Vol.2013-DBS-157 No.8
Vol.2013-IFAT-111 No.8
2013/7/22

地域に偏りのあるトピックを用いた Twitter ユーザの生活に関わる地域推定

堂前 友貴[1,a)]　関 洋平[2,b)]

概要：本研究では，ツイート中のトピックには，地域に偏りがあるものと，共通で現れるものがあるという仮定のもと，トピックを利用し，Twitter ユーザの生活に関わる地域を推定する手法を提案する．まず，訓練データとして，ロケーション項目に地域名を記述しているユーザのツイートを使用し，LDA を用いて地域ごとにトピックを作成する．次に，各トピックを比較し，地域に偏りのあるトピックに地域ラベルを付与する．そして，地域ラベルが付与されたトピックに対応するツイートを使用し，ツイートをトピックに分類する．各ユーザの生活に関わる地域は，ユーザの ツイート に割り当てられたトピックに対して，付与された地域ラベルに基づき推定する．都道府県を，生活にかかわる地域の単位とし，16 の都道府県を対象として，ユーザの生活に関わる地域の推定実験を行ったところ，精度 0.59，再現率 0.54，F 値 0.56 となった．

キーワード：Twitter，地域推定，トピックモデル．

Twitter User's Life Area Estimation Using Biased Topics Reflecting Area

Abstract: In this study, we propose Twitter user's life area estimation method using topics, by the assumption that the topics in the tweet appeared differently according to the area. Topics are created with LDA in each area by using the tweet sets from users who describe the area names in the Twitter's location field as training data. Then, we compare each topics, and grant regional label to the biased topic reflecting area. We implement classifiers to categorize tweets into the topics relevant to the area. Twitter user's life area is estimated based on the majority area labels in the topics assigned to the user's tweets. Through the user's life area estimation experiment by defining prefecture as area unit, and select 16 unit. The result was precision 0.59, recall 0.54, F-measure 0.56.

Keywords: Twitter, Area estimation, Topic model.

1. はじめに

現在，マイクロブログと呼ばれる，自身の状況や雑記などを短い文章で気軽に投稿・共有することができるサービスの利用が活発に行われている．中でも，代表的なマイクロブログの一つである Twitter[*1]は，ツイートと呼ばれる短文の投稿が，1 日に 3 億 4 千万件行われる人気サービスである．多くの人が利用するサービスである Twitter のユーザを対象とした，ユーザ支援や情報推薦の研究，アプリケーションの開発などは，数多く行われている．この際，ユーザが生活している地域などの属性情報は，ユーザを検索する際の指標や，適切な情報の推薦に有用なものである．しかし，マイクロブログにおいては，地域・年代等のユーザの属性となるデータを記述する項目が，bio と呼ばれる自由記述の自己紹介文と，ロケーションという場所情報に限られている．日本語 Twitter ユーザに対する，伊藤ら [8] の調査では，約 4,600,000 ユーザに対する属性が推定できる単語を利用した分析で，記述率は性別で 7.62%，年齢（年代）で 3.34%，職業で 13.62%，地域で 24.98%という結果が報告されている．この調査結果にある，地域を

[1] 筑波大学大学院 図書館情報メディア研究科
〒 305-8550 茨城県つくば市春日 1-2
[2] 筑波大学 図書館情報メディア系
〒 305-8550 茨城県つくば市春日 1-2
[a)] doumae@slis.tsukuba.ac.jp
[b)] yohei@slis.tsukuba.ac.jp
[*1] https://twitter.com/

明示的に記述していない約 75%のユーザについて，生活に関わる地域を推定できれば，地域を指定した世論調査や，ユーザ同士の交流の支援などへの応用が期待できる．

そのため，本研究では，ユーザの潜在的な属性の一つとして，ユーザの生活に関わる地域の推定に取り組む．ここで，ユーザの生活に関わる地域とは，居住地や勤務地など，日常生活で関わることの多い地域とする．地域など，Twitter ユーザの属性には，ツイートのデータを手がかりとした分類器を用いた手法 [10,12] がある．これらの手法は，ユーザのツイート集合を 1 つの文書として扱い，タスクごとに収集した訓練データとなるユーザのツイート集合から語彙的な情報などを学習し，推定を行う．これらの研究では．クラスごとのツイート集合に対する単語の出現傾向に基づき，推定の手がかりとなる素性を選択する．

