情報処理学会論文誌　データベース　Vol.7　No.3　1–13 (Sep. 2014)

半教師ありトピックモデルにより選択した地域特徴語を用いたTwitterユーザの生活に関わる地域の推定

堂前 友貴[1]　関 洋平[2,a)]

受付日 2014年3月20日, 採録日 2014年7月7日

概要：Twitter において，ユーザの生活に関わる地域は，社会行動の分析において重要な属性の 1 つであるが，プロファイルに明示的に記述されていることは少ない．本研究では，Twitter ユーザを対象として，半教師ありトピックモデルを利用した地域特徴語の選択に基づく，生活に関わる地域属性の推定手法を提案する．本研究では，半教師ありトピックモデルにより地域に特徴的な語を選択する．具体的には，地域情報サイトから収集した地域特徴語を含むツイートを教師データとした，半教師ありトピックモデルにより，地域に特徴的なトピックを抽出する．そして，トピックから選定した地域特徴語を使用し，ツイートごとに地域ラベルを付与する．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地域ラベルに基づき推定する．提案手法に基づき，都道府県を，生活に関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 0.65，再現率 0.67，F 値 0.66 の評価値が得られた．

キーワード：Twitter，地域推定，半教師ありトピックモデル

Estimation of Twitter User's Life-area Using Area Related Terms Selected by Semi-supervised Topic Model

Yuki Doumae[1]　Yohei Seki[2,a)]

Received: March 20, 2014, Accepted: July 7, 2014

Abstract: In Twitter, the life area of a user is an important attribute that is used for social behavior analysis. In most cases, information regarding a user's life area is not explicitly published in their Twitter profiles. We propose a method to identify the nature of a user's life area using area clue terms selected by a semi-supervised topic model. We extracted area-oriented topics by semi-supervised learning using terms collected from an area information website as supervision. We assigned an area label to each tweet using area-oriented terms from the extracted topics. The nature of a Twitter user's life area is identified as the area label that is most frequently used for topics identified in the user's tweets. We have evaluated our approach using 1,600 users from 16 Japanese prefectures. The result for precision, recall, and F-measure were 0.65, 0.76, and 0.66, respectively.

Keywords: Twitter, area estimation, semi-supervised 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気軽に投稿・共有できるサービスの利用が活発である．なかでも，代表的なマイクロブログサービスである Twitter[*1]は，ツイートと呼ばれる 140 字以内のメッセージ[*2]が，1 日に約 5 億件行われている人気サービスである．そのため，それらのサービスを対象とした，ユーザ支援や情報推薦などの研究が数多く行われている．この際，ユー

[1] 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[2] 筑波大学図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba, Ibaraki 305–8550, Japan
[a)] yohei@slis.tsukuba.ac.jp

[*1] https://twitter.com/
[*2] https://support.twitter.com/articles/247765

ザの性別や生活する地域などの，属性情報は，ユーザを検索する際の指標や，適切な情報の推薦などに有用なものである．その中で，本研究では，様々な応用が期待できるという理由から，地域を選択し，ユーザの生活に関わる地域の推定に取り組む．ユーザの生活に関わる地域が推定できれば，地域ニュースや広告のより適切な推薦，地域ごとのユーザの分析 [19] などに利用できる可能性がある．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．

Twitter においては，地域・年代など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 160 文字以内の自由記述の自己紹介項目と，ロケーションという場所を記述する項目に限られており，ユーザの属性情報が明示されないことが多い．日本語 Twitter ユーザに対する，伊藤ら [12] の調査では，約 4,600,000 ユーザに対する，属性が推定できるキーワードを利用した分析で，地域の記述率は 24.98%という結果が報告されている．この調査結果からも分かるように，多くのユーザが属性を記述しておらず，これらのユーザの属性を推定することで応用が期待できる．

Twitter ユーザの潜在的な地域属性を推定するためには，一般的な戦略として，地域を示すラベルを利用した教師あり学習と，地域内の類似度を利用したクラスタリングが考えられる．教師あり学習に基づく戦略の関連研究としては，文献 [6], [16], [17] があげられる．この戦略には，地域属性との対応関係を考慮した推定が行えるメリットがあるが，ツイートのような短文の投稿を対象とした場合，手がかりを素性として適切に反映できない可能性がある．一方，クラスタリングに基づく戦略を採用すると，地域属性との対応関係を明確に反映した学習がむずかしい．

本研究では，地域属性に依存したトピックをふまえつつ，手がかりとなる素性をラベルなしデータから獲得することを重視し，半教師ありトピックモデル [4] を採用した地域属性の推定手法を提案する．Twitter のツイートは，短文の雑談であるため，明確に地域名を記述しない場合がある．したがって，事前に地名辞典などから定義した手がかりだけを利用するのではなく，その地域でつぶやかれているトピックを利用して，地域に特有のイベントを含む手がかりなどを拡張することで，多くのユーザを対象とした地域属性の推定を実現でき，再現率の向上が見込める．

さらに，Twitter におけるツイートには，表記揺れの特徴がある．各地域のユーザの全ツイートを対象として，地域ユーザのツイート集合を文書と見なした TF-IDF のような特徴量 [16] に基づき，手がかりを選択してしまうと，表記揺れのある手がかりは，その頻度が，他の特徴と比較して相対的に疎となり，適切に抽出できない可能性がある．一方で，地域に関連するトピックに絞り込んで手がかりを抽出できれば，地域特有の手がかりは相対的に密となることから，上記の問題は解消されることが期待できる．本研究では，地域情報サイトから収集した地域特徴語を含むツイートを教師データとした半教師ありトピックモデル [4] に基づき，地域に特徴的なトピックを抽出し，地域特徴語を選択する．また，選択した地域特徴語を利用して，ツイートに地域ラベルを付与し，ユーザの生活に関わる地域を推定する．

本論文の構成を以下に示す．2 章で属性推定とトピック生成に関する関連研究を，Twitter を対象とした研究を中心とした紹介を行い，3 章で提案手法の詳細について述べる．4 章では，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．5 章では，4 章の実験結果への考察を述べる．最後に，6 章でまとめと今後の課題について述べる．

2. 関連研究

2.1 ユーザの属性推定と地域特徴的語の選択

本研究は，ユーザの属性推定の 1 つと位置づけられる．本節では，Twitter ユーザの属性推定について，属性推定の手がかりの獲得に関する研究を中心に紹介する．また，地域判定特有の課題として，地域特徴語の獲得を行っている研究についても紹介する．

Twitter ユーザの属性については，そのプロフィール情報である bio における記述が少ないことから，属性を推定する研究がさかんに行われており，性別 [2], [8], [18]，年代 [8], [18]，政治的指向 [6], [8], [18]，地域 [3], [16], [17], [18], [20]，属性全般 [13] などが対象となっている．推定にあたっては，その属性特有の手がかりをどのように獲得するかが課題となる．推定の手がかりは，ユーザのツイート，bio，また，ユーザの振舞い（リスト，フォロー関係，メンション情報）などのデータが利用される．

推定手法は，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 [6], [16], [17] が用いられる場合がある．分類器を用いた手法は，ユーザのツイート集合を 1 つの文書として学習を行い，対象とするユーザの推定を行う研究が多い．この際，教師データから，推定の手がかりをどのように選択するかが，推定の精度に大きく寄与する．ユーザ属性の手がかりを表層的特徴から求める場合，属性のクラスに応じた統計的な偏り [16], [17] や，人手で選択した手がかり [8] を利用する．

