

マイクロブログのジオタグと 発言コンテキスト解析による行動予測手法

齊藤 裕樹^{1,a)} 高山 翼^{2,†1} 山上 慶^{3,†2} 戸辺 義人⁴ 鉄谷 信二²

受付日 2013年5月14日, 採録日 2013年10月9日

概要: GPS 機能を備えた携帯端末などによる位置情報取得技術の普及により, 人々の訪れた場所どうしを結ぶ行動履歴を解析する研究が活発に行われている. 経路情報の蓄積によって得られる行動履歴は, 人々の興味と移動との関係を解析する手段として注目されている. 一方, Twitter をはじめとするマイクロブログはテキストとともに GPS センサによる位置情報を付与し, 公開することが可能なサービスとして世界的に利用されている. 本論文では, マイクロブログ上に蓄積された人々の行動履歴を基に利用者の行動とコンテキストを解析し確率過程モデルに適用させることで, 将来の利用者の行動を予測する手法の提案を行う. また, マイクロブログの実データを用いた行動予測精度の評価実験から, 単純統計を用いた従来手法より高い予測精度が得られることを確認した.

キーワード: マイクロブログ, コンテキスト解析, 位置情報サービス, 行動解析, Web マイニング

People Behavioral Prediction by Mining Geo-tags and Its Contexts from Micro-blog Services

HIROKI SAITO^{1,a)} TSUBASA TAKAYAMA^{2,†1} KEI YAMAGAMI^{3,†2} YOSHITO TOBE⁴
NOBUJI TETSUTANI²

Received: May 14, 2013, Accepted: October 9, 2013

Abstract: The advance of GPS-enabled portable devices such as PDAs and smart phones facilitates people to record their location histories. Location trajectories imply human behaviors and preferences related for their interests. On the other hand, microblog services such as Twitter enable us to publish text messages (e.g., Tweets) and location-tags (e.g., Geo-tags) to subscribers. This paper proposes a schema for predicting user behavior by analyzing location trajectories and contexts by applying a stochastic model. And, we confirm the effectiveness of our schema through experiment using the actual data obtained from microblog service.

Keywords: micro-blogs, context analysis, location-based services, behavioral analysis, Web mining

¹ 明治大学総合数理学部先端メディアサイエンス学科
Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164-8525, Japan

² 東京電機大学未来科学部情報メディア学科
Department of Information Systems and Multimedia Design, School of Science and Technology for Future Life, Tokyo Denki University, Adachi, Tokyo 120-8551, Japan

³ 東京電機大学大学院未来科学研究科情報メディア専攻
Graduate School of Science and Technology for Future Life, Tokyo Denki University, Adachi, Tokyo 120-8551, Japan

⁴ 青山学院大学理工学部情報テクノロジー学科
Department of Integrated Information Technology, School of Science and Technology, Aoyama Gakuin University, Sagami-hara, Kanagawa 252-5258, Japan

1. はじめに

GPS 機能を備えた携帯端末などによる位置取得技術の普及により, 人々が訪れる場所どうしを結ぶ行動履歴を解析する研究が活発に行われている [1], [2], [3], [4]. 経路情報を蓄積することによって得られる行動履歴の解析は,

^{†1} 現在, 株式会社富士通システムズ・イースト
Presently with Fujitsu Systems East Limited

^{†2} 現在, 株式会社ジェイアール東日本情報システム
Presently with JR East Japan Information Systems Company

^{a)} hrksaito@meiji.ac.jp

人々の興味と移動との関係を抽出する手段として注目が集まっている。一方、Twitterをはじめとするマイクロブログサービスは、テキストとともに GPS センサによる位置情報を付与し公開することが可能なサービスとして世界的に利用されている。

本論文では、マイクロブログサービスで発信されるテキストと位置情報から、テキストと移動の関係を確率過程モデルに適用し行動解析を行う。また、解析により得られた情報を蓄積し行動モデル化を行い、利用者の未来の行動を予測する手法を提案をする。提案手法の有効性を示すために Twitter 上の実データを用いた行動予測精度の評価実験の結果、単純な行動履歴から求めたものよりも予測精度が高いことを示す。

本論文の構成は次のとおりである。1 章に引き続き、2 章でマイクロブログサービスを用いた行動予測の概要について示す。次に 3 章で提案の中核となる確率過程モデルへの適用、移動確率計算手法、行動予測手法について述べる。4 章では、提案手法の有効性を確認するために行ったマイクロブログサービスの実データを用いた評価実験とその結果について述べる。さらに 5 章では、関連研究と本研究の位置づけを示し、6 章では、本論文の内容をまとめ結論づける。

