスタと呼ぶ部分集合にデータの集合を分割すること」と定義している [41]．機械学習には，事前に学習用データを必要とする教師あり学習と，学習用データを必要としない教師なし学習が存在するが，クラスタリングは後者に属する．クラスタリングは統計やパターン認識，データベース，データマイニング，人工知能などで広く用いられる手法である．

クラスタリング手法は，階層的クラスタリングと非階層的クラスタリングに分かれる．さらに，階層的クラスタリングは分枝型と凝集型に分かれるが，一般に用いられるのは凝集型である．階層的クラスタリングは，何らかの式を用いてクラスタ間の距離を計算し，最も距離の近いクラスタ同士を結合させることでクラスタを生成していく．また，非階層的クラスタリングは，階層的クラスタリングとは異なる方法でクラスタを生成すクラスタリング手法の総称である．

表 4.1 にそれぞれの特徴をまとめた．なお，この表はデータマイニングの基礎 [42] より引用したものである．

4.2　代表的なクラスタリング手法の紹介

ここから，代表的なクラスタリング手法を紹介する．階層的クラスタリング手法として最近隣法・群平均法・重心法・ウォード法・Newman 法を，非階層的クラスタリング手法として k-means 法，x-means 法について簡単に説明する．また，本研究で用いる位置情報，時間情報，テキス

Fig. 4.1　ジオタグ付き Tweet の分布例 1

Fig. 4.2　ジオタグ付き Tweet の分布例 2

Table 4.1　階層的クラスタリングと非階層的クラスタリングの比較

	階層的クラスタリング	非階層的クラスタリング
方法	階層的併合法	K-means ファジィクラスタリング 混合密度分布法
入力	事例間の類似度	クラスタ数および事例 (座標) の初期配置
事前知識	類似度の計算法および併合後の クラスタ代表点の計算方法	目的関数 (あるいは尤度)
計算法	類似度を最大化する事例同士を 順次結合 (類似度行列の縮約)	目的関数を最大化 するように事例を配置
代表点	事例間の類似度から計算	重心あるいは重み付き重心
出力	結合順と結合レベルを木で表した 樹形図 (dendrogram) として出力	事例の最適な配置
長所	重心を使わない方法であれば さまざまな類似度が利用可能 カテゴリカル変数のデータも処理可能	多数の事例を一度にグループ分け することが可能．全体的な配置が鳥瞰可能
欠点	多数の個体に対しては 樹形図が理解困難	各事例間の類似性の議論が困難

ト情報を用いたクラスタリングに関連する，位置情報と時間情報を用いたクラスタリングについて紹介する．なお，これらの手法については神嶌 [41] の論文を参照して記述している．

最短距離法

最短距離法は，2 つのクラスタのそれぞれの中から 1 個ずつ個体を選んで個体間の距離を求め，それらの中で，最も近い個体間の距離をこの 2 つのクラスタ間の距離とする方法である．最短距離法は，空間濃縮という性質のため，外乱に弱く，実データではいい結果が出にくいという欠点がある．

群平均法

群平均法は，2 つのクラスタのそれぞれから 1 個ずつ個体を選んで個体間の距離を求め，それらの距離の平均値を 2 つのクラスター間の距離とする方法である．

重心法

重心法は，クラスタに含まれる要素から重心を求め，その重心間の距離をクラスタの間の距離とする方法である．重心を求める際には，クラスターに含まれる個体数が反映されるように，個体数を重みとして用いる．

ウォード法

ウォード法は，2 つのクラスタを融合した際に，群内の分散と群間の分散の比を最大化する基準でクラスタを形成していく方法である．ウォード法は，階層的クラスタリングを行う際に一般的によく持ちいられる計算方法である．

k-means 法

k-means 法は，非階層的クラスタリングの代表的な手法である．k-means 法では，予めクラスタの数 k を決定する．次に，k 個の点をランダムに設置し，その点を中心点とし，全 k 点より最短距離にある要素を同じクラスタとする．その後，中心点を全要素の重心に移動させ，計算を繰り返す．重心が移動しなくなったら，計算を終了する．

これらの手順をまとめると，以下のようになる．

（1）k 個の代表点 c_1〜c_k をランダムに選択

（2）要素 x ∈ X を min_i D(x, c_i) なる代表点に割り当て

（3）代表点をその代表点に含まれる全要素の重心に変更する

（4）代表点への割り当てが変化しなければ終了．変化すればステップ 2 へ

階層的クラスタリングにくらべ，k-means 法は計算時間が短いことが大きな特徴として挙げられる．ただし，事前にクラスタ数を決定しなければならないという制限がある．また，k-means 法の特徴として，すべての点をクラスタの要素数が一定になるようにクラスタ分けするという点があげられる．

x-means 法

k-means 法では，事前にクラスタ数を決定しなければならないという制限があるため，あるデータ集合のクラスタ数を事前に決定できないときに利用できないという欠点がある．そこで，クラスタ数を自動的に決定する手法が必要となる．

x-means 法 [43] は，k-means 法を k=2 から繰り返し，ベイズ情報量規準 (BIC) がある閾値を下回るまで試行を繰り返すことで，最適なクラスタ数を発見する．なお，BIC は以下で示される値である．

$$BIC = -2 * \ln L + k * \ln n \tag{4.1}$$

ただし，L はクラスタの尤度関数，n は標本数，k は独立変数の数である．

x-means 法を利用する際には，x-means 法を改良させた石岡の手法 [44] が使われる事が多い．

このようなクラスタ数の自動決定手法は x-means 法以外にも JainDubes 法，UpperTail 法など様々存在する．クラスタ分析におけるクラスタ数自動決定法の比較を行った研究では [45] ，データセットに応じて適切なクラスタ数決定手法を使い分ける必要があるとしており，クラスタ数決定のための唯一の手法が確立されていない状態であるといえる．

Newman 法

Newman 法 [46] は階層的クラスタリングのうち，凝集法の一種である．最近開発された手法で，Web マイニングの研究で広く使われている．元々はソーシャルグラフのクラスタリング手法として考案されたもので，密に結合したクラスタを発見するのに適した手法である．

位置情報と時間情報を用いたクラスタリング

本研究では，位置情報，時間情報，テキスト内容を用いてクラスタリングを行う．この 3 つの要素を用いた研究例は少ないが，位置情報と時間情報を用いたクラスタリングは「時空間クラスタリング」と呼ばれ，病理学や犯罪学などの研究分野で用いられている．

たとえば，犯罪発生地域を時空間クラスタリングし，3 次元地図として作成した研究 [47] や，観光客の行動履歴から観光地の特性を分析した研究 [48] などがあげられる．時空間クラスタリングは，多くの場合 3 次元カーネル密度推定を用いることが多い．また，要素間の重み付けを行う際，時間と距離をどのようなバランスで重みとするかを決定する必要があるが，これは今までの研究事例から適切と思われる値を用いることが多い．たとえば，上述の [48] では時間 1 時間を距離 500m と等価であるとして要素間の重み付けを行なっている．

本研究の関連研究として，位置情報とテキスト情報を用いたクラスタリングを行なっている研究 [38] が挙げられるが，この研究でも位置情報とテキスト情報をどのようなバランスで用いるべきかは考察されていない．そこで，本研究では新しくどのようなバランスで位置情報・時間情報・テキスト情報を用いるべきかを考察する必要がある．

4.3 本研究で用いるクラスタリング手法の考察

ここまで，クラスタリングについて紹介してきた．本項では，研究目的を達成するためにどのようなクラスタリングを行うべきか考察する．

まず，要件として以下の 3 点があげられる．

- 大量のユーザデータを扱うため，計算量を減少させる必要がある
- ユーザごとの活動地点の数は未知であるので，クラスタ数を自動で決定する必要がある
- 密なノードの集合を抽出する必要がある

まず本研究では，大量のデータを用いてユーザの活動を分析することを目的としている．そのため，できるだけ計算量が少ない手法を用いるべきである．計算量を少なくするために，最もシンプルな方法である階層的クラスタリングの最短距離法・群平均法・ウォード法などは利用しない．

また，各ユーザごとに日常的に行動する場所の数には違いがあり，クラスタ数を事前に決定することはできない．そのため，計算量は少ないがクラスタ数を与えなければならない k-means 法は利用できない．

さらに，上述したようにジオタグはユーザの活動する場所に密に分布する傾向がある．クラスタリング手法としては，このような密なノード集合を抽出できる手法が適切であると考えられる．x-means 法はクラスタ数を自動で決定できるが，k-means 法の「すべての点をクラスタのサイズが一定になるようにクラスタ分けする」という特徴を同様に持っている．そのため，分散分布した要素のクラスタリングを行う目的に向いている．一方で，ユーザの Tweet は，密

