学術研究論文

旅行者の行動分析のための旅行ブログエントリの属性推定

藤井一輝[†]，難波英嗣[†]，竹澤寿幸[†]，石野亜耶[††]，奥村学[‡]，倉田陽平[‡‡]
[†]広島市立大学大学院情報科学研究科
[††]広島経済大学ビジネス情報学科
[‡]東京工業大学科学技術創成研究院
[‡‡]首都大学東京都市環境学部

あらまし　旅行者のニーズを知ることは，観光事業を計画する上で非常に重要である．これまで，アンケートの実施などにより，こうした調査が行われてきたが，コストも時間も非常にかかるという問題があった．近年では，アンケートの代わりに旅行ブログエントリが分析に使われるようになってきている．旅行ブログエントリには旅行者の様々な経験が記述されているため，旅行に関する情報を得る上で有益な情報源であると考えられる．我々は，旅行者の行動分析に関する調査を行っており，その第一歩として，旅行ブログエントリの属性を推定する手法を提案する．提案手法の有効性を確認するため，実験を行った．ブログ著者の性別推定では，提案手法のひとつである SSL+TF 手法において 0.877 の推定精度を得た．ブログ著者の使用言語の推定では，精度 0 .972，再現率 0.797 を得た．観光タイプの推定では，提案手法のひとつである IG+MT 手法において，精度 0.597，再現率 0.337 を得た．

キーワード　旅行ブログ，行動分析，属性

Automatic Identification of Attributes of Travel Blog Entries for Travellers' Behaviour Analysis

Kazuki Fujii[†]，Hidetsugu Nanba[†]，Toshiyuki Takezawa[†]，Aya Ishino[††]，
Manabu Okumura[‡]，Yohei Kurata[‡‡]
[†]Graduate School of Information Sciences, Hiroshima City University
[††]Department for Information Systems in Business, Hiroshima University of Economics
[‡]Institute of Innovative Research, Tokyo Institute of Technology
[‡‡]Graduate School of Urban Environmental Sciences, Tokyo Metropolitan University

It is important for tourism planning to be aware of travellers' needs. Traditionally, such analyses were conducted using questionnaires, but these are costly and time-consuming. Recently, in the light of social media, travel blog entries have been used instead of questionnaires.

In travel blog entries, various travellers' experiences and opinions are described, and they can help identify travellers' needs. We are investigating towards travellers' behaviour analysis, and as a first step, we propose a method for identifying attributes of travel blog entries. To confirm the effectiveness of our method, we conducted some examinations. To identify gender, our method, SSL+TF, obtained an accuracy score of 0.877. To identify languages, we used the langdetect program, and confirmed that it obtained precision of 0.972 and recall of 0.797. To identify the content type of each blog entry, our method, IG+MT, obtained precision of 0.597 and recall of 0.337.

Keywords: travel blog, behaviour analysis, attributes

１．はじめに

ある観光地における旅行者のニーズや行動パターンを知ることは，観光事業を計画する上で非常に重要である．こうした情報を得るためには，旅行者に対してアンケートを実施することが一般的であった。しかしこの手法では，非常に時間とコストがかかるという問題を抱えていた．この問題に対し，Web 上で公開されている旅行記，すなわち旅行ブログエントリを収集，分析するという方法が近年では広まりつつある．本研究では，旅行ブログエントリを対象に, 旅行者の行動を分析するための技術を開発する．

旅行ブログエントリを用いた分析にはいくつかの先行研究や事例がある．例えば，Wenger[1]は，オーストリアを訪れた旅行者が書いた旅行ブログエントリを対象に，旅行者の嗜好や行動の傾向を分析し，女性の旅行者は，男性の旅行者に比べて食事に興味があることを明らかにしている．しかし，Wenger の分析は，そのすべてを人手で行っているため，限られた地域の少量のブログエントリしか分析対象にできないという問題点を抱えている．そこで本研究では，世界中のあらゆる地域について，従来の人手による分析では不可能であった大量の旅行ブログエントリを分析可能にするための技術を実現する．

本研究の構成は以下の通りである．2 節では関連研究を紹介する．3 節では旅行ブログエントリからの属性の自動推定について，4 節では属性の自動推定における実験と考察，5 節では本稿のまとめについて述べる．

２．関連研究

2.1. 観光に関する調査・分析

観光施策を展開するためには，観光地のマーケティングが必要不可欠であり，一般的にアンケート調査が用いられてきた．アンケート調査により旅行者の分析を行っている研究として，林ら[2]の研究がある．林らは，関西空港国際線出発ロビーにて，出国待ちの旅行者 1,014 名を対象に, 訪問国や旅行日数, 観光目的などの項目に対してアンケート調査を行い, 旅行動機を明らかにすることを試みた．また, Xia ら[3]はフィリップ島に訪れた旅行者 464 名を対象に性別, 観光地, 居住地などの項目のアンケートを行い, 決定木を用いて旅行者の行動分析を行っている．Jonsson ら[4]も同様にアンケート調査により, 163 名の性別や旅行の動機などの情報を収集し，分析を行っている．しかし，アンケート調査は，人手により行われており，時間や労力といったコストが掛かってしまう問題点がある．さらに，調査が行われている場合でも，調査の実施時期が何年も前であるといった場合，必ずしも最新の情報を反映した調査結果になっていない可能性がある．そこで，本研究ではソーシャルメディアの 1 つである旅行ブログエントリを用いて分析を行う．頻繁に投稿されるソーシャルメディアを対象にすることにより，常時，最新の情報を反映した分析が可能になると考えられる．