本研究では，地域の推定に有用な手がかりとして，トピックを用いる．トピックを用いることで，単語単体で比較する時には得られなかったが有用な手がかりを利用できる．具体的な手法としては，まず，教師なし機械学習の手法であるトピックモデルを用いて，地域に偏りのあるトピックの抽出を行う．そして，地域ラベルが付与されたトピックに対応するツイートを使用して，ツイートをトピックに分類し，各ユーザの生活に関わる地域は，ユーザのツイートに割り当てられたトピックに対して，付与された地域ラベルに基づき推定する．

本論文の構成を以下に示す．2 節で Twitter ユーザの属性推定に関する関連研究の紹介を行い，3 節で提案手法の詳細について述べる．4 節では，生活に関わる地域の単位として，都道府県を設定し，提案手法の検証のための評価実験を紹介する．最後に，5 節でまとめと今後の課題について述べる．

2. 関連研究

Twitter ユーザの属性推定に関する研究について，特に，地域を対象として行っているものを中心に示す．地域など，Twitter ユーザの属性の推定は，ユーザのツイート，bio，また，ユーザのふるまい（リスト，フォロー関係など）等のデータを元として行われる．推定手法は，タスクごとに収集したユーザを訓練データとした，機械学習による手法，特に，SVM などの分類器を使用した手法 [4,5,10,12] が主流である．分類器を用いた手法のベースとしては，ユーザのツイート集合を 1 つの文書として扱い，タスクごとに収集した訓練データとなるユーザのツイート集合から語彙的な情報などを学習し，推定を行うものが多い．Hecht ら [4] は，居住地についての研究で，36%ユーザが一切の場所情報を提供していない状態であっても，潜在的な単語分布の偏りにより，適切な推定が行えることを示している．彼らは，単語の出現頻度を特徴量とした分類器による実験で，Country で 88.9%，State で 30.3%の分類精度で推定を行った．また，クラス間に出現するの語彙の違いを利用した素性選択による手法には，池田ら [12] や西村ら [10] の提案がある．池田ら [12] は，AIC (Akaike's Information Criterion) の考え方を用いて，あるクラスに偏って出現するキーワードリストを作成し，それを素性とした SVM による多クラス分類器を作成し，居住地（8 地方区分）で 0.71 など比較的高い精度で推定を行っている．西村ら [10] は，地域特徴語を利用した SVM による多クラス分類による，ユーザの居住地推定を行っている．地域を単位とした TF-IDF に基づき獲得した地域特徴語を素性とし，47 都道府県を推定単位とした居住地推定を行い，F 値 0.34 を達成している．

このようなツイート集合などを用いる手法では，コンテンツの量が性能に影響するという報告もなされている．Burger ら [2] は，語彙的な特徴を利用した分類器の研究で，ツイートの量や bio の有無が性能に影響していると報告している．

また，ソーシャルグラフ上のユーザ属性伝搬による評価値の向上を図る研究 [6,11] もある．これらは，同一属性をもつユーザ同士が交流を行いやすいという仮説に基づき，行われている．Zamal ら [6] は，年代，性別，政治的指向についての研究で，語彙的な素性を用いた分類器による推定結果に，フレンド関係にあるアカウントのデータを使うことで推定精度が向上することを報告している．

本研究では，機械学習によりクラスごとの特徴の獲得を行うが，ツイートごとにトピックを付与し，トピックに付与されたラベルに基づいて属性値を推定する点で従来手法と異なる．トピックを用いることで，単語単体で比較するときには得られなかった有用な手がかりを獲得できる．

3. ユーザの生活に関わる地域の推定手法

トピックには，地域に偏りがあるものと，共通で現れるものがあるという仮定のもと，トピックを利用したユーザの選別を行う．例えば，その地域にある球団に関するトピックや，地域イベントに関するトピックは偏りがあり，朝の挨拶や，全国放送のテレビ番組のトピックなどはどのような地域でもほぼ出現すると考えられる．

