

マイクロブログを用いたユーザの訪問目的と動向の推定

野沢 悠哉[†] 遠藤 雅樹^{†,††} 江原 遙[†] 廣田 雅春^{†††} 横山 昌平^{††††}
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

^{††} 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

^{†††} 大分工業高等専門学校 情報工学科 〒870-0152 大分県大分市大字牧 1666

^{††††} 静岡大学 情報学部 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: [†]{nozawa-yuya,endo-masaki}@ed.tmu.ac.jp, ^{††}{ehara,ishikawa-hiroshi}@tmu.ac.jp,

^{†††}m-hirota@oita-ct.ac.jp, ^{††††}yokoyama@inf.shizuoka.ac.jp

あらまし 日本国内の旅行者の数の増加傾向に伴い、ビジネスや観光といった旅行者の訪問目的・訪問動向などの観光マーケティング情報の重要性が増している。現在、政府は旅行者の誘致活動や受け入れ体制整備のため、市場調査を用いてマーケティング情報を収集・公開している。しかし、これらの情報の粒度は大きく、旅行全体の訪問目的などはわかっていても、実際のスポットの訪問目的・訪問動向といった、実際に求められる詳細な情報を取得する事は難しい。一方、近年のマーケティング調査では、Twitterのようなソーシャルメディアからデータ分析を行う手法が利用されている。ソーシャルメディアでは、各ユーザが、即時にその場で投稿を行うため、粒度の細かい情報を得ることができる。そこで、本研究では、各ユーザの訪問におけるマーケティング情報を取得するため、Twitterの各投稿に対して、ユーザの訪問目的と訪問動向の推定、分類を行うことで、より粒度の細かい情報を取得する手法を提案する。

キーワード Twitter, 属性推定, 観光調査

1. はじめに

近年、観光は、基幹産業として重要な地位を占めている。観光による消費活動は、運輸・宿泊・旅行サービス・飲食・農林水産・小売・製造などの産業全体への高い波及効果が期待されている^(注1)。そのため、観光客数を増加させることは重要な課題であり、現在、様々な取り組み^(注2)が行われている。

観光客を増加させることを目的とした取り組みを効果的に行うためには、訪問先やその目的などの観光客の情報を正確に把握する必要がある。観光客に関する情報の中でも、観光客の動向を得るために、政府は、旅行・観光消費動向調査^(注3)や、訪日外国人消費動向調査^(注4)などの市場調査を行っている。これらの調査では、たとえば、都道府県などの単位で観光客が訪問した地域をアンケートによって調査している。アンケートによる調査の利点として、対面で調査を行っているなどの理由からアンケート結果の信頼性の高さがあげられる。また、調査を行う側が設問の内容を設定することができるため、設問に即した情報を得やすいことも、アンケートによる調査の利点である。しかし、アンケート調査の欠点として、調査員などのコストの

制限により、得られるデータの量には限界がある。また、調査を行う側が分析したい内容を表現する回答を正確に得ることが可能な設問を設定することは難しい。さらに、たとえ、正確な設問が設定が出来たとしても、得られるデータは、設問の設定に従うため、たとえば、都道府県単位や、月単位の設問で得たアンケート結果を観光地単位などのそれより細かい粒度で分析することが出来ないことや、分析する対象に対して柔軟に用いることが出来ないなどのデータの質にも限界がある。

そのため、アンケートによる調査に代わる方法として、観光地や商品に対する感想や意見などが投稿されている、ソーシャルメディアから観光情報を抽出する研究が盛んである[1], [2], [3], [4]。特に、観光情報の抽出に適した情報源のひとつとして、Twitter^(注5)があげられる。Twitterでは、ユーザはツイートと呼ばれる短い文章を投稿する。一度に投稿可能な文字数が少ないことや、携帯端末での投稿が可能なことから、観光地などで体験した内容やその感想をユーザがその場で投稿することが多い。Twitterから観光客に関する情報を抽出する場合、次のような利点があげられる。まず、調査のために、調査員などを派遣する必要がないため、少ないコストで、大量のデータを集めることができる。また、アンケートによる対面調査と比較して、分析する内容を事前に決定する必要がないため、分析するニーズに合わせて、時間や空間などの粒度を決定することが可能である。

