社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

ブログからのユーザの 行動経路の自動抽出と可視化

小田原　周平　　石野　亜耶　　難波　英嗣　　竹澤　寿幸

広島市立大学大学院　情報科学研究科　〒731-3194 広島市安佐南区大塚東 3 丁目 4 番 1 号

E-mail: {odawara, ishino, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし　人間の行動経路の分析は，都市計画，建築計画，カーナビゲーション，観光行政，防犯，インフルエンザなどの感染経路の特定など，様々な分野において重要な課題として位置付けられている．本研究では，旅行者の行動経路，および災害時の被災者の避難経路をとりあげ，ブログを対象にしたこれらの自動分析を試みる．我々は，旅行者の行動経路の抽出として旅行ブログエントリ，災害時の被災者の避難経路の抽出としてマイクロブログ(Twitter)を利用し，これらのブログから自動的にユーザの行動経路の抽出を行う手法を提案する．本研究では，ブログデータベースの中からユーザの行動経路を自動抽出する実験を行い，提案手法の有効性を示した．

キーワード　ブログ，情報抽出，行動経路，Twitter

Automatic Extraction and Visualization of Transportation Information From Blogs

Shuhei ODAWARA　Aya ISHINO　Hidetsugu NANBA　and　Toshiyuki TAKEZAWA

Graduate School of Information Sciences, Hiroshima City University

3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

E-mail: {odawara, ishino, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

Abstract　The analysis of human's transportation information is considered as an important issue in various fields, such as city planning, architectural planning, car navigation, sightseeing administration, crime prevention, and specification of infection route of the epidemic. In this paper, we focus on travelers' transportation information and victim's escape route at the disaster, and attempt to extract them from blogs. We use the travel blog entries as an information source for the extraction of traveler's transportation information, and micro blog (Twitter) for the extraction of victim's escape route at a disaster. We propose a method for extracting user's transportation information from these blogs, and experimentally confirmed the effectiveness the method.

Keyword　Blog, Information Extraction, Transportation Information, Twitter

1. はじめに

人間の行動経路の分析は，都市計画，建築計画，カーナビゲーション，観光行政，防犯，インフルエンザなどの感染経路の特定など，様々な分野において重要な課題として位置付けられている．このような分析には，目的に応じて多種多様なデータが利用されるのが一般的である．本研究では，旅行者の行動経路，および災害時の被災者の避難経路をとりあげ，ブログを対象にしたこれらの自動分析を試みる．

まず，旅行者の行動経路の抽出として，ブロガーが日記形式で綴った旅行記である旅行ブログエントリに焦点を当てる[1]. 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，日本では今，「観光」を 21 世紀の基幹産業と位置付けた多様な取り組みが，国や地方公共団体，民間で積極的に推進されている．観光を支援する媒体としては地方公共団体や旅行会社などが運営する観光ポータルサイトや，旅行情報雑誌「るるぶ」などの観光情報データベースが既にいくつか作成されており，Web 上で公開されているものも少なくない．これらの媒体は，旅行者が，ある観光地の観光スポットや名所をどのような順序で訪れるかという計

画を立てる際に利用する機会が多い．しかし，観光スポットや名所をどのような順序で訪れれば良いのかは，旅行者が観光地に訪れる時期，旅行者の年齢，旅行者が利用可能な交通手段（例えば，運転免許を持っていれば，自動車やバイクでの移動が可能）など，様々な条件に依存するため，上述の情報源で紹介されている経路が旅行者本人にとって最適なものであるという保証はない．また，このような旅行者の様々な条件を考慮した経路情報を人手で作成・提供するには多大なコストを要するという問題がある．

そこで本研究では，旅行記が記述されているブログエントリ（以下，旅行ブログエントリ）から旅行者の行動経路を自動的に抽出することで低コストでのデータベース生成を目指す．近年，ブログの著者の属性（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研究が進んでいるが[2，3，4]，これらの技術と本研究の成果を組み合わせることにより，旅行者に最適な経路を自動的に提案することが可能になると考えられる．また，網羅性の高さや最新の観光情報を素早く獲得できる点などで既存のデータベースよりも有用なものになることが期待できる．

次に，災害時の被災者の避難経路の抽出としてマイクロブログ(Twitter)に焦点を当てる．現在，東北関東大震災に関して様々な情報が飛び交っている．その中でも，特に被災された方々の避難経路情報は重要な情報であり，できるだけ正確な情報を大量に整理する必要がある．しかしながら実際は，情報は大量でかつ様々なところに分散して存在しているために探しにくいだけでなく，情報を提供する側と情報を必要とする側で適切に必要な情報を共有できていないと考えられる．

