マイクロブログのジオタグを用いた訪問地の違いに着目したユーザ性別推定手法の提案

三浦　理緒† 　廣田　雅春†† 　加藤　大受††† 　荒木　徹也† 　遠藤　雅樹†,†††† 　石川　博†

† 首都大学東京大学院　システムデザイン研究科　〒191–0065 東京都日野市旭が丘 6–6
†† 岡山理科大学　総合情報学部　〒700–0005 岡山県岡山市北区理大町 1–1
††† ウイングアーク１ｓｔ株式会社　〒150–0031 東京都渋谷区桜丘町 20–1 渋谷インフォスタワー
†††† 職業能力開発総合大学校　基盤ものづくり系　〒187–0035 東京都小平市小川西町 2–32–1
E-mail: †miura-rio@ed.tmu.ac.jp, ††hirota@mis.ous.ac.jp, †††kato.d@wingarc.com, ††††endou@uitec.ac.jp, †{araki, ishikawa-hiroshi}@tmu.ac.jp

あらまし　近年の訪日外国人観光客の増加を受け，インバウンド市場の活性化のためのマーケティングが重要視されている．効果的なマーケティングを目的とした，マイクロブログユーザの属性推定を行う研究は盛んに行われている．既存研究では，1 言語を対象としたユーザのプロフィール文や投稿内容のテキストを利用しているが，訪日外国人には様々な言語を使用するユーザが存在し，これらの手法では対応が困難である．本稿では，マイクロブログの投稿に付与された位置情報に基づいて，ユーザの属性を推定する手法を提案する．「新橋はサラリーマンが多い」，「原宿は女子高生が多い」というように，ある地域に訪れる人の属性が偏って存在する場合がある．本手法ではこれに倣い，「ある属性の人は，ある地域によく訪れる」という仮定に基づき，マイクロブログユーザの投稿の位置情報に基いた特徴量を生成し，機械学習により分類を行うことで，訪れた場所に着目してユーザの属性を推定する手法を提案する．Twitter のデータを用いて性別推定実験を行い，分類性能の評価を行なった．

キーワード　マイクロブログ，ジオタグ，属性推定，性別

1. はじめに

近年，訪日外国人観光客は増加傾向にある (注1)．また，訪日外国人の日本における消費額は年々増加しており，2017 年 7 月〜9 月期の消費額が過去最高を更新した (注2)．これを受け，さらなるインバウンド市場の活性化のための，外国人観光客向けのマーケティングが重要視されている．外国人観光客の消費活動を促進するような効果的なマーケティングを行なうためには，人々の特徴を分析することが重要である．たとえば，「訪日外国人は何に興味を持って来日したか」や「外国人は日本を訪れて何に興味を持ったか」といった，訪日外国人を対象とした興味・関心の分析が必要になっている．

人々の興味・関心を分析するうえで，人の特徴の分析を行なうことが必要不可欠である．人の特徴を分析する際に，有用な情報源の一つとして，Twitter(注3) に代表されるマイクロブログが挙げられる．スマートフォンの普及により，マイクロブログの利用者は著しく増加しており，そのデータ量は膨大で，商品や観光地の口コミなどの情報も多く含まれる．これらの情報は，従来の収集方法であった政府や企業が行なうアンケート調査に比べ，より人々の意見や感情を直接的に，かつリアルタイムに反映している．また，得られるデータ量も大規模であり，同等の量のデータをアンケート調査によって収集するのに比べてコストが低い．この膨大なマイクロブログデータを解析することで，効果的なマーケティングのための，ユーザ属性の分析を行えると考えられる．しかし，マイクロブログユーザが自らのプロフィール情報に属性を明記してない場合が多く，そのようなユーザにおいては，明示的な情報のみでユーザ属性を判定することは不可能である．