そこで、本研究では、観光客に関する情報を抽出する研究の

(注1)：観光庁 観光立国の実現に向けて：

<http://www.mlit.go.jp/common/000138664.pdf>

(注2)：観光庁 観光地域づくり

<http://www.mlit.go.jp/kankocho/shisaku/kankochi/index.html>

(注3)：観光庁 旅行・観光消費動向調査

<http://www.mlit.go.jp/kankocho/siryou/toukei/shouhidoukou.html>

(注4)：観光庁 訪日外国人消費動向調査

<http://www.mlit.go.jp/kankocho/siryou/toukei/syouthityousa.html>

(注5)：<https://twitter.com/>

表 1 「観光」と「ビジネス」に該当する行動

観光	周遊観光, 温泉, 自然鑑賞, 博物館・美術館, コンサート映画演劇鑑賞, イベント参加, 植物園, テマパーク, トレッキング山登り, 海水浴, スポーツ観戦, エコツアー, キャンプ・体験学習, 保養・リゾート, 街歩き, 動物園保養・リゾート, 水族館, 釣り・クルーズ, 修学旅行, ホームステイ, アマチュアのスポーツ活動, 新婚旅行等
ビジネス	本社・支店・工場などの訪問, 取引先の訪問, 建設・土木等の監督, 研修・セミナーへの参加, 会議・大会・学会・コンベンションへの参加, 見本市・展示会への参加, プロのスポーツ活動, 接待旅行・講演や演奏会の開催出演, 有給の研究・教育・調査活動・ボランティア活動, 視察取材, インセンティブツアー, ツアー添乗, その他業務目的のすべての活動等

ひとつとして、Twitter のユーザから訪問者とその訪問期間を抽出し、訪問者の観光地での訪問動向や、その訪問目的を推定する。提案手法では、ユーザの各ツイートに対して訪問者の訪問動向を訪問時におけるツイート群に対してユーザの訪問目的を分類する。ここで、本論文において、日頃は、特定の地域でツイートをしているが、観光や、ビジネスなどの理由により、短期間のみ別の地域でツイートするユーザを訪問者とする。訪問期間とは、訪問者がその地域に滞在していた期間を表す。次に、訪問目的とは、ユーザが訪問を行う理由を表し、「観光」と「ビジネス」の 2 種類を本研究の推定対象とする。また、本研究において、「観光」と「ビジネス」に該当する行動は、観光庁の旅行・観光消費動向調査の旅行目的を参考に作成した表 1 に従うとする。訪問動向は、ユーザがそのツイートの時点で何を行っているかを表す。本研究では、「観光」、「ビジネス」、「食事」、および「購買」の 4 種類の行動を推定対象とする。訪問動向を別に推定するのは、訪問目的のみでは、訪問者の行動の詳細がわからないためである。たとえば、ビジネスで出張したユーザが仕事の合間に観光や食事に該当する行動を行うことなどが予想される。しかし、観光を目的としたユーザと比較すると時間の制約が考えられるため、ビジネスを目的としたユーザには短時間で楽しむことができる観光や食事を勧める必要があると考えられる。そのため、本研究では、訪問目的より詳細な訪問動向も推定する。

本論文の構成は以下の通りである。2 章では、関連研究について述べる。3 章では、Twitter を用いた訪問者の訪問目的と訪問動向を推定する手法について述べる。4 章では、実験を行う手法について述べる。5 章では、提案手法により分類された結果を基に、評価実験を行い、比較、考察を行う。6 章では本研究のまとめを述べる。

2. 関連研究

マーケティングなどに応用するため、ユーザの属性を推定する研究が盛んである。推定の対象となる主な属性は、性

別 [5], [6], [7], [8], 年齢 [6], [7], [8], [9], 政治的指向 [6], [7], [10], 居住地域 [7], [8], [11], [12], [13], 職業 [14], [15], 立場 [16] などがげられる。本研究で取り組む、訪問者の訪問動向と、訪問目的の分類も、ユーザの属性の推定に関連する研究のひとつである。

また、Web 上の情報から観光情報を抽出する研究も、近年、盛んに行われている。中嶋ら [17] は、位置情報付きツイートを用いて、観光ルートを推薦する手法を提案している。その際、収集したツイートを手がかり語や、品詞の特徴から、「食事」、「景観」、「行動」の 3 つのカテゴリに分類している。また、新井ら [18] は、日本語と英語で記述された旅行ブログエントリを対象に、「買う」、「食べる」、「体験する」、「泊まる」、「見る」、「その他」の 6 種類のタイプに分類する手法を提案している。その際、情報利得を用いて各カテゴリ特有の手がかり語を収集し、その有無と転移学習を用いて SVM を行っている。