ソーシャルメディアの観光マーケティングへの利用は，その最初期には，ソーシャルメディアがそもそも情報源としてどの程度有用であるのか，信頼できるものであるのか議論されてきた[5,6]．その後，ソーシャルメディアが数多くの分析に用いられるようになってきた．Li ら[7]は，中国のポータルサイトの旅行ブログエントリを用いて台湾の観光客から見た中国のイメージ分析を行っており,「景観」や「買い物」,「宿泊」などのカテゴリごとに分類し，旅行ブログエントリの記述された内容を分析している．分析により，温泉に関する旅行ブログエントリは非常に少ないが，それについて書かれている記事全てが好印象であり，温泉に力を入れていく必要であることを明らかにしている．同様に神田ら[8]は，世界遺産登録された「石身銀山遺跡とその文化的景観（島根県大田市）」に関する旅行ブログエントリを対象とし，記述内容を「見る」や「食べる」,「泊まる」な

どのカテゴリに人手で分類し，分析している．その結果,「食べる」において，出雲そばや海の幸といった単語が頻繁に出現しており，地域の伝統的な食材に力を入れる必要があることを明らかにしている．村上ら[9]は，東京，北海道，石川を訪れた訪日外国人旅行者のブログエントリを収集し，TTM(Tiny Text Miner)という分析ツール [10] を用いて各地域に頻出する名詞，動詞，形容詞を比較し，各地域のイメージ分析を行っている．

前節で述べた Wenger[1]は, オーストリアを訪れた旅行者のブログエントリについて, 旅行者の属性(年齢，性別，出生国)を用いた分析を行っている．旅行者の属性は，旅行者本人がブログ・サイトのプロフィール欄に記載したものを利用している．ただし，すべての旅行者がプロフィール欄に年齢，性別，出生国を書いているわけではないため，分析対象となるブログエントリは上記の属性情報をすべて記述している旅行者に限定される．さらに分析そのものも人手で行っているため, Wenger の分析で用いたブログエントリは 188 件にとどまっている．一方で，Wenger は, 旅行者の属性がプロフィール中に明示的に記載されていなくても，ブログエントリ本文を読めば分かる場合もあると指摘している．この作業を人手で行うのは大変なコストがかかることになってしまう．もし属性情報を自動的に推定できるようになれば，本節で言及した様々な分析を，より大規模なブログデータを用いて実施できるようになると考えられる．次節では，属性推定に関する研究について述べる．

2.2. ソーシャルメディアにおける属性推定に関する研究

これまでに, ブログや Twitter などの様々なソーシャルメディアを対象にした属性推定に関する研究が行われている．その目的のひとつは，どのような消費者がどのようなニーズを持っているのかを企業が把握するためのマーケティングの調査対象としてソーシャルメディアが利用される場合である．これまでの研究では，ブログ著者の属性(性別，年齢，居住域)などを文体や記載内容から自動的に推定する手法が提案されている[11, 12, 13]．本研究でも，ブログ著者の属性のひとつとして著者の性別の自動推定を行う．本研究では，Ikeda らが提案する半教師有り学習の手法を用いる．ただし，学習の際に男性あるいは女性のブログ著者が頻繁に使う単語も素性として用いることにする．これにより，Ikeda らの手法を改善できることを示す．

観光に関するソーシャルメディアを対象にした属性推定に関する研究として佐伯ら[14]と石野ら[15]のものが挙げられる．佐伯らは，様々な言語で記述された Twitter を利用し, 訪日外国人の使用言語と訪問先の関係について分析を行っている. 本研究でも，旅行ブログエントリを対象にブログ著者の使用言語を自動推定する．石野ら[15]は，日本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」の 5 種類の観光タイプに自動分類する手法を提案している．本研究でも，英語で記述された旅行ブログエントリを，石野らと同じ 5 種類の観光タイプに分類する. 石野らは，手がかり語を素性とし，機械学習に基づいた分類方法を提案している．これに対し，本研究では，機械翻訳を用いた言語横断的な機械学習を行う．まず，機械翻訳器を用いて英語ブログエントリを日本語に翻訳し，次に，この翻訳結果を，石野らの分類器を用いて分類する．この結果を，機械学習の際の素性のひとつとして用いることで，単純な(英語の)手掛かり語を素性とした学習手法よりも分類精度が向上することを実験により示す．

2.3. 自然言語処理技術を用いた旅行ブログエントリの分析に関する研究

徳久らは，ある観光地の観光開発を行う際，類似する観光地に関するブログエントリから，観光開発のためのヒントとなる文を，自動で分類する手法を提案している[16, 17]．ヒント文とは，例えば,「山陰海岸」の開発を行う場合，類似観光地である「三陸海岸」に関するブログに出現する「遊歩道から断崖絶壁を登った」といった文を指す．これがヒント文になりうるのは，三陸海岸では遊歩道を整備することで観光客の満足度を高めることができたと解釈することができ, これが山陰海岸の開発でも行うべき，という発想につながるためである．もし，本提案手法の技術を用いて旅行ブログエントリやブログ著者の属性を推定することができれば，例えば食に関するブログエントリのみを収集し，食に関する分析を集中的に行うことが可能になると考えられる．

