広島経済大学研究論集
第39巻第３・４号　2016年12月
http://dx.doi.org/10.18996/kenkyu2016390305

旅行ブログエントリの自動抽出

石　野　亜　耶*

1. はじめに

2007年１月に「観光立国推進基本法」が施行され，2008年10月には国土交通省の外局として観光庁が設置されるなど，日本では「観光」を21世紀の基幹産業と位置付け，観光を支援する多様な取り組みが積極的に推進されている。

観光庁の発表[1)]によると，2015年には，日本人海外旅行者数は1,600万人，訪日外国人旅行者数は1,900万人を超えている。また，2020年の東京オリンピック・パラリンピックを開催するにあたり訪日外国人旅行者数2,000万人を目標としており，今後も訪日外国人旅行者は増加していくと考えられる。このような中，どのようにして旅行者に対し観光支援を行うかは重要な問題である。

2015年に観光庁から発表された訪日外国人消費動向調査[2)]によると，出発前に得た旅行情報で役に立ったものは，ガイドブックの「ロンリープラネット」が6.8%，「ミシュラン」が1.1%，「その他ガイドブック」が12.2%，旅行会社の「パンフレット」が11.4%，「ホームページ」が11.6%，ソーシャルメディアでは，「個人のブログ」が24.1%，「YouTube」が6.5%，「Twitter」が1.7%である。

個人のブログには，どこを訪れたのか，何を食べたのか，お土産に何を買ったのか，また，旅行を経験してどのように感じたのかなどの旅行記が記載されている。本研究では，このような旅行記が記載されたブログエントリを，旅行ブログエントリと呼ぶ。旅行ブログエントリには，様々な観光情報が含まれているため，旅行する際の有益な情報源であるといえる。

そこで，著者は，旅行者への観光支援を行うために，旅行ブログエントリを利用することで，旅行者に観光情報を提示するシステムの構築［1］に取り組んでいる。本研究では，上記のシステムを構築するための第一歩として，旅行ブログエントリを自動抽出し収集する手法を提案する。

本論文の構成は以下の通りである。２章では関連研究，３章では提案手法，４章では実験と結果について述べる。５章では結論を述べる。

2. 関連研究

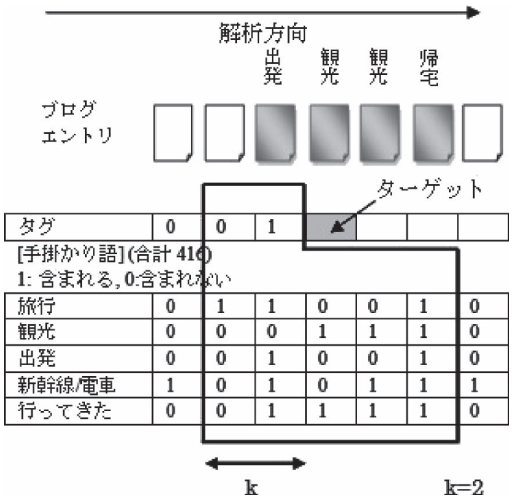
旅行ブログやその旅行ブログエントリを登録したポータルサイトとしては，Travel Blog[3)]，旅行・観光ブログ村[4)]などがある。しかし，このようなポータルサイトに登録されていない一般のブログエントリの中にも旅行ブログエントリが多数存在する。様々な層のより多くの旅行ブログエントリを収集するため，本研究では，一般のブログエントリから，旅行ブログエントリを自動抽出し収集する手法を提案する。

旅行ブログエントリを自動抽出するための手法を提案した先行研究として石野ら［2］の研究がある。石野らの研究では，ターゲットとなるブログエントリが旅行ブログエントリかどうかを判定するために，ターゲットとなるブログエントリの情報のみを利用するのではなく，同一のブログ著者により前後に投稿されたブログエントリに含まれる情報にも注目し，旅行ブロ

* 広島経済大学経済学部助教

グエントリの判定を系列ラベリング問題として機械学習を用いて解決する手法を提案している。機械学習には CRF を使用し，旅行ブログエントリを精度0.867，再現率0.381で自動判定している。旅行ブログエントリを自動判定するために CRF に与えた素性とタグを図１に示す。

石野らの手法では，ターゲットとなるブログエントリが，旅行ブログエントリかどうかを判定するために，ブログ著者が投稿した一連のブログエントリを収集する必要がある。また，人手で収集した手掛かり語を使用しているため，手掛かり語の網羅性が低く，再現率低下の原因となっていると考えられる。そこで本研究では，ターゲットとなるブログエントリのみに対し，自動で収集した手掛かり語を利用することで，旅行ブログエントリを自動判定する手法を提案する。