ユーザの属性の推定の際に用いられる手法は多岐に渡る。その中でも、教師あり学習を用いて属性を分類する場合は、特に、SVM (Support Vector Machine) が用いられることが多い。西村ら [12] は、Twitter のユーザに対し、プロフィールの Location に含まれる都道府県名をラベルとし、ツイート内の名詞に基づいて作成した特徴量によるマルチクラス分類を用いてユーザの居住地の推定を行っている。池田ら [8] は、Twitter のユーザのプロフィールに含まれる情報を用いて、AIC (赤池情報量基準) に基づいてプロフィール推定のためのキーワードリストを自動的に構築し、さらに、SVM を用いて、プロフィールごとのキーワードの出現傾向を学習することで、ユーザの性別、年齢、居住地の推定を行っている。榊ら [14] は、Twitter のユーザの自己紹介文、投稿内容、被リスト名から、tf-idf を特徴量として SVM により、職種別にユーザの職業の推定を行っている。野呂ら [16] は、ブログの記事本文の単語の出現頻度に基づく特徴量を用いて SVM により分類する手法と能動態・受動態の出現頻度に基づいて作成した行為極性辞書を用いて分類する手法から、著者がブログの記事に記載されたイベントにおいて、主催者が参加者かの推定を行っている。

これらの研究に対し、本研究では、Twitter を対象に、SVM を用い、訪問者の訪問目的を「観光」、「ビジネス」に、訪問動向を「観光」、「ビジネス」、「食事」、「購買」にそれぞれ分類する手法を提案する。1 度の訪問に対して、観光とビジネスといった訪問目的で分類し、さらに、その中でも訪問中に訪問者が行っている動向にも着目し、分類をすることで、ビジネス目的の訪問者が観光をしていること等を発見できる点に新規性がある。また用いる SVM に関しても、ツイートの本文には、複数のクラスに属するものがあると予想し、マルチラベル SVM を用いる点でも区別できる。

3. 提案手法

本章では、Twitter を用いた訪問者の訪問目的と訪問動向を推定する手法について述べる。本研究では、Twitter から取得したツイートから、訪問期間のツイートを抽出し、訪問先での訪問動向の推定と訪問目的の推定を行う。

提案手法において、訪問動向の推定では、訪問中のそれぞれのツイートごとに推定を行う。訪問動向のクラスは、前述した4つのクラス「観光」、「ビジネス」、「食事」、「購買」と、それ以外のツイートを表す「その他」の5つである。

訪問目的の推定では、ユーザの訪問全体に推定を行う。また、訪問目的の推定では、訪問動向の推定の結果を用いた推定手法と、SVMを用いた推定手法の2種類を提案する。訪問目的のクラスは、1つの訪問期間に対して、「観光」、「ビジネス」と、それ以外の「その他」の3つである。

3.1 前処理

本節では、ツイートの取得や、ツイートに対する前処理について述べる。Twitterからのツイートの取得には、Twitter Streaming API^(注6)を用いた。その際に、ツイートに付与された位置情報に基づいて、日本国内で投稿されたツイートを取得するように設定した。また、取得したツイートの中で、位置情報共有サービスのFourSquare^(注7)を経由して投稿されたと思われる、本文中に「I'm at 地名 url」を含むツイートは削除した。また、リプライ、リツイートも排除した。それ以外のツイートについては、取得したツイートの本文のurl部分と絵文字を削除した。

3.2 訪問先でのツイートの抽出

本節では、ユーザのツイートから、訪問者とその訪問の期間のツイートを抽出する手順について述べる。本論文では、日頃、東京都以外に滞在しており、観光や、ビジネスなどにより東京都に訪れた訪問者を推定し、推定された訪問者を対象とする。

はじめに、ユーザのツイートを投稿時間順にソートする。次に、訪問先の範囲の緯度経度を設定し、それぞれのツイートの位置情報に基づいて、ユーザが訪問先で投稿した全ツイートを取得する。その後、訪問先で投稿した全ツイート数と、ユーザの全ツイート数から、訪問先でのツイート数の割合を算出する。その割合が閾値未満である場合、そのユーザを訪問者とする。閾値は、0.1, 0.2, ... 0.5のそれぞれを設定し、目視で確認して、0.3とした。また、訪問者が訪問先で投稿したツイートを訪問ツイートとする。

次に、訪問者の全ツイートの中で、訪問ツイートが連続している場合、その連続しているツイートの集合を1度の訪問期間における訪問期間ツイート集合とする。訪問期間ツイート集合のそれぞれのツイートの本文を1つにまとめたものを訪問期間文書とする。訪問ツイートは、訪問動向の推定に用いる。また、訪問期間文書は、訪問目的の推定のSVMに用いる。

3.3 ベクトル化

本節では訪問ツイートの本文をベクトル化する手法について述べる。はじめに、3.2節で抽出された、全ての訪問ツイート本文に対し、形態素解析を適用し、名詞、動詞、および形容詞を抽出する。本研究では、形態素解析器としてMeCab^(注8)を用いた。しかし、ツイートの本文には、新語が多く含まれてい