群ら[18]は, ブログ中に出現する「京都駅へ行く」や「到着したのは銀閣寺です」といった文から旅行者(ブログ著者)が実際に訪れた場所(地名)を抽出し，その抽出された地名列に多頻度パターン抽出手法のひとつである PrefixSpan[19]を適用することで, 旅行者の行動パターンを抽出する手法を提案している．さらに，それらを地図上にマッピングすることにより，集約して提示するシステムを実現している．中嶋ら[20]も群らと同様の手法を用い，ブログを対象に旅行者が訪れた場所を自動抽出している．中嶋ら

は，さらに，抽出された各場所について，以下に示す 4 種類の付随情報を抽出する手法を提案している．

- ブログ著者の体験情報
- ブログ著者の評価表現に基づく感想
- ブログ著者がその場所で得た客観的情報
- その場所に関する説明記述

ここで，もし，本研究で推定する属性と組み合わせることができれば，例えば女性と男性の行動経路を比較するといった多様な分析が可能になると考えられる．

Hao ら[21]は，トピックモデルを用いて旅行ブログからある場所に関する情報を抽出する手法を提案している．具体的には，次の 3 つのモジュールを開発している．

(1) 自由な形式で記述されたクエリに対する目的地の推薦
(2) 各目的地の特徴を表した要約の生成
(3) 旅行ブログエントリ中で情報量のある個所および関連画像の抽出

Hao らの開発したシステムにおいても，本研究で開発する技術を用いることで，例えば，男性が記述したブログエントリのみを対象にすることで，男性向けの情報の推薦が可能になる．また，本研究の技術を用いて各ブログエントリに付与された食，宿泊などのタイプを考慮し，タイプごとに要約を生成するといった応用も考えられる．

３．旅行ブログエントリからの属性の自動推定

3.1. 旅行ブログエントリの属性

2.1 節で述べた通り，これまでに旅行者(ブログ著者)や旅行ブログエントリの属性に基づいた観光に関する様々な調査・分析が行われてきたが[1, 7, 8]，これらはブログ著者本人が記載しているプロフィール情報や人手で判定した属性情報を用いていたため，大規模なブログエントリを対象にした分析ができなかった．この問題に対し，本研究では，以下に示す 3 種類の属性情報の自動推定を目指す．この技術を用いることにより，先行研究[1, 7, 8]で行われた調査・分析の対象データの大規模化が可能になると考えられる．

- 性別
  ブログ著者の性別に関する情報であり，ブログ著者が記述した旅行ブログエントリの集合から推定する．
- 使用言語
  ブログ著者が使用する言語に関する情報であり，ブログ著者が記述した旅行ブログエントリの集合から推定する．
- 観光タイプ
  旅行ブログエントリに記述されている観光目的に関する情報であり，旅行ブログエントリから観光の主な目的となる 5 種類のタイプを推定する．

以上の 3 種類の属性を自動的に推定し，推定された属性に基づいて分析を行っていく．これにより，人手では困難であった大量の旅行ブログエントリを利用した分析が可能となる．

本研究では，旅行ブログエントリが登録されている TravelBlog[1]を利用する．TravelBlog とは，2002 年からはじまった Web 上で最も古い旅行コミュニティのひとつで，約 500,000 件のブログエントリ，7,000,000 件の画像が投稿されている．TravelBlog では，旅行ブログエントリを投稿する際に，訪問地に関する情報をあらかじめ決めて投稿する仕様となっている．この訪問地は，例えば「日本の東京都千代田区」の場合，“Asia/Japan/Tokyo/Chiyoda”といったように階層構造を持っている．なお，この地域階層構造は，最上位で 10，最下層で約 10,000 の地域から構成されている．そのため，訪問地の情報を容易に取得できる．なお，TravelBlog に投稿された旅行ブログエントリの数は，上述の階層構造の最上位レベルで集計すると，表 1 に示す内訳となっている．

[1] http://www.travelblog.org

表 1　TravelBlog に投稿された旅行ブログエントリ数(最上位レベル)

地域	ブログエントリ数
Europe	124,390
Middle East	11,097
North America	85,015
Oceania	75,329
Oceans and Seas	1,120
South America	55,179
Asia	151,250
Antarctica	356
Africa	30,907
Central America Caribbean	20,096

また，TravelBlog には，ブログ著者のプロフィールページが設けられており，自由に記述することができる．プロフィールページには，性別などの情報が記述されているが，それらの情報を載せているブログ著者はごく一部であり，無記入のブログ著者も少なくない．そのため本研究では，旅行ブログエントリに着目して属性の推定を行う．属性の自動推定については，3.2 節では性別の推定， 3.3 節では使用言語の推定，3.4 節では観光タイプの推定について，それぞれ説明する．

3.2. 性別の自動推定

本研究では，英語で記述されたブログエントリを用いブログ著者の基本情報である性別に基づいて分析を行う．ブログ著者の性別による訪問地の違いを明らかにすることにより，男性と女性のどちらにプロモーションを行えばよいかの判断が容易になる．例えば，ある訪問地において，男性の旅行者が少ないと分かれば，男性に対してのプロモーションが必要であることがわかる．