提案手法の概要図を図 1 に示す．ユーザの生活に関わる地域の推定は，以下の手順で行う．

(1) トピックの生成
　LDA により，地域ごとにトピックを生成する．
(2) トピックへの地域ラベルの付与
　作成されたトピックの比較を行い，地域ごとの偏りを抽出し，地域名のラベルを付与する．
(3) ユーザの生活に関わる地域の推定
　(a) ツイートの地域性推定
　　ツイートごとに，地域に偏りがあるトピック，または地域に偏りのある共通語が出現するかを推定する．

(b) ユーザの生活に関わる地域の推定
最終的な，ユーザの地域を決定する．
各段階の詳細な手法について，以降の節に記述する．

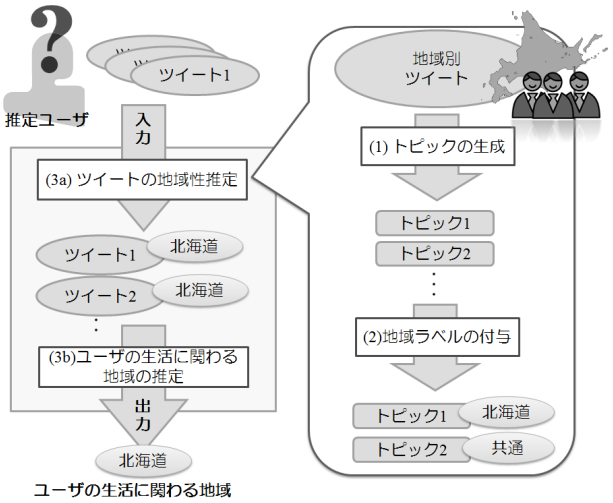

図 1 ユーザの生活に関わる地域の推定

3.1 トピックの生成

本手法では，文書の確率的な生成モデルである LDA(Latent Dirichlet Allocation) [1] を用いて，いくつかの地域ごとにトピック分析を行い，その結果の比較により，地域に偏りのあるトピックを抽出する．

短い文書であるツイートに対し LDA を適用するモデルとして，Zhao ら [7] の提案した Twitter-LDA がある．このモデルの特徴は，(1) ユーザのツイートをまとめて 1 つの集合として扱う点，(2) 1 ツイート 1 トピックとしてトピックを割り当てる点の二点である．本研究でも，このモデルを参考としたトピックモデルを用いる．また，固有名詞を除くキーワードに関しては，Diao ら [3] の研究で提案されている，トピック間で共通に出現するような語（共通語）はトピックの推定からのぞいて使用する手法を適用する．サンプリング方法には，崩壊型ギブスサンプリングを用いた．ツイートのトピックの更新式を式 (1)，キーワードのバックグラウンドかどうかの判定の更新式を，式 (2) に示す．

$$p(z_i) \propto (N_{c|u}+\alpha)\cdot\frac{\Pi_{v=1}^{V}\Pi_{k=0}^{E(v)-1}(n_{v|c}+k+\beta)}{\Pi_{k=0}^{E(v)-1}(n_{.|c}+k+V\beta)} \quad (1)$$

$$p(x_{i,j}) \propto \frac{n_{q|p}+\gamma}{n_{.|p}+2\gamma}\cdot\frac{n_{\omega_{i,j}|l}+\beta}{n_{.|l}+V\beta} \quad (2)$$

ここで，C はトピック数，V は異なりキーワード数，k は総キーワード数，N はツイート のカウント，n はキーワードのカウント，c はトピック，u はユーザをそれぞれ表し，l はスイッチである．なお，α,β,γ はパラメータであり，本実験では，予備実験により最適な値を求め，$\alpha=0.1$，$\beta=0.03$，$\gamma=1.0$ で行った．

この時得られた共通語の集合には，特定地域内でのみどのようなトピックにも共通で現れるような語というのが存在すると考えられる．そのため，全ての地域の共通語の生起確率を正規化し，上位の語の比較を行う．特定の地域に偏って出現する語に関しては，トピックとは別に，特定地域に偏って出現しやすい語として扱う．