（石野［1］より抜粋）

図１　旅行ブログエントリを判定するために CRF に与えた素性とタグ

観光を支援するためにブログエントリを利用する研究［3, 4, 5, 6, 7］がある。岡本ら［3］は，一般のブログ検索エンジンを利用することで，地名を含むブログエントリを収集し，それらのブログエントリから，地域イベント情報を抽出する手法を提案している。

郡ら［4］は，ブログエントリからユーザの行動時の代表的な経路とその文脈を抽出し，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提案している。

藤井ら［5］は，旅行ブログエントリからブログ著者の性別や職業などの属性や使用言語，観光目的を自動で判定することで，旅行者の行動分析を行っている。

徳久ら［6］は，旅行ブログエントリから，観光開発のヒントとなる文を抽出する手法を提案している。

新田ら［7］は，旅行ブログエントリから訪日外国人旅行者の訪れた場所，食べたもの，購入したものなどの行動情報を抽出する手法を提案し，行動情報を地図にマッピングして提示するシステムを構築している。

上記の研究のように，ブログエントリから観光に関連する情報を抽出する研究の他に，ブログエントリを人手で分析し，観光地の観光イメージを分析する研究［8］や，観光行動と満足度を分析する研究［9］もある。

このように，旅行ブログエントリを情報源として利用する研究も活発に行われており，旅行ブログエントリを大量に自動収集するための手法が必要である。

石野ら［1］は，旅行ガイドブックのページに対し，関連する旅行ブログエントリや質問応答コンテンツを自動的に対応付ける手法を提案し，旅行ガイドブックの情報を拡張する手法を提案している。石野らの提案手法では，旅行ガイドブックに関連する旅行ブログエントリを高い精度で対応付けるために，旅行ガイドブックと旅行ブログエントリのタイプを自動分類し，その結果を利用している。旅行ガイドブックや旅行ブログエントリのタイプの自動分類には，タイプごとの特徴的な単語を自動収集し手掛かり語として使用している。本研究では，タイプ

分類のための手掛かり語の自動収集手法を利用して，旅行ブログエントリの自動判定に使用する手掛かり語を自動収集する。自動で収集した手掛かり語を利用することで，旅行ブログエントリの判定漏れを減少させることができると考えられる。

3. 旅行ブログエントリの自動抽出

本章では，旅行ブログエントリの自動抽出手法について説明を行う。旅行ブログエントリの抽出手法は，以下の2つのステップに分かれている。Step1については3.1節，Step2については3.2節で説明する。また，Step2に利用する手掛かり語の収集手法については，3.3節で説明する。

●Step1：ブログエントリの収集
●Step2：旅行ブログエントリの自動判定

3.1 ブログエントリの収集

まず，旅行に関連すると考えられるブログエントリを収集する。ブログエントリの収集は，「旅行」というキーワードをクエリとし，Yahoo! ブログ検索[5]で検索を行うことで収集する。検索条件は，更新日時順とする。

3.2 旅行ブログエントリの自動判定

本研究では，観光情報が記述されたブログエントリを，旅行ブログエントリであると判定する。観光の定義とは，1995年に観光政策審議会によって定義された「余暇時間の中で，日常生活圏を離れて行う様々な活動であって，触れ合い，学び，遊ぶということを目的とするもの」[6]とする。

3.2節で収集したブログエントリには，「旅行」というキーワードは含まれているが，旅行ブログエントリでないブログエントリも含まれている。例として，「旅行に行きたい」という願望を記載したブログエントリや，旅行番組の感想を記述したブログエントリがある。そこで本研究では，3.1節で収集されたブログエントリに対し，旅行ブログエントリであるかどうかを，機械学習を用いて自動で判定する。機械学習にはサポートベクトルマシン（Support Vector Machine, SVM）［10］を使用する。

本研究では，機械学習に以下の素性を使用することで，旅行ブログエントリの自動判定を行う。単語分割は，MeCab[7]により行う。なお，本研究では，ターゲットとなるブログエントリが旅行ブログエントリかどうかを判定するために，ターゲットとなるブログエントリ内の情報のみを利用する。