このような問題を解決するために，マイクロブログのデータを用いてユーザの属性を推定する研究が行われている [1], [2], [3], [4], [5], [6]．これらの研究では，ユーザの投稿内容やプロフィール文などのテキストから，性別や年代，職業などの属性を推定している．しかし，テキストを用いる手法は，ある特定の言語にのみ適用可能であり，言語依存性が高い．外国人ユーザのデータでは，プロフィールやテキストのほとんどは外国語で記述されており，言語の種類も多いので，それらの複数の言語を同時に解析するコストは高く，実現が困難である．そのため，外国人ユーザのデータを対象とした場合においても適用可能である，言語依存性の低い分析手法が必要である．

本研究では，人の属性と訪れた地域の関係性に着目する．たとえば，「女子学生は原宿に多い」「サラリーマンは新橋に多い」

(注1)：日本政府観光局 訪日外客統計：https://www.jnto.go.jp/jpn/statistics/since2003_tourists.pdf
(注2)：観光庁 訪日外国人消費動向調査：http://www.mlit.go.jp/kankocho/siryou/toukei/syouhityousa.html
(注3)：https://twitter.com/

というように，ある特定のユーザは特定の地域によく訪れることがある．つまり,「ある属性の人は，ある地域によく訪れる」という仮定を置くことができる．この仮定に倣い，ユーザが訪れた地域を示すジオタグに基づいた特徴量を生成し，教師あり機械学習手法を用いて分類，推定を行なう手法を提案する．また，マイクロブログの一つである Twitter のデータを用いて属性推定実験を行なった．今回は，位置情報を用いた属性推定の有効性を評価するために，正解の作成が容易な性別を推定対象とする属性とした．そして，本実験の推定精度をもとに，提案する手法により性別の分類が可能か考察を行なった．本手法では位置情報のみを用いるため，言語依存性が低い．また，位置情報を扱っている多くのサービスで適用可能であり，複数のサービスを同時に取り扱うことができることが利点として挙げられる．

本論文の構成は次の通りである．まず 2. 章で関連研究について述べる．次に，3. 章で，本研究で提案する手法を述べ，そして 4. 章では，実際のデータを用いて提案手法の性能を評価する実験を行い，結果を示し 5. 章で考察を述べる．最後に，6. 章で，本研究のまとめと今後の課題を述べる．

2. 関 連 研 究

Web 上の情報を使って，ユーザ属性を抽出する研究は，盛んに行われている．マイクロブログの投稿やプロフィールなどのテキストを用いてユーザの属性を抽出する研究 [1], [2], [3], [4], ユーザ間のつながりによって構成されるソーシャルグラフを用いてユーザの属性を抽出する研究 [5], [6] などがある．また，抽出したユーザ属性をもとに，ユーザの観光行動の分析を行う研究 [7] がある．

池田ら [1] は，Twitter ユーザの投稿内容のテキストから，属性のクラスごとに特徴的な単語を，赤池情報量基準における出現単語の偏り度合いを算出することで抽出し，それらを素性として，サポートベクターマシンで学習・属性推定を行なう手法を提案している．また，年代・性別・居住地域を推定対象属性とし，推定実験を行なっている．Miller ら [2] は，Twitter ユーザの投稿内容のテキストを，N-gram によって分割したうち，特徴的なものを抽出し，ナイーブベイズおよびパーセプトロンを用いて分類することで，属性推定を行なう手法を提案している．平野ら [3] は，年代と職業などの属性間の依存関係に着目し，Markov Logic を用いることで属性の全てを同時に考慮し，Twitter ユーザの投稿内容のテキストからユーザ属性を推定する手法を提案している．Burger ら [4] は，Twitter ユーザの投稿内容だけでなく，スクリーンネーム (ユーザが自由に記述できる名前) の，属性のクラスごとの特徴を抽出し，ユーザ属性を推定する手法を提案している．これらの研究では，ウェブ上のテキストを解析し，ユーザ属性を推定しているが，テキストに用いられている言語ごとに異なる言語処理を行なう必要があるため，言語依存度が高い．そこで，本研究では言語依存度の低い情報として，マイクロブログの投稿に付与されたジオタグを用いて，ユーザ属性を推定する手法を提案する．