池田ら [17] 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 により，居住地（8 地方区分）について，推定精度 0.71 を達成している．西村ら [16] は，地域特徴語を利用した，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 SVM を利用して行い，F 値 0.34 を達成している．これらの手法では，キーワードの出現頻度の比較を

行っているため，特有なキーワードであっても，表記揺れなどにより同じ意味の出現傾向が一貫して出現せず，疎になる場合，手がかりとして獲得できない可能性がある．一方で，機械的に偏りを抽出するのではなく，人手で選択した特徴を利用して推定する研究もある．Rao ら [8] は，性別・年代・地域の起源，政治的指向の属性推定に，社会言語学的な特徴を使用している．社会言語学とは，年齢，性別，社会階級の条件による言語特徴の違いを研究した学問で，たとえば，女性の方が「ええ」などのチャネル応答が多いなどの特徴がある．彼らは，それぞれを 2 値分類問題とした SVM で，これらの社会言語学的な特徴と，unigram と bigram の語彙的な特徴を使用し実験を行い，性別と年代において N-gram のみを使用したものよりも，高い評価値が得られることを確認した．

本研究では，トピック単位で，属性クラスに偏りのある手がかりを選択する．ツイートは，各ツイートに含まれる単語が少なく，表記揺れも多い．トピックを利用すれば，出現傾向の類似した語の集合を，手がかりとして獲得することができる．Pennacchiotti ら [6] は，ユーザ中心の情報（プロフィール，語彙，振舞い，ソーシャル性）と，ソーシャルグラフの情報を統合し，ユーザ間の関係を考慮してグラフのアップデートを行う．彼らは，SVM に用いる語彙の素性選択の 1 つに LDA を用いているが，教師なし LDA を使用している点が本研究とは異なる．本研究では，地名や施設名など，地域属性に特有の手がかりを教師データとして利用して，地域特徴語を獲得する．

ユーザの属性推定には，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 [10], [18] や，メンション情報を利用した研究 [13] も行われている．これらは，同一属性を持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．これらの手法は，ソーシャルグラフによってのみ推定を行うのではなく，分類器などによる手法と組み合わせ，その推定精度の向上に用いられる．本研究では，これらの研究に基づく精度向上は今後の課題とする．

また，地域属性の手がかりとしては，地名やその地域にしか存在しない施設名，あるいはその位置情報など，地域属性に特有の手がかりを利用できる．これに関連して，地域に特徴的な語を獲得する研究について紹介する．奥ら [14] は，対象となる空間全体に対し，対象地域における出現頻度が相対的に高い語句を地域性の高い語句と定義し，IDF と市区町村名との共起頻度を考慮し，スコアを計算する方法を提案している．今井ら [15] は，ブログ記事内の共起回数と，地理的な距離に基づくスコアを同時に使用することで，POI（Point of Interest）に関連する地名を抽出している．また，Cheng ら [3] は，ジオタグ付きツイートから，地域に特有に出現する単語を判別し，近隣地域を考慮してスムージングを行うことにより，地域属性判定のための手がかりを獲得している．これらの研究から，地名や施設名などを獲得し，その語と共起する語から特徴的なものを選択することで，地域に特有な語が選択できることが分かる．

本研究では，機械学習で使用する語彙の選択に，トピックを用い，その作成に地域情報サイトから収集した語彙を用いた半教師あり学習を適用する．また，ユーザのツイート集合を 1 文書として一括でクラスに分類するのではなく，ツイートごとに地域ラベル付与し，属性値を推定する．これにより，話題のまとまりを考慮することで得られる地域特有の手がかりを利用できるようになり，また，地域情報に関係するツイートと関係しないツイートを区別することが可能となり，より精度の高い地域情報の推定が実現できる．

2.2 Twitter を対象としたトピックモデルと半教師あり学習

本手法では，Twitter のツイート集合からトピックを生成するのに，文書の確率的な生成モデルである LDA（Latent Dirichlet Allocation）[1] を利用する．そのため，Twitter を対象とした LDA や，教師あり LDA について，関連研究を紹介する．

LDA は短い文書に対しては適用が困難であることが指摘されており，通常のモデルをそのまま短い文書であるツイートに適用しても，適切な結果を得ることは難しい．Zhao ら [11] は，短い文書であるツイートに対するモデルを提案し，既存の LDA よりツイート集合に対し優れた性能を実現することを示している．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 2 点である．(1) は，ユーザのツイートは，それぞれのユーザのトピック分布に基づいて生成されると仮定している．本研究でも，LDA をツイートに対し適用するため，これらの仮定を参考とする．

本研究では，トピックを作成した後，地域に特徴的なトピックを選択する必要がある．あらかじめ獲得したいトピックのラベルが定まっており，教師データが獲得できる場合，ラベル情報を教師データとして与える教師あり LDA のモデルがいくつか提案されている．Ramage ら [7] の提案した Labeled LDA は，各文書のトピックが教師データとして与えられたラベル（ハッシュタグ（hashtag）[*3] など）による制約を受け，ラベルごとにトピックが割り当てられる．ここで，ツイート集合はラベルと潜在トピックの混合としてモデル化される．しかし，このモデルでは，1 つのラベルに対し 1 つのトピックしか割り当てられず，地域をラベルとした教師データとする場合，その地域内のトピックが 1 つだけであることは考えにくいため，適切なトピッ

[*3] ツイートのキーワードまたはトピックの印付けに使用される文字列

クとラベルの関係を得られないと考えられる．

さらに，地域に適切なトピック数がどの程度になるかということを事前に予測することは難しい．Teh ら [9] は，HDP（Hierechical Dirichlet Process）-LDA で，トピック数をあらかじめ決めなくても，パラメータを階層化し，トピック数を自動的に最適化するモデルを提案している．

この Teh らのモデルに，ラベルを付与する制約を加えたモデルとして，Kim ら [4] は，DP-MRM（Dirichlet Process with Mixed Random Measures）を提案した．このモデルでは，ラベルとトピックの対応が階層化され，その対応が自動的に獲得される．また，ラベルに複数のトピックが割り当てられるだけでなく，あるトピックに複数のラベルを付与することも可能となる．このモデルにより，地域内に存在する複数のトピックにラベルを付与でき，また，地域間で共有されるトピックについても獲得できると考える．そのため，本研究では，この DP-MRM のモデルを採用する．この際，対象の性質を考慮し，Zhao ら [11] の Twitter–LDA を参考として，ユーザごとに投稿したツイート集合をまとめて，1 文書として扱う．

3. ユーザの生活に関わる地域の推定手法

本研究では，半教師ありトピックモデルを利用した，Twitter ユーザの生活に関わる地域の推定を提案する．Twitter ユーザの生活に関わる地域の推定は，推定の前処理としてのトピックの生成による地域特徴語の選択と，対象ユーザの生活に関わる地域の推定の 2 段階に大きく分けられる．推定の主な手順は，以下のとおりである．

(1) 地域に特徴的なトピックの選択
　半教師あり LDA により，地域ラベルの付与されたトピックを選択し，推定に使用する地域特徴語を獲得する．
(2) 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，地域性の判定を行う．
　(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．

3.1 地域に特徴的なトピックの選択

半教師あり LDA により，ツイート集合から，地域ラベルの付与されたトピックを生成する．ここでは，半教師あり LDA として，Kim ら [4] の提案した DP-MRM（Dirichlet Process with Mixed Random Measures）を適用する．