2. マイクロブログサービスを用いた行動予測

GPS 機能を備えた携帯端末の普及により、多くの位置情報サービス (Location Based Service: LBS) が iPhone や Android といった様々なプラットフォーム上で実現されている。初期の位置情報サービスでは、地図上でのナビゲーションを行うことや、位置を指定した最近傍の施設やサービスの検索などが提供されたが、位置情報は利用者の携帯端末内で用いられるのみで、他の利用者と共有は行われていなかった。しかし近年、位置情報サービスでは、人やものの位置情報を蓄積し利用者間で情報共有することで、位置情報に新たな価値を生み出す動きが活発に行われている。また、GPS によって取得された人々が訪れた場所どうしを結んだ行動履歴を蓄積することで、街中の人々の動線を解析することやナビゲーションシステムなどへの応用が期待されている。

一方、Twitter をはじめとするマイクロブログサービスは、短いテキストを発信するサービスであり、その場の状況に関するメッセージを即座に発信できるという特徴がある。特に、スマートフォンなどの携帯端末上で用いられることにより、即時性と臨場感の高い情報発信が可能である。また、テキストを発信する際に GPS による位置情報を付加し、実世界と直接リンクした情報を扱うことも可能である。このように、人がセンサとなり現在の状況や行動に関するテキストを位置情報とともに情報発信すること可能である。これにより、人々の日常活動や目標指向の行動を知

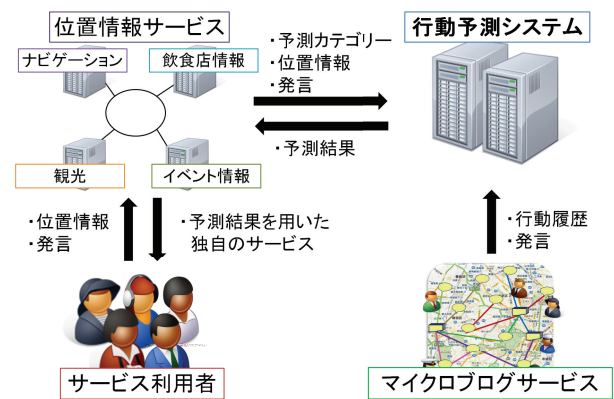


図 1 マイクロブログの発言と位置情報を用いた行動予測システム
Fig. 1 People behavioral predicting system among text messages and location histories in micro-blog services.

ることが可能になることから、新たな価値のある情報を作り出す研究が活発である [5], [6].

本研究では、まず人々がマイクロブログサービスに発信したメッセージと位置情報から、発言コンテキストと移動経路を抽出し行動履歴として蓄積を行う。蓄積した行動履歴を基に利用者のコンテキストから未来の行動を予測する手法を検討する。具体的には、人の行動において、移動先は移動元の情報のみに影響されるという仮定に基づき、人の行動にマルコフ連鎖を適用し行動モデルを作成する。また、人々がマイクロブログ上で発信するメッセージと行動には一定の関係があるとの仮定に基づき、メッセージと行動の組合せを条件付き確率による定式化を行う。これらにより、利用者の未来の行動予測を行う。次に、予測の結果から利用者に次の行動に対する適切な情報提示を行う推薦システムの提案を行う。図 1 に提案システムのシステムモデルを示す。本システムは、2つの機能を核として設計する。1つ目は、行動履歴をマイクロブログサービスから取得し、行動解析を行い蓄積する機能である。2つ目は、利用者の発言と位置情報から行動予測を行う機能である。

3. 確率過程モデルを用いた行動予測手法

本章では、行動推定の基本的方針を述べ、人の行動に対して確率過程モデルの適用を行い、移動確率の計算手法と行動予測手法について述べる。

3.1 行動予測の基本方針

マイクロブログは、その場その瞬間に見たことや感じたことを気軽に短いテキストとして発信できることから、利用者の興味関心を反映するメディアであるといえる。また、マイクロブログの発言に付与された位置情報により、利用者が興味関心を持った場所を知ることができる。これらにより、利用者の位置情報を時間軸上に並べることによって抽出した移動経路は、単なる移動軌跡だけではなく、その場その状況を反映するコメントや写真をともなうため、位

置情報に関する豊富なコンテキスト情報を含むこととなる。以上のように、移動中に行った発言を収集することは、人間の行動とその意味や目的を集合知として集約するものと考えられる。

たとえば、鎌倉を観光で訪れた人々の場合、鶴岡八幡宮や建長寺、高德院などの位置情報とともに、鶴岡八幡宮でお参りしたことや、高德院で鎌倉大仏を見たこと、建長寺の半僧坊からの景色に関する発言をした可能性が高いと考えられる。このような行動履歴を多くの人から集約することにより、移動経路とその移動の意味を知ることが可能であると考える。