そこで，現在では ANPI NLP[1]という自然言語処理研究者を中心としたプロジェクトが開始されており，マイクロブログ，特に近年注目を集めている Twitter などから個々に述べられている人の安否情報に注目し，Google 社の‘Google Person Finder’上のデータと照合しながら，最新の安否情報をできるだけ整理する事を目的として活動を行っている．本研究では，ANPI NLP で提供されている震災情報に関連する Twitter のデータから避難経路の抽出を行うことにより，被災者の避難経路情報を整理し，避難経路に関するデータベースを作成する． 避難経路の抽出を行う事によって，避難に困っている被災者や，車などで輸送物資を配送する援助者に，有益な情報を提供できることが期待できる．

最終的には，得られた行動経路を地図上にマッピングする事により，集約してユーザに提示し，閲覧できるようなシステムの開発を目指す．

本論文の構成は以下のとおりである．2 節では関連研究について述べる．3 節では提案手法，4 節では提案手法の有効性を調べるために旅行ブログエントリを用いて行った実験，5 節では 3 節で構築したモデルを適用し，Twitter からの避難経路の自動抽出について述べる．また結論については 6 節で述べる．

2. 関連研究

本節では，「地理情報検索」，「行動経路の抽出」，に関する研究について述べる．

2.1. 地理情報検索

近年，地理情報検索に関する様々な研究が行われている．Web を用いた地理情報検索に関連する研究について述べる．本研究と同様に，Web から地域情報を自動収集しようとする研究がある．大槻ら[5]は，地域情報ウェブディレクトリを自動編集するシステムを提案している．地域情報ウェブディレクトリは地域情報検索に利用される．

本研究では，地域情報を収集するにあたって，旅行ブログエントリを対象としているが，大槻らは，自治体が提供する地域情報サイトと，そのリンク先の地域サイトを対象としている点で異なる．

本研究と同様に，ブログを情報源とし，地域情報を自動抽出する研究がある．岡本ら[6]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．また，藤坂ら[7]は，Twitter に代表されるマイクロブログから地域イベントを発見し，その特性を検証するためのシステムを提案している．しかし，ブログの中には観光と関係ないものも存在するため，ブログエントリを全て使うと，十分な精度で観光情報が抽出できない可能性がある．本研究では，まず，ブログ集合から旅行ブログエントリを自動検出し，次に，そこから観光情報の抽出をすることにより，高い精度での抽出を目指す．

また，マイクロブログから地震の発生地を自動抽出する研究がある．坂茂ら[8]は，Twitter のツイートを観測することによって，即座に地震発生地の検出を行う手法を提案している．坂茂らはこの研究の有効性を示し，地震発生地検出アプリケーションとして，地震リポーティングシステムを開発している．坂茂らは，地震の発生地の抽出を行うシステムの開発を行っており，本研究での避難経路の抽出とは異なる．

[1] http://trans-aid.jp/ANPI_NLP/index.php/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8

2.2. 行動経路の抽出

近年，Web やブログで自らの行動を日記として発信することが盛んになってきている．郡ら[9]は，ブログ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．システムとしてはまず，ブログ内に現れる各地名が実際にビジターがその場所を訪れたという文脈で使用されているかを判定し，訪れていると判定された場合はその地名を破棄するという地名フィルタを作成する．次に，作成した地名フィルタにより取得した各ブログエントリの「訪れた地名」に順序づけを行い，地名によるシーケンシャルパターンを生成し経路抽出を行っている．しかし，郡らは，研究対象を京都に関するブログに限定としているため，本研究では経路情報を収集するにあたって，全国の旅行ブログエントリを対象としている点で異なる．

また，Davidov ら[10]は，Web から交通手段や経路の地理的なネットワークを見つける手法を提案している．Davidov らは，種となる特定の地域を表す項の小さな集合が与えられると,その地域の場所の名前とともに，連結性と交通に基づいたグラフを発見するアルゴリズムを提案している．このアルゴリズムでは，‘[Transport] from A to B’のような‘A から B へ’という表層パターンで経路抽出を行っている．Davidov らは‘from A to B’という限られたパターンで抽出をおこなっているが，本研究では機械学習を利用して行動経路抽出を行うため，より多くのパターンでの行動経路抽出が可能であると考えられる．