本研究では，(1) 地域ラベルを付与できる，半教師あり学習のモデルである点，(2) トピック数を事前に指定する必要のないノンパラメトリックなモデルである点を考慮し，DP-MRM を採用する．(1) に関しては，ツイートには，様々なトピックが存在することから，生成したトピックの中から，その地域に特徴的な複数のトピックを選択する必要がある．地域に特徴的なトピックを選択するにあたっては，地域をラベルとした教師データを活用する手法が有効である．ただし，ツイートには，あらかじめ地域に相当するラベルが付与されていないため，地名や施設名などの地域に特徴的な語を含むツイートにラベルを付与し，教師データとして使用する．また，ツイートは短い文であり，文中に地名などを含んでいないが地域に特徴的なものもあるため，地域特徴語を含んでいないツイートデータも活用する必要がある．これらのデータを活用した半教師あり学習に基づきラベル付けを行い，あるラベルに対する複数のトピックを対応づける DP-MRM は，本研究に最適なモデルであると考える．(2) に関しては，各地域ラベルに対応するトピック数は，事前に予測することは難しいが，DP-MRM はノンパラメトリックモデルであり，適切なトピック数を決定できる．以上の理由から，本研究では，DP-MRM を採用し，地域に特徴的なトピックを選択する．

モデルについての説明を簡潔に行う．なお，本研究では，このモデルを，ツイート集合に対し適用する際，Zhao ら [11] の Twitter–LDA を参考とし，ツイート中に含まれる同一ユーザのツイート集合を 1 文書と見なす．以降では，適用後のデータへの対応を分かりやすくするため，DP-MRM [4] において文書に相当するキーワード集合を「ユーザのツイート集合」，文書数を「ユーザ数」と置き換えて説明する．

j 番目のユーザの投稿したツイート集合における，i 番目のキーワード x_{ji} は，ディリクレ過程（Dirichlet Process）に基づき，トピック k とラベル l が選択された後，多項分布パラメータ θ_{ji} により生成される．以降では，ディリクレ分布を Dir，ディリクレ過程を DP，多項分布を Mult として表す．モデルの生成過程は以下のとおりとなる．

(1) $H|\beta = Dir(\beta)$
　ここで，H は基底測度，β はディリクレ事前分布を表す．
(2) トピックに対するディリクレ過程から G_0^k をサンプリングする．
　$G_0^k|\gamma_k, H \sim DP(\gamma_k, H)$
　ここで，G_0^k は，トピック k に対する基底測度，γ_k はその集中度を表す．
(3) λ_j を，ラベル $r_j = (l_{k \in label(j)})$ からサンプリングする．
　$\lambda_j \sim Dir(r_j \eta)$
　ここで，η は，λ_j の密度を制御するパラメータを表し，$label(j)$ は，ユーザ j にあらかじめ付与されているラベルを返す関数である．
(4) ユーザ j に対するディリクレ過程から G_j をサンプリングする．
　$G_j \sim DP(\alpha, \sum_{k \in label(j)} \lambda_{jk} G_0^k)$
　ここで，G_j は，ユーザ j に対する，各ラベルの確率密

度の混合分布であり，α は集中度，λ_{jk} は，G_0^k の混合割合である．

(5) ユーザ j の投稿したツイート集合におけるキーワード x_{ji} が，多項分布パラメータ θ_{ji} から生成される．

$\theta_{ji}|G_j \sim G_j$

$x_{ji}|\theta_{ji} \sim Mult(\theta_{ji})$

上記の生成過程を，U をユーザ数，K をトピック数，N_j をユーザ j の投稿したツイート集合における単語の数として，グラフィカルモデルで表現したものを，**図 1** に示す．

以上のモデルを実装し，地域に特徴的なトピックの判別に使用する．サンプリングは，崩壊型ギブスサンプリング [5] を採用し，キーワードとトピック k およびラベル l を結ぶ中間的なテーブル t に対してと，そのテーブル t に対して付与されるトピック k およびラベル l に対してそれぞれ行う．モデルの説明でも述べたとおり，ツイート集合に対して適用する際には，Zhao ら [11] の Twitter–LDA を参考とし，ツイート中に含まれる同一ユーザのツイート集合を 1 文書と見なす．なお，同様に，1 ツイート 1 トピックの考え方 [11] を参考とし，各ツイート中のキーワードは，そのツイートに対応する同一のテーブル t に割り当てられるようにする．

更新式は後述のとおりである．なお，ここで，n はキーワードのカウント，m は 1 つ以上キーワードが割り当てられているテーブルのカウント，K は 1 つ以上キーワードが割りてられているトピック数である．なお，α，β，η，γ はハイパーパラメータである．

j 番目のユーザが投稿したツイート集合の i 番目のキーワード x_{ji} を割り当てるテーブル t のサンプリングは，下記に示す更新式で行う．この際，各ツイート中に含まれるキーワードを，そのツイートに対応する同一のテーブルに割り当てる制約を与える方法として，キーワードが 2 つ以上含まれるツイートに対しては，ツイート中のすべてのキーワードの $p(t_{ji} = t|t_{-ji}, rest)$ の総乗値 [11] でサンプリングする．式 (1) は既存のテーブルに割り当てられる確率，式 (2) は新しいテーブルを作成して割り当てられる確率である．なお，以下で示す式 (1)–(6) は，DP-MRM [4] において提案された既存の式である点に注意されたい．また，ハイパーパラメータなどの，分布 p の条件の要素は，文献 [4] にならい，以下では，$rest$ として省略する．さらに，式中の，$n_{jt.}$ は，j 番目のユーザが投稿したツイート集合の中で t に割り当てられているキーワード数，$n_{j..}$ は，j 番目のユーザが投稿したツイート集合の総キーワード数である．

$$p(t_{ji} = t|t_{-ji}, rest) = \frac{n_{jt.}}{n_{j..} + \alpha} f_{k_{jt}l_{jt}}(x_{ji}) \quad \textbf{if } t \textbf{ is existing} \tag{1}$$

$$p(t_{ji} = t|t_{-ji}, rest) = \frac{\alpha}{n_{j..} + \alpha} \Gamma(x_{ji}) \quad \textbf{if } t \textbf{ is new} \tag{2}$$

ここで，$\Gamma(x_{ji})$ は下記の式で表される．なお，式中の $m_{j..}$ は，j 番目のユーザが投稿したツイート集合のテーブル数[*4]，$m_{.kl}$ は，全体でトピック k でラベル l のテーブル数，$m_{.k.}$ は，全体でトピック k のテーブル数，$m_{jk.}$ は，j 番目のユーザが投稿したツイート集合の中でのトピック k のテーブル数である．

$$\Gamma(x_{ji}) = \sum_{k=1}^{K} \frac{m_{jk.} + \eta}{m_{j..} + K\eta} \left(\sum_{l=1}^{L} \frac{m_{.kl}}{m_{.k.} + \gamma_k} f_{kl}(x_{ji}) + \frac{\gamma_k}{m_{.k.} + \gamma_k} f_{kl_{new}}(x_{ji}) \right) \tag{3}$$