蔵内ら [5] は，Twitter ユーザのフォロー関係からソーシャルグラフを構築し，マルコフ確率場を用いてグラフ上のユーザ属性をモデル化し，最適化問題として真の属性を推定する手法を提案している．奥谷ら [6] は，Twitter ユーザ間でやり取りされるメンションに基づいてソーシャルグラフを構築し，クラスタリングを行いユーザのプロフィールを推定する手法を提案している．これらの研究では，ユーザ間のつながりによって構成されるソーシャルグラフを解析することで，ユーザの属性を推定している．しかし，ユーザ間のつながりは，職業や趣味など，様々なコミュニティにおいて形成されるため，正しい属性を推定するのが困難な場合がある．また，性別のような，つながりを持つ際に比較的関係性が低い属性については，適用が困難である．そこで，本研究では，ユーザの推定対象の属性に対する依存度の低い手法として，訪れた場所をあらわすジオタグを用いて推定を行なう．

佐伯ら [7] は，日本国内でジオタグ付きツイートを投稿した外国人観光客ユーザを在日外国人，訪日外国人の 2 つの属性に分類し，ユーザが投稿した場所を利用して，2 つの属性における観光行動の比較・分析を行った．佐伯らが分類を行なった 2 つの属性だけでなく，他の様々な属性を考慮することで，より細かな分析を行なうことができると考えられる．本研究では，様々なユーザ属性を考慮した観光行動の分析を行えることを目指し，マイクロブログの投稿に付与されたジオタグを用いた手法を提案する．

3. 提 案 手 法

本章では，Twitter ユーザが投稿したツイートの位置情報に基づいた特徴量を用いて機械学習手法に適用し，属性推定を行なう手法について述べる．3.1 節では特徴量の生成を，3.2 節で機械学習による分類を行なう手法について述べる．

3.1 位置情報に基づいた特徴量の生成

本節では，ツイートに付与された位置情報を用いて，ユーザごとに訪れた場所をあらわす特徴量を生成する方法について述べる．

はじめに，ツイートの位置情報に基づいて，対象となるエリアで投稿されたツイートを抽出し，ツイートの投稿者のユーザ ID をもとに，重複しないようにユーザ $U = \{u_1, u_2, \cdots, u_n\}$ を決定する．対象となる領域全体を重複しない任意のメッシュ $P = \{p_1, p_2, \cdots, p_m\}$ に区切る．今回は，500m 四方の大きさのメッシュになるように，緯度と経度の範囲を設定する．$T_u(u_i)$ をあるユーザ u_i によるすべてのツイートの集合とする．また，$T_p(p_j)$ をある緯度経度の範囲で表せる場所 p_j 内で投稿されたすべてのツイートの集合とする．あるユーザ u_i を，そのユーザがどこでツイートしたかによって表した初期ベクトル $\boldsymbol{v}(u_i)$ は，

$$\boldsymbol{v}(u_i) = (t(u_i, p_1), t(u_i, p_2), \cdots, t(u_i, p_m)) \qquad (1)$$

と表せる．ここで，$t(u_i, p_j)$ は，ユーザ u_i が場所 p_j で一度でもツイートを投稿したことがあるかを表し，その値は 0 または 1 となる:

$$t(u_i, p_j) = \begin{cases} 1 & (|T_u(u_i) \cap T_p(p_j)| > 0) \\ 0 & (otherwise). \end{cases}$$

このユーザごとに訪れた場所をあらわしたベクトルを特徴量とし，教師あり機械学習の入力として用いる．また，機械学習に学習させるために，各属性に数値を割り当てたラベルも同時に入力する．今回は性別を推定対象属性としたため，男性を 1，女性を 0 とした．