●単語：単語の有無
●写真：写真の枚数
●旅行に関連する手掛かり語：3.3節で説明する手法により収集した手掛かり語の有無

本手法により，旅行ブログエントリであると判定されたブログを抽出し，旅行ブログエントリとして収集する。

3.3 旅行ブログエントリの自動判定のための手掛かり語の収集

旅行ブログエントリを自動判定するために使用する，旅行ブログエントリのタイプ分類の結果を利用した手掛かりの収集方法について説明する。

石野ら［1］は，旅行ブログエントリを，記載内容をもとに，旅行者の観光目的（タイプ）を以下の5種類に自動分類する手法を提案している。

・見る：観光名所などの見て楽しめる物やイベントについての情報を記載されている。
・体験する：○○体験や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が記載されている。
・買う：土産物に関する情報が記載されている。
・食べる：飲食に関する情報が記載されている。
・泊まる：宿泊施設に関する情報が記載されている。

石野らは，情報利得を用いて手掛かり語を収集し，平均で精度0.659，再現率0.510で旅行ブログエントリの自動分類を行うことに成功している。

情報利得とは，「ある単語の出現の有無」の情報が，クラスに関する曖昧さ（エントロピー）をどれくらい減少させるかを示す値である。情報利得を用いて収集した単語は，クラス分類において，有効な単語であると考えられる。任意の単語 w における情報利得 $IG(w)$は，次のように定義される。

$$IG(w) = H(C) - \Big(P(X_w = 1)H(C \mid X_w = 1) + P(X_w = 0)H(C \mid X_w = 0)\Big)$$

エントロピー $H(C)$は，次のように求める。

$$H(C) = -\sum_c P(c)\log P(c)$$

$$H(C \mid X_w = 1) = -\sum_c P(c \mid X_w = 1)\log P(c \mid X_w = 1)$$

$$H(C \mid X_w = 0) = -\sum_c P(c \mid X_w = 0)\log P(c \mid X_w = 0)$$

なお，C はクラス，$P(c)$は全クラスにおける $c \in \{P,N\}$の確率，P は正例，N は負例，X_w は単語 w に対する確率変数である。単語 w が出現する場合 $X_w = 1$，出現しない場合は $X_w = 0$である。石野らは，「見る」，「体験する」，「買う」，「食べる」，「泊まる」のタイプごとに情報利得を求め，値の高い単語を，タイプ分類の手掛かり語として使用している。

タイプ分類に有用な単語は，旅行ブログエントリの自動判定においても有用な単語であると考えられる。そこで本研究は，タイプ分類が行われた旅行ブログエントリから，情報利得を用いてタイプごとに特徴的な単語を200件収集し，旅行ブログエントリの自動判定に使用する。収集した手掛かり語の例を表１に示す。

4. 実　　験

３章で述べた提案手法の有効性の確認するため，実験を行った。

表１　旅行ブログエントリの自動判定に使用する手掛かり語の例

タイプ	手掛かり語の例
見る	見る，眺め，公園，美術館，都市，建築，植物園，紅葉，開花，花，蝋梅，試合，野球，広島カープ，メジャーリーグ，ライトアップ，教会，鳥居，聖堂，広場
体験する	温泉，露天風呂，温泉郷，銭湯，泉質，塩サウナ，風呂上り，アトラクション，絶叫，ミッキー，USJ，カラオケ，大会，自転車，登山道，釣り，海
買う	買う，買い込む，買い物，販売，土産，お土産屋さん，ショッピングモール，免税店，スーパー，レジ，Ｔシャツ，雑貨，お酒，チョコレート，ブランド，化粧品
食べる	食べる，お腹いっぱい，美味しい，ランチ，味，風味，食感，食べ物，料理，レストラン，注文，メニュー，ドリンク，デザート，飲茶，ラーメン，お好み焼き
泊まる	泊まる，宿泊，滞在，ホテル予約，部屋，シングル，ホテル，ビジネスホテル，国民宿舎，リゾート，オーシャンビュー，オンザビーチ，ロビー，浴場，バスルーム

4.1 実験手法

データセット

実験用データとして，3.1節で収集したブログエントリ1,000件に対し，被験者3名により旅行ブログエントリかどうかを人手で判定した結果を用いる。被験者2名以上により旅行ブログエントリであると判定されたブログエントリを旅行ブログエントリとする。人手で判定を行った結果を表2に示す。

表2 旅行ブログエントリの人手での判定結果

旅行ブログエントリである	旅行ブログエントリではない	合計
475	525	1,000

比較手法

旅行ブログエントリの自動判定を行うために，機械学習に与える素性として，先行研究［2］において人手で収集した手掛かり語の有無を利用した場合をベースラインとする。また，3.2節で説明した素性を利用した場合を提案手法とする。