キーワードのサンプリングを行った後，テーブル t のサンプリングを行う．j 番目のユーザが投稿したツイート集合のテーブル t に割り当てられるトピック k およびラベル l のサンプリングは，下記に示す更新式で行う．ここで，式 (4) は既存のラベルまたはトピックに割り当てられる確率，式 (5) は新しいラベルまたはトピックを作成して割り当てられる確率である．

$$p(k_{jt} = k, l_{jt} = l|k_{-jt}, l_{-jt}, rest) \propto \frac{m_{jk.} + \eta}{m_{j..} + K\eta} \cdot \frac{m_{.kl}}{m_{.k.} + \eta_k} f_{kl}(x_{jt}) \quad \textbf{if } l \textbf{ is existing} \tag{4}$$

$$p(k_{jt} = k, l_{jt} = l|k_{-jt}, l_{-jt}, rest) \propto \frac{m_{jk.} + \eta}{m_{j..} + K\eta} \cdot \frac{\gamma_k}{m_{.k.} + \gamma_k} f_{kl}(x_{jt}) \quad \textbf{if } l \textbf{ is new} \tag{5}$$

また，ここで $f_{kl}(x_{jt})$ は以下のように定義される．

$$f_{kl}(x_{ji}) = \frac{\int f(x_{ji}|\phi_l^k) \Pi_{x_{j'i'} \in x_{kl}} f(x_{j'i'}|\phi_l^k) h(\phi_l^k) \mathrm{d}(\phi_l^k)}{\int \Pi_{x_{j'i'} \in x_{kl}} f(x_{j'i'}|\phi_l^k) h(\phi_l^k) \mathrm{d}(\phi_l^k)} \quad where \quad x_{kl} = \left\{ x_{jt}; k_{jt_{ji}} = k, l_{jt_{ji}} = l \right\} \tag{6}$$

DP-MRM [4] では，それぞれのラベル l は，トピックの集合と対応づけられると同時に，それぞれのトピック k に

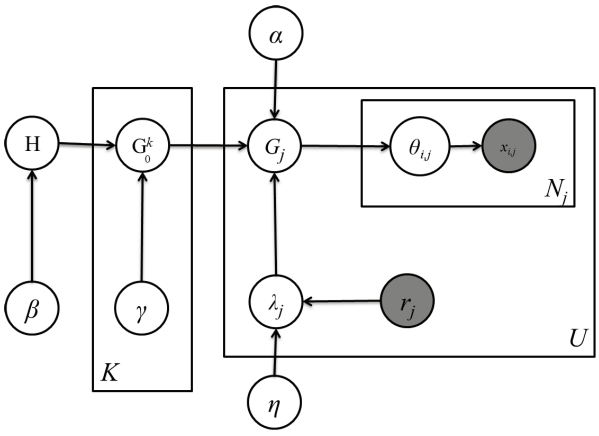

図 1　DP-MRM [4] のグラフィカルモデル
Fig. 1　Graphical model of DP-MRM [4].

[*4] テーブルとは，中華料理店フランチャイズ [9] のメタファに基づく概念であり，各ユーザのツイート集合中で，単語（客）とトピック・ラベル（料理）を対応づけるものである．

ついて，複数のラベルを付与することもできる．本手法では，生成されたトピック集合のうち，トピック k の全テーブル数に対して，トピック k に対して地域ラベル l が付与されたテーブル数の割合が，一定以上の比率のものを，付与されたラベルの比率が偏っている地域に特有のトピックとして，選択する．実験で使用した閾値については，4.2.2 項で述べる．

この手法の適用には，ラベルつきデータとして与えるツイートが必要となる．地域名のラベルを与える手がかりとしては，Twitter ユーザのロケーション項目の利用が考えられる．しかし，ロケーション項目に地域名が記述されているツイートすべてをラベルつきデータとして扱うのは，ツイートの内容を考慮していないため不適切と思われる．ロケーション項目は，ユーザに対して付与されているものであり，すべてのツイートが地域に関連するトピックなわけではなく，地域には依存しないトピックが多く含まれている．また，旅行などの外出により，他の地域に関係の深いトピックを投稿する場合も考えられる．

そのため，ロケーション項目に地域名を記述したユーザにより投稿され，かつ，地域特有と思われる語句を含むツイートを，ラベルつきデータとする．ここで使用する，地域特有と思われる語句の抽出には，地域情報サイトである Yahoo!ローカルサーチ API[*5]を使用する．具体的な抽出データの条件などについては，4 章で述べる．

3.2　ユーザの生活に関わる地域の推定

ユーザの生活に関わる地域を推定するために，3.1 節で説明したとおり，トピックに対して地域ラベルが付与されたテーブル数の割合に基づき，その地域に特有のトピックを選択する．ここで，地域に特有のトピック内で生起確率が高い語を，地域特徴語として扱う．なお，複数地域において，生起確率が高い場合は，複数の地域で共通のトピックとして扱う．これにより，同じトピックに対して複数地域のラベルが付与される場合があることに注意されたい．また，多くの地域に共通のトピックにも現れやすい語が地域特徴語として選択されると，推定の際にノイズになる．そのため，あらゆる地域ラベルの付与率が，すべて一定比率以下となるトピックの集合を，地域に依存しないトピック集合 k_c として定義し，ツイートの地域ラベルの付与の際に使用する．実験で使用する具体的な付与率の閾値については，4.2.2 項で述べる．なお，本研究では，各ユーザのプロフィール情報は，プロフィール上に明記されていない属性を推定することを想定しているため用いない．推定の手順は下記のとおりである．

(1) ツイート t の地域ラベル l の付与
　ツイート t に対して，式 (7) で求めた値が最大となるトピックのラベルを，地域ラベルとして付与する．なお，すべてのトピックにおいて，$tScore(t,k) = 0$ または，$tScore(t,k_c)$ が最大となるツイートには，ラベルを付与しない．ここで，$p(w,k)$ はトピック k におけるキーワード w の生起確率である．

$$tScore(t,k) = \sum \ln p(w,k) \tag{7}$$

(2) ユーザ u の各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，ラベルづけされた地域ごとに重みを付与していく．ここで，$N(u,l)$ は，地域 l がラベルづけされたツイート数である．

$$Score(u,l) = N(u,l) \tag{8}$$

(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 $Score(u,l)$ が最大の地域を，ユーザの生活に関わる地域として採用する．

4.　ユーザの生活に関わる地域の推定実験

4.1　目的

半教師あり LDA により獲得した地域特徴語が，ユーザの生活に関わる地域の推定に有用なものであるかを検証するため，日本語 Twitter データを使用し，評価実験を行う．

また，比較手法としては，文献 [16], [17] が考えられるが，ここでは，採用している地域の粒度の近さを考慮し，地域特徴語の素性選択に基づいた分類器を用いて地域を推定する手法 [16] を参考にして，比較実験を行う．さらに，素性選択を行わない，bag-of-words による分類器を用いた手法をベースラインとすることにより，問題の難しさを明らかにする．

以降，実験方法，データ，結果について述べる．

4.2　実験方法

ユーザの生活に関わる地域の推定の評価については，評価尺度に，評価データに対する，推定の精度・再現率・F 値を採用し，全体の推定結果と，地域ごとの推定結果について検証する．

ここでは，生活に関わる地域の単位として都道府県を採用する．本手法では，生活に関わる地域を推定することにより，Twitter を介した人と人とのつながりを支援することを目的の 1 つとしている．一例として，Twitter 上で震災からの復興ボランティアを募集する際には，そのアクセスを考えて，生活にかかる地域として，都道府県などの粒度で区別をできれば，Twitter で呼びかける対象を絞り込むことができる．同一県内でも実際にアクセスできるかどうかの判断は，最終的にユーザに任せるにしても，どうあってもアクセスできないような地域のユーザを排除することで，人のつながりを効果的に支援できると考える．また，兵庫県で居住し，大阪府に勤務するなど，生活に関わ