に集中した要素以外は外れ値として各地に分布する．そのため手法として x-means 法は適していないと考えられる．

これらの点から，本研究では Newman 法をクラスタリング手法として採用する．

4.4 Newman 法を用いたクラスタリング

本節では，Newman 法の特徴について示し，具体的なクラスタリング手順を考察する．

4.4.1 Newman 法の計算式

Newman 法は評価関数 modularity Q という指標によって結合させる 2 つクラスタを選択する．Newman 法では，この Q の値が大きくなるようなクラスタ構造が最もらしいとされている．Q は以下の式で表される．

$$Q = \sum_i (e_{ii} - a_i^2) \tag{4.2}$$

ただし，

$$a_i = \sum_j e_{ij} \tag{4.3}$$

である．

式 4.2 において，e_{ij} はクラスタ i とクラスタ j を結ぶエッジ数の総エッジ数に対する割合を表し，e_{ii} はクラスタ i 内のエッジ数の割合である．

実際のクラスタリングでは，クラスタリング前後の Q の差分を取ることでより少ない計算量で同様の結果を導く手法を用いる．

$$\Delta Q = 2(e_{ij} - a_i a_j) \tag{4.4}$$

クラスタリングを行う際は，Δ Q の値が大きいクラスタ同士を結合させていく．この時，すべてのクラスタ対の Δ Q の値が負になったとき（すなわち，Q が最大の時），クラスタリングを終了し，残ったクラスタを採用する．

4.4.2 ノード間の重み付け手法の検討

Newman 法はノード間に重み付けを行うことができる．本研究では，ジオタグ付き Tweet をひとつのノードとし，ノード間の距離，時間差，投稿内容の類似度によりノード間の重み付けを算出し，クラスタリングを行う．

ノード間の距離

物理的な距離が近いノード同士の関連度は高い．そして，距離が遠くなればなるほど関連度は下がる．そこで，緯度経度から得られる距離をまず重み付けに利用する．本研究では，ノード n(i),n(j) に対して以下のような式で重み付けを行う．ただし，距離はメートル単位で扱う．

距離を計算する際には, [17] で示されているとおり，緯度経度距離を実際の距離に直して計算する．

$$w_l(i,j) = \begin{cases} 1 - \frac{|n_l(i) - n_l(j)|}{L} & (|n_l(i) - n_l(j)| \leq L) \\ 0 & (|n_l(i) - n_l(j)| > L) \end{cases} \tag{4.5}$$

GPS の誤差や同じ場所内での移動を考慮しても，あまりにも距離が離れたノード同士は無関係であると考えられるため，本研究では $L = 1000m$ 以上のノードについては重み付けを 0 とする．

最大円弧が 1000m という理由は，

- GPS 精度の誤差 [26] （端末によるが，屋外で 10 から 20m 程度，屋内だと 200m 程度，最大で数キロずれることもある）
- twitter のクライアントで生じる誤差（前回位置からの移動距離が少ないと，ＧＰＳデータを新規取得しないで，前回のものを使いまわす）
- ひとつの場所，例えば新宿や渋谷といった場所の領域の大きさ（今回のターゲットである東京都市部において，自分のいる地域を示す時，人は渋谷や原宿，池袋，代官山といったように駅の名前を用いることが多い．例として，東京都の中心を走る山手線の駅の間隔は，平均で 1.2km）

といった要素を勘案して決定している．

ノード間の時間差

ユーザは，ひとつの場所には同じような時間帯に滞在することが多い．例えば，家には朝と夜の時間帯に滞在し，仕事場や学校には昼の時間帯に滞在している．そこで，クラスタリングにおいて時間の要素を追加することで，より実際の活動に即したクラスタの抽出が行えると考えられる．

本研究では，Tweet の投稿時刻を元として，以下の式でノード同士の重み付けを行う．ただし，時刻は分単位で扱う．

$$w_t(i,j) = \begin{cases} 1 - \frac{|n_t(i) - n_t(j)|}{T} & (|n_t(i) - n_t(j)| \leq T) \\ 0 & (|n_t(i) - n_t(j)| > T) \end{cases} \tag{4.6}$$

ただし，時間差を算出するときは，日付情報を用いず単純に時刻を比較して計算を行う．例えば，2011/08/31 12:00:00 と 2011/08/29 11:00:00 の差は 60 分である．これは，日々の行動を取得するため，どの時間帯にその場所に滞在しているかでクラスタを作る意図があるからである．

本研究では，T = 180 分 とし，それ以上離れた 2 点については関連性がないものと考え，重みを 0 とする．

Fig. 4.3　クラスタリング結果の例

ノード間の投稿内容の類似性

さらに，投稿内容の類似度もクラスタリングの重み付けとして用いる．同じ場所での投稿は完全ではないにせよ類似性があると考えられる．このため，時刻や位置だけでなく，投稿内容も考慮に入れることで，より良いクラスタリングができると考えられる．

本研究では，投稿内容に含まれる単語を単語ベクトル化し，コサイン距離を取ることでノード同士の重み付けを行う．ここで，$N(i), N(j)$ はノード n(i),n(j) の単語ベクトルである．

$$w_c(i,j) = \frac{N(i)N(j)}{\sqrt{|N(i)||N(j)|}} \tag{4.7}$$

これらをまとめると，ノード i,j 間の重み付けの式は以下で表すことができる．

$$w(i,j) = w_l(i,j) + \alpha w_t(i,j) + \beta w_c(i,j) \tag{4.8}$$

これら位置情報，時間情報，投稿内容の重み付けのバランスを決定する α,β の値については，評価実験を行うことで適切な値について考察する．

図 4.3 は，上記の手法を用いて実際にクラスタリングを行った例である．図の小さいアイコンが各ジオタグ付き Tweet を，大きいアイコンがクラスタの中心点を表している．

第5章　ラベリングによるユーザの活動内容の推定

次に，クラスタリングにより取得された各クラスタに対して，家や仕事場といったラベル付けを行い，その場所でユーザがどのような活動を行なっているかを推定する．

5.1　ラベリング手法の調査と検討

まず，クラスタに対して適切にラベリングを行う手法について検討する．

ユーザの Tweet は，場所との相関関係があると考えられる．例えば，食事を行う場所であれば食事の感想を投稿し，職場では仕事に関する投稿をする可能性が高い．一方で，一つ一つの Tweet は必ずしも場所に依存したものに限らず，ユーザは同じ場所でも様々な投稿を行うと考えられる．そこで，あるクラスタ内での投稿を集約し，傾向を発見することで，ラベリングを行うことが妥当であると考えられる．

このような目的を達成するためには，統計に基づく機械学習の手法が有効である．機械学習の手法には，統計的な傾向からデータを数式に当てはめる回帰的手法と，統計的な傾向から幾つかのパターンにデータを分類する分類的手法が存在する．本研究で用いるべき手法は，後者の分類的手法である．

分類的手法とは，まず正しいクラスが与えられている訓練データからデータセットの特徴を発見し，その後テストデータを解析し，妥当であると思われるクラスに分類することで分類を行う手法のことを指す．分類的手法には，Naive Bayes・決定木・K 近傍法・SVM[49] ・ニューラルネットワークなどの手法がある．それぞれの手法には計算時間や精度の点で特徴がある．

これらの手法のうち，本研究では Naive Bayes を用いる．Naive Bayes は分類的手法の中で最も基本の手法であり，計算時間も短く，また多クラス分類も容易であることが特徴である．Naive Bayes は，文章分類で多く用いられ，メールのスパムフィルタリングやニュース記事の自動ジャンル分けなどに用いられている．

Naive Bayes について簡単に説明する．Naive Bayes では，まず訓練データを元にしてベイズ確率によりある文章中の単語がそれぞれのカテゴリに属する確率を求め，その後テストデータの文章に含まれる単語のカテゴリの確率をかけ合わせ，最終的にある文章がどのカテゴリに含まれるかを決定する．

Naive Bayes の ”Naive” とは，事後確率を求める際にすべての組み合わせは独立であることを仮定していることに由来する．例えば,「机」と「椅子」という単語は，関連性があり，お互

いの出現確率は独立ではないと考えられる．しかし，Naive Bayes ではこのような単語間の関連性を無視し，それぞれの単語が独立に出現することを仮定している．この仮定はやや乱暴に思えるかもしれないが，実用上では精度の低下は見られず，分類問題を解くにあたっては効果的であることが知られている．

また，本研究では Naive Bayes を用いて多クラス分類を行うため，Naive Bayes の多項モデルを利用している．加えて，上記のように文章に含まれる単語のカテゴリ確率をかけあわせて文章全体のカテゴリ確率を求める場合，一つでも出現確率が 0 の単語があると文章全体の確率が 0 となってしまう．そこで，MAP 推定と呼ばれる手法を用いることにより，値を調整することでこの問題に対処する．