そこで、本論文では、過去に蓄積された行動履歴に対して、確率過程モデルを適用することにより行動解析を行い、移動経路と発言を考慮した行動モデルを作成する。また、作成した行動モデルを用いて未来の利用者の行動を予測する手法を提案する。提案手法は、具体的には以下のような方針に基づく。ある移動経路を通った人々のうち特定の発言をした人の割合を、条件付き確率で計算すると、移動経路に対して関係の強いキーワードが高い割合を占めることとなる。これより、発言から未来の移動経路の確率を求める。たとえば、過去に建長寺から若宮大路を経て鶴岡八幡宮に行くルートを通った人は、高い割合で歴史的な建物や風情や景色に関する発言をしたと考えられる。このような割合を基に、現在建長寺で風情や景色を楽しんでいる発言をする人に対して、予測される次の目的地である鶴岡八幡宮の観光情報を提示することや目的地へのナビゲーションを行う。

以下、確率過程を用いた行動のモデル化および行動予測の手法について述べる。

3.2 人の行動の確率過程モデルへの適用

利用者はGPS機能を備えた携帯端末を持ち、マイクロブログサービスに対して発言とともに現在地の緯度、経度の情報を送信する。発言のテキスト情報は単語ごとに分解し重み付けしベクトル空間上に表し、タイムスタンプの組合せ $(T_k, time_k)$ を記録する。また、「現在地の緯度・経度・タイムスタンプ」の組合せ $(L_k, time_k)$ を記録する。これらを蓄積しユーザごとに発言コンテキスト T の時系列 $T_1, T_2, \dots, T_n$ および、移動にともなう位置 L の時系列 $L_1, L_2, \dots, L_n$ を保持する。

人の行動において、位置 L_k から位置 L_{k+1} への移動は、蓄積された過去の移動情報のみに依存していると仮定すると、2地点間の移動確率はマルコフ過程によって表すことが可能である。3地点間以上の移動も2地点間の移動の連なりとして表し、多地点の行動もマルコフ連鎖により表すことが可能である。したがって、 N 地点間の移動確率 $P(L_1, \dots, L_N)$ は式 (1) で得ることができる。

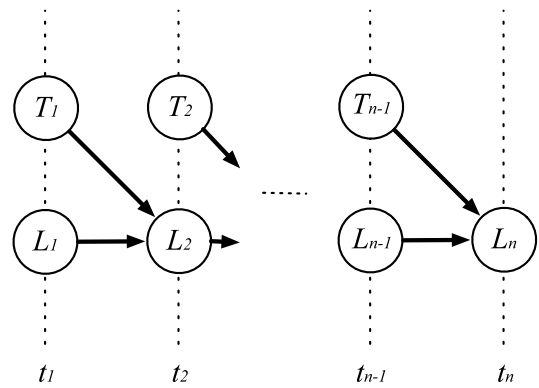


図 2 移動と発言コンテキストの状態遷移モデル

Fig. 2 State transition model based on movements and contexts.

$$P(L_1, \dots, L_N) = P(L_1) \prod_{n=2}^N P(L_n | L_{n-1}) \quad (1)$$

次に、位置 L_k で発生した発言コンテキスト T_k について考える。図 2 は、滞地点を 1 つの状態と見なし移動を状態遷移として表した図である。位置 L_k で発生した発言コンテキスト T_k は、位置 L_{k+1} への移動に関係性があり、移動確率に影響すると仮定する。これにより、 N 地点間の位置 L と発言コンテキスト T の同時分布確率は式 (2) で得ることができる。

$$P(L_1, \dots, L_N) = P(L_1) \prod_{n=2}^N P(L_n | L_{n-1}, T_{n-1}) \quad (2)$$

3.3 条件付き確率を用いた 2 地点間移動確率の計算

次に、これまでの行動履歴が既知であり、現在地から次の移動先への移動確率を求める方法を検討する。前節で得られた式 (2) を漸化式で表すと式 (3) のようになる。

$$\begin{aligned} P(L_1, \dots, L_N, T_1, \dots, T_N) \\ = P(L_1, \dots, L_{N-1}, T_1, \dots, T_{N-1}) P(L_N | L_{N-1}, T_{N-1}) \end{aligned} \quad (3)$$