るという特徴があるため、MeCabの辞書に、新語に対応するためにipadicを拡張したmecab-ipadic-NEologd^(注9)を利用した。次に、形態素解析された本文の集合に対して、tf-idfを適用し、ベクトル化する。そして、tf-idfの次元数を削減するために、LSA (Latent Semantic Analysis) [19]を用いた。同様の手順で、訪問期間文書もベクトル化する。

3.4 「その他」のフィルタリング

本節では、クラスとラベルの中でノイズとなる「その他」の処理を行う手法について述べる。本研究において、訪問動向の分類では、「観光」、「ビジネス」、「食事」、「購買」に属さない訪問ツイートと、訪問目的の分類では、「観光」、「ビジネス」に属さない訪問期間文書を「その他」としている。そして、実際の、訪問期間中のツイートの集合にも、「その他」扱いとなるツイートが含まれていると考えられる。これ以降の手法では、訪問動向の分類は4クラスを対象に行うため、事前に、「その他」に該当するツイートをフィルタリングする必要がある。また、訪問目的の分類のSVMは2クラスを対象に行うため、事前に、「その他」に該当する訪問期間文書をフィルタリングする必要がある。

ここでは、「その他」をフィルタリングするため、SVMを用いて、2クラス分類を行う。訪問期間ツイート、または、訪問期間文書に対して、人手により「その他」と「その他以外」にクラスを付与したものをトレーニングデータとして、SVMを適用する。以降は、「その他以外」に属すると推定されたものを対象に提案手法の説明を記す。

3.5 訪問動向の推定手法

本節では、それぞれの訪問ツイートを「観光」、「ビジネス」、「食事」、「購買」の4つのクラスに分類する手法について述べる。分類には、3.3節で作成したベクトルを特徴量としたマルチラベルSVMを適用する。訪問期間ツイートの中には「仕事おわたり。風で散り始めてきたから板橋行って桜見に行こ。そいえば現場でこんなの居た(； ；)」のように、1つのツイートの中に、「観光」や、「ビジネス」などの複数の訪問動向を含むツイートも存在する。そのため、訪問動向を推定するための方法として、マルチラベル分類が有効であると考え、1つのツイートに対して、「観光」、「ビジネス」、「食事」、「購買」、それぞれが適切であるかを分類した。マルチラベルによる分類は、各クラスに該当するか否かの2値分類のSVMを組み合わせてることによって、実装した。

3.6 訪問目的の推定手法

本節では、訪問者の訪問期間における、訪問目的を推定する手法について述べる。訪問目的のクラスは、「観光」、「ビジネス」、それ以外の「その他」の3つである。本研究の訪問目的の推定では、訪問動向の推定の結果を用いた推定手法と、訪問期間文書とマルチラベルSVMを用いた推定手法の2種類の手法を用いて実験を行う。

まず、訪問動向の推定の結果を用いた訪問目的の推定について述べる。訪問動向推定の推定結果を用いて、訪問期間のクラ

(注6) : <https://dev.twitter.com/overview/documentation>

(注7) : <https://ja.foursquare.com/>

(注8) : <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

(注9) : <https://github.com/neologd/mecab-ipadic-neologd>

表 2 訪問ツイートと訪問期間文書のそれぞれのクラス、ラベルの件数

	観光	ビジネス	食事	購買	その他	合計
訪問ツイート	1,073	702	928	659	1,754	5,107
訪問期間文書	381	468			349	1,198

スを以下の手順で設定する。

(1) 訪問期間ツイート集合の中で、訪問動向の分類の結果が「ビジネス」のものが存在する場合、訪問期間のクラスは、「ビジネス」とする。これは、観光が目的で訪れた訪問者は、ビジネスに該当する行動をしないが、ビジネスが目的で訪れた訪問者は、仕事が終わった後などに、観光に該当する行動をする可能性があると考えられるためである。

(2) 1 以外の場合のビジネスと分類されなかった訪問期間について、訪問動向の分類の結果が「観光」のものが存在する場合、訪問期間のクラスは、「観光」とする。

(3) 1, 2 の以外の場合、訪問期間のクラスは「その他」とする。

その後、設定した訪問期間全体の推定クラスと正解クラスから訪問目的の推定結果を計算する。

次に、訪問期間文書とマルチクラス SVM を用いた推定手法について述べる。訪問期間文書は 3.2 節で述べたものであり、3.3 節においてベクトル化されている。それらを「観光」、「ビジネス」、「その他」でマルチクラス SVM により訪問目的の分類を行う。