そこで本研究では，ブログ著者の性別の推定を行うため，2 つの手法を用いる．1 つ目の手法は，Ikeda ら [12] が提案した半教師有り学習(SSL: Semi-Supervised Learning)による手法を用いる．Ikeda らは，ブログ著者ごとにライティングスタイルがあると仮定している．例えば，男性のブログではアクティブな活動が多く記述され，女性のブログではコスメなどの美容に関する話題が多いかもしれない．こういったライティングスタイルのようなブログ著者の特徴を教師無しブログから得て，ブログ分類を行っている．つまり，教師無しブログが多いほど，様々な側面から評価したブログの特徴を得るこができ，教師有りブログが少量でも，大量の教師無しブログを用いることにより，教師有りブログの数を補える学習が可能となる．Ikeda らは，教師有りブログの数が少ない条件下でブログ著者の性別の推定実験を行っている．その結果，教師有りブログのみの学習では正解率約 0.760，一方，提案手法である半教師有り学習を用いた手法では正解率約 0.890 と高い結果を得ている．そのため本研究でも，教師有りブログの特徴を教師無しブログから得た半教師有り学習を用いて性別の推定を行う．

2 つ目の手法は，単語の出現頻度(TF：Term Frequency)を用いた手法である．人手により，男性と女性に推定された旅行ブログエントリから，それぞれ TF 値を算出し，以下に示す式(1)と(2)を用いて男性と女性のそれぞれに頻繁に出現する単語を収集し，収集された単語を機械学習の素性として，性別の推定を行う．

$$\text{男性に関する語}(w) = \frac{\text{全男性ブログエントリ中の単語 w の出現頻度}}{\text{全男性ブログエントリの総単語数}} - \frac{\text{全女性ブログエントリ中の単語 w の出現頻度}}{\text{全女性ブログエントリの総単語数}} \quad (1)$$

$$\text{女性に関する語}(w) = \frac{\text{全女性ブログエントリ中の単語 w の出現頻度}}{\text{全女性ブログエントリの総単語数}} - \frac{\text{全男性ブログエントリ中の単語 w の出現頻度}}{\text{全男性ブログエントリの総単語数}} \quad (2)$$

なお，対象とする単語 w は，TreeTagger[2]を用いて判定された名詞，動詞，形容詞とし，出現頻度(TF)が 1 回，単語長が 15 文字以上，単語長が 1 文字以下のものは対象外とした．

収集された単語として，男性の場合，人の名前や地名などの固有表現が多く収集された．女性の場合，「pm」や「am」などの時間を表す単語が多く収集された．

3.3. 使用言語の自動推定

TravelBlog に投稿されている旅行ブログエントリでは，英語によるものが多い．一方，フランス語やドイツ語など様々な言語で記述された旅行ブログエントリも存在する．そこで，旅行ブログエントリに使用された言語を自動推定する．本研究では，ブログ著者がどこの国の言語を使用しているかを推定す

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

るため，Java のライブラリである言語推定器(langdetect[3])を用いる．ある言語で記述されたテキストに対して，53 言語について 99%以上の精度で使用言語の推定が可能であり，入力をテキストデータとし，使用言語とその確率を出力する．本研究では，使用言語の推定精度を検討するため，2 つの手法により，ブログ著者の使用言語の推定を行う．

1 つ目の手法は，langdetect により得られた確率が最も高い使用言語をブログ著者の使用言語とする手法(Top)である．具体的な流れを図 1 に示す．まず，ブログ著者が投稿した旅行ブログエントリに対して langdetect を使用する．ブログ著者が投稿した旅行ブログエントリの集合から，langdetect により得られた使用言語とその確率の平均を求める．そして，平均の確率が最も高かった使用言語をブログ著者の使用言語とする手法である．図 1 の場合，ブログ著者の使用言語は英語となる．

2 つ目の手法は，閾値を設けた手法(Threshold)である．具体的には，各言語の平均の確率に対して，閾値を設ける手法である．この手法では，ブログ著者は複数の言語を使用することとなる．例えば，図 1 の場合，閾値を 0.1 以上と設定した際，ブログ著者の使用言語は英語とドイツ語となる．

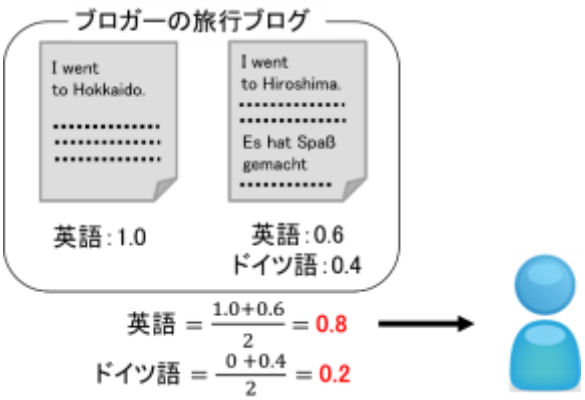

図 1: ブログ著者の使用言語の推定

3.4. 観光タイプの自動推定

同じ訪問地であっても，旅行者によって訪れた目的は様々である．ある旅行者は食事を目的に訪れ，また，ある旅行者は景観を目的に訪れている．旅行者が訪れた目的を明らかにすることにより，各訪問地の特徴を活かした観光施策を行うことができると考えられる．