3.2 トピックへの地域ラベルの付与

3.1 節で生成された各トピックに，それぞれの地域のラベルを付与する．しかし，3 節冒頭で述べたように，どのような地域でも共有されるトピックが存在することが予測される．そのため，トピック間の類似度の分析を行い，地域ごとに特有なトピックの抽出を行い，偏りがあると判断されるものに地域ラベルを付与する．

類似度の計算には，トピックモデルで生成されたトピックの類似度の計算に用いられることが多い，JS–divergence (Jensen–Shannon divergence) [9] を用いた．JS–divergence は，2 つの確率分布間の距離尺度であり，2 つの確率分布間の類似性が高いほど値は小さくなる．2 つの確率分布，P，Q に対して，類似度 $JS(p,q)$ は以下のように求める．

$$JS(p,q)=\frac{1}{2}(KL(p,r)+KL(q,r)) \quad (3)$$

$$KL(p,q)=\sum p(x)\log\frac{p(x)}{q(x)} \quad (4)$$

$$r(x)=\frac{p(x)+q(x)}{2} \quad (5)$$

また，地域特有であるが近隣の地域などと共有されるトピックがあると考えられる．そのため，過半数以上の地域で共有されるトピックは，推定に寄与しないトピックとして扱うが，過半数以下で共有されるトピックは，複数の地域ラベルを付与する．それぞれの地域の組み合わせに対し，トピックの類似度 $JS(p,q)$ が閾値以下となる地域数が，比較を行った地域数の半分以下のものに地域ラベルを付与する．

3.3 ユーザの生活に関わる地域の推定

前節で獲得した，ラベル付きトピックおよび，地域特有の共通語を用いてユーザの生活に関わる地域を推定する．推定の手順は下記の通りである．

(1) ツイート t のトピック z の算出
ツイート t のトピック z は，下記の式で求めた値が最大となるトピックである．なお，全てのトピックにおいて，$tScore(t,z)=0$ となるツイートには，トピックを付与しない．ここで，$p(w,z)$ はトピック z における単語 w の生起確率である．

$$tScore(t,z) = \sum p(w,z) \quad (6)$$

(2) ユーザ u 地域に対しての重み $Score(u,l)$ の算出

あるユーザ u のツイート集合 T のうち，ラベル付けされたトピックが付与されるか，また，地域特有共通語が含まれるかを判定し，地域ごとに重みを付与していく．ここで，$N_{top}(l)$ は，地域 l がラベル付けされたツイート数，$N_{bg}(l)$ は地域特有の共通語を含むツイート数である．

$$Score(u,l) = N_{top}(u,l) + N_{bg}(u,l) \quad (7)$$

(3) ユーザ u の生活に関わる地域 $location(u)$ の選択

ユーザの地域に対しての重み $Score(l)$ が最大のものを選択する．

4. 評価実験

提案手法の有効性を検証するために，評価実験をおこなった．以降，実験方法，データ，結果，考察について述べる．

4.1 実験方法

生活に関わる地域の単位として都道府県を採用し，従来研究 [10] を参考とした手法との比較評価実験を行う．

推定を行う地域は，近隣地域の傾向と離れた地域間の傾向を分析するため，関東地方の 1 都 6 県（東京都，茨城県，栃木県，群馬県，埼玉県，千葉県，神奈川県）と，近畿地方の 2 府 5 県（京都府，大阪府，三重県，滋賀県，兵庫県，奈良県，和歌山県），福岡県，北海道の 16 の都道府県を選択した．

トピックを作成する際の実験設定は，下記のように行った．トピック数は 100，繰り返し回数は 100 とし，使用する語は，形態素解析の結果，名詞，動詞，形容詞のいずれかに判定されたキーワードである．ツイートの形態素解析には，MeCab[*2]を使用した．この際，動詞や形容詞に関しては表記ゆれの影響を考慮し，全て基本形として扱った．また，トピックの比較に用いる JS–divergence の計算や，ツイートのトピック推定では，各トピック生起確率上位 1,000 語までを用いた．