前述の図 2 の位置 L_k の発言コンテキスト T_k と位置 L_{k+1} の関係から、これまでの行動履歴が既知であり、位置 L_k 上に存在する利用者の位置 L_{k+1} への移動確率は、 L_n と T_n が独立事象であると仮定すると式 (4) によって得られる。

$$P(L_{k+1} | L_k, T_k) = P(L_{k+1} | L_k) P(L_{k+1} | T_k) \quad (4)$$

式 (4) のうち、位置 L_k を条件とした位置 L_{k+1} への移動確率 $P(L_{k+1} | L_k)$ と、発言コンテキスト T_k を条件とした位置 L_{k+1} への移動確率 $P(L_{k+1} | T_k)$ をそれぞれ求める方法を以下に示す。

まず、位置 L_k を条件とした位置 L_{k+1} への移動確率は、位置 L_{k+1} が位置 L_k の状態に依存するため、式 (5) のとおりベイズの定理によって求めることができる。

$$P(L_{k+1}|L_k) = \frac{P(L_{k+1})P(L_k|L_{k+1})}{P(L_k)} \quad (5)$$

次に、発言コンテキスト T_k は L_k と同様にベイズの定理を用い式 (6) として

$$P(L_{k+1}|T_k) = \frac{P(L_{k+1})P(T_k|L_{k+1})}{P(T_k)} \quad (6)$$

のように求められるが、実際には発言は複数の単語の集合 $message_k = \{word_1, \dots, word_n\}$ で構成される。本研究では、複数の単語の生起確率は独立したものとし bag-of-words の手法 [7] をとる。よって、単純ベイズ確率モデルにより式 (6) の $P(T_k|L_{k+1})$ は以下のように計算できる。

$$P(T_k|L_{k+1}) = P(word_1 \wedge \dots \wedge word_n|L_{k+1}) \\ = \prod_i P(word_i|L_{k+1}) \quad (7)$$

以上の計算手法は、移動直前の 1 地点の情報から移動確率を求めるものであるが、高階マルコフ連鎖の考え方をを用い N 地点前の情報を用いるように拡張が可能である。高階マルコフ連鎖は、複数の状態履歴に依存して決まる確率過程であり、複数の状態履歴の組を 1 つの状態空間とすることにより、単純マルコフ連鎖と同様に表現することが可能である。すなわち、式 (4) を拡張し、式 (8) のとおり N 個の状態組を条件に用いることにより N 地点前の情報から移動確率を求めることができる。

$$P(L_{k+1}|L_{k-N+1}, L_{k-N+2}, \dots, L_k, T_{k-N+1}, T_{k-N+2}, \dots, T_k) \\ = P(L_{k+1}|L_{k-N+1}, L_{k-N+2}, \dots, L_k) \\ P(L_{k+1}|T_{k-N+1}, T_{k-N+2}, \dots, T_k) \quad (8)$$

3.4 人の移動先の予測手法

前節の移動確率計算を行い、行動履歴 $\{L_k, T_k\}$ から最も移動する可能性の高い地点を選択する。そのために、すべての位置 $\mathbf{L} = \{l_1, \dots, l_n\}$ についての移動確率を計算し、式 (9) に示す移動確率行列 $M_{L_k \rightarrow L_{k+1}}$ を生成する。

$$M_{L_k \rightarrow L_{k+1}} = \begin{pmatrix} P(l_1|L_k, T_k) \\ P(l_2|L_k, T_k) \\ \vdots \\ P(l_n|L_k, T_k) \end{pmatrix} \quad (9)$$

式 (9) より位置 $\mathbf{L} = \{l_1, \dots, l_n\}$ への移動確率を得たうえで、最も移動確率の高い位置を選択し、予測結果とする。

また、複数地点の一連の移動を 1 つの状態と見なすことにより、経路と経路上の発言コンテキスト $\mathbf{L}, \mathbf{T} = \{L_1, \dots, L_k, T_1, \dots, T_k\}$ を用いた行動予測が考えられる。これを式 (9) に適用すると、移動確率行列 $M_{\{L_1, \dots, L_k\} \rightarrow L_{k+1}}$ は以下のとおりとなる。

$$M_{\{L_1, \dots, L_k\} \rightarrow L_{k+1}} = \begin{pmatrix} P(l_1|L_1, \dots, L_k, T_1, \dots, T_k) \\ P(l_2|L_1, \dots, L_k, T_1, \dots, T_k) \\ \vdots \\ P(l_n|L_1, \dots, L_k, T_1, \dots, T_k) \end{pmatrix} \quad (10)$$

同様にして、位置 $\mathbf{L} = \{l_1, \dots, l_n\}$ への移動確率のうち最も移動確率の高い位置が予測結果となる。

4. 評価

提案手法の有効性を示すために評価実験を行った。本実験では、マイクロログサービスより得た行動履歴を基に、従来手法と提案手法で予測した移動先について、正解である実際に利用者のとった行動と比較することで正答率を求める。