4. 実験方法

本章では、Twitter のデータに対して、提案手法である訪問動向と訪問推定の複数の組み合わせについて評価実験を行う。

4.1 実験条件

実験に用いる Twitter のデータは、2015 年 3 月 11 日から 2015 年 10 月 28 日までの期間内に日本国内で投稿された、105,610,964 件のツイートを用了。3.1 節、3.2 節で述べた処理により、得られた東京への訪問者は 68,588 人、訪問ツイートは 706,221 件、訪問期間文書は 218,052 件である。実験に用いる正解データを作成するため、訪問ツイートの「観光」、「ビジネス」、「食事」、「購買」、「その他」のラベルを訪問期間の文書に「観光」、「ビジネス」、「その他」のクラスを人手により付与した。それぞれの件数を表 2 に示す。

その後、3.3 節で述べた処理により、訪問ツイートと、訪問期間文書をそれぞれベクトル化した。その際、LSA の次元数は、100 次元から 500 次元までを 100 次元ずつ増やして、マルチクラス SVM を適用し、Accuracy が最も高かった 400 次元を以下の実験では用いる。

実験に用いる全ての SVM は、カーネル関数にガウスカーネルを用いた。また、ハイパーパラメータである c と、 γ は、5 交差検定を行い、それぞれの実験で、最も Accuracy が高いものを用いた。分類結果を比較する為、評価指標として、Accuracy, Precision, Recall, F 値を用いる。

4.2 実験パターン

本節では、本論文で行った実験のパターンについて述べる。

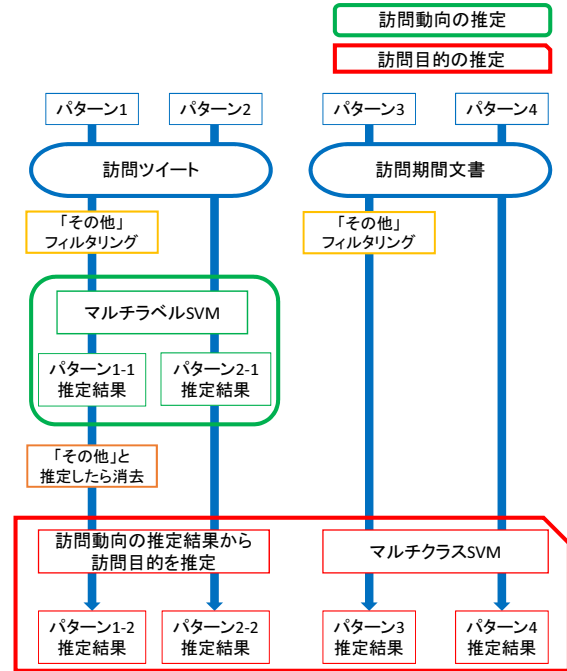


図 1 実験全体図と実験のパターン

はじめに、本論文で行った実験の全体図を図 1 に示す。訪問動向の推定では、訪問ツイートに対し「その他」のフィルタリングを適用し、マルチラベル SVM を用いて訪問動向の推定を行うパターン 1-1 と、適用せずにマルチラベル SVM を用いて訪問動向の推定を行うパターン 2-1 の 2 種類の実験を行った。

訪問目的の推定では、訪問動向の推定で行ったパターン 1-1 の結果から、「その他」とであると分類された訪問ツイートを消去し、訪問目的の推定を行うパターン 1-2 と、パターン 2-1 の結果から訪問目的の推定を行うパターン 2-2、訪問期間文書に対し「その他」のフィルタリングを適用し、マルチクラス SVM を用いて訪問目的の推定を行うパターン 3、適用せずにマルチクラス SVM を用いて訪問目的の推定を行うパターン 4 の 4 種類の実験を行った。

5. 実験結果

本章では、提案手法の有効性を検証するため、各実験パターンの結果を述べ、比較、考察を行う。

5.1 「その他」のフィルタリングについて

はじめに「その他」のフィルタリングについて述べる。「その他」のフィルタリングは、パターン 1-1、パターン 3 で用い、パターン 1-1 では、訪問ツイートに対して、パターン 3 では、訪問期間文書に対して適用される。訪問期間ツイートと訪問期間文書に対する「その他」のフィルタリングの分類結果を表 3 に、分類によって消去された訪問ツイート、訪問期間文書の数とその割合を表 4 に示す。

表 3 より、訪問期間ツイート、訪問期間文書のどちらに対しても、「その他」の Recall は低いことがわかる。これにより、「その他」フィルタリングでは、「その他」消去が十分でないことを示している。また、「その他以外」の Recall は高いため、誤消去が少ないこともわかる。また、表 4 より、訪問ツイートの「そ