3. 旅行ブログからの行動経路抽出

本研究では，旅行ブログエントリから行動経路を抽出するシステムを作成する．そして，作成したモデルを Twitter にも適用させ，ANPI NLP で提供されている震災情報に関連する Twitter のデータから避難経路を抽出する．本節では，旅行ブログエントリから旅行者の行動経路を自動抽出する手法について説明を行う．

旅行ブログエントリには，移動した経路，その際に利用した移動手段，移動時間の情報が多く書かれている．そこで，ユーザの行動経路のほか，利用した手段や移動にかかった時間を抽出するために，移動経路に関する 5 種類のタグを定義することとした．また，これらのタグを旅行ブログエントリに付与した例を以下の図 1 に示す．

- ○ from：移動元
- ○ to：移動先
- ○ via：経由地，経由道路など
- ○ method：移動手段
- ○ time：移動に要した時間

<from>広島</from>から<to>大阪<to>まで<time>5 時間前後</time>かけて，<method>バス</method>で行った．

図 1：旅行ブログエントリにタグを付与した例

本研究では機械学習として CRF を使用した．CRF 基本手法は与えられた文に含まれる語を分類するのに使用した．素性とタグは以下のように CRF に与える．

(1) ターゲットとなる単語から，CRF に与える前後の単語数 k
(2) ターゲットとなる単語の前に存在する，ターゲットからの距離が k 以内に現れる単語
(3) ターゲットとなる単語の後に存在する，ターゲットからの距離が k 以内に現れる単語

我々は予備実験の結果から，k=4 と定めた．また，機械学習には以下の素性を使用した．

- ○ 単語
- ○ 品詞
- ○ 括弧（「」,『』など）
- ○ from：‘から’,‘を出発’など from の手掛かり語かどうか（40 語）
- ○ from_to1：‘博物館’,‘温泉’など観光名所の後に付きそうな単語（45 語）
- ○ from_to2：‘観光’,‘駅’など行き先の後に付きそうな単語（11 語）
- ○ from_to3：観光名所リスト（13,779 件）
- ○ from_to4：駅名・道の駅名・空港名（9,437 件）
- ○ to：‘まで’,‘に到着’など to の手掛かり語かどうか（271 語）
- ○ via：‘経由’,‘を通って’など via の手掛かり語かどうか（43 語）
- ○ via2：高速道路（101 件）
- ○ method：‘飛行機’,‘自動車’など method の手掛かり語かどうか（148 語）
- ○ method2：乗り物リスト(128 件)
- ○ method3：列車，高速バスの愛称・路線名（2,033 件）
- ○ time：‘分’,‘時間’など time の手掛かり語かどうか（77 語）

4. 実験

本研究では旅行ブログエントリからのユーザの行動経路の自動抽出に関する実験を行った．

実験に用いるデータ

旅行ブログエントリであると判定されたエントリから約 10,000 文(193 エントリ)を抽出し, from, to，via，method，time のタグを人手で付与したデータを，機械学習に用いた．人手で付与したタグの数を表 1 に示す.

表 1：人手で付与したタグの数

	訓練用（件）	テスト用（件）
from	136	30
to	384	126
via	58	15
method	245	55
time	87	27

機械学習と評価尺度

機械学習には CRF を用いた．また，精度と再現率を用いて評価を行った.

実験結果と考察

実験結果を表 2 に示す．表 2 より，精度において高い数値を記録することができた．‘via’の精度が低いのは，訓練用データに 58 件，テスト用データに 15 件しか存在せず，件数が少ないためである.

表 2：行動経路の抽出の実験結果

	精度	再現率
from	75.0	30.0
to	75.0	45.2
via	55.6	33.3
method	94.9	66.0
time	87.6	50.0
平均	80.3	46.8

人手では‘タグをつけていない’が，システムでは‘タグをつけた’ものが 26 件存在した．この検出誤りの主要な原因を以下に示し，説明を行う.

(1) 手掛かり語の曖昧性（69.6%）
(2) 実際には移動していないもの（17.4%）
(3) その他（13.0%）

以下に，それぞれの検出誤りについて説明する.

(1) 手掛かり語の曖昧性（69.6%）

26 件のうち 16 件(69.6%)が‘to’や‘via’では地名や観光名所，経由した道路，‘time’では時間になっておらず，それぞれのタグの定義に合っていないものであった．抽出の失敗例を表 3 に示す.

表 3：抽出誤りの例

via	〈via〉**結局お昼**〈/via〉を過ぎても動かなかったのか.