4.1 実験データセット

本実験では、マイクロログサービスである Twitter より、神奈川県鎌倉市内の 9km 四方の空間でアップロードされた位置情報付き発言の取得を行った。取得には Twitter の StreamingAPI を用いた。この API は、実世界の緯度、経度を指定することにより、指定領域内でアップロードされた発言を取得することができる。発言は、テキスト、位置情報、ユーザ ID、タイムスタンプの 4 つで構成される。本実験では、2012 年 7 月 6 日から 2013 年 1 月 25 日までに 56,583 件の発言を取得した。取得した発言を MeCab [8] を用いて形態素解析を行い、名詞と固有名詞を抽出し予測に用いる。本実験では、280,284 個の単語を抽出した。同一ユーザが 1 日で位置情報を付与した発言を複数行った場合、 k 回目の発言 T_k および発言された位置 L_k に対して、「移動元 L_{k-1} 、移動先 L_k 、発言 T_{k-1} 」を一組の情報として 1 回の移動があったと見なす。これらにより、延べ 13,941 ユーザから 43,210 件分の移動をとまう発言を抽出した。なお、本研究では、同一日に同一ユーザからの発言を移動と見なしているため、同一ユーザであっても異なる日に発言している場合は、別ユーザと見なすこととした。発言のあった位置を地図上に点画したものを図 3 に示す。

4.2 実験方法

実験対象である鎌倉市内 9km 四方の領域を格子状に分割し、同一格子状にある発言は同一の地点のものとして見なすようクラスタリングを行い、位置 L を設定する。本実験では、本実験では、1 つの格子サイズを 100m 四方から 1,000m 四方まで 100m 刻みの 10 種の領域のデータセットを作成した。1 領域あたりの移動をとまう発言の平均回数と、平均単語数を表 1 に示す。以下、1 つの格子領域の大きさを変化させたときの正答率の変化を確認する。

実験データセットからランダムに抽出した 90% のデータ

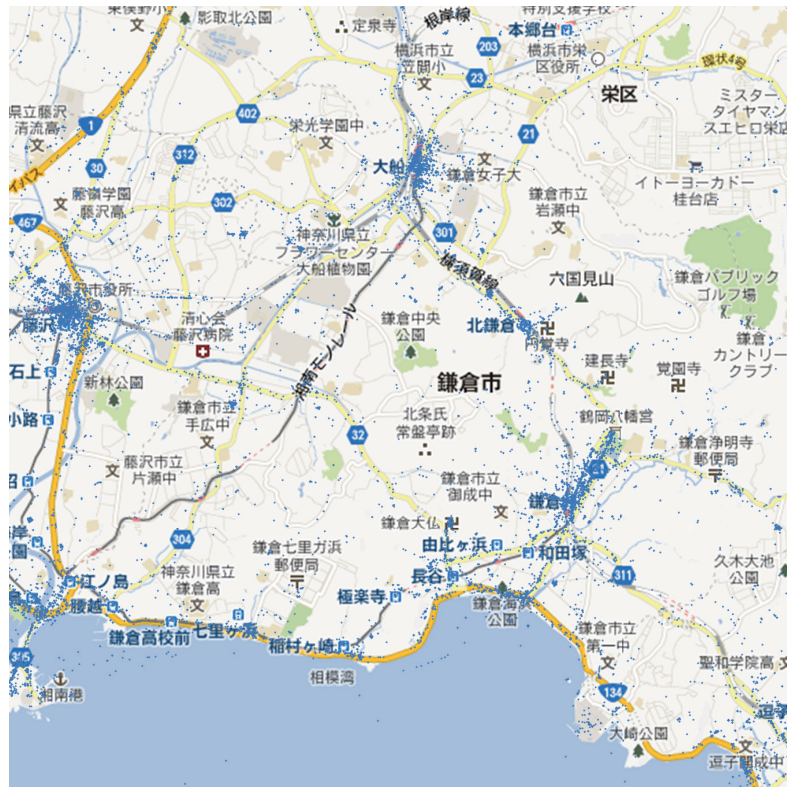


図 3 地図上にプロットしたジオタグ付き発言*1

Fig. 3 Map-plotted tweets with geo-tags*1.

表 1 格子領域サイズと取得されたデータ数
Table 1 Grid area size and acquired data.