そこで，訪問地へ訪れた目的を明らかにするため，個々の英語旅行ブログエントリについて，表 2 に示す石野ら[15]が提案している観光タイプ「買う」，「食べる」，「体験する」，「泊まる」，「見る」の 5 種類の観光タイプを自動推定する．

表 2: 観光タイプの判定基準

タイプ	判定基準
買う	観光地で購入したお土産などの物品に関する情報．また，その物品に対する旅行者(ブログ著者)の評価．
食べる	飲食に関する情報．レストランに関する評判情報．
体験する	ものづくり体験やスキー，スキューバダイビングなど，自分の体を使って楽しめる物についての情報．
泊まる	旅行者(ブログ著者)が宿泊した施設に関する情報．
見る	観光名所やパレードなど，見て楽しめる物やイベントについての情報．
その他	上記の観光タイプに該当しない情報．

石野らは，日本語旅行ブログエントリの観光タイプを推定する際，情報利得を用いて，各タイプに分類する上で有用と思われる手がかり語を自動的に収集している．これらの手がかり語が旅行ブログエントリ中に出現するか否かを素性とし，サポートベクターマシン(SVM)を用いて観光タイプの推定を行っている．本研究では，英語旅行ブログエントリの観光タイプを推定する 3 種類の方法を提案する．第一の方法は，石野らの手法と同様，情報利得を用いて，各タイプに分類する上で有用と思われる手がかり語を自動的に収集し，それらをサポートベクターマシン(SVM)の素性に用いる．なお，手がかり語の一部を表 3 に示す．

[3] https://code.google.com/p/language-detection/

表 3　旅行ブログエントリから情報利得により収集した手がかり語の例

タイプ	手がかり語の例
買う	buy, shop, accessory, leather, trading , ad , souvenir , shopping, night-time
食べる	dinner, food, fish, delicious, sashimi, beef, soup, seaweed, taste
体験する	climb, dive, summit, hike, mountain, foot, jump, sky, snorkel
泊まる	hotel, relax, room, breakfast, night, guest, 9pm, luxury, bed, fare, bathroom
見る	museum, see, castle, temple, building, bridge, shrine, century, tale, carving

第二の方法は，英語旅行ブログエントリを，機械翻訳器[4]を用いて日本語に翻訳し，石野らの日本語ブログエントリ用の分類器を用いてその翻訳結果の観光タイプを推定する方法である．第三の方法は，第二の方法で得られた観光タイプそのものを，第一の手法で学習する際の素性とは別に新たに別のひとつの素性として加える方法である．すべての単語の有無を素性として SVM により分類器を構築する手法をベースライン手法とし，これらの 3 種類の方法がベースライン手法を上回ることを実験により確認する．

4. 実験

本研究では，提案した手法の有効性を確認するため，性別の推定，使用言語の推定，観光タイプの推定の 3 種類の実験を行った. 実験の詳細については，それぞれ，4.1 節，4.2 節，4.3 節で述べる．

本実験で使用する旅行ブログエントリでは，3 節で述べた TravelBlog を用いた．TravelBlog では，自由記述によるブログ著者のプロフィールページが設けられており，性別などの正解データの作成には，プロフィールページを用いて行った．また，性別の推定および観光タイプの推定実験では，英語により記述された旅行ブログエントリのみを対象に行った．

4.1 性別の推定実験

4.1.1　実験条件

【実験に用いるデータ】

無作為に選択した 228 人のブログ著者を使用した．このデータに対し，人手により性別推定を行った結果を実験に使用した．なお，228 人のうち，男性は 77 人，女性は 151 人であった．

【機械学習と評価尺度】

機械学習を用いて性別の推定を行った．機械学習には TinySVM[5]を用いた．線形カーネルを使用し，2 分割交差検定を行った．評価尺度には，正解率を使用した．

【実験手法】

提案手法の有効性を確かめるため，以下に示す提案手法について実験を行った．

- **Baseline**：全てのブログ著者を女性と推定した場合．
- **SSL**：教師無しブログから教師有りのブログの特徴を捉える半教師有り学習を用いた手法．
- **TF**：男性・女性のそれぞれに頻繁に出現した単語を素性として与える．
- **SSL+TF**：半教師有り学習の手法(SSL)に頻繁に出現する単語(TF)を素性として与える．

4.1.2 実験結果と考察

提案手法により，得られた実験結果を表 4 に示す．半教師有り学習と単語の出現頻度を用いた SSL+TF 手法では，全ての手法の中で最も高い正解率を得た．

表 4: 性別の推定結果

手法	正解率
Baseline 手法	0.662 (151/228)
SSL 手法	0.667 (152/228)
TF 手法	0.776 (177/228)
SSL+TF 手法	**0.877** (195/228)

性別の推定実験での SSL 手法，TF 手法，SSL+TF 手法の実験結果について考察を行う．SSL 手法について， Ikeda らも同様に性別の推定実験を行っていたが，その際の正解率は約 0.890 であった．しかし，本実験では，期待していた結果を得ることが出来なかった．これは，対象としたデータの違いがその原因として考えられる．Ikeda らが実験の対象としたブ

[4] Microsoft Translator API を翻訳器として用いる．(https://datamarket.azure.com/dataset/bing/microsofttranslator)