本研究で Naive Bayes を用いる理由は以下のとおりである．

（1）大量のデータを扱うため，計算時間の短い手法が有効であると考えられるため

（2）既存研究に同様の研究がなく，精度の基準値として基本的な手法で分類した値を示すため

（3）多クラス分類を容易に行うことができるため

まず，1 については，本研究では多数のジオタグユーザの Tweet を計算し，分類することが求められる．SVM などの手法は，精度の面では Naive Bayes よりも勝るが，パターン認識のための学習に時間がかかる．

2 についてだが，Naive Bayes はクラス分類問題の最も基本的な手法であり，分類精度の基準となることが多い．また，既存研究の中で本研究と同じ動機で研究を行なっている例はない．そこで，精度の基準としての値を出すという点でも，Naive Bayes を用いる理由付けとなる．

3 についてだが，Naive Bayes は複数クラスに分類することが可能な手法である．他の手法は，2 クラスに分類するものが多い．もちろん，SVM などの 2 クラス分類手法でも，ペアワイズ法などの手法を用いることで多クラス分類は可能であるが，計算量が爆発的に増加してしまうという問題点もある．

これらの理由から，本研究では Naive Bayes をクラス分類手法として用いる．また，Naive Bayes に加えて，Naive Bayes を改良した Complement Naive Bayes[50] も手法として用いる．この手法は，Naive Bayes で確率を計算する際に，補集合を用いることで精度の向上を狙った手法である．補集合を用いることで，単語数が増えた際に発生するバイアスを減らすことができる．

5.2 ラベリングアルゴリズム

本節では，Naive Bayes・Complement Naive Bayes を用いた具体的な分類手法について説明する．

クラスタに含まれる Tweet 群はひとつのドキュメントであると捉えることができる．本研究では，クラスタをドキュメントと捉え，分類を行う．ドキュメントの分類では，ドキュメント

Table 5.1　ラベル一覧

	ラベル一覧
家	ユーザの自宅
仕事場	ユーザの職場
学校	ユーザの通う学校
娯楽	買い物や食事など，余暇を過ごす場所
移動	鉄道や飛行場など交通機関に関する場所
その他	その他分類外の場所

に含まれる単語とその出現回数を特徴量として用いる事が多い．これを Bag of words という．本研究でも Bag of words を作成し，分類を行う．

日本語の文から単語を抽出することは簡単ではない．英語など単語がスペースで区切られた言語と比較して，単語の切れ目がわかりずらく，また同じ表記でも複数の意味を持つ単語や，品詞の異なる単語が存在するためである．そこで，文から単語を抽出するための手法として，形態素解析が用いられる．

本研究では，形態素解析ツールとして MeCab[51] を用いる．MeCab は，条件付き確率場に基づく高い解析精度を誇る形態素解析ツールである．さらに，他の ChaSen や KAKASI といった形態素解析ツールに比べ，高速で動作することも特徴である．

単語の抽出手順としては以下のとおりである．

（1）Tweet から，RT や@付き投稿など Twitter 特有の表現記法を消去する．

（2）Tweet を形態素解析し，単語を抽出する

（3）単語のうち，名詞・形容詞・動詞・副詞に該当する単語を抽出する．

ここで，単語として名詞・形容詞・動詞・副詞のみを利用する理由としては，他の助動詞や助詞などの品詞は一般的なものであり，場所によって出現頻度が変わるとは考えにくいからである．

このような手順で得られた単語を特徴量とし，Naive Bayes・Complement Naive Bayes を用いてラベリングを行う．

5.3　利用するラベルの選択

本節では，活動内容を適切に表すためのラベルの選択について説明する．

先行研究では，ラベルについて「家」「仕事場」の二種類とすることが多いが，本研究では，その 2 つ以外の人の行動，例えば映画や買い物，駅での移動などの行動についても調査する．

Table 5.2　交通センサスの項目一覧

No.	項目	説明
1	出勤	通勤のため会社へ行く場合
2	登校	就学先への登校，校外活動
3	家事買物	業務での買物は含まない
4	食事社交娯楽	日常生活圏内の私的なつきあい
5	観光	観光名所，旧跡などへの観光
6	保養	温泉，家族知人との交流などの保養
7	スポーツ	ハイキング，ゴルフ，運動会などスポーツ
8	体験型	レジャー遊園地ドライブ名産品の飲食等
9	その他私用	通院習い事など
10	送迎	送迎（業務での送迎は含まない）
11	荷物の運搬を伴わない業務	業務目的で荷物を運搬しない場合
12	荷物の運搬を伴う業務	業務目的で車を利用した場合
13	帰社	業務が終わって会社へ戻るための運行
14	帰宅	勤務先通学先，買物，外出先から自宅
15	その他	上記以外のその他
16	不明	移動目的が不明な場合

本研究では，ユーザのその地点での活動内容を表すラベルとして，クラスタに対して以下の 6 つラベルを割り当てる（表 5.1）.

これらのラベルは，国道交通省が行なっている交通センサス (表 5.2) の行動分類項目を参考に，日常的な活動に当てはまるものを利用している．交通センサスの項目のうち，旅行や保養など，日々の行動に当てはまらない項目については除外している.

第6章　評価実験

6.1　評価実験の概要

4 章，5 章で提案した手法について，評価実験を行い，手法の有効性の確認を行った．実験は，25 人の被験者に対しておこなった．被験者は，日常的に Twitter とジオタグを利用しているユーザである．

実験には，実験のために特別に投稿した Tweet ではなく，被験者が普段投稿しているジオタグ付き Tweet を収集し，利用した．

まず，被験者に対して事前にアンケートを行い，被験者が日常的に訪問している場所についての回答を得た (図 6.1，6.2)．アンケートでは，自宅や勤務先，休日過ごす場所など，被験者が訪問している場所について任意の数だけ回答を得た．被験者に対するアンケートの結果を，以下の表 6.1 に示す．

これらの情報と，ジオタグから提案手法を用いて推定した被験者の行動特性とを比較することで，提案手法がどれだけ被験者の行動を適切に推定できているかを評価する．

6.2　クラスタリングに関する評価

まず，4 章で示したクラスタリング手法の評価を行う．被験者のジオタグ付き Tweet を元に，被験者の日常的な行動について把握を行う．

クラスタリングは，式 4.8 の変数 α,β の値を変えながら様々なパターンで行った．クラスタリングには被験者が投稿した最新 500 件のジオタグ付き Tweet を利用した．

ユーザの活動地点の推定を行う先行研究 [28] [29] では，多くの例でユーザの実際の活動地点と推定した地点との距離を取ることで手法の評価を行っている．本研究でも同様に，クラスタリングにより取得されたクラスタの中心点と，アンケートによって得られた位置との物理的な距離を比較することで，適切なクラスタリングが行えているかを評価する．

具体的には，以下の式 6.1 のように，両者の緯度経度距離の総和を算出することで一致度を計算する方法を用いる．

$$D_u = \frac{1}{n} \sum_i |p_i - c_x| \tag{6.1}$$

なお，p_i はアンケートにより取得された地点の位置であり，c_x は p_i に最も近いクラスタである．また，n はアンケートにより取得された地点の数である．クラスタリングにより取得され


あなたがよく行く場所についてのアンケート

東京大学大学院修士1年の酒巻と申します。(@makisaka)

あなたがよく行く場所について、アンケートをお願いします。
このアンケートの利用方法は、場所の特性とtwitterの投稿内容に関係性があるかを調べるものです。
いただいた情報は研究目的にのみ利用し、それ以外の用途で利用することは一切ございません。

あなたの性別を教えてください *
男性
女性

あなたの年齢を教えてください *
10代

あなたの身分について教えてください *
学生

あなたのtwitter idを教えてください *
解答例: @makisaka

あなたが普段よく行く場所をあげ、その場所でどのようなことをしているか教えてください。(月に1回は行く場所) *
解答例: 渋谷 買い物、居酒屋 池袋 大学 品川 食事、デート のように、場所と行っていることをペアにして回答してください。思いつく限り回答していただけると助かります。

Fig. 6.1 被験者アンケートの例1


あなたがよく行く場所についてのアンケート-2

東京大学大学院修士1年の酒巻と申します。(@makisaka)

あなたがよく行く場所について、アンケートをお願いします。
このアンケートの利用方法は、場所の特性とtwitterの投稿内容に関係性があるかを調べるものです。
いただいた情報は研究目的にのみ利用し、それ以外の用途で利用することは一切ございません。

まずは、こちらのWebページをご覧下さい。
http://sezakilab.dip.jp/test/k.php
user名: sakamaki
password: tomohiro