また，比較手法として，地域特徴語により素性選択を用いた分類器により地域を推定する手法 [10] を参考にして実験を行った．この手法では，都道府県名をロケーション項目に記述したユーザの投稿を 1 文書とみて TF-IDF を使用することで，特徴的な単語を素性として選択している．素性は各都道府県の TF-IDF 値上位 5,000 件を選択し，分類器には LibSVM-3.12[*3]を用いた．使用する品詞は，提案手法と同じく，名詞，形容詞，動詞ある．

評価尺度は，推定の精度・再現率・F 値を採用した．

[*2] http://mecab.sourceforge.net/
[*3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

4.2 データ

トピック生成に用いるデータは，2012 年 6 月 15 日から 7 月 14 日の一カ月間に投稿された日本語ツイートのうち，投稿時のユーザのロケーション項目に，本実験で対象とする都道府県名が明記されているツイートである．なお，このツイートはいずれも，名詞，動詞，形容詞のいずれかを含んでいること，プログラムにより自動的に投稿を行う bot と呼ばれるアカウントをある程度除外するために，ユーザ名に “Bot”等を含んでいないことなどを条件としている．

このうち，提案手法のトピックの生成および，比較手法の地域特徴語の選択に使用するデータとして，各都道府県で 1 日最大 10,000 件，期間合計 300,000 件，16 都道府県の合計 4,800,000 件をランダムに抽出した．

また，比較手法での分類器への訓練データとして，ユーザの投稿数が機械学習の質に影響を与えることと，地域ごとの訓練データの量の偏りが精度に影響を与えることを考慮し，期間中に 100 件以上ツイートを行い，素性に選択された語を 1 回以上使用しているユーザ，各地域 4,000 件，合計 64,000 件に対し，最大 200 件のツイートを抽出した．

評価データには，人手で収集したユーザ，各地域 100 件，合計 1,600 件を用いた．このユーザは，bio またはロケーションに複数地域を記述しておらず，生活圏であることが確認できる記述（“xx 在住” など）があること，100 件以上の取得できる投稿があることなどを条件として収集を行った．評価に用いるユーザのデータは，ユーザ 1 件あたり収集時点から最新のツイート最大 200 件である．

4.3 結果

トピックの生成と地域ラベルの付与，ユーザの生活に関わる地域の推定についてそれぞれ結果を述べる．

4.3.1 トピックの生成と地域ラベルの付与

実験において生成されたトピックの分析や傾向については，考察で述べる．トピックの類似度の比較において，地域に偏りのあるトピックとして抽出されたトピックの一部を，表 1 に示す．1 は地域にある施設に関するトピック，2 は地域でのイベントに関するトピック，3 と 4 は首都圏の交通状況に関するトピックで，関東の一部地域間で共通のトピックである．

共通のトピックに関しては，時事的なニュースやニコニコ動画の話題などが共通の話題として正しく抽出された．表 2 は共通の話題として抽出されたトピックで，「高橋容疑者逮捕」に関するトピックの一部である．

4.3.2 ユーザの生活に関わる地域の推定

生活に関わる地域の推定結果のマクロ平均を，表 3 に示す．データ量や使用する素性に制約などがある状況ではあるが，提案手法が比較手法を上回ることができた．また，各地域の F 値の比較を図 2 に示す．北海道や福岡県といった，近隣の地域が推定対象に含まれない地域の評価値

表 1 地域に偏りのあるトピックの一部

	作成地域	ラベル	上位語（10 件）
1	兵庫県	アンパンマンミュージアム	マン，神戸，アンパン，ランド，ミュージアム，アニメ，世 界，新聞，ウォーク，キャラクター
2	神奈川県	自衛隊一般公開	海軍，公開，護衛，自衛隊，東京，海上，ふ頭，晴海，指揮，統制
3	千葉県	首都圏の交通状況	駅，電車，分，運転，京成，バス，運休，京葉線，常磐線，総武線
4	茨城県	首都圏の交通状況	電車，車，運転，常磐線，帰宅，分，気，上野，バス，東京

表 2 共通のトピックの一部

	作成地域	ラベル	上位語（10 件）
1	東京都	高橋容疑者逮捕	高橋，者，容疑，克也，逮捕，男，確保，買取，漫画，身柄
2	千葉県	高橋容疑者逮捕	高橋，者，容疑，警察，逮捕，めん，時間，克也，店，君
3	栃木県	高橋容疑者逮捕	者，高橋，容疑，逮捕，男，ら，克也，オウム，警察，相談