機械学習と評価尺度

旅行ブログエントリの自動判定の機械学習にはSVMを用いた。2次の多項式カーネルを使用し，5分割交差検定を行った。評価尺度として，以下に示す精度・再現率を用いた。精度は検出誤りの少なさを表す評価指標，再現率は検出洩れの少なさを表す評価指標である。本研究では，手掛かり語を自動収集することで，再現率の向上を目指している。

$$\text{精度} = \frac{\text{システムが検出した正解件数}}{\text{システムが検出した件数}}$$

$$\text{再現率} = \frac{\text{システムが検出した正解件数}}{\text{人手で判定した正解件数}}$$

4.2 実験結果

実験結果を表3に示す。表3のベースラインの結果が，先行研究［2］の結果と値が異なるのは，先行研究では，旅行ブログエントリ判定のターゲットとなるブログエントリのみではなく，ターゲットの前後のブログエントリの情報も利用しているからである。

表3 旅行ブログエントリの自動判定結果

手法	精度	再現率
ベースライン	0.809	0.400
提案手法	0.794	0.568

表3より，ベースラインに比べ，提案手法では，再現率は0.015ポイントとわずかながら低下したが，精度は0.168ポイント改善することができた。

ベースラインでは，人手で収集した手掛かり語を使用しているため，手掛かり語の網羅性が低く，再現率が低いという問題があった。提案手法では，自動で収集した手掛かり語を利用しているため，人手で収集するよりも多くの手掛かり語を収集することができ，再現率を改善することに成功している。

よって，旅行ブログエントリのタイプ分類の結果を利用して自動収集した手掛かり語を，旅行ブログエントリの自動判定に利用する提案手法の有用性を確認することができた。

5. ま と め

本研究では，旅行ブログエントリを自動抽出する手法を提案した。提案手法は，以下の2つのステップに分かれている。

●Step1：ブログの収集
●Step2：旅行ブログエントリの自動判定

提案手法の有効性を確認するために，実験を

行った。実験の結果，精度0.794，再現率0.568という結果を得ることができ，自動収集した手掛かり語を機械学習に利用する提案手法の有効性を確認することができた。

本研究では，日本語で記述したブログエントリを対象としている。今後は，収集した手掛かり語を翻訳することで，多言語の旅行ブログエントリを収集する予定である。また，収集した旅行ブログエントリを利用し，旅行者へ観光情報を提示するシステムの構築を行う予定である。

謝辞：本研究は，JSPS 科研費 JP16K16679の助成を受けたものです。

注

1） http://www.mlit.go.jp/kankocho/siryou/toukei/in_out.html
2） http://www.mlit.go.jp/common/000146049.pdf
3） http://www.travelblog.org/
4） http://travel.blogmura.com/
5） http://blogs.yahoo.co.jp/
6） http://www.mlit.go.jp/singikai/unyusingikai/kankosin/kankosin39.html
7） http://mecab.sourceforge.net/

参 考 文 献

［1］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.
［2］石野亜耶，難波英嗣，竹澤寿幸，“旅行ブログエントリからの観光情報の自動抽出”，日本知能情報ファジィ学会誌，Vol. 22, No. 6, pp. 667–679, 2010.
［3］岡本昌之，菊池匡晃，“ブログからの地域イベント情報抽出”，情報処理，Vol. 51, No. 1, pp. 14–17, 2010.
［4］郡宏志，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.
［5］藤井一輝，難波英嗣，竹澤寿幸，石野亜耶，“旅行ブログエントリの属性情報に基づいた旅行者の行動分析”，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），2015.
［6］徳久雅人，奥村秀人，村田真樹，“観光開発支援のためのブログ記事からの評判分析”，観光と情報，Vol. 7, No. 1, pp. 85–98, 2011.
［7］新田崇人，難波英嗣，石野亜耶，竹澤寿幸，“外国人旅行者の行動分析および地域性の判定”，第８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 2016），2016.
［8］村上嘉代子，川村秀憲，“外国人から見た日本旅行―英語ブログからの観光イメージ分析―”，人工知能学会誌，Vol. 26, No. 3, pp. 286–293, 2011.
［9］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.
［10］Corinna Cortes, Vladimir Vapnik, “Support-Vector Networks”, Journal of Machine Learning, pp. 273–297, 1995.