[*5] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/localsearch.html

る地域が複数にまたがる場合があるが，上記の目的においては，1 つの生活地域が推定できればよいものとしており，複数の地域の推定は今後の課題とする．

また，提案手法は，ある程度のデータが得られれば地域の粒度は関係なく推定できるが，本実験では，(1) 学習・評価ともに一定数のデータが得られ，(2) 全国に適用できる区分という観点から，都道府県を採用した．なお，これ以上細かい地域を推定する必要がある場合は，段階を分けて，都道府県を推定した後に，その中での細かい地域を推定することが有効と考える．

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．

4.2.1　半教師あり LDA

地域に偏りのあるトピックには，行事や季節的なものなど，時期に依存するものがある．そのため，時期ごとのトピックを選択するために，1 カ月を単位とし，LDA の適用を行う．なお，3.2 節で述べたユーザの生活に関わる地域の推定については，予備実験により各月のデータに限定して適用するよりも，区別を行わない方が良い結果が得られたことから，各月のデータを区別しない点に注意されたい．半教師あり LDA に初期値として与える値は，トピック数 300，ラベル数 16（推定を行う生活に関わる地域数），各ユーザ内に仮想的に作成されるテーブル数[*6]は，最大 10[*7]とした．

使用する語は，形態素解析の結果，名詞と判断された語である．なお，ツイートの形態素解析には，MeCab[*8] Version 0.98 を使用し，辞書は，NAIST Japanese Dictionary[*9]を使用した．ただし，適用データ内での出現回数が，50,000 回以上の語と，1 回の語に関しては，ストップワードとして推定から取り除いた．また，各パラメータの値は，$\alpha = 0.5$，$\beta = 0.5$，$\eta = 0.1$，$\gamma = 0.1$ である．

4.2.2　地域特徴語の選択

LDA の実行結果として得られるトピックのうち，特定地域のラベルの比率が閾値 σ 以上のものを，地域に特有なトピックとして選択する．また，多くの地域に共通するトピック集合（3.2 節で述べた地域に依存しないトピック集合）として，すべての地域でのラベルの比率が閾値 τ 以下となるものを選択する．ここで，閾値は σ は 0.3，τ は 0.1 とした．選択された各トピック内の生起確率上位のうち，最大 N 語以内のキーワードを，地域特徴語とする．ここで，N は 1,000 語とし，トピック内のキーワードの出現回数が 5 回以上のキーワードとした．

4.2.3　比較手法

地域特徴語による素性選択に基づく分類器を用いて，地域を推定する手法 [16] を参考にした，比較手法を実現した．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 値を計算することで，特徴的な単語を素性として選択している．ここでは，素性は各都道府県の TF-IDF 値上位 10,000 件を選択し，分類器には LibSVM-3.12[*10]を用いた．LibSVM のカーネルは RBF を，タイプとしては nu-SVC を，比較した中で最良の結果が得られたことから選択した．また，パラメタは，デフォルトの値（degree を 3，nu を 0.5）とした．使用するキーワードは，提案手法と同様に，MeCab による形態素解析の結果，名詞と判断された語であり，ストップワードも同様に，1 カ月ごとのデータ内で，出現回数 50,000 回以上の語と 1 回となる語である．

なお，ベースラインである bag-of-words では，素性選択を行わずに，分類器での学習を行う．分類器の条件は比較手法と共通であり，使用するキーワードは MeCab による形態素解析の結果，名詞と判断された語である．

4.3　データ

本節では，地域特徴語を選択するための半教師あり LDA に使用するデータとするツイートと，評価に使用するユーザの収集方法，条件について説明する．また，比較手法のデータについても説明する．

4.3.1　半教師あり LDA への適用データ

LDA によるトピック作成に使用するデータには，2012 年に投稿された日本語ツイートを用いる．ツイートは，Twitter の Search API[*11]を使用して収集されたものである．また，日本語で記述されたツイートを収集するため，言語に “ja”（日本語）と，日本全域をカバーする位置情報[*12]とを条件として指定した．

収集されたツイートのうち，実験で使用するのは，下記に示す特定の条件を満たしたツイートである．

(1) 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．
(2) ツイート本文に，名詞を含んでいる．
(3) ユーザのスクリーンネームに “bot”，“公式” などの特定の語を含んでいない[*13]．

また，上記の条件を満たしたツイートのうち，地域特有

[*6] ここでは，各ユーザ内のテーブルの数を，テーブル数とする．
[*7] 各ツイートは同一のテーブルに割り当てられるため，ツイート数が 10 に満たないユーザは，ツイート数がテーブル数の上限となる（例：1 ツイートしかないユーザの場合は 1）．
[*8] http://mecab.sourceforge.net/
[*9] http://sourceforge.jp/projects/naist-jdic/
[*10] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[*11] http://search.twitter.com/search.json
[*12] 円形で日本全域を囲む場合，中心地点となる，兵庫県西脇市を中心とする半径 2,000 km 圏内
[*13] 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や，店舗や企業などの組織アカウントをある程度除外するため

の語を含むツイートを，ユーザごとにまとめて 1 文書として，ラベルつきデータとして与える．地域特有と思われる語句の抽出には，Yahoo!ローカルサーチ API[*14]を使用する．獲得している情報は，店舗（施設，企業）名，住所に含まれる地名（市区町村以下の町丁にあたる地名），最寄駅名，沿線名である．各都道府県の市区町村ごとに住所と業種の大分類[*15]を指定し，API により最大 3,000 件の情報を獲得した．ここでの語句は抽出時の表記をそのまま使用して，形態素に分割するなどの処理は行っていない．この中で，他の都府県との重複がない語句を含むツイートを，各地域のラベルつきデータとして与える．

LDA は 1 カ月ごとに適用を行うため，データ数はそれぞれ，ラベルが付与された 1,600,000（各地域 100,000）件のツイートと，ラベルが付与されていない各地域 10,000 件/日の 1 カ月分のツイート（4,640,000 件から 4,960,000 件）である．

4.3.2　評価データ

評価データには，人手で判定を行ったユーザ（各地域 100 人，合計 1,600 人）を用いた．これらのユーザは，一定条件を満たすユーザを機械的に収集し，それらのユーザに対し人手での判定を行った．機械的に獲得したユーザの条件は下記のとおりである．

(1) ロケーションに，対象となる地域名を 1 つだけ記述している．
(2) 2012 年 1 月から 12 月の 1 年間，日本語を含む投稿を毎月 100 件以上 10,000 件以下で行っている．
(3) 町丁以下の住所を記述していない[*16]．
(4) ユーザのスクリーンネームに，“bot” や “公式” など，非個人アカウントである可能性の高い特定の語句を含んでいない．

これらの条件などから獲得したユーザに対し，第一著者が人手で判定を行ったものを最終的な評価データとした．人手での判定では，下記の条件について目視で確認を行った．

(1) 対象となる地域が，生活に関する地域であることを確認できる記述[*17]が，ロケーションや bio などにある．
(2) 飲食店や自治体，企業などの組織アカウントや，“bot” と呼ばれる自動投稿が中心のアカウントではない．
(3) 複数地域での生活が分かるような記述があった場合に関しては，評価データから取り除く．