表 3 実験結果

地域	精度	再現率	F 値
提案手法	**0.59**	**0.54**	**0.56**
地域特徴語選択（比較手法）	0.52	0.42	0.47

が高くなっている．また，東京都が最も低いといった結果になった．これは，東京都単体で特有なトピックというものが少なく，近隣の地域で共有されるトピックが多いことや，他の地域に比べ時事的な話題に関するトピックが特に多かったことが原因として考えられる．

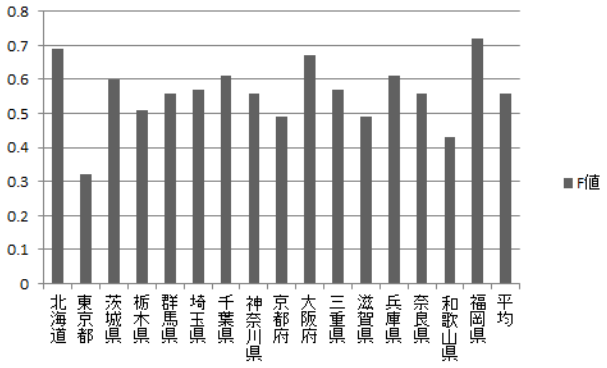

図 2 各地域における F 値の結果

4.4 考察

全ての地域で共有されはしないが，複数の地域で共有されるトピックには，近隣の地域のトピックが他の地域でも多く現れた．首都圏の交通状況に関するトピックや，イベントのトピックなどである．これは，千葉や神奈川などに住んでいて東京に仕事で通うといった，複数の地域間を移動して生活している人が多いことが関係しているのではないかと考えられる．

分析をおこなったところ，人手でラベルを付与すると共通の話題を示すものになると思われるトピックが類似のトピックとして判断されないケースが存在した．表 4 で示す大阪と京都で祇園祭りに関するトピックでは，話題としては同一のものであるが，出現する語が大きく異なり，JS–divergence の値も，閾値の 0.054 よりも，顕著に大きな 0.163 となり，類似と判定されなかった．京都での上位語は，代表的な行事に関する語が集中した大阪と異なった結果となった．このような傾向の違いは，推定に有用なものだと考えられる．しかし，地域に特有でない話題に関して，共通と思われる話題で類似でないと判定され，地域特有のトピックとなってしまう可能性もある．表 2 で示した「高橋容疑者逮捕」に関してのトピックは，京都府でも生成されたが，他の政治ニュースと同一のトピックとして生成され，他と傾向の異なる語が多く上位に現れ，他の地域の「高橋容疑者逮捕」のトピックと類似とは判断されなかった．

また，地域的な偏りがあるようには見えない話題に関して，ツイート数が著しく多く，話題が詳細化しているトピックも見られた．一例として，群馬では政治関係のトピックが多く見られた．期間前後で首長選挙や議会選挙などに該当するものが，7 月 15 日の甘楽町長選挙だけであるにも関わらず，このような結果となったのは，これが，実際に政治に関心が強い人が多いのか，期間に依存するものか，収集したデータが偶然偏っていたのか，といったことは，今後他の期間の異なるデータセットでの実験や分析で明らかにしていきたい．

また，地域特有ではあるが，再現性が不明なトピックの扱いについても検証が必要である．事前にトピックを作成する場合，再現性のないトピックは，他の期間のデータの推定には有用でない．一例として，自然災害に関するトピックが挙げられる．これらのトピックは，地名なども含まれる点や，沖縄や九州などでは毎年特定の時期に台風が来やすい点などを考慮すると，有用な情報ととるとこもできるが，他の地域にいつ現れるかわからないといった問題もある．再現性が期待できないようなトピックに関しては，今後，異なる期間でのトピックの作成などを通し，周期性・季節依存性があるかの抽出とともに分析を行っていきたいと考えている．トピックを作成するデータの期間と，推定を行うユーザのデータの期間の設定の仕方により，再現