一人のユーザが訪れる場所は，メッシュの総数に比べ非常に少ないため，それぞれのベクトルはスパースな場合が多い．そのため，過度に疎なベクトルを入力することによる誤分類を起こす可能性が高い．そこで，前処理としてどのユーザも範囲内で投稿していないメッシュを取り除く．生成するベクトルの概略図を図 1 に示す．

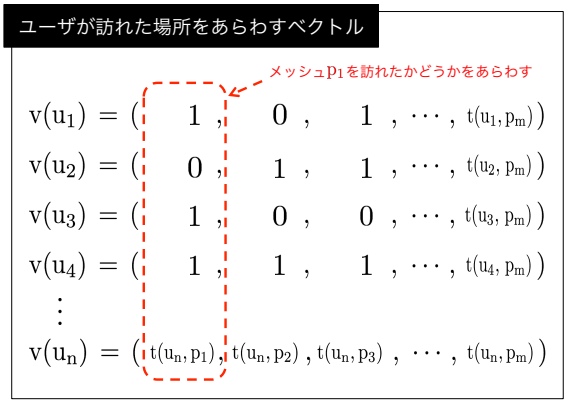

図 1　ベクトルの概略図

3.2　機 械 学 習

本研究では教師あり機械学習手法として，サポートベクターマシン (SVM) [8]，Random Forest [9]，XGBoosting [10] を使用した．SVM については，グリッドサーチを行ない各パラメータを決定した．また，Stratified K-Fold 法により任意の個数 K にデータを分割して，交差検定を行ないその平均値を推定精度とした．SVM および RandomForest の実装には，scikit-learn [11] の SVC，RandomForest を使用した．

4.　評 価 実 験

本章では，実際に Twitter から収集したツイートを用いて，3. 章で提案した手法により性別推定を行ない，性能を評価する．

4.1　データセット

実験のために，Twitter Streaming API(注4) を用いてジオタグ付きツイートをランダムに収集した．収集期間は 2017 年 1 月 1 日から 2017 年 4 月 30 日の全 120 日間であり，この期間に東京 23 区内で投稿されたツイートを用いた．これらのジオタグ付きツイートには，多くは日本人のツイートであったが，海外の利用者のツイートも含まれた．日本でツイートをする外国人利用者には多くの旅行客が含まれる．日本人と外国人利用者との訪れる場所の傾向が異なる可能性がある．そのため本実験では言語設定が日本語以外に設定されたツイートをデータセットから取り除いた．収集したジオタグ付きツイートの投稿者 8578 名のユーザ ID の一覧を作成したのち，プロフィールやテキストを人手で確認して，性別を判別可能なユーザを抽出した．収集したツイートの投稿者の中には，収集期間のうちの投稿数が極端に少ないユーザや，特定の場所でのみ投稿する BOT なども含まれた．これらのユーザは，過度に疎なベクトルとなるため，投稿した場所が多い順番にユーザを並べてプロフィールの確認を行なった．8578 名のうち 1282 名のプロフィールを確認し，男性ユーザ 400 名，女性ユーザ 100 名を抽出した．合計 500 名のユーザのジオタグ付きツイートを用いて実験を行なった．

抽出した 500 名のユーザのツイートを用いて，3. 1 節の方法でベクトルを生成した結果，次元数が 1677 のベクトルがそれぞれ生成された．次元数は，東京 23 区を 500m 四方のメッシュに区切った際の，メッシュの個数をあらわす．これらのベクトルを特徴量として，機械学習に入力した．なお，男性と女性のユーザ数が異なるため，男性ユーザ 400 名を 100 名ずつのグループ M_1，M_2，M_3，M_4 に無作為に分割し，男女各 100 名ずつの 4 つの組み合わせに機械学習を適用した．また，Stratified K-Fold を用いて，$K = 10$ に設定して 4 つの組み合わせをそれぞれ，各性別のラベルのユーザ数が等しくなるようにデータを 10 分割し，1 つをテストデータ，残りの 9 つを学習データとした．