表 3 「その他」のフィルタリングの適用結果：評価指標

適用対象	Accuracy	属性名	Precision	Recall	F 値
訪問ツイート	0.742	その他	0.852	0.326	0.472
		その他以外	0.722	0.97	0.828
		avg / total	0.770	0.744	0.700
訪問期間文書	0.730	その他	0.442	0.228	0.266
		その他以外	0.752	0.940	0.834
		avg / total	0.664	0.730	0.670

表 4 「その他」のフィルタリングの適用結果：消去数

適用対象	属性名	適用前	適用後	消去数	消去割合 (%)
訪問ツイート	ビジネス	702	665	37	5.270
	観光	1,073	1,038	35	3.262
	購買	659	643	16	2.428
	食事	928	917	11	1.185
	その他	1,745	1,177	568	32.550
訪問期間文書	ビジネス	468	447	21	4.487
	観光	381	351	30	7.874
	その他	349	269	80	22.922

の他」は、約 33%消去されている。それ以外のクラスの誤消去は、数%にとどまっている。これらのことより、完全ではないものの、「その他以外」の数を保持したまま、「その他」の数を減らすことができているといえる。

次に、表 3 において、「その他」の Recall が低い原因について考察する。本実験において、「その他」に属するのは、訪問ツイートでは、「観光」、「ビジネス」、「食事」、「購買」以外の訪問ツイート、訪問期間文書では、「観光」、「ビジネス」以外の訪問期間文書である。そのため、「その他以外」に比べ、多様な内容のツイートが含まれており、特徴量のベクトルにばらつきがあることが原因と考えられる。

「その他」を消去することで、「その他」を含むデータセットは、本研究の目的に適した「その他」を含まないデータセットに近づくと考えられる。一方、「その他以外」を誤消去しているが、誤消去は僅かである。これらのことより、「その他」のフィルタリングを行うことで、これ以降の推定の性能が向上することが期待される。

表 3、表 4 より、「その他」のフィルタリングでは、「その他」の全てが消去されてはいないので、これ以降の実験でも、分類に「その他」を含めている。

5.2 訪問動向の推定について

次に、訪問動向の推定の性能について述べる。訪問動向の推定は、4.2 節で述べた 2 種類の実験パターンがある。それぞれのパターンについての実験結果を表 5 に示す。

表 5 において、パターン 1-1 とパターン 2-1 の全クラスの F 値の平均はいずれも約 0.65 である。それぞれの属性ごとの F 値を比較すると、パターン 1-1、パターン 2-1 いずれも、観光の F 値が低いことがわかる。ここで、表 1 において、ビジネスに比べ、観光に該当する行動の種類が多いことがわかる。そのため、観光は、他の属性に比べ、訪問ツイートのベクトルにばらつきがあることが考えられる。結果として、分類の際に、観

表 5 訪問動向の分類結果

パターン	Accuracy	属性名	Precision	Recall	F 値
1-1	0.544	ビジネス	0.758	0.646	0.698
		観光	0.642	0.556	0.596
		購買	0.788	0.666	0.722
		食事	0.744	0.652	0.694
		その他	0.616	0.572	0.592
		avg / total	0.698	0.612	0.648
2-1	0.570	ビジネス	0.718	0.650	0.676
		観光	0.638	0.536	0.582
		購買	0.724	0.676	0.696
		食事	0.728	0.644	0.684
		その他	0.710	0.706	0.710
		avg / total	0.704	0.646	0.672

表 6 パターン 1-1 の提案手法によるラベルの推定結果とその正解

正解属性 \ 推定属性	ビジネス	観光	購買	食事	その他	ラベルなし
ビジネス	432	53	16	70	88	79
観光	42	579	52	78	206	155
購買	19	42	430	49	49	90
食事	71	89	44	598	85	113
その他	78	175	42	81	671	212

表 7 パターン 2-1 の提案手法によるラベルの推定結果とその正解

正解属性 \ 推定属性	ビジネス	観光	購買	食事	その他	ラベルなし
ビジネス	452	47	25	79	115	82
観光	48	577	57	76	232	169
購買	23	49	444	52	68	94
食事	85	82	49	597	105	127
その他	64	186	76	92	1,230	213