4.3.3　比較手法において使用するデータ

比較手法における SVM の素性とする地域特徴語の選択には，半教師あり LDA に用いたデータと同様に，4.3.1 項で述べた地域特有の語を含むラベル付きデータと，ラベルが付与されていないデータを用いている．ラベルが付与されていないデータに関しては，4.3.1 項に記述した収集条件 (1) に基づき，ユーザのロケーション項目に地域名が記述されているため，その地域名を地域ラベルとして使用する．評価データも，共通のものを使用している．

ただし，比較手法では，地域特徴語の選択を行った後，ユーザごとに特徴を学習する必要がある．そのため，下記の条件を満たす，ユーザアカウント各地域 1,100 人[*18]，合計 17,600 人を選択した．なお，ベースラインである bag-of-words に用いるデータは，素性選択を行わないため，ユーザアカウント合計 17,600 人のデータのみを使用する．

(1) ロケーションに，対象となる地域名を 1 つだけ記述している．
(2) 2012 年 1 月から 12 月の 1 年間，日本語を含む投稿を毎月 100 件以上 10,000 件以下で行っている．
(3) ロケーションに，“出身” など，ロケーションに書かれた地域名が現在の生活に関わる地域ではない可能性の高い特定の語句を含んでいない[*19]．
(4) 町丁以下の住所を記述していない．
(5) ユーザのスクリーンネームに，“bot” や “公式” など，非個人アカウントである可能性の高い特定の語句を含んでいない．

なお，学習を行ったユーザのツイート数の合計は 73,183,560 件（約 610 万件/月）であり，各ユーザの平均ツイート数は約 4,158 件（約 347 件/月）である．

4.4　結果

ユーザの生活に関わる地域の推定結果について述べる．提案手法および比較手法の評価結果を，**表 1** に示す．データ量や使用する素性に制約がある状況ではあるが，提案手法がすべての値で比較手法を上回ることを確認した．また，提案手法，比較手法ともに，ベースラインである bag-of-words による結果を大きく上回り，手法の有効性を示した．提案手法と比較手法との評価結果に対し，t 検定（有意水準 5%，両側検定）を行ったところ，再現率と F 値で有意差が確認できた．

また，各地域ごとの推定結果を，**表 2**（提案手法），**表 3**（比較手法）に示す．F 値は，和歌山の 0.82 が最も高く，東京の 0.36 が最も低い結果となった．さらに，東京，大阪，京都などの大都市を除く地域で，比較手法より推定精度が

[*14] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/localsearch.html
[*15] 01：グルメ，02：ショッピング，03：レジャー・エンタメ，04：暮らし・生活
[*16] 町丁以下の詳細な住所を記述している場合，店舗や企業などの組織アカウントである場合が多いため
[*17] “xx 在住”，“xx 大生” など
[*18] 予備実験では，各地域の学習するユーザ数が同一である方が，高い評価値が得られた．そこで，保有データ中から条件下で収集できるユーザ数が最も少ない地域で 1,100 人となったことから，この件数に揃えた．
[*19] 本条件がなく，訓練データに含む場合，ノイズを含んだ学習となり，比較手法の精度が低下する．なお，4.3.2 項の評価データにおいては，目視により人手判定する段階で生活に関わる地域でないと判定しているため，本条件がないことによる問題は生じない．

向上している場合が多い．これは，地域に特有のトピックが地方で有効に機能している可能性がある一方で，比較手法においては，トピックに反映されにくい素性が大都市において有効であることによると考えられる．また，F 値が 0.7 を超える地域が，提案手法については 16 地域中 7 地域ある（比較手法は 3 地域）ことから，提案手法は，地域に特有のトピックがうまく見つけられた地域については，高い推定精度を実現できる可能性があると考えられる．最後に，再現率に着目すると，16 地域中 11 地域において，提案手法が比較手法を上回っており（3 地域が下回り，2 地域が同点），全体の平均についての有意差もあることから，地域特有のトピックを利用することにより，概して再現率が向上しているといえる．

地域による評価値の違いに関しては，5.3 節で考察を行う．

表 1　ユーザの生活に関わる地域の推定
Table 1　Evaluation results for estimation of twitter user's life area.

手法	精度	再現率	F 値
提案手法	**0.65**	**0.67***	**0.66***
地域特徴語選択（比較手法）	0.62	0.60	0.61
bag-of-words（ベースライン）	0.31	0.12	0.17

*：比較手法に対して t-検定（有意水準 5%，両側検定）で，有意に向上．

表 2　地域ごとの評価結果（提案手法）
Table 2　Evaluation results by prefectures (proposed method).

地方	地域	精度	再現率	F 値
北海道	北海道	0.83	0.79	0.81
関東	茨城	0.58	0.64	0.61
	栃木	0.89	0.59	0.71
	群馬	0.86	0.63	0.72
	埼玉	0.55	0.81	0.66
	千葉	0.82	0.79	0.80
	東京	0.42	0.31	0.36
	神奈川	0.51	0.57	0.54
近畿	三重	0.56	0.76	0.65
	滋賀	0.51	0.64	0.57
	京都	0.53	0.62	0.57
	大阪	0.50	0.61	0.55
	兵庫	0.51	0.67	0.58
	奈良	0.77	0.64	0.70
	和歌山	0.74	0.92	0.82
九州	福岡	0.82	0.77	0.80

表 3　地域ごとの評価結果（比較手法）
Table 3　Evaluation results by prefectures (comparison method).

地方	地域	精度	再現率	F 値
北海道	北海道	0.82	0.80	0.81
関東	茨城	0.59	0.54	0.56
	栃木	0.67	0.59	0.62
	群馬	0.77	0.63	0.69
	埼玉	0.55	0.51	0.53
	千葉	0.64	0.71	0.67
	東京	0.46	0.47	0.46
	神奈川	0.51	0.54	0.52
近畿	三重	0.57	0.55	0.56
	滋賀	0.38	0.52	0.44
	京都	0.63	0.66	0.64
	大阪	0.64	0.57	0.60
	兵庫	0.59	0.47	0.53
	奈良	0.57	0.62	0.60
	和歌山	0.79	0.79	0.79
九州	福岡	0.80	0.68	0.74

5. 考察

4 章の実験結果をふまえた分析に基づき，考察を述べる．ここでは，地域に特徴的なトピックとして選択されたトピックの数や，トピックの内容，それによる生活に関わる地域の推定への影響などについて説明する．

5.1 地域に特徴的なトピック数

地域に特徴的なトピックとして選択されたトピック数を，**表 4** に示す．数に偏りがある[*20]ものの，すべての期間ですべての地域に対し，特徴的なトピックを選択することができた．最も多いのは，和歌山の 174 件，最も少ないのは，兵庫の 39 件である．また，この結果には，複数地域に出現するトピックも含まれる．

首都圏である千葉・東京・神奈川は，いずれも 50 件以下と比較的少ない数となった．また，千葉・東京・神奈川に

表 4　地域に特徴的なトピック数
Table 4　Number of area-oriented topics.

地方	地域	トピック数
北海道	北海道	138
関東	茨城	72
	栃木	189
	群馬	135
	埼玉	87
	千葉	48
	東京	45
	神奈川	48
近畿	三重	84
	滋賀	69
	京都	114
	大阪	69
	兵庫	39
	奈良	162
	和歌山	174
九州	福岡	81

[*20] 初期値としては，4.2.1 項で述べた 300 のトピックを 16 の地域について等しい確率でランダムに割り当てる．したがって，初期トピック数の地域別分布に偏りはない．

表 5　関東地域のユーザに対する駅名の出現数とその割合（%）
Table 5　Frequency and rate of station names in tweets for Kanto region users (%).