[5] http://chasen.org/~taku/software/TinySVM/

ログエントリは，スポーツや政治などの様々なトピックを含んでおり，このトピックの傾向なども有益な情報である．一方，本研究では，旅行や観光に限定したブログエントリを対象に性別の推定を行った．そのため，ブログエントリの特徴を正確に捉えることができず，性別推定の正解率が向上しなかったと考えられる．そのため，TF 手法では，旅行ブログエントリのみを対象とし，男性と女性に頻繁に出現する単語を性別推定の素性として用いているため，正解率の向上が見られたと考える．

4.2 使用言語の推定実験

4.2.1 実験条件

【実験に用いるデータ】

ブログ著者 109 人に対し，人手によりブログ著者の使用言語の推定を行った結果を実験に用いた．人手により推定された使用言語とブログ著者数を表 5 に示す．ただし，複数の言語を使用するブログ著者も存在する．

表 5: 人手による使用言語とブログ著者数

使用言語	人数	使用言語	人数
英語	83	ポルトガル語	1
ドイツ語	10	スウェーデン語	1
スペイン語	9	アフリカーンス語	1
オランダ語	9	ハンガリー語	1
フランス語	6	フィンランド語	1
デンマーク語	5	スロベニア語	1
イタリア語	2	ルーマニア語	1
日本語	2		

【評価尺度】

評価尺度には，精度と再現率を用い，本研究では再現率よりも精度を重視する．推定精度が低い場合，正確な分析が行えないためである．また，旅行ブログエントリは，日々作成され膨大に存在するため，対象とする旅行ブログエントリを増やすことにより，再現率の低さを補うことができる．

【実験手法】

既存のライブラリ langdetect を用いて，使用言語の推定を行っていくが，推定精度を検討するため，以下の手法により実験を行った．

- **Baseline** : langdetect により推定された全ての使用言語をブログ著者の使用言語とした場合．
- **Top** : langdetect により推定された使用言語において，最も高い確率を持つ言語をブログ著者の使用言語とした場合．
- **Threshold** : langdetect により推定された使用言語において，その使用言語の確率に閾値を用いた場合．なお，予備実験により別途用意したデータを用いて F 値が最も高くなった時の値を閾値とした．このため，一人のブログ著者が複数の言語を使用する，と判定される場合もありうる．

4.2.2 実験結果と考察

提案手法により得られた使用言語の推定結果を表 6 に示す．Baseline 手法の精度に比べ，Top 手法と Threshold 手法の精度は，それぞれ，0.472 ポイント，0.387 ポイント向上した．特に，Top 手法では，最も高い精度を得ることが出来た．

表 6: 使用言語の推定結果

手法	精度	再現率
Baseline 手法	0.500	0.925
Top 手法	**0.972**	0.797
Threshold 手法	0.887	0.887

使用言語の推定結果について考察を行う．3.2 節で述べたが，langdetect の言語推定の精度は 99%である．しかし，Baseline 手法である langdetect により推定された全ての使用言語をブログ著者の使用言語とした場合の精度は 0.500 であった．この理由は，文章量に関係している．langdetect では，ある程度の長さの文章に対して，高精度で推定することができる．しかし，本研究で対象としたデータでは，写真をメインとした短い文書の旅行ブログエントリも少なくない．上記の理由により，Baseline 手法の精度は 0.500 であった．一方，Top 手法や Threshold 手法では，高い精度を得ることができた．

4.3 観光タイプの推定実験

4.3.1 実験条件

【実験に用いるデータ】

660 件の旅行ブログエントリを実験に用いた．各ブログエントリに対し，人手で観光タイプを付与し，これらを機械学習の際の訓練用および評価用データとして用いた．観光タイプの内訳を表 7 に示す．なお，ひとつの旅行ブログエントリには複数の観光タイプが付与されても良いという設定にしているため，表 6 に示す旅行ブログエントリ数の総和は 660 を超えている．

表 7: 観光タイプの推定実験に用いたデータの詳細

観光タイプ	旅行ブログエントリ数
買う	30
食べる	97
体験する	143
泊まる	61
見る	316
その他	155

【機械学習】

機械学習手法として SVM を，カーネル関数は線形カーネルを用いた．学習の際，2 分割交差検定を行った．ま

た，評価には再現率と精度を用いた．

【比較手法】

- Baseline 手法：英語の旅行ブログエントリに出現する全単語を素性とした手法.
- IG 手法：英語の旅行ブログエントリを対象に情報利得を利用する．情報利得により収集された手掛かり語を素性として用いる手法.
- MT 手法：英語の旅行ブログエントリを，Microsoft Translator API を用いて日本語に翻訳し，石野ら[15]の手法を用いて観光タイプを推定する手法.
- IG+MT 手法：MT 手法の出力結果を IG 手法の素性のひとつとして加える手法.
-