1 領域あたりの面積 (m ²)	1 領域あたりの移動をともなう発言の平均回数	平均単語数
1×10^4	3.36	15.92
4×10^4	13.43	63.69
9×10^4	30.21	143.30
16×10^4	54.41	254.56
25×10^4	83.91	398.07
36×10^4	120.83	573.22
49×10^4	175.14	818.67
64×10^4	217.64	1,018.25
81×10^4	271.87	1,289.74
100×10^4	335.64	1,592.27

を学習データとし、残り 10% のデータは正解データとし評価を行う。正答判定は、予測結果が正解データの移動先終点と一致していたか否かによって行うこととし、正答率の比較には、以下の従来手法と 2 通りの提案手法の結果を用いた。

- 従来手法

発言を考慮しない単純統計による予測結果を求める。この手法では、ある位置 L_k からみた次の移動先 L_{k+1} を、過去に人々がとった行動の人数比のみに基づいて計算を行う。

- 提案手法

提案手法を用い、移動とそのコンテキストを用いた予測結果を求める。なお、次の移動先 L_{k+1} について、条件付き確率の計算に、式 (4) を用いた 1 地点前のみの行動履歴 $\{L_k, T_k\}$ を用いる手法 (1 Location) と、式 (8) を用いた 2 地点前までの行動履歴 $\{L_{k-1}, L_k, T_{k-1}, T_k\}$ を用いる手法 (2 Locations) の 2 通りの比較を行う。

4.3 結果と考察

4.3.1 予測精度の評価

単純統計による従来手法の正答率 (Conventional) と、提案手法のうち 1 地点前の行動から予測した正答率 (1 Location)、および 2 地点前の行動から予測した正答率 (2 Locations) を図 4 に示す。格子領域の面積が増えるに従い、すべての手法の正答率が高くなるのが分かる。これは、1 領域あたりの情報量が増え学習の精度が高まったからと考えられる。

また、従来手法と比べ、1 地点前の行動からの予測結果の精度が高く、さらに 2 地点前の行動からの予測精度が高いことから、予測に用いる情報量に応じて精度が高まるのが分かる。

なお、実用面においては、格子の 1 領域あたりの面積が広い場合、位置に関する精度が低下するとともに、本来移動と見なすべき離れた場所の発言であっても同一地点のもの

*1 地図データ ©2013 Google, ZENRIN.

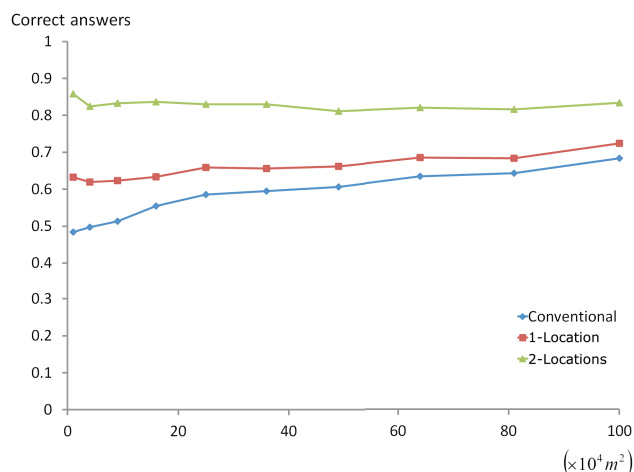


図 4 格子領域面積と正答率の比較

Fig. 4 Comparison of grid area and correct answers.

表 2 発言件数, 出現単語の総数, 単語セットの大きさ

Table 2 Messages, total words and size of word set.

	発言件数	出現単語の総数	単語セット
鶴岡八幡宮を含む領域	223	807	157
由比ヶ浜を含む領域	133	650	217

のとして扱われることが増えると考えられる。適切な目的地の予測のためには、発言の分布を基にクラスタリングを行い、発言が集中している領域からランドマークを抽出する手法などが考えられる。これらについては、今後の課題である。

4.3.2 学習データの特徴

地点ごとに学習された単語ベクトルの特徴を示すため、鎌倉駅を含む領域を起点とした移動のうち、(1) 鶴岡八幡宮を含む領域と (2) 由比ヶ浜を含む領域の 2 つの移動先について学習された単語の比較を行う。鎌倉駅からこれらの地点への移動は、多くの観光客が行うものであり、神社と海岸という異なる性質の目的地で背景にあるコンテキストが異なることから、この 2 地点を比較することで学習データの特徴が明確になると考えられる。