光の F 値が他の属性に比べ、低くなっていると考えられる。

また、パターン 1-1、パターン 2-1 の提案手法によるラベルの推定結果とその正解を対応させた結果を表 6、表 7 に示す。表 6 と表 7 では、正解属性は人手によって付与されたラベルを表し、推定属性は、提案手法によって推定されたラベルを表す。表 6、表 7 の両方において正解属性が「購買」の訪問ツイートが、他の属性に比べ、誤分類が少ないことがわかる。これは、「購買」に属する訪問ツイートの本文には「買う」や「おみやげ」などの単語が多く出現することが要因となり、他の属性に比べ、訪問ツイートのベクトルの多様性が少ないと考えられる。また、正解属性が「観光」の訪問ツイートが「その他」の属性に多く推定されていることもわかる。これは、「観光」と「その他」に属する訪問ツイートのベクトルが、他の属性に比べ、どちらもばらつきがあることが原因と考えられる。この問題については、観光を複数の属性に細分化するなどの対処が必要と考えられる。さらに、それぞれのラベルにおいて、「ラベルなし」と推定される結果も多いことがわかる。本実験でのマルチラベル SVM は、各属性において二値分類の SVM が構成されるため、どの属性にも属さない、「ラベルなし」の推定結果が生じる。たとえば、本来、「ビジネス」である訪問ツイートが、マルチラベル SVM では「ラベルなし」と推定される可能性がある。このことから、マルチラベルの構成を「ラベルなし」が出現しない構成に再構築する必要があると考えられる。

表 8 訪問目的の分類結果

パターン	Accuracy	属性名	Precision	Recall	F 値
1-2	0.760	ビジネス	0.830	0.770	0.800
		観光	0.740	0.720	0.730
		その他	0.700	0.780	0.740
		avg / total	0.760	0.760	0.760
2-2	0.720	ビジネス	0.790	0.740	0.760
		観光	0.750	0.640	0.690
		その他	0.620	0.770	0.690
		avg / total	0.730	0.720	0.720
3	0.654	ビジネス	0.704	0.796	0.744
		観光	0.664	0.640	0.650
		その他	0.554	0.452	0.482
		avg / total	0.654	0.654	0.646
4	0.652	ビジネス	0.654	0.852	0.742
		観光	0.682	0.546	0.602
		その他	0.634	0.504	0.556
		avg / total	0.658	0.654	0.642

次に、表 5 においてパターン 1-1 とパターン 2-1 を比較すると、「その他」をフィルタリングした結果であるパターン 1-1 の評価指標の値が低下していることがわかる。また、パターン 2-2 に比べ、「その他」をフィルタリングしたパターン 1-1 では、「その他」の全ての評価指標の値が低いこともわかる。一方、それ以外の属性については「その他」フィルタリングを適用したパターン 1-1 の方が評価指標の値が高い。この結果は、「その他」のフィルタリングの分類性能が良く、その分類で、「その他」と推定したものを消去したことが要因となっている。ここで、訪問動向の推定で用いる「その他」以外の 4 つのクラスのトレーニングデータを 1 つにしたものをフィルタリングの「その他以外」のトレーニングデータとして SVM を行っている。そのため、訪問推定による SVM とフィルタリングの SVM による「その他」の分類能力に大きな差はないと考えられる。結果として、フィルタリングによって、誤って「その他以外」と判定されたものを訪問動向の推定において改めて「その他」と分類するのは難しい。この問題については、今後の課題とするが、フィルタリングと、訪問動向の推定に用いる手法のどちらかを変更することで改善すると考えられる。

5.3 訪問目的の推定について

次に、訪問目的の推定の性能について述べる。訪問目的の推定は、4.2 節で述べた 4 種類の実験パターンがある。それぞれのパターンについての実験結果を表 8 に示す。

パターン 1-2 とパターン 3、パターン 2-2 とパターン 4 をそれぞれ比較すると、訪問期間文書を使い、マルチクラス SVM で訪問目的を推定する手法よりも、訪問動向の推定結果を用いた手法が分類性能が良いことがわかる。これは、訪問期間文書を用いた手法では、訪問期間中の全てのツイートを 1 つの文書にまとめ、その文書からクラスを推定するのに対し、訪問動向の推定結果を用いた手法では、訪問期間中の全てのツイートの属性からクラスを推定していることが要因であると考えられる。訪問期間文書を用いる場合も、訪問動向の推定結果を用いる場

合も、その期間の訪問目的を決定する要因は 1 つのツイートである。そのツイートを発見することが、訪問期間文書を用いる手法に比べ、訪問ツイートを用いる手法の方が容易であるため、訪問動向の推定結果を用いる手法が分類性能が良いことが考えられる。

次に、パターン 1-2 とパターン 2-2、パターン 3 とパターン 4 をそれぞれ比較すると、「その他」を消去することで、訪問目的の推定の分類性能が向上することがわかる。これは、「その他」を消去することで、他のクラスの分類性能が向上することと、訪問期間中に「ビジネス」と「観光」を行っていない訪問を「その他」の訪問目的としていることが要因であると考えられる。5.2 節では、あらかじめ「その他」を消去することで、他の属性の分類性能が向上することを確認した。同時に、「その他以外」と推定されたものを改めて「その他」と推定することは難しいことも確認した。しかし、訪問目的における「その他」は訪問動向における「購買」、「食事」と「その他」の集合であると考えられるため、相対的に、訪問目的における「その他」の分類性能が向上したと考えられる。このことから、訪問目的の推定においては、「その他」を消去することが有用であると言える。