このページを見ながら、各赤いマーカーをクリックして以下の問いに答えてください。

あなたのtwitter idを入力してください
例: @makisaka

各マーカーの地点について、あなたはこの場所の周辺によく行きますか？あなたが「よく行く場所」の番号をすべて記入してください。 *
「この場所によく行くかどうか」は自分の感覚でかまいません。

以下の質問は、上の質問であなたがその場所によくいくと答えた場所のみ回答してください。

0番の地点について。その地点をクリックして表示された単語は、あなたのその場所での行動をどの程度表現していますか？
例: あなたにとって「大学」の場所であれば、「研究」「授業」といった単語があれば自分の行動と一致している。

	1	2	3	4	5	
自分の行動と良くあっている単語群だ	○	○	○	○	○	自分の行動と全くあっていない単語群だ

1番の地点について。その地点をクリックして表示された単語は、あなたのその場所での行動をどの程度表現していますか？

	1	2	3	4	5	
自分の行動と良くあっている単語群だ	○	○	○	○	○	自分の行動と全くあっていない単語群だ

2番の地点について。その地点をクリックして表示された単語は、あなたのその場所での行動をどの程度表現していますか？

	1	2	3	4	5	
自分の行動と良くあっている単語群だ	○	○	○	○	○	自分の行動と全くあっていない単語群だ

Fig. 6.2 被験者アンケートの例2

Table 6.1　被験者に対するアンケート

項目	属性
被験者の人数	25 人
被験者の年齢	20〜40 歳の男女
被験者の職業	学生，社会人など
被験者の募集方法	Twitter 上での募集
被験者の居住地区	関東圏 30km 以内に在住していること
被験者が訪問している場所の数（平均）	4.48 箇所

たクラスタは，クラスタにふくまれるジオタグ付き Tweet の数で並び替え，上位からアンケートにより取得された地点の数と同じだけ利用することとする．

すべてのユーザについて同様の計算を行い，すべてのユーザの平均の値により評価を行う．

Pattern A

まず，式 4.8 において，$\alpha = 0, \beta = 0$ という条件でクラスタリングを行った．つまり，緯度経度距離のみを用いてクラスタリングを行うという条件である．

この条件では，平均誤差距離 D は 1893.5m という結果になった．

Pattern B

次に，$\alpha = \mathrm{x}, \beta = 0$ という条件でクラスタリングを行った．つまり，時空間情報を用いてクラスタリングを行った．x は 0 から 0.5 まで，0.005 刻みで変化させ，計算を繰り返した．結果を図 6.3 にまとめる．

この場合，$\alpha = 0.1$ の時，距離の誤差が 1736.9m と，最も距離の誤差が短くなった．これは具体的には，距離 1000m に対して時間 10 分を同等の重みとした場合である．

Pattern C

次に，$\alpha = 0, \beta = \mathrm{x}$ という条件でクラスタリングを行った．つまり，緯度経度距離とテキスト情報を用いてクラスタリングを行った．x は 0 から 0.5 まで，0.005 刻みで変化させ，計算を繰り返した．結果を図 6.4 にまとめる．

この場合，$\beta = 0.3$ の時距離の誤差が 1748.9m と最も距離の誤差が短くなった．

Pattern D

最後に，すべての要素を含めた条件として，$\alpha = \mathrm{x}, \beta = \mathrm{y}$ という条件でクラスタリングを行った．ここでは，x をそれぞれ 0.1, 0.2, 0.3 とし，y は 0 から 0.5 までそれぞれ 0.01 刻みで変化させ，計算を行った．

この条件では，$\alpha = 0.1$ $\beta = 0.3$, の時平均誤差距離 D は 1665.7m という最も誤差が少ない結果になった．Pattern D の実験結果を図 6.5 にまとめる．

これらの実験の結果を表 6.2 にまとめる．誤差距離が少ない順に，Pattern D, Pattern B, Pattern C, Pattern A となった．