表 3 では，「結局お昼」など場所を表す単語ではないものが抽出されている．これは後にある「を過ぎて」という手掛かり語が素性として入っているため誤って抽出された.「お昼」などの不要語リストを作成し素性に入れることで解消できると思われる.

(2) 実際には移動していないもの（17.4%）

26 件のうち 4 件(17.4%)が‘from’や‘to’や‘method’の定義に当てはまっている単語を抽出しているが，実際には移動しておらず，実際には訪れていない場所や利用していない手段であった．抽出の失敗例を表 4 に示す.

表 4：抽出誤りの例

from	<from>**熱海**</from>から快速のグリーン車でゆっくり帰って寝ていくことも考えましたが，結局は<from>三島</from>から<method>新幹線</method>に乗ってました.
to	2008 年 1 月の相方の誕生日に<to>**平湯温泉**</to>へ行ったときに約束したもの

表 4 のように‘こう考えたが，違うルートにした’という表現や,「行ったときに約束した」など，実際にそのときに移動していないものも抽出されている．これは提案手法が from では「から」, to では「へ行った」などの手掛かり語のタグ周辺に出現する表現，または「平湯温泉」など，タグの中の表現にしか注目していないためと考えられる．この抽出誤りは，タグ周辺の表現ではなく，タグ周辺よりも後の文脈の表現にも注目することで解決できると思われる.

人手では‘タグをつけた’が，システムでは‘タグをつけていない’ものが 110 件存在した． この検出誤りの主要な原因を以下に示し，説明を行う.

(1)タグ周辺に出現する手掛かり語が短いもの（59.1%）
(2)手掛かり語の不足（17.3%）
(3)その他（23.6%）

以下に，それぞれの検出誤りについて説明する.

(1) タグ周辺に出現する手掛かり語や，一文が短いもの（59.1%）

98 件のうち 58 件（59.1%）がタグ周辺に出現する手掛かり語が形態素解析されたときに 1 単語のみの手掛かり語であったり，一文が短いために手掛かり語がないものであった．システムで再現できなかったタグの例を表 5 に示す.

表 5：再現できなかったタグの例

to	岩国錦帯橋へ。
method	<time>１１：４４</time>発 あさま５２１号

表 5 において，下線部分が再現できなかったタグの部分である．この例のように，to では「へ。」という短い手掛かり語で素性が立った際にあまり特徴がなかったり，タグを再現したい行にタグの部分しか記述されておらず，タグ部分の周辺に手掛かり語が存在しないためタグを付けなかった．このようなものを抽出するには，1 文だけに注目するのではなく，周辺の文にも焦点を当て，文脈を考慮することにより，解決できると思われる．

(2) 手掛かり語の不足（17.3%）

98 件のうち 17 件（17.3%）がタグ周辺に出現する表現，またはタグの中の文字列が手掛かり語に含まれていないものであった．システムで再現できなかったタグの例を表 6 に示す．

表 6：再現できなかったタグの例

from	<method>新幹線</method>に乗る前に駅弁買い込んで東京駅**に別れを告げる**.
to	**さらに進んで，**大黒寺

表 6 において，下線部分が再現できなかったタグの部分，太字部分が不足していたと思われる手掛かり語である．from では「に別れを告げる」，to では「さらに進んで,」など，あまり出現しないめずらしい表現で手掛かり語には含まれていなかったため，システムで再現できなかったと思われる．これは，それらの単語を手掛かり語としてそれぞれの素性に入れることで解決できると思われる．

5. Twitter からの避難経路の自動抽出

4 節で作成したモデルを使用し，Twitter からの避難経路の自動抽出を行う．本研究では，ANPI NLP で提供されている震災情報に関連する Twitter のデータ 49,263 件を使用した．タグの自動付与を行った例を以下の図 2 に示す．

2011-0316-20:12:33 Ramah2010m RT@kei4nakamura：【安否情報求む】谷田部覚心の情報ありませんか？13 日の夜に<from>河北町</from>から<method>徒歩</method>で<to>湊</to>に向かいましたが，その後の消息が分かりません．#ishinomaki#anpi

図 2：Twitter のタグの自動付与の例

本研究ではタグを付与した後，以下の 4 パターンを避難経路として自動抽出する．

(1) <from>A</from>と直近の<to>B</to>を‘→’で結合
(2) <from>A</from>と直近の<via>B</via>を‘→’で結合
(3) <to>A</to>と直近の<to>B</to>を‘→’で結合
(4) <via>A</via>と直近の<via>B</via>を‘→’で結合