2 つの移動先について移動をともなう発言件数、出現単語の総数、単語の重複を省いた単語セットの大きさを表 2 に示す。

それぞれの領域の学習データにおける単語ごとの上位 10 位までの頻度を図 5 と図 6 に示す。鶴岡八幡宮を含む領域では「八幡宮」、「鶴岡」、「鎌倉」の順に頻度が高く、名称そのものが学習データに高い割合で存在していることが分かる。また、由比ヶ浜を含む領域では「鎌倉」、「駅」、「線」と「横須賀」の順に頻度が高く、「江」と「電」がそれに続いている。「江」と「電」は鶴岡八幡宮を含む領域では存在しない単語である。由比ヶ浜は、江ノ電の路線上の代表的な観光地であることから、学習データには場所に関するコンテキストが存在することが分かる。

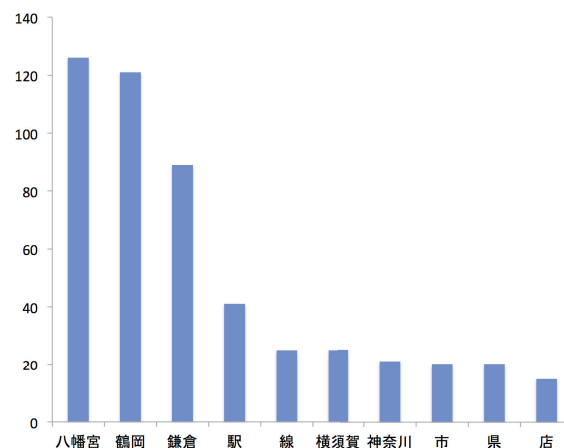


図 5 単語ごとの頻度分布：鶴岡八幡宮を含む領域

Fig. 5 Frequency distribution of words: Tsurugaoka-hachiman-gu.

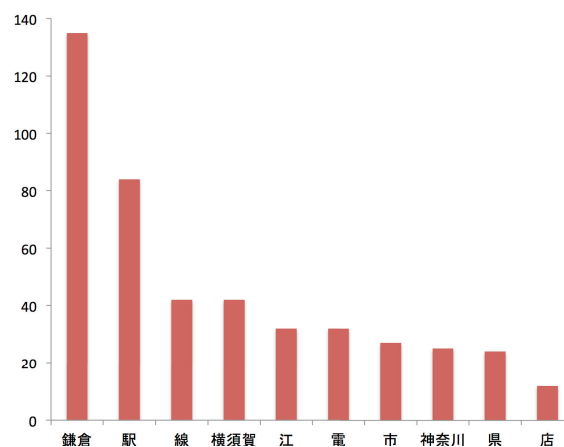


図 6 単語ごとの頻度分布：由比ヶ浜を含む領域

Fig. 6 Frequency distribution of words: Yui-ga-hama.

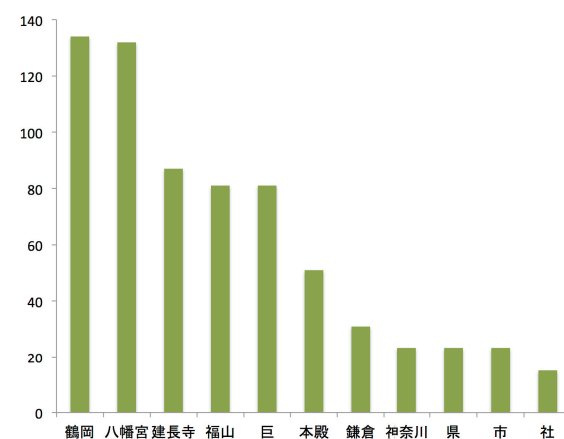


図 7 単語ごとの頻度分布：鎌倉駅を含む領域

Fig. 7 Frequency distribution of words: Kamakura station.

次に、2 地点間の双方向の移動における方向別の学習データの特徴を示す。前述の (1) 鎌倉駅から鶴岡八幡宮への移動と、(3) 鶴岡八幡宮から鎌倉駅への移動について学習された単語の比較を行う。図 7 は、鎌倉駅を含む領域の学習データの上位 10 位の頻度分布である。図 5 の鎌倉駅から

鶴岡八幡宮へ行く場合と比較すると、図 7 には「鶴岡」「八幡宮」のほかに、「巨」「福山」「本殿」といった鎌倉の神社内部を説明する単語が存在することが分かる。これらは、観光を終え鎌倉駅に戻る人々から、移動の過程で得られた知識や感想の一部として発せられたものであると解釈できる。このことから、直接場所を説明する単語以外にも、移動のコンテキストとなる単語が学習データに含まれることが分かる。