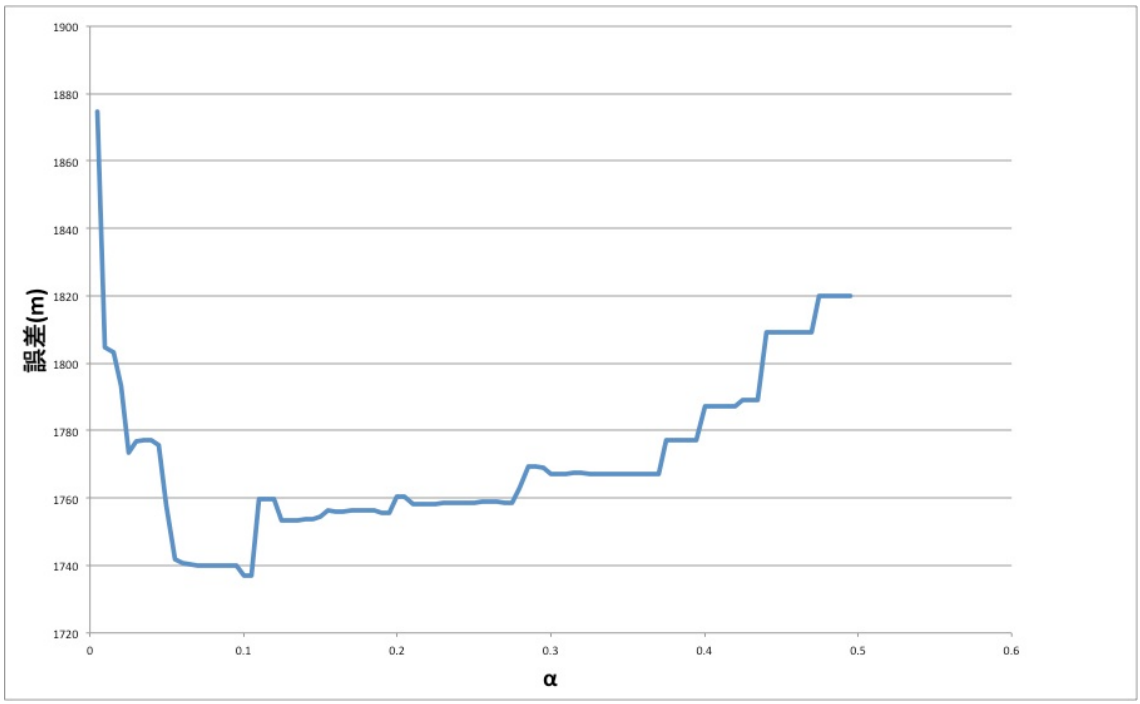

Fig. 6.3　Pattern B の結果

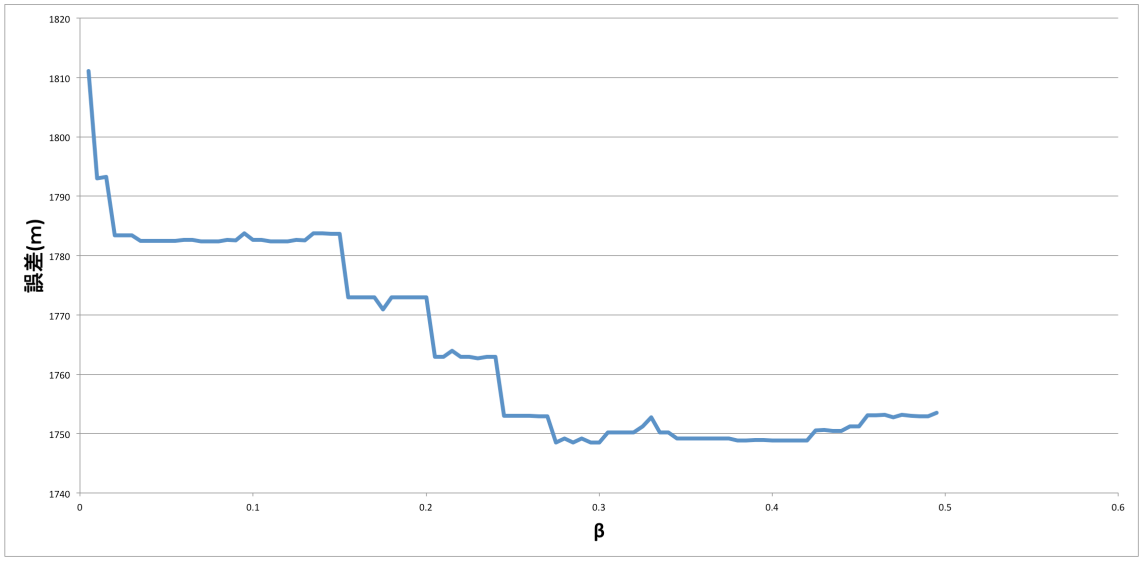

Fig. 6.4　Pattern C の結果

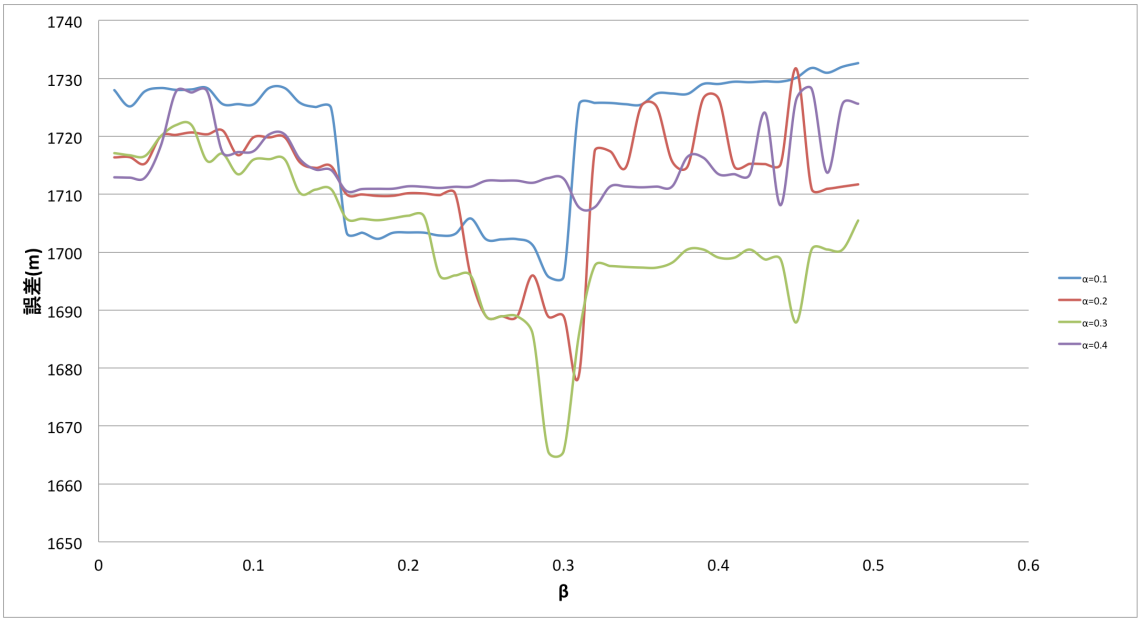

Fig. 6.5　Pattern D の結果

Table 6.2　実験結果

	平均誤差距離 D(meter)
Pattern A(位置情報)	1893.5
Pattern B(位置情報・時間情報)	1736.9
Pattern C(位置情報・テキスト情報)	1748.9
Pattern D(位置情報・時間情報・テキスト情報)	1665.7

6.3　本手法に必要となる Tweet 数の推定

次に，ユーザの行動を推定するために，どの程度のジオタグ付き Tweet が必要であるかについて評価実験を行った．

前節で行った評価実験において，利用するユーザのジオタグ付き Tweet 数を変化させて行動推定を行った結果が図 6.6 である．なお，クラスタリングの重み付けは上記の実験で最適となった Pattern D の重み付けを利用している．

このグラフより，利用するジオタグを増やせば増やすほど精度が上がるが，ジオタグ数が 700 を超えたあたりからジオタグ数を増やしても誤差があまり変化しないことがわかる．

6.4　ラベリングに関する評価

次に，5 章で示したラベリング手法の有効性について評価を行う．評価実験では，上記の実験で取得されたクラスタを用いる．

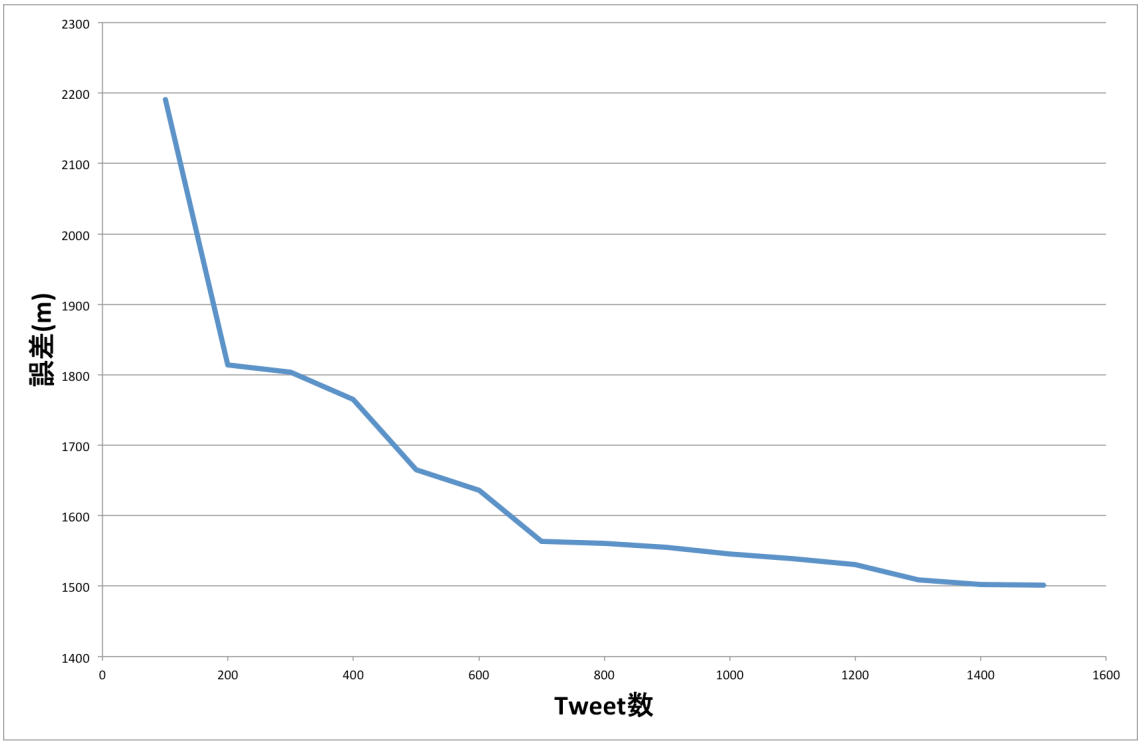

Fig. 6.6　The Geotagging tweets

まず，取得したクラスタに対して正解データを設定する．アンケートにより取得した情報を元に，クラスタに対して手動で表 5.1 の 6 種類の正解データを与える．

正解データを与えたクラスタを用いて，leave-one-out cross-validation による評価を行った．leave-one-out cross-validation とは，機械学習で得られた分類器を評価する目的で用いられる方法である．leave-one-out cross-validation では，まずサンプルデータを k 個に分け，そのうちの一つを正解データ（分類器を用いて分類するデータ）とし，残りを教師データ（分類器を作成するためのデータ）とする．その後，k 個に分割したサンプルデータを正解データとして k 回計算を行い，全結果の正答率の平均値を用いて評価を行う方法である．k の値については，一般にサンプルデータ数と同じ，つまり正解データが 1 つとなるようにすると最も誤差が少なくなると言われている [52]．ただし，k の数を増やすほど，計算量が増えるため，k の値についてスタージェスの公式を用いる例も多い．

本論文では，サンプルデータ数がそれほど多くないため，k = クラスタ数 とし，評価を行った．実験結果を，以下の表 6.3 にまとめる．15 人の被験者から獲得した全クラスタ数は 112 であった．これらのクラスタに手動でラベル付けを行い，cross-validation を行った結果，Naive Bayes を用いた場合の適合率は 0.643，Complement Naive Bayes を用いた場合は 0.696 となった．

また，ラベルごとの適合率を表 6.4 にまとめる．表 6.4 を見ると，家や学校などは比較的高い適合率を示しているが，仕事場の適合率は低いという結果になった．これは，大学や学校では特徴的な単語をつぶやきやすい一方で，職場ではユーザによって投稿内容がバラバラである

Table 6.3　cross-validation の結果

全被験者から獲得したクラスタの総数	112
適合率 (Naive Bayes)	0.643
適合率 (Complement Naive Bayes)	0.696

Table 6.4　cross-validation の結果 (ラベル毎の結果)

		Naive Bayes		Complement Naive Bayes	
ラベル名	ラベル数	正解数	適合率	正解数	適合率
家	15	12	0.80	12	0.80
仕事場	10	4	0.40	4	0.40
学校	21	17	0.81	17	0.81
娯楽	54	41	0.76	41	0.76
移動	25	10	0.40	10	0.40
その他	25	8	0.24	8	0.24

ことが原因であると考えられる.