男女で訪れる場所の傾向にどの程度差があるかを確認するために，特徴量選択として，範囲内でユーザが投稿したツイート数が多いメッシュを，多い順に 5 個，10 個，20 個，50 個，100 個，200 個，400 個，600 個，800 個，1000 個，1200 個，1400 個，1677 個に絞り込んだ場合のベクトルをそれぞれ作成し，分類結果を比較する．

4.2　実 験 結 果

SVM，Random Forest，XG Boost で分類を行なった結果を表 1 に示す．正答率と F 値は，男性 4 グループそれぞれのデータを使用した結果，および Stratified K-Fold を用いて交差検定を行なった結果を全て平均した値を示している．

分類実験を行なった結果，最も精度が良かったときの正答率と，男性・女性の F 値は，SVM ではそれぞれ 85.4%，0.866，0.831 となり，Random Forest ではそれぞれ 84.1%，0.844，0.831，XG Boost ではそれぞれ 82.5%，0.811，0.834 となった．SVM では，メッシュの個数が少ない場合では Random Forest，XG Boost に比べ分類性能が低いが，メッシュの個数を増やした場合では，SVM が最も分類性能が高い．また，どの手法においてもメッシュの数が 5 個の場合ではうまく分類ができていないが，メッシュの数を増やすに従って精度が大きく向上し，メッシュの数が 100 個以上の場合，正答率が 80%前後，F 値が 0.80 前後の精度で分類を行えた．メッシュの数が 100 個以上の場合では，精度がなだらかに向上した．どの手法でも，計算時間はメッシュの個数を増やすに従い増加しており，特に XG Boost では著しく増加した．

(注4)：https://dev.twitter.com/streaming/overview

表 1　性別推定結果

機械学習手法		SVM				Random Forest				XG Boost			
精度		正答率 (%)	F 値		時間 (ms)	正答率 (%)	F 値		時間 (ms)	正答率 (%)	F 値		時間 (ms)
			男性	女性			男性	女性			男性	女性	
	5	61.0	0.560	0.633	1.95	67.0	0.623	0.697	12.2	67.5	0.625	0.704	8.20
	7	65.1	0.617	0.666	2.34	73.5	0.705	0.755	11.4	69.5	0.671	0.710	9.38
	10	65.9	0.612	0.685	1.17	74.8	0.733	0.757	11.8	73.4	0.719	0.742	12.1
	15	70.5	0.669	0.725	1.17	78.0	0.774	0.781	11.6	78.3	0.767	0.793	14.5
	20	71.9	0.694	0.728	2.34	81.5	0.813	0.813	12.0	80.9	0.797	0.818	19.1
	50	79.3	0.780	0793	2.73	82.8	0.828	0.823	12.0	82.5	0.810	0.835	30.5
	100	80.8	0.805	0.803	4.69	82.1	0.823	0.815	12.0	81.6	0.801	0.827	50.8
個数	200	84.1	0.844	0.832	6.25	82.5	0.831	0.811	12.9	82.9	0.820	0.834	88.7
	400	85.4	0.863	0.836	10.1	80.1	0.839	0.817	13.2	82.5	0.812	0.831	162
	600	85.3	0.862	0.835	14.0	83.0	0.832	0.820	15.0	82.1	0.809	0.829	228
	800	85.3	0.865	0.834	19.9	83.0	0.834	0.816	13.8	82.0	0.809	0.828	299
	1000	85.4	0.865	0.833	24.9	82.0	0.827	0.806	14.3	82.5	0.811	0.834	360
	1200	85.4	0.866	0.831	28.9	83.4	0.839	0.823	14.7	81.5	0.803	0.822	425
	1400	85.4	0.866	0.831	33.2	83.2	0.837	0.822	14.6	82.4	0.811	0.832	486
	1677	85.4	0.866	0.831	39.4	84.1	0.844	0.831	15.0	82.4	0.810	0.832	571