表 4　同一の話題で出現語が異なるトピック

	作成地域	ラベル	上位語（10 件）
1	京都府	祇園祭（京都）	手，女の子，顔，浴衣，ノリ，姿，体温，接近，的，関
2	大阪府	祇園祭（京都）	日本，祇園祭，水，充，大阪，久保，よね，日，京都，手段

性のないイベントのトピックを除外することや，通年で有用なトピック，季節的に有用なトピックを判別し，推定するユーザのツイートの投稿時期によって，推定に用いるトピックを変えることで，推定の精度の向上が期待できる．

5. おわりに

本論文では，トピックモデルを利用してユーザの生活に関わる地域を推定手法を提案した．ユーザのツイートを一つの集合としてではなく，個々のツイートで地域性の判別を行う点と，その推定にトピックを利用する点が従来手法とは異なる．具体的には，まず，地域ごとにトピックモデルを適用しトピックを作成し，その比較による地域に偏りのあるトピックの抽出を行った．そして，ユーザのツイートごとに地域性の判別を行い，最終的なユーザの生活に関わる地域を推定した．評価実験において，データ量や使用する素性に制約などがある状況ではあるが，提案手法が，素性選択を行い，ツイートを１つの集合として扱う分類器による従来研究を参考とした手法を上回ることができた．

今後の課題としては，まず，より適切に地域を表すトピックの抽出が挙げられる．これには，トピックモデルの改良の他，異なる期間でトピックモデルの適用を行い，周期性・季節依存性のあるトピックの抽出や，再現性のない時事的なトピックの排除などを行う予定である．

次に，クラス間の語彙の比較による特徴語の選択といった，現在利用していない手法との併用の検討である．クラス間で出現する語彙の違いを利用した地域の推定は，8 地方区分などの粒度の大きな地域区分では，比較的高い精度が行えることが報告されている [12]．そのため，段階を踏み，はじめに大きな粒度の地域を推定したあと，その中の粒度の小さな地域に推定を行うことなども考えられる．

謝辞

本研究の一部は，科学研究費補助金基盤研究 C（課題番号 24500291），基盤研究 B（課題番号 25280110）ならびに萌芽研究（課題番号 25540159）の助成を受けて遂行されたものです．

参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309, Edinburgh, Scotland, 2011.

[3] Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, pp. 536–544, Stroudsburg, PA, USA, 2012.

[4] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber ' s heart: The dynamics of the location field in user profiles. In *Proceedings of the 17th ACM International Conference on Human Factors in Computing Systems*, pp. 237–246, Vancouver, BC, Canada, 2011.

[5] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: User classification in twitter. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge and Discovery Democrats*, pp. 430–438, 2011.

[6] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the 6th International Conference on Weblogs and Social Media*, pp. 388–390, 2012.

[7] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Sonference on Advances in Information Retrieval*, pp. 338–349, Berlin, Heidelberg, 2011.

[8] 伊藤淳, 西田京介, 星出高秀, 戸田浩之, 内山匡. Twitter と Blog の共通ユーザプロフィールを利用した Twitter ユーザ属性推定. 情報処理学会研究報告．情報基礎とアクセス技術研究会, Vol. 210, No. 4, pp. 1–8, 2013.

[9] 横山正太朗, 江口浩二, 大川剛直. 潜在トピックを用いたブログ空間からの情報伝搬ネットワーク抽出. 電子情報通信学会論文誌. D, 情報・システム, Vol. 93, No. 3, pp. 180–188, 2010.

[10] 西村駿人, 数原良彦, 鷲崎誠司. 地域特徴語選択を用いたマルチクラス分類による twitter ユーザの居住地推定. 電子情報通信学会技術研究報告：信学技報, Vol. 112, No. 367, pp. 23–27, 2012.

[11] 蔵内雄貴, 内山俊郎, 内山匡. マルコフ確率場を用いたソーシャルネットワークからのユーザ属性推定. 第 5 回 Web とデータベースに関するフォーラム WebDB Forum 2012 論文集, pp. C–1, 2012.

[12] 池田和史, 服部元, 松本一則, 小野智弘, 東野輝夫. マーケット分析のための Twitter 投稿者プロフィール推定手法. 情報処理学会論文誌 CDS, Vol. 2, No. 1, pp. 82–93, 2012.