第7章　考察

ここでは，今回の評価実験で得られた結果について考察する．

7.1　クラスタリングに関する考察

最も精度が良かったのは位置情報，時間情報，テキスト情報すべてを利用した Pattern D であった．この結果より，ジオタグ付き Tweet ように，位置情報と時間情報，テキスト情報を持ったデータセットからクラスタを生成する際は，位置情報のみを利用するのではなく，位置情報，時間情報，テキスト情報を利用することで精度を向上させることができるといえる．

また，これら 3 つのパラメータをどのように利用するべきかについてだが，本研究では，Pattern D の実験結果より 位置情報 : 時間情報 : テキスト $= 1 : 0.1 : 0.3$ の割合で利用すると最も精度が良かった．時空間クラスタリングの分野では，位置情報と時間情報をどのような割合で利用するかについて経験を基に判断することが多い．ジオタグ付き Tweet のようなデータセットのクラスタリングを行う際も，同様にどのようなバランスでこれらの情報を利用すればよいかを考察する必要がある．本研究での最適なバランスは，上記の結果となったが，今後研究例が増えるに伴って，最適なバランスが確立されていくと考えられる．

また，ユーザごとの実験結果をみると，ユーザによって，テキストを重視したほうが精度の良いクラスタリングが可能なユーザ，時間を重視したほうが良いユーザがわかれた．例えば，あるユーザ A は，位置情報のみでクラスタリングを行った場合の誤差は 3358m，一方でテキストを用いた場合は 2525m という結果になった．これは，ユーザによって，Tweet 内容が位置に依存せず，どの場所でも同じような内容を投稿するユーザ，その場所で起きていることを投稿するユーザと分かれるからであると考えられる．

つまり，位置と Tweet 内容に相関があるユーザやそうでないユーザなど，はじめにユーザの分類を行い，その後ユーザごとにクラスタリングの重み付けを帰ることで，より精度よくクラスタリングができると考えられる．

また，今回の手法では，日常的な行動を取得することを目的としてクラスタリングを行い，ノードが密に集中している場所の抽出を行ったが，本来は外れ値からも重要なユーザの行動を取得できる．例えば，数日間だけある場所に旅行や出張をしていたことがジオタグから抽出できると考えられる．このような情報を活用することも，今後の重要な課題である．

7.2 必要なジオタグ数についての考察

図 6.5 を見るに，ジオタグ数が 700 以上では大きく精度は変わらないことが分かる．つまり，700 以上のジオタグ付き Tweet を投稿しているユーザは，行動調査の対象とすることができるということである．行動調査に必要なジオタグ数は，行動調査が可能なユーザ数と分析を行うための計算時間に影響する．ここでは，700 というジオタグ付き Tweet の意味について考察する

3 章で示したとおり，本研究では，ジオタグ収集システムを用いてジオタグを収集した．このデータを元に，ジオタグ付き Tweet を行なっているユーザと，ジオタグ付き Tweet 数の関係について，調査を行った．

図 7.1 のヒストグラムは，ユーザの投稿した Tweet 数を横軸に，ユーザ数を縦軸に取ったものである．これらのデータは，3 章で示したジオタグ収集システムにより集められたジオタグを元に，投稿しているジオタグ数が 100 以上のユーザを抽出したものである．ヒストグラムは 100 刻みで，100 以下の値を切り捨てている．また，図 7.2 は，同様のデータの累積頻度分布である．この図の y 軸は，x 軸の値以上のジオタグ付き Tweet を投稿しているユーザの割合を表している．

図 7.1 と図 7.2 より，ジオタグ投稿数のピークは，400 から 500 Tweets であることがわかる．また，400 から 500 Tweets 以上の投稿を行なっているユーザ数は全体の 80 %弱，また，700 から 800 Tweets 以上の投稿を行なっているユーザは全体の約半分であることがわかる．

つまり，ジオタグ付き Tweet を利用しているユーザのうち，約半数に対しては本手法を有効に利用できるということになる．もちろん，それ以下の Tweet 数のユーザも本手法の性能を完全に利用することはできないが，ある程度の精度で行動調査が可能である．

また，クラスタリングの計算量は，Newman 法の場合，エッジ数を m，ノード数を n とした時，O((m+n)n) である．このため，行動調査に必要なジオタグ数を減らすことは計算時間の点からも重要である．より少ないジオタグ付き Tweet 数で計算を行えるようにすることは今後の重要な課題であるといえる．

7.3 ラベル付けに関する考察

今回の評価実験では Naive Bayes を用いた場合の適合率が 0.643，また Completemt Naive Bayes を用いた場合の適合率が 0.696 という結果になった．この値は，6 値分類問題としては決して低いものではないが，実用性を考えるとより高い精度で分類を行う必要があると考えられる．

より精度を上げるために，まず考えられるのはクラスタ内の Tweet の特徴を利用することである．今回用いた手法では，クラスタ内に含まれるすべての Tweet を用いてラベリングを行った．しかし，ユーザの投稿する Tweet は自身の行動を表すものだけではない．例えば，ユーザが今考えていたこと，ニュースに対するコメント，本やテレビの感想などである．このような Tweet は，ユーザの行動を推測するためには利用するべきではない．そこで，ユーザの Tweet からユーザの行動を表す Tweet のみを抜き出す手法が必要となる．