6. おわりに

本論文では、観光客に関する情報を抽出する研究の 1 つとして、Twitter に投稿されたツイートの位置情報と本文を用いることで、ユーザの観光地での訪問動向と訪問目的を推定する手法を提案した。提案手法では、ユーザの中から観光地を訪れる訪問者を発見し、訪問者のツイートを tf-idf と LSA に基づく特徴ベクトルを用いて、マルチラベル SVM により、訪問動向を「観光」、「ビジネス」、「食事」、「購買」、「その他」に分類した。その後、SVM の推定結果より訪問目的を「観光」、「ビジネス」、「その他」に分類した。評価実験では、訪問動向、訪問目的の F 値は共に約 0.65 となった。

今後の課題として、訪問動向の推定の性能の改善がけられる。one-class-SVM などの、外れ値検知の手法を「その他」フィルタリングに用いることや、マルチラベル SVM の「ラベルなし」の推定結果に対処し、必ずラベルが振られるように SVM を構築すること、「観光」に属する行動の種類を細分化し、「観光」のサブクラスを作ることなどから、訪問動向の推定の性能が改善することが考えられる。また、提案手法による分類結果を可視化することもけられる。訪問目的別に、それぞれの訪問動向と動向時の位置を可視化し、分析することで、観光客数の増加に有効な情報を得られることが考えられる。

謝 辞

本研究（の一部）は傾斜的研究費（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」による

- [1] 上原尚, 嶋田和孝, 遠藤勉. Web 上に混在する観光情報を活用した観光地推薦システム. 電子情報通信学会技術研究報告 信学技報, Vol. 112, No. 367, pp. 13–18, 2012.
- [2] 樽井勇之. 協調フィルタリングとコンテンツ分析を利用した観光地推薦手法の検討. 上武大学経営情報学部紀要, Vol. 36, pp. 1–14, 2011.
- [3] 倉田陽平, 相尚寿, 真田風. 写真共有サイト投稿データを利用した新たな観光マップの構築. 観光科学研究, 2015.
- [4] 倉田陽平, 奥貫圭一, 貞広幸雄. 個人嗜好に応じた観光コース自動作成システムの開発. 地理情報システム学会平成 12 年度研究発表大会梗概集, Vol. 9, pp. 199–202, 2000.
- [5] John D Burger, John C Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309. Association for Computational Linguistics, 2011.
- [6] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44. ACM, 2010.
- [7] 蔵内雄真, 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.
- [8] 池田和史, 服部元, 松本一則. 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 論文誌トランザクション, Vol. 2011, No. 2, pp. 82–93, 2012.
- [9] John D Burger and John C Henderson. An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 15–20, 2006.
- [10] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks aficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 430–438. ACM, 2011.
- [11] 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. ACM, 2010.
- [12] 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 NLC 言語理解とコミュニケーション, Vol. 112, No. 367, pp. 23–27, 2012.
- [13] 堂前友貴, 関洋平. 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. 情報処理学会論文誌. データベース, Vol. 7, No. 3, pp. 1–13, 2014.
- [14] 榊剛史, 松尾豊. ソーシャルメディアユーザの職業推定手法の提案. 知能と情報, Vol. 26, No. 4, pp. 773–780, 2014.
- [15] 田中成典, 中村健二, 加藤諒, 寺口敏生. マイクロブログの投稿時間に着目したユーザの職業推定に関する研究. 情報処理学会論文誌 データベース, Vol. 6, No. 5, pp. 71–84, 2013.
- [16] 野呂勇太, 廣田雅春, 野澤浩樹, 横山昌平. 能動態・受動態の出現回数に基づくブログ記事毎の立場推定手法の提案. 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 2015), 2015.
- [17] 中嶋勇人, 新妻弘崇, 太田学. 位置情報付きツイートを利用した観光ルート推薦. 情報処理学会研究報告. データベース・システム研究会報告, Vol. 2013, No. 28, pp. 1–6, 2013.
- [18] 藤井一輝, 石野亜耶, 藤原泰士, 前田剛, 難波英嗣, 竹澤寿幸. 多言語旅行ブログエントリを用いた観光情報提示システム. 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 (DEIM2014), 2014.
- [19] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. In-

dexing by latent semantic analysis. *JAsIs*, Vol. 41, No. 6, pp. 391–407, 1990.