5. 考　　察

4. 章で作成したベクトルと，機械学習による推定結果について考察を行なう．作成したベクトルは，前処理を行なかった場合では次元数が 2792 であったが，前処理後の次元数は 1677 であった．つまり，範囲内で一人もツイートを投稿していないメッシュが 1115 箇所存在し，全メッシュの約 3 分の 1 であり，非常に多くのメッシュを前処理によって取り除いたことになる．今回の実験では，正解に用いるユーザ数が 500 名であり，これらのユーザが訪れていない場所が多かったことが原因として挙げられる．ユーザの行動分析を行なうためには，今回取り除いた場所も重要な情報になりうる．そのため，正解に用いるユーザをより増やす必要があると考えられる．

機械学習による推定結果を考察するために，実験に使用したデータセットに使った男女それぞれのツイートの投稿場所を地図上にマッピングし，分析を行なった．各性別のユーザのツイート投稿場所を図 2 にあらわす．図 2 では，男性を青色，女性を赤色でツイートの投稿場所をあらわしている．また，ツイートの投稿者数の割合の差を色の濃度であらわしており，色が濃いほど男性と女性の投稿者数の割合の差が大きいことを示す。マッピングされた場所において，男性と女性の投稿者数の割合の差が大きいエリアが存在した．たとえば，六本木周辺では，男性の割合が大きい場所と，女性の割合が大きい場所が混在した．これらの場所では，男性は国立新美術館や東京ミッドタウンなどのランドマークや飲み会，女性はフレンチ，エステについてのツイートを投稿していた．また，男性の割合が大きい御茶の水周辺では，カレーやラーメン，丼ぶり，飲み屋など，男性が好むと思われる食事などのツイートが多く投稿されていた．女性の割合が大きい恵比寿周辺では，お洒落なランチやフレンチ，焼肉屋などの写真を載せたツイートが多く見られ，お洒落なものを好む女性のツイートが投稿されていた．これらのように，男性と女性とでは興味の対象が異なることが，男性と女性の訪れる場所に影響した可能性がある．これらから，一部のメッシュでは男女で投稿者数の割合に差異があり，男女で訪れる場所の傾向が異なることがわかった．これにより,「ある属性の人は，ある地域によく訪れる」という仮定が正しいことが確認できた．最も分類性能が高かった SVM の場合，正答率が 85.5%，F 値が男性・女性でそれぞれ 0.866，0.831 の精度で分類することができ，本手法はユーザの性別分類に有効であることがわかった．しかし，範囲内でユーザが投稿した数が多い順にメッシュを絞り込んでベクトルを生成した場合，メッシュの数が 10 以下では十分な精度で分類することができない．これより，メッシュの数はある程度必要であることがわかった．ユーザの数が多いとベクトルの次元数も増えるため，計算時間の削減のためには，メッシュの数を絞り込む以外の次元削減の方法を試みることが必要である．

6. おわりに

本論文では，マイクロブログの投稿に付与されたジオタグを用いて，ユーザ属性を推定する手法を提案した．また，実際のデータを用いて，提案手法を適用した性別推定実験を行ない，分類性能の評価，考察を行なった．その結果，最も分類性能が高かったもので正答率が 85.5%，男性と女性それぞれの F 値 0.866，0.831 の精度で分類することができた．また，男女それぞれのツイートの投稿場所を地図上にマッピングした結果，男性ユーザと女性ユーザの投稿数の割合の差が大きい地域があることがわかり，男性と女性とでは投稿場所に偏りがあることが確認できた．これらのことより，ジオタグを用いた性別推定手法がある程度有用であることがわかった．

本手法では，ユーザがある場所を訪れたかどうかに基づいたベクトルを生成したが，各メッシュは地図上での位置関係を考慮しておらず，各次元が独立している．そのため，駅の周辺