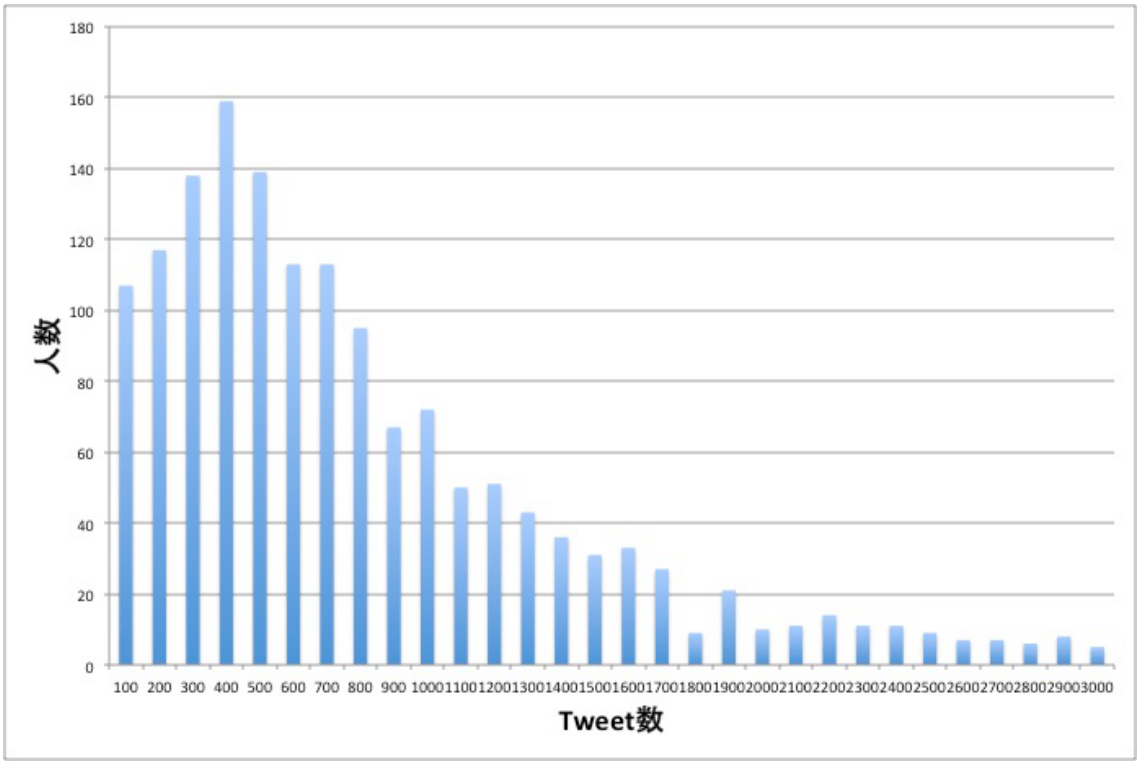

Fig. 7.1　ジオタグ付き Tweet とユーザ数の関係を示すヒストグラム

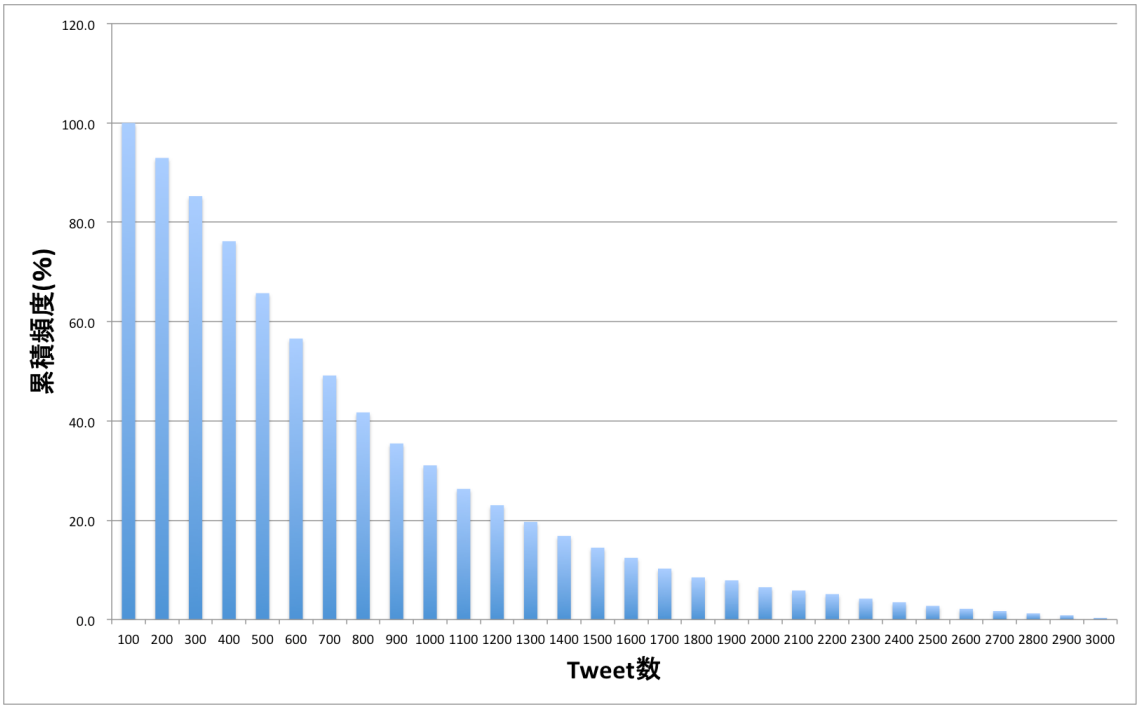

Fig. 7.2　ジオタグ付き Tweet とユーザ数の関係 (累積頻度)

2 章でも述べたように，田中らの Twitter の利用法に関する研究 [14] では，ユーザによって Twitter の利用方法は異なり，特にニュースを頻繁に投稿するユーザや意見や感想を投稿するユーザなどがいることを指摘している．田中らのユーザ分類指標を元に，ユーザもしくは Tweet の特徴を捉えることで，分類器による分類を行うことができるのではないかと考えられる．

また，Tweet の分類だけでなく，単語レベルで取捨選択を行うことも考えられる．今回用いた手法では名詞・形容詞・動詞・副詞のすべてを利用しているが，この中にも分類に不適合な単語が含まれることがある．岡山大学の研究グループでは行動辞書 [33] を開発しており，この辞書を用いることで，単語群から行動に関係する単語のみを抜き出すことができる．行動辞書は発展途上であるようだが，精度の改善に役立つと考えられる．

さらに，そもそも分類する項目数を減らすことも考えられる．Sibren らの研究 [28] など，先行研究の多くは，家と仕事場を仕事場の 2 値分類問題とすることで場所へ対する意味づけを行なっている．これらの研究のように，クラスタを「家」「仕事場」そして「その他」の 3 値分類問題とすることも考えられる．

第8章　結論

1章でも触れたが，パーソントリップなどの社会全体の行動調査，また個人の行動アシストのための行動調査，どちらの分野においても人の行動情報は重要である．本論文では，行動情報抽出を目的としたジオタグ付きTweetを用いたデータマイニング手法を用いた研究について議論を行った．ジオタグ付きTweetは属性として位置情報・時間情報・テキスト情報を持っている．これらの属性を用い，重み付けを行ったクラスタリング手法について考察した．これにより,「ユーザが定期的にどこで活動しているか」の推定を行った．

25人のユーザを対象とした被験者実験を行ったところ，平均誤差1665.7mで被験者の位置を推定することができた．また，位置情報，時間情報，テキスト情報をどのように組み合わせれば精度よくクラスタリングが出来るかについて考察を行った．その結果，位置情報だけでなく時間やテキストを利用することで，より精度の良いクラスタリングができることを確認した．ただし，ユーザによって重視すべき項目が異なると言え，より精度を上げてクラスタリングを行うためにはユーザの分類が必要であると考えられる．さらに，700以上のジオタグ付きTweetがあれば，ユーザの行動推定をある程度の誤差で行うことができることがわかった．

また，クラスタに対してNaive Bayes, Complement Naive Bayeesを用いてラベル付けを行うことで，その場所でユーザがどのような活動を行っているかという知識を抽出した．これにより,「ユーザはその場所でどのような活動をしているか」の推定を行った．

特に，ユーザの自宅や学校などは高精度でラベル付けを行うことができた．しかし，勤務先などはユーザによって投稿内容が異なるため，精度は低かった．

本研究を通して，マイクロブログサービスのデータを用いることで，イベントや災害など社会全体からの知識抽出けでなく，行動調査のようなプライベートな知識も抽出可能であると示すことができた．ただし，本研究で示せたのは，ジオタグ付きTweetを利用しており，かつある程度のTweetを行なっているユーザに対してのみ，行動調査が可能であるという点である．今後，より少ないTweeetで行動調査を可能にする研究に取り組む必要がある．

また，技術的な面では，ジオタグ付きTweetのように，位置情報と時間情報，テキスト情報を持つ対象に関して，どのようにクラスタリングを行うべきかの考察を行った．ジオタグ付きTweetのようなデータセットを扱った先行研究が少ない中，位置情報・時間情報・テキスト情報の重み付けのバランスについて考察し，示すことができた．

今後の課題としては，クラスタリング・ラベリング共に精度の向上が必要であると考えられる．また，本研究で抽出した行動情報を利用するアプリケーションの開発を行いたいと考えている．

Webとリアル・ワールドの距離が近くなった今，Webからの情報抽出によるリアル・ワールドの拡張・改善に関する研究はこれから発展していくだろう．本論文で，Webとリアル・ワールドをつなぐ橋渡しに少しでも貢献できたら幸いである．

謝辞

本論文を執筆するにあたり，多くの方々にお世話になりました．

学部とは異なる研究分野に取り組むにあたり，研究に対する助言はもちろん，学問に対する取り組み方，思考方法など様々な面で指導して頂いた瀬崎薫教授，研究テーマの決定から結論のまとめ方まで，日常的に研究に対してアドバイスを頂き，また私の研究に関して様々な方とお会いする機会を提供していただいた瀬崎研究室岩井将行助教，研究室の先輩として研究に対する姿勢を常に示して下さった瀬崎研博士課程ハンセギョンさん，石塚宏紀さん，Asif Hossain Khan さん，党聡維さん，研究室で共に時間を過ごし，楽しく思い出に残る時間を共に過ごした同期の何斌斌君，澤上佳希君，高い意識を持ち，多くの刺激を与えてくれた後輩の李晨超君，汪少哲君，奥野淳也君，中山俊平くん，清水和人くん，Jose Pablo Alvarez Lacasia くん，研究活動や出張の手配で大変お世話になった秘書の松本夏穂さん，坪野明日香さん，多くの時間を共に過ごした空間情報学系研究室はじめ社会文化環境学専攻の友人たち，また，同じ研究グループとして学会や研究会の際にお世話になった東京電機大学，法政大学，富山県立大学のみなさまに，感謝の言葉を述べさせていただきたいと思います．学部時代を過ごした国立情報学研究所の KasM グループの皆様には，研究室を卒業した後も，研究に関して多大なアドバイスをいただきました．ここに御礼申し上げます．

最後に，生まれてから今に至るまで育ててくれた両親への敬愛と感謝の言葉で，本論文を締めくくりたいと思います．

発表文献

- 酒巻智宏, 岩井将行, 瀬崎薫. マイクロブログのジオタグを用いた行動調査の可能性に関する一考察. 第 4 回電子情報通信学会ヒューマンプローブ研究会, 2010
- 酒巻智宏, 岩井将行, 瀬崎薫. マイクロブログのジオタグを用いたユーザの行動パターンの推定に関する研究. 第 2 回集合知シンポジウム, 2011
- 酒巻智宏, 岩井将行, 瀬崎薫. マイクロブログのジオタグを用いたユーザの行動パターンの調査に関する研究. 第 73 回情報処理学会全国大会, 2011
- 酒巻智宏, 岩井将行, 瀬崎薫. 位置情報付き投稿におけるテキスト解析を用いたラベル付け手法の検討. FIT2011 第 10 回情報科学技術フォーラム, 2011
- 酒巻智宏, 岩井将行, 瀬崎薫. Twitter のジオタグを用いたユーザの活動地点の推定. 第 5 回電子情報通信学会ヒューマンプローブ研究会, 2011
- Tomohiro Sakamaki, Masayuki Iwai, Kaoru Sezaki. User Behavior Analysis using Twitter and Geo-tag. International Conference on Human Probes and Smartphone Sensing, 2011
- 酒巻智宏, 岩井将行, 瀬崎薫. マイクロブログのジオタグを用いたユーザの行動分析. 情報処理学会論文誌：データベース, 第 53 号（投稿中）

参考文献

[1] 大向一輝, 武田英明, 松尾豊. リアルワールドとしての web. 人工知能学会誌, Vol. 21, No. 4, pp. 403–409, 2006.