ユーザ	駅名													
	栃木県		群馬県		茨城県		埼玉県		千葉県		東京都		神奈川県	
	出現数	割合	出現数	割合	出現数	割合	出現数	割合	出現数	割合	出現数	割合	出現数	割合
栃木県	31128	71.9	1960	6.7	1956	7.4	3420	8.2	4103	9.8	13070	8.1	4852	5.6
群馬県	4523	10.5	19958	68.6	1709	6.5	6263	15.1	3903	9.3	15889	9.9	4711	5.4
茨城県	1875	4.3	1229	4.2	14195	53.6	1887	4.5	4743	11.3	15156	9.4	4908	5.6
埼玉県	1974	4.6	2049	7.0	2524	9.5	19985	48.0	5324	12.7	30163	18.8	7079	8.1
千葉県	1409	3.3	1258	4.3	1730	6.5	2216	5.3	15696	37.4	16808	10.5	4653	5.3
東京都	1167	2.7	1278	4.4	1725	6.5	5087	12.2	4288	10.2	42418	26.4	6497	7.5
神奈川県	1192	2.8	1366	4.7	2641	10.0	2738	6.6	3867	9.2	27153	16.9	54490	62.5

表 6　地域に特徴的なトピックの一部（地名など地域一般）
Table 6　Example of area-oriented topics (place-name).

	地域ラベル	上位語（10 件）
1	茨城	牛久，茨城，つくば，土浦，村，阿見，ミニ，駅前，取手，行き
2	北海道	白石，発寒，新札幌，琴似，札幌，苗穂，桑園，北広島，み，青葉

表 7　地域に特徴的なトピックの一部（イベント）
Table 7　Example of area-oriented topics (event).

	地域ラベル	上位語（トピック内の生起確率の順位）
1	大阪	大阪 (1)，市 (2)，梅田 (5)，マラソン (6)，前 (9)，市役所 (17)，テックス (20)，イン (24)，大阪城公園 (44)
2	神奈川	艦 (2)，一般 (3)，海軍 (4)，護衛 (9)，自衛隊 (10)，海上 (11)，晴海 (16)，船 (21)，艦隊 (24)，旗艦 (25)，艦船 (71)
3	東京	選挙 (2)，石原 (9)，都知事 (10)，東京 (16)，都議会 (18)，票 (21)，慎太郎 (26)，猪瀬 (32)，投票 (36)，区 (52)

表 8　複数地域に特徴的なトピックの一部
Table 8　Example of multiple areas oriented topic.

地域ラベル	上位語（トピック内の生起確率の順位）
大阪，奈良，兵庫	試合 (1)，野球 (2)，阪神 (3)，勝 (5)，プロ (6)，オリックス (7)，選手 (9)，大阪 (10)，甲子園 (13)，阪神タイガース (15)

加え，大阪，兵庫などの地方の中で中心的な地域は，特徴的なトピックの選択数が少ない結果となった．これは，都心部では，生活圏が複数の地方にまたがる人や，他の地方から訪れる人が多いことが原因の 1 つとして考えられる．この根拠として，**表 5** に，関東各地域のユーザ各 100 人と，そのツイートに出現したその地域の駅名[*21]の対応数を示す．東京の駅名は，どの地域でも出現数が多く，千葉や埼玉では，同一地域の駅名よりも多い．この結果から，東京への移動や東京の話題が周辺地域において多いことが示唆される．また，東京に関わりが深いトピックについても，他の地域のユーザがそのトピックに関する投稿を行っているとすれば，3.2 節で述べた手法に基づき，付与されるラベルの相対的な比率が減ることから，東京の地域ラベルが付与されることは少なくなる．さらに，東京の駅名は他の地域での出現数が多いことから，地域特有でない話題においても，東京の語が出現しやすいことがうかがえ，教師データのラベルが，特定のトピックに偏りにくいことも原因の 1 つと考えられる．

[*21] Yahoo!Loco により収集した，都道府県別の鉄道駅

5.2　選択されたトピック内容とラベル

5.2.1　選択されたトピックの内容

地域に特徴的なトピック，または共通として選択されたトピックの内容について例示する．

地域に特徴的なトピックとして選択されたトピックの一部を，**表 6** と**表 7** に示す．表 6 の 1 は茨城，2 は北海道のラベルが付与されたトピックであり，上位に地域内の地名が多く現れていることからも，適切にラベルが付与されていることが分かる．表 7 は，地域でのイベントに関するトピックである．1 は，大阪のラベルが付与された，大阪マラソンに関するトピックである．スタート会場である大阪城公園前や，ゴールである大阪市役所前，インテックス大阪前などに関するキーワードが選択されている．2 は海上自衛隊の実播公開に関するトピックであり，3 は東京都知事選に関するトピックである．このように，時事的なイベントに関しても，地域に特徴的なトピックを選択できた．また，複数地域に特徴的なトピックとしては，**表 8** に示すトピックなどが選択された．このトピックは，野球チームの阪神とオリックスを中心としたトピックである．最後に，多くの地域に共通のトピックとして選択されたトピッ

表 9　共通のトピックの一部
Table 9　Example of common topics for many areas.

	上位語（10 件）
1	イラスト，大学，制作，成安，造形，ゼミ，卒業，クラス，展，配信
2	降水，確率，最低，曇，晴，発表，予報，入試，雨，のち
3	外来，専門，病院，治療，ケア，アレルギー，緩和，病気，手術，股関節

表 10　ラベルが不適切と思われるトピックの一部
Table 10　Example of miss-labeled topic.

地域ラベル	上位語（10 件）
神奈川	入荷，商品，早め，少量，在庫，完売，ファミリー，品，切れ，化粧

クの一部を，表 9 に示す．共通のトピックでは，大学や天気，少数の地域に限定されないトピックが選択されている．

5.2.2　不適切な地域ラベルの付与

地域ラベルが付与されたトピックの一部は，地域に特徴的でない語が多く含まれているトピックも存在する．ただし，3.2 節で説明した，共通のトピック k_c を使用することで，ユーザの生活に関わる地域の推定には，影響を与えないものがあることが分かった．

表 10 に示しているトピックは，神奈川のラベルが付与されているが，上位語のほとんどは買い物一般に関する語句であり，地域特徴語とはいい難い．このようなラベルが付与された原因は，神奈川のラベルつきデータに，商業施設でのセールに関するツイートが偏って存在したからであった．このトピックの場合，他の期間で共通の話題となって選択されたトピックに買い物に関するトピックがあり，評価ユーザのツイートに対し，このトピックによる地域ラベルの付与は行われなかった．時期に依存しないようなトピックの場合，誤ったラベルが付与されても，共通のトピックを用いることで影響を小さくできることが分かる．

5.3　地域ごとの推定精度の傾向と分析

推定精度は，表 4 に示した，地域に特徴的なトピックが多く選択された地域の評価結果が高くなる傾向が見られた．和歌山，北海道，栃木，群馬は選択されたトピック数が多く，評価結果も高い地域である．

一方で，トピック数とは異なる傾向を見せる地域がいくつか存在した．福岡や千葉は，トピック数が少ないのに評価値が高くなり，京都や奈良は，トピック数が多いがあまり評価値が高くないといった結果が見られた．