【実験結果と考察】

実験結果を表 8 に示す．情報利得を用いた IG 手法により，タイプ「食べる」,「体験する」,「見る」では，それぞれ精度 0.728，0.726，0.744 を得ることができ，大幅な精度向上が見られた．これらの結果に比べ，タイプ「買う」と「泊まる」では，それぞれ精度 0.222，0.452 であった．いずれの提案手法もベースライン手法(Baseline)を上回ることができたが，タイプ「食べる」,「体験する」,「見る」と比べ，精度の向上は見られなかった．この理由として，実験に使用したデータ件数が少ないためだと考えられる．英語により記述された旅行ブログエントリのデータ件数を補うため，IG+MT 手法では，MT 手法を素性のひとつに用いた．その結果，タイプ全体での平均精度において，MT 手法は IG 手法に比べ低下したが，データ件数の少なかったタイプ「買う」と「泊まる」において，それぞれ，精度 0.028，0.086 向上させることができた．また，IG+MT 手法では，MT 手法から得られた結果を素性として加えることにより，IG 手法の平均精度と比べ，精度を 0.023 ポイント向上することができた．この結果より，機械翻訳および日本語用分類器を用いることにより，データ数が少ない問題を解決でき，精度を向上させることができた．

5. おわりに

本研究では，大量の旅行ブログエントリを用いて分析をするための 3 種類の属性「性別」,「使用言語」,「観光タイプ」の自動推定を行う手法を提案した．性別の自動推定では，半教師有り学習と単語の出現頻度を用いた手法を組み合わせた SSL+TF 手法により，正解率 0.877 を得た．使用言語の自動推定では，言語推定器により最も高い確率を持つ言語をブログ著者の使用言語とする Top 手法により，精度 0.972，再現率 0.797 を得ることができた．また，観光タイプの自動推定では，IG+MT 手法により，精度 0.597，再現率 0.327 が得られた．2.1 節でも述べたが，これまでに旅行者(ブログ著者)や旅行ブログエントリの属性に基づいた観光に関する様々な調査・分析が行われてきた[1, 7, 8]．同様の調査・分析を他の地域で行う際に本研究の成果が利用可能であると考えられる．

参考文献

[1] A. Wenger, “Analysis of Travel Bloggers’ Characteristics and their Communication about Austria as a Tourism Destination”, Journal of

表 8: 観光タイプの推定実験に用いたデータの詳細

	評価尺度	買う	食べる	体験	泊まる	見る	平均
Baseline	精度	0.011	0.122	0.140	0.087	0.381	0.148
	再現率	0.375	0.676	0.970	0.527	0.961	0.702
IG	精度	0.222	0.728	0.726	0.452	0.744	0.574
	再現率	0.125	0.342	0.311	0.083	0.617	0.296
MT	精度	**0.250**	0.389	0.535	**0.538**	0.580	0.458
	再現率	0.200	0.577	0.266	0.115	0.873	0.406
IG+MT	精度	**0.250**	**0.810**	**0.741**	0.410	**0.773**	**0.597**
	再現率	0.094	0.473	0.295	0.149	0.672	0.337

Vacation Marketing, Vol. 14, No. 2, pp. 169-176 (2008)

[2] 林 幸史, 藤原 武弘, “訪問地域, 旅行形態, 年齢別にみた日本人海外旅行者の観光動機”, 実験社会心理学研究, 日本グループ・ダイナミックス学会, Vol. 48, No. 1, pp. 17-31 (2008)

[3] J. Xia, V. Ciesielski and C. Arrowsmith, “Data Mining of Tourists’ Spatio-temporal Movement Patterns -- A Case Study on Phillip Island”, Proc. of the 8th International Conference on GeoComputation, pp. 1-5 (2005)

[4] C. Jonsson and D. Devonish, “Dose Nationality, Gender, and Age Affect Travel Motivation? A Case of Visitors to the Caribbean Island of Barbados”, Journal of Travel and Tourism Marketing, Vol. 25, No. 3-4, pp. 398-408 (2008)

[5] R. W. Mack, J. E. Blose and B. Pan, “Believe it or not: Credibility of Blogs in Tourism”, Journal of Vacation Marketing, Vol. 14, No. 2, pp. 133-144 (2008)

[6] G. Akehurst, “User Generated Content: the Use of Blogs for Tourism Organizations and Tourism Consumers”, Journal of Service Business, Vol. 3, No. 1, pp. 51-61 (2009)

[7] Y. R. Li and Y. Y. Wang, “Exploring the Destination Image of Chinese Tourists to Taiwan by Word-of-Mouth on Web”, Proc. of World Academy of Science Engineering and Technology Vol. 7, pp. 977-981 (2013)

[8] 神田 佑亮, 藤原 章正, 張 峻屹, “ブログ情報を用いた観光行動と満足度の分析に関する一考察”, 土木計画学研究, 講演集, Vol. 39 (2009)

[9] 村上 嘉代子, 川村 秀憲, “外国人からみた日本旅行 –英語ブログからの観光イメージ分析-“, 人工知能学会誌, Vol. 14, No. 2, pp. 169-176 (2011)

[10] 松村 真宏, 三浦 麻子, “人文・社会科学のためのテキストマイニング”, 誠信書房 (2009)

[11] N. Yasuda, T. Hirao, J. Suzuki, and H. Isozaki, “Identifying Bloggers' Residential Areas”, Proc. of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp. 231-236 (2006)

[12] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, “Effects of Age and Gender on Blogging”, Proc. of AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp. 199-205 (2006)

[13] D. Ikeda, H. Takamura and M. Okumura, “Semi-Supervised Learning for Blog Classification”, Proc. of the 23rd AAAI Conference on Artificial Intelligence, pp. 1156-1161 (2008)