[2] 大向一輝. 3.web2.0 と集合知. 情報処理, Vol. 47, No. 11, pp. 1214–1221, 2006.

[3] 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.

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *WWW2010*, 2010.

[5] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 119–128, 2010.

[6] 森尾淳, 中野敦. パーソントリップ調査の実態調査上の問題点と改善手法. *IBS Annual Report*, pp. 86–88, 2006.

[7] 松本修一. Gps 携帯を活用した行動調査に関する基礎的研究. *KEIO SFC JOURNAL*, Vol. 9, No. 1, 2009.

[8] 長尾光悦, 川村秀憲, 山本雅人, 大内東. Gps ログからの周遊型観光行動情報の抽出. 情報処理学会研究報告, Vol. 1, No. 78, pp. 23–28, 2005.

[9] Pan Hui, Richard Mortier, Michal Piórkowski, Tristan Henderson, and Jon Crowcroft. Planet-scale human mobility measurement. In *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement*, pp. 1:1–1:5, 2010.

[10] Patrik Floreen, Michael Przybilski, Petteri Nurmi, Johan Koolwaaij, Anthony Tarlano, Matthias Wagner, Marko Luther, Fabien Bataille, Mathieu Boussard, Bernd Mrohs, and Sianlun Lau. Towards a context management framework for mobilife. In *In IST Mobile and Wireless Communications Summit*, 2005.

[11] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *In Proceedings of the first workshop on Online social networks*, pp. 19–24, 2008.

[12] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 2007.

[13] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768, 2010.

[14] 田中淳史, 田島敬史. twitter のツイートに関する分類手法の提案. *DEIM 2010*, 2010.

[15] semiocast. Only 30 % of messages on twitter are from the u.s.

[16] Eran Toch, Justin Cranshaw, Paul Hankes Drielsma, Janice Y. Tsai, Patrick Gage Kelley, James Springfield, Lorrie Cranor, Jason Hong, and Norman Sadeh. Empirical models of privacy in location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 129–138, 2010.

[17] 渡邉裕子, 小林一郎, 和泉憲明, 橋田浩一. イベント構造の抽出に基づく画像管理法. セマンティックウェブとオントロジー研究会, 2008.

[18] Tom Lovett, Eamonn O'Neill, James Irwin, and David Pollington. The calendar as a sensor: analysis and improvement using data fusion with social networks and location. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 3–12, 2010.

[19] Soren Auera, Jens Lehmann, and Sebastian Hellmann. Linkedgeodata - adding a spatial dimension to the web of data. In *8th International Semantic Web Conference (ISWC2009)*, 2009.

[20] Alexei Pozdnoukhov and Christian Kaiser. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pp. 8:1–8:8, 2011.

[21] Vivek K. Singh, Mingyan Gao, and Ramesh Jain. From microblogs to social images: event analytics for situation assessment. In *Proceedings of the international conference on Multimedia information retrieval*, pp. 433–436, 2010.

[22] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE'10*, pp. 201–212, 2010.

[23] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pp. 10:1–10:9, 2011.

[24] 江島啓介, 鈴木秀幸, 合原一幸. 東京都市圏パーソントリップ調査データに基づく新型インフルエンザ感染伝播の数理モデリング. 運輸と経済, Vol. 70, No. 1, pp. 54–62, 2010.

[25] 尾川春香, 小林廉毅. Gis を用いた都内分娩施設のアクセス評価. 日本公衆衛生学会総会抄録集, pp. 572–, 2009.

[26] 小高佑樹, 平野研人, 因雄亮, 北爪繭子, 樋口政和, 川崎秀二, 村上仁己, Omiya Yasuhiro, Kawata Keizo, Murakami Hitomi. 携帯電話 gps の特性評価 : Gps 誤差の群特性. 映像情報メディア学会技術報告, Vol. 34, No. 10, pp. 5–8, 2010.

[27] NTT ドコモ. モバイル空間統計に関する情報.

[28] Sibren Isaacman, Richard Becker, Stephen Kobourov3, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *Pervasive Computing*, pp. 133–151, 2011.

[29] Jialiu Lin, Guang Xiang, Jason I. Hong, and Norman Sadeh. Modeling people's place naming preferences in location sharing. In *Ubicomp '10*, pp. 75–84, 2010.

[30] 鈴木信雄, 津田和彦. 移動要求抽出のための移動目的発言分類法に関する一考察. 人工知能学会全国大会論文集, Vol. 24, No. 3, pp. 31–32, 2010.

[31] Nguyen MinhThe, 川村隆浩, 中川博之, 田原康之, 大須賀昭彦. Cgm からの自己教師あり学習と条件付き確率場を用いた人間行動マイニング. 人工知能学会第 24 回全国大会, 2010.

[32] 佐々木健太, 長野伸一, 長健太, 川村隆浩. Web 上のライフストリームからのユーザ行動情報の抽出. 人工知能学会第 25 回全国大会, 2011.

[33] 竹内孔一. 意味の包含関係に基づく動詞項構造の細分類. 言語処理学会第 14 回年次大会発表論文集, pp. 1037–1040, 2008.

[34] 青木政勝, 瀬古俊一, 西野正彬, 山田智広, 武藤伸洋, 阿部匡伸. ライフログのための位置情報ログデータからの移動モード判定の検討. 電子情報通信学会技術研究報告. OIS, オフィスインフォメーションシステム, Vol. 108, No. 156, pp. 7–12, 2008.

[35] 吉井英樹, 白井隼人, 佛圓俊一朗, 小松尚久. 区間速度を用いた位置情報履歴データの分析手法に関する検討. 電子情報通信学会技術研究報告. LOIS, ライフインテリジェンスとオフィス情報システム : IEICE technical report, Vol. 109, No. 39, pp. 109–112, 2009.

[36] 青木政勝, 瀬古俊一, 西野正彬, 山田智広, 武藤伸洋, 阿部匡伸. Gps 未計測区間における移動手段判定手法の検討. 情報処理学会研究報告. UBI, Vol. 2008, No. 110, pp. 39–44, 2008.

[37] 小林亜令, 岩本健嗣, 西山智. 釈迦：携帯電話を用いたユーザ移動状態推定・共有方式. 電子情報通信学会技術研究報告. MoMuC, Vol. 108, No. 44, pp. 115–120, 2008.

[38] Peter Diggle, Barry Rowlingson, and Ting-li Su. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, Vol. 16, No. 5, pp. 423–434, 2005.

[39] 山中努, 田中祐也, 土方嘉徳, 西田正吾. 時空間情報を伴うテキストデータを用いた状況把握支援システム. 知能と情報：日本知能情報ファジィ学会誌：journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 22, No. 6, pp. 691–706, 2010-12-15.

[40] 酒巻智宏, 岩井将行, 瀬崎薫. マイクロブログのジオタグを用いた行動調査の可能性に関する一考察. 電子情報通信学会 HPB 研究会, 2010.

[41] 神嶌敏弘. データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう！ －. 人工知能学会誌, Vol. 18, No. 1, pp. 59–65, 2003.

[42] 元田浩, 山口高平, 津本周作, 沼尾正行. データマイニングの基礎. オーム社, 2006.

[43] Dan Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 727–734, 2000.

[44] 石岡恒憲. x-means 法改良の一提案：k-means 法の逐次繰り返しとクラスターの再併合. 計算機統計学, Vol. 18, No. 1, pp. 3–13, 2006.

[45] 志津綾香, 松田眞一. クラスター分析におけるクラスター数自動決定法の比較. アカデミア, 情報理工学編, Vol. 11, pp. 17–34, 2011.

[46] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, Vol. 38, No. 2, pp. 321–330, 2004.

[47] 中谷友樹, 矢野桂司. 犯罪発生の時空間 3 次元地図：ひったくり犯罪の時空間集積の可視化. 地學雜誌, Vol. 117, No. 2, pp. 506–521, 2008.

[48] 古谷知之. 観光行動データの時空間データマイニング. *Keio SFC journal*, Vol. 9, No. 1, pp. 41– 51, 2009.

[49] 平博順, 向内隆文, 春野雅彦. Support vector machine によるテキスト分類. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 98, No. 99, pp. 173–180, 1998.

[50] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *In Proceedings of the Twentieth International Conference on Machine Learning*, pp. 616–623, 2003.

[51] 工藤拓. Mecab : Yet another part-of-speech and morphological analyzer.

[52] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

付 図

付 表