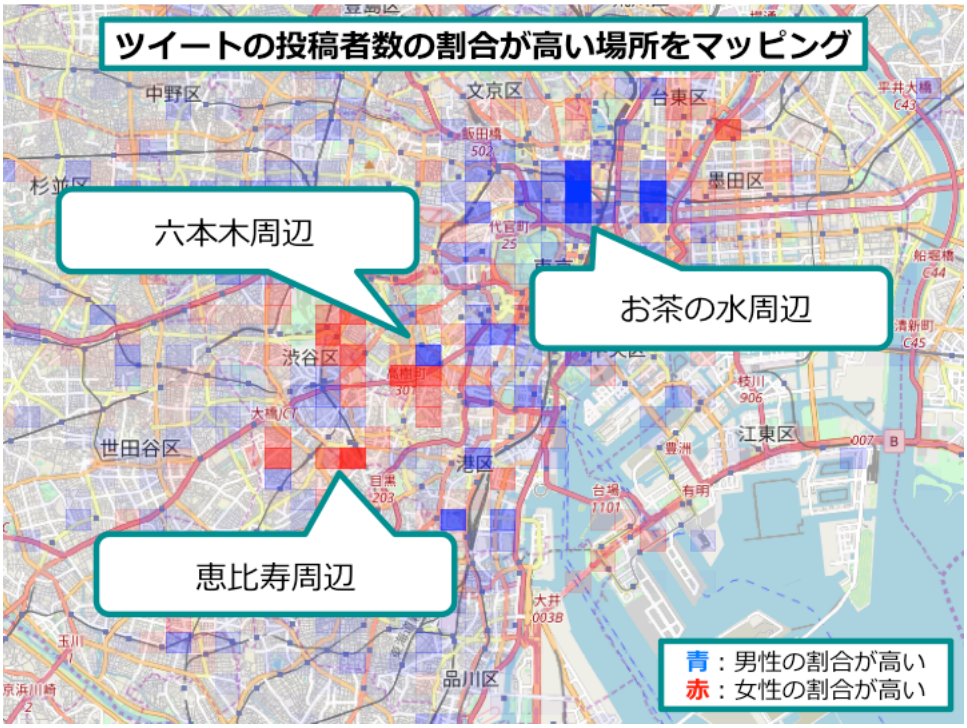

図 2　各性別のユーザのツイート投稿場所

や，観光地の周辺のメッシュを一つの場所として扱うなど，一つの場所を不均一のセルにすることで，推定精度の向上や計算時間の削減が見込める．また，本論文では，推定対象属性として，比較的正解の作成が容易な性別を用いて実験を行なったが，ユーザの観光行動の分析のために，年代や職業を用いて実験を行なうこともまた今後の課題として挙げられる．

謝　　辞

本研究は，首都大学東京傾斜的研究 (全学分) 学長裁量枠戦略的研究プロジェクト戦略的研究支援枠「ソーシャルビッグデータの分析・応用のための学術基盤の研究」，および JSPS 科研費 16K00157, 16K16158 による．

文　　献

[1] 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 (CDS), Vol. 2, No. 1, pp. 82–93, 2012.

[2] Zachary Miller, Brian Dickinson, and Wei Hu. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science*, Vol. 2, No. 04, p. 143, 2012.

[3] 平野徹, 牧野俊朗, 松尾義博. Markov logic を用いたテキストからのユーザ属性推定. 人工知能学会全国大会論文集, Vol. 27, pp. 1–4, 2013.

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

[5] 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. 電子情報通信学会論文誌 D, Vol. 96, No. 6, pp. 1503–1512, 2013.

[6] 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, 2014.

[7] 佐伯圭介, 遠藤雅樹, 廣田雅春, 倉田陽平, 横山昌平, 石川博. 外国人 twitter ユーザの観光訪問先の属性別分析. 第 7 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 2015), C4-3, 2015.

[8] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[9] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, Vol. 2, No. 3, pp. 18–22, 2002.

[10] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, Vol. 12, No. Oct, pp. 2825–2830, 2011.