[14] 佐伯 圭介, 遠藤 雅樹, 廣田 雅治, 倉田 陽平, 石川 博, “Twitter データを利用した訪日外国人の訪問先の言語別分析”, 観光情報学会誌 観光と情報, Vol.11, No.1, pp. 45-56（2015）

[15] 石野 亜耶, 藤井 一輝, 藤原 泰士, 前田 剛, 難波 英嗣, 竹澤 寿幸, “旅行ブログエントリと質問応答コンテンツを利用した旅行ガイドブックの情報拡張”, 人工知能学会論文誌, Vol. 29, No. 3, pp. 328-342 (2014)

[16] 徳久 雅人, 村田 真樹, “観光開発のヒントをブログ記事から得るための支援技術 ～SVM を用いる場合～”, 第 8 回観光情報学会全国大会発表概要集, pp. 44-45 (2011)

[17] 謝花 博, 徳久 雅人, 村田 真樹, “観光開発のヒントをブログ記事から得るための支援技術 ～能動学習を用いる場合～”, 言語処理学会第 18 回年次大会, pp. 1324-1327 (2012)

[18] 群 宏志, 服部 峻, 手塚 太郎, 田島 敬史, 田中 克己, “ブログからのビジターの代表的な経路とそのコンテキスト抽出”, 情報処理学会研究報告データベースシステム研究会, Vol. 2006, No. 78, pp. 35-42 (2006)

[19] J. Pei, J. Han, B. Mortazavi-Asi, and H. Pinto, “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”, Proc. of the 17th International Conference on Data Engineering (2001)

[20] 中嶋 勇人, 太田 学, “旅行ブログ記事からの名所とその付随情報の抽出”, 第 5 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 2013) (2013)

[21] Q. Hao, R. Cai, C. Wang, R. Xiao, J. M. Yang, Y. Pang, and L. Zhang, “Equip Tourists with Knowledge Mined from Travelogues”, Proc. of World Wide Web Conference 2010 (2010)

謝辞

本研究の一部は，総務省による戦略的情報通信研究開発推進制度(SCOPE)の支援を受けて行われた．

藤井 一輝（非会員）1990 年生．2013 年 広島市立大学情報科学部知能工学科卒業．2015 年 広島市立大学大学院情報科学研究科知能工学専攻 博士前期課程修了．2015 年 株式会社エネルギア・コミュニケーションズに入社．現在に至る．

難波 英嗣 （正会員）1972 年生．1996 年東京理科大学理工学部電気工学科卒業．2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士(情報科学)．2001 年日本学術振興会特別研究員．2002 年東京工業大学精密工学研究所助手．同年広島市立大学情報科学部講師．2010 年広島市立大学大学院情報科学研究科准教授．現在に至る．テキストマイニング，情報検索，テキスト要約に関する研究に従事．観光情報学会，情報処理学会，人工知能学会，言語処理学会会員．

竹澤 寿幸（正会員）1961 年生．1984 年早稲田大学理工学部電気工学科卒業．1989 年早稲田大学大学院理工学研究科博士後期課程修了．工学博士．1987 年早稲田大学情報科学研究教育センター助手．1989 年（株）ATR 自動翻訳電話研究所研究員．音声対話翻訳の研究開発に従事．2007 年より広島市立大学大学院情報科学研究科教授．現在に至る．音声対話や観光情報学の研究と教育に従事．観光情報学会，電子情報通信学会，情報処理学会，人工知能学会，日本音響学会，言語処理学会会員．

石野 亜耶（正会員）2009 年広島市立大学情報科学部知能情報システム工学科卒業．2011 年広島市立大学大学院情報科学研究科博士前期課程修了．2014 年同大学大学院情報科学研究科博士後期課程満期退学．同年同大学大学院にて博士号（情報科学）取得．同年広島経済大学経済学部ビジネス情報学科助教．2017 年同大学経済学部ビジネス情報学科准教授．現在に至る．テキストマイニング，観光情報処理に関する研究に従事．観光情報学会，言語処理学会，情報処理学会，人工知能学会，言語処理学会会員．

奥村 学（非会員）1962 年生．1984 年東京工業大学工学部情報工学科卒業．1989 年同大学院博士課程修了．同年，東京工業大学工学部情報工学科助手．1992 年北陸先端科学技術大学院大学情報科学研究科助教授，2000 年東京工業大学精密工学研究所助教授，2009 年同教授，現在に至る．2017 年より，ホンダ・リサーチ・インスティチュート・ジャパン客員研究員を兼務．工学博士．自然言語処理，知的情報提示技術，語学学習支援，テキスト評価分析，テキストマイニングに関する研究に従事．情報処理学会，電子情報通信学会，人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．

倉田陽平（正会員）1977 年生．2000 年東京大学工学部都市工学科卒業．2002 年同修士課程修了．2007 年米国・メーン州立大学空間情報理工学科博士課程修了(Ph.D)．ドイツ・ブレーメン大学研究員を経て，2010 年首都大学東京大学院都市環境科学研究科観光科学域准教授，現在に至る．同大学ソーシャルビッグデータ研究センター員．観光情報学会，IFITT，地理情報システム学会，情報処理学会，サービス学会，日本観光研究学会，各会員．