福岡に関しては，近隣の地域がなく，他と重複しない地域に特徴的なキーワードが多く選択されたことや，近隣の地域がないことで，他の地域からの移動が少なく，ノイズとなる情報がデータに少ないことが理由として考えられる．同様に評価値が高い北海道は，選択されたトピック数も多いが，同じ傾向が見られる．

京都は，トピック数が多いが，特徴的なキーワードに関して，他の地域の評価ユーザが観光で訪れるなどして京都に関する投稿を行ったり，ノイズとなる情報が多くなったりしたと考えられる．一方で，地域に特徴的な語であっても，京都の評価ユーザの中では，観光地となるような場所に関する投稿は，全体の中でそれほど多く見られなかった．奈良に関しては，同様の傾向とともに，和歌山への推定の誤りが多く，近隣の地域で似たキーワードを含み，かつ生起確率が和歌山の方が高いトピックが選択された影響が考えられる．また，都道府県名で唯一，「奈良」という語がストップワードに選択された影響もあると考えられる．

これらの誤りの傾向をふまえ，評価ユーザの使用するデータから，長期休暇中に投稿されたデータを取り除くことや，平日のデータのみを使用することを，検討したい．

6.　おわりに

本研究では，Twitter ユーザを対象として，半教師ありトピックモデルを利用した地域特徴語の選択に基づく，生活に関わる地域属性の推定手法を提案した．具体的には，地域情報サイトから収集した地域特徴語を含むツイートを教師データとした，半教師ありトピックモデルにより，地域に特徴的なトピックを抽出する．次に，抽出した地域特徴語を使用し，ツイートごとに地域ラベルを付与する．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地域ラベルに基づき推定する．推定実験においては，データ量や使用する素性に制約があるものの，提案手法が，地域特徴語による素性選択に基づく分類器を用いた比較手法を上回ることが確認できた．

今後の課題としては，(1) ラベルつきデータの選定方法の見直しや，トピックの時期による重みの変更などによる推定精度の向上や，(2) どの地域にも分類できないユーザの判定などがあげられる．

ラベルつきデータの選択方法の見直しに関しては，テレビ番組に関する内容など，時事的なキーワードが多く含まれるトピックに誤ったラベルが付与された場合，これらのトピックが推定のノイズとなることが確認された．地域に特徴的な語でも，テレビなどにより時事的に他の地域のユーザも投稿するような場合があり，そのような場合にツイートに付与されたラベルが，トピックのラベルの付与に

影響を与えている場合がある．現在は，地域情報サイトから収集した地域に特徴のある語については，語句どうしが他の地域と共起しないかは調べているが，他の地域のツイート中にどの程度含まれているかのチェックは行っていない．そのため，ラベルの付与に用いる地域に特徴的な語句のうち，推定期間中に他の地域でも多く投稿が行われている語句を除くことによる，ラベル付与の精度の改善についても検討していきたい．

さらに，推定に用いるトピックを，各月のデータに限定して適用を行った場合は予備実験において評価値が低下したが，数カ月単位でトピックを利用する，あるいは，各月単位でトピックを重みづけするなどの工夫により，評価値を向上させることも検討の余地がある．

また，実用においては，対象外地域のユーザや，地域に関する情報をいっさい発しないユーザの存在から，「不明」との判定の必要性が生じてくる．それらのユーザに対する対応としては，地域ラベルの付与率が一定の割合を超えなければ，ユーザの生活に関わる地域としては「不明」と出力することが考えられる．ただし，地域ラベルを正確に判定するためには，トピックから地域に特徴的でない語を取り除くことが必要となる．現在は，地域に特徴的なトピックから，出現確率の高い語を地域特徴語として使用しているが，多くの地域に共通するトピックに出現確率の高い語を，ユーザの推定の際には取り除くなどの工夫が考えられる．

謝辞　本研究で使用した Twitter データの収集に際しては，筑波大学の佐藤哲司教授と研究室のメンバに協力をいただいた．ここに深く感謝する．

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），萌芽研究（課題番号 25540159）の助成を受けて遂行された．

参考文献

[1] Blei, D.M., Ng, A.Y., Jordan, M.I. and Lafferty, J.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[2] Burger, J.D., Henderson, J., Kim, G. and Zarrella, G.: Discriminating Gender on Twitter, *Proc. 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP'11*), Edinburgh, Scotland, pp.1301–1309 (2011).

[3] Cheng, Z., Caverlee, J. and Lee, K.: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, *Proc. 19th ACM International Conference on Information and Knowledge Management* (*CIKM'10*), Toronto, Canada (2010).

[4] Kim, D., Kim, S. and Oh, A.: Dirichlet Process with Mixed Random Measures: A Nonparametric Topic Model for Labeled Data, *Proc. 29th International Conference on Machine Learning*, Edinburgh, Scotland (2012).

[5] Liu, J.S.: The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem, *Journal of the American Statistical Association*, Vol.89, No.427, pp.958–966 (1994).

[6] Pennacchiotti, M. and Popescu, A.-M.: Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter, *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, pp.430–438 (Aug. 2011).

[7] Ramage, D., Dumais, S., Liebling, D. and Manning, C.D.: Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp.248–256 (2009).

[8] Rao, D., Yarowsky, D., Shreevats, A. and Gupta, M.: Classifying Latent User Attributes in Twitter, *Proc. 2nd International Workshop on Search and Mining User-generated Contents*, Toronto, ON, Canada, pp.37–44 (2010).

[9] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: Hierarchical Dirichlet Processes, *Journal of the American Statistical Association*, Vol.101, No.476, pp.1566–1581 (2006).

[10] Al Zamal, F., Liu, W. and Ruths, D.: Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors, *Proc. 6th International Conference on Weblogs and Social Media*, Dublin, Ireland, pp.388–390 (2012).

[11] Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X.: Comparing Twitter and Traditional Media Using Topic Models, *Proc. 33rd European Conference on Advances in Information Retrieval*, Berlin, Heidelberg, pp.338–349 (2011).

[12] 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).

[13] 奥谷貴志，山名早人：メンション情報を利用した Twitter ユーザプロフィール推定，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2014），p.B2–3 (2014).

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

[15] 今井良太，数原良彦，戸田浩之，鷲崎誠司：Web 文書を利用した POI に関連する地名の抽出，情報アクセスシンポジウム 2013，pp.9–13 (2013).

[16] 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，Vol.112, No.367, pp.23–27 (2012).

[17] 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 コンシューマ・デバイス&システム，Vol.2, No.1, pp.82–93 (2012).

[18] 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，p.C–1 (2012).

[19] 李　龍，若宮翔子，角谷和俊：Tweet 分析による群衆行動を用いた地域特徴抽出，情報処理学会論文誌 データベース，Vol.5, No.2, pp.36–52 (2012).

[20] 三木翔平，新田直子，馬場口登：単語の地理的局所生の経時変化を考慮したツイートの発信位置推定，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2014），p.B3–1 (2014).

堂前 友貴

2014 年筑波大学大学院図書館情報メディア研究科博士前期課程修了．

関 洋平　（正会員）

1996 年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．2005 年総合研究大学院大学情報学専攻博士後期課程修了．博士（情報学）．同年豊橋技術科学大学情報工学系助手．2008 年コロンビア大学客員研究員．2010 年筑波大学図書館情報メディア系助教，現在に至る．自然言語処理，意見分析，情報アクセスの研究に従事．ACM，ACL，電子情報通信学会，言語処理学会，日本データベース学会，人工知能学会各会員．

（担当編集委員　奥 健太）