例えば，図 2 の例では，(1)のパターンが適用され，避難経路は‘河北町→湊’となる．

実験結果と考察

抽出結果の例を以下の表 7 に示す．

表 7：抽出した避難経路の例

ツイートした時間	ユーザ名	避難経路
2011-0313-00:06:19	youigarasi	石巻→一関
2011-0314-07:34:08	toshix141	池袋→練馬高野台
2011-0314-18:33:57	”tasukusuta”	宮城→山形
2011-0314-21:31:15	kyo0v0	八戸→陸前高田
2011-0315-09:00:20	edisymyb	仙台→山形

避難経路を抽出した数は 49，2363 件中 50 件という結果となった．抽出数が少なかった主要な原因を以下に示し，説明を行う．

(1) 震災情報に関連する手掛かり語の不足
(2) Twitter と旅行ブログエントリとの違い

(1) 震災情報に関連する手掛かり語の不足

本研究の評価実験では手掛かり語を，旅行ブログエントリをもとに作成した．そのため，避難経路抽出の際，震災情報に関連する語を手掛かり語としなかった事が原因と考えられる．タグ付けを行わなかった例を以下の図 3 に示す．

0313202604 JKS_KIK RT@radio_rfc_japan：安否確認情報．富岡町東洋愛セイエンの父兄の皆さんへ，園のみんなは三春町**に避難**しています．三春で偶然会いました先生方のおかげでみんな元気出した．田村市の園生の父兄より##life_anpi

図 3：Twitter でタグの自動付与を行わなかった例

図 3 において，下線部が再現できなかったタグの部分，太字部分が不足していたと思われる手掛かり語である．ここでは震災情報に関連する「に避難」が手掛

かり語に含まれていなかったため，システムで再現できなかったと思われる．これは，震災情報に関連する語を手掛かり語として素性に入れる事で解消できると思われる．

(2) Twitter と旅行ブログエントリとの違い

Twitter は 1 件のツイートの字数制限が 140 字以内と制限されているため，1 件あたりの情報量が少ない．そのため今回使用した ANPI NLP で提供されている震災情報に関連する Twitter のデータは，旅行記を綴った長い文章で書かれた旅行ブログエントリよりも情報量が少ないため，手掛かり語があまり含まれていないと考えられる．

今回使用した Twitter のデータはユーザ毎や時間順などで整理されていなかったため，同じユーザの複数のツイートを，1 件のツイートとして時間の経過ごとに結合させ，1 件あたりの情報量の増加をはかった．結合させたデータを使用して避難経路の自動抽出を行ったところ，避難経路が 16 件から 50 件まで増加した．しかしながら抽出数としてはまだ少ないため，Twitter と旅行ブログエントリとの文量以外での違いを調査し，解決方法を見つける必要があると考えられる．

6. 結論

本研究では，旅行ブログエントリから行動経路を抽出するシステムを作成し，評価実験を行った．そして，作成したモデルを Twitter にも適用させ，震災情報に関連する Twitter のデータからの避難経路の自動抽出を行った．評価実験では，実験の結果，精度 80.3%，再現率 46.8%となり，高い精度が得られた．また，作成したモデルを Twitter に適用させて避難経路の自動抽出を行った結果，50 件の避難経路が得られた．今後は，得られたデータをどのようにユーザに提示し，閲覧できるようにするのかを課題としたい．

文　献

[1] 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．知能と情報．Vol.22，No.6，pp.667-679，2010.

[2] Norihito Yasuda,Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki. Identifying Bloggers' Residential Areas. Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.231-236, 2006.

[3] Daisuke Ikeda, Hiroya Takamura, and Manabu Okumura. Semi-supervised Learning for Blog Classification. Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161, 2008.

[4] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. Effects of Age and Gender on Blogging. Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp.199-205, 2006.

[5] 大槻 洋輔，佐藤 理史．地域情報ウェブディレクトリの自動編集．情報処理学会論文誌，Vol.42，No.9，pp.2310-2318，2001.

[6] 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.

[7] 藤坂 達也，李 龍，角谷 和俊．地域イベント発見および特性検証のための実空間マイクロブログを用いたユーザ移動パターン分析システム．情報処理学会創立 50 周年記念(第 72 回)全国大会，pp.845-846，2010.

[8] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo. Earthquake Shakes Twitter Users:Real-time Event Detection by Social Sensors. 18th International World Wide Web Conference (WWW2010), 2010.

[9] 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.

[10] Dmitry Davidov, Ari Rappoport. Geomining : Discovery of Road and Transport Networks Using Directional Patterns. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.267-175, 2009.