以上から、学習データには移動に関するコンテキストが反映されていることが分かる。

4.3.3 領域計算量の評価

提案手法は、予測対象とする領域の格子領域上で識別される行動履歴ごとに状態を保持し、状態ごとにベイズ分類器を生成し処理を行うものである。領域計算量はベイズ分類器の数に比例するため、1 格子あたりの面積と行動履歴に用いる地点数を変化させ、システム上で保持される状態数の評価を行った。図 8 は、格子サイズに対する状態数の変化を 1 地点前の情報のみを用いた場合 (1 Location) と、2 地点前の情報を用いた場合 (2 Locations) について示したものである。格子サイズが小さいと、短い距離間で行われる発言であっても識別されるため状態数が多くなり、格子サイズが大きいと、離れた地点の発言も同一の行動履歴と見なされることが増えるため状態数が少なくなることが分かる。一方、行動予測に用いる地点数が 2 地点の場合、とりうる 2 地点の組合せの数の増加に従い 1 地点のものより多くなった。2 地点の際の状態数は、最大 1 地点のときの 19 倍であるが、累乗分の増加とはならないことが分かる。これは、人々の移動する経路は、隣接する場所や道路をとることが多く、組合せが限定されたためと考えられる。

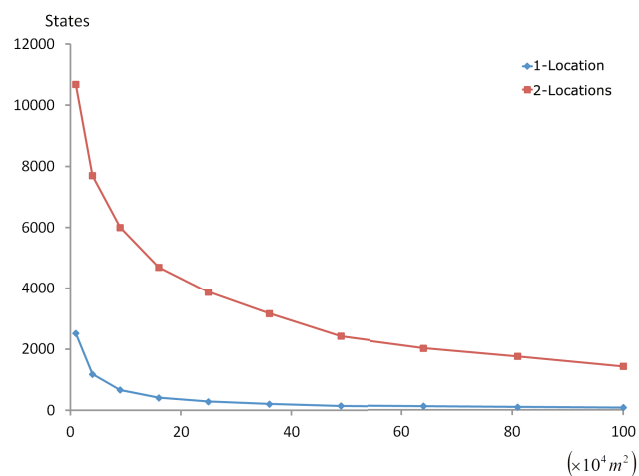


図 8 1 地点の履歴と 2 地点の履歴を用いた際の、格子サイズに対する状態数

Fig. 8 Grid size and states using history of one point and two points.

5. 関連研究

GPS などの位置計測技術の発展により、位置とタイムスタンプからなる位置履歴を記録することが可能である。この位置履歴を、ユーザの興味や趣向に関する手がかりと見なして分析する研究が行われている。Counts ら [1] は、センサによって取得された情報に位置タグをつけ複数ユーザで共有するシステムを提案している。平井ら [2] は、GPS ナビゲーションツールを利用して目的地へたどり着く過程を分析し、実世界との知覚的な相互関係を分析している。また、位置履歴からユーザの行動を分析し複数のユーザ間のソーシャルメディアを提案する研究 [3] や、データマイニングによってユーザ間のコミュニティを抽出する研究 [4] などが存在する。しかし、本研究では、位置の履歴だけを分析するのではなく、マイクロブログ上で人々が送信した発言とジオタグを蓄積し、位置履歴だけでなく同時に発言から得られるコンテキストを用いる。これにより、行動を起こす背景や目的を含めた分析を行うことが可能である。

また、ブログや SNS, Twitter などは、その瞬間に見たことや感じたことを気軽に Web 上に公開し共有するなど、世の中の関心を即時的に反映する新たなメディアであることから、これらから有用な実世界情報の抽出を目的とした研究が行われている。Hiruta ら [9], [10] は、場所誘因型ジオタグ付きツイートを抽出するため、位置情報メタデータと、発言内容と位置情報の対応付けを解析する手法を提案している。この研究は発言位置を説明するような情報を直接抽出することを主眼としているが、本研究は移動先の予測を目的としているため、発言内容が場所を直接説明する内容ではない場合もコンテキストとなる情報が存在すれば予測が可能となる。しかし、4 章の実験での「鎌倉」のような地域全体を表す単語や、どの発言であっても登場するような出現頻度の高い単語は、本研究の手法であっても予測に有益ではないと考えられる。ノイズとなる単語の除去や、単語ごとの重み付けなどによる精度向上は今後の課題である。

Lee ら [11] は、Twitter 上の群衆の振舞いをクラスタ化することにより、実世界のイベントを抽出する試みを提案している。一方、言語解析的なアプローチとしては、ソーシャルメディアでの複数の発言状況から注目する出来事の発生を推定するための、テキストの類似度に基づく情報伝搬経路解析手法 [5] や、Twitter の複数発言間の構造を HITS アルゴリズムにより解析し、authority と hub の 2 つの尺度を用いて情報取得に有用なユーザを抽出する手法 [6] などが存在する。Twitter は発言の文字数が短く、目的のトピックを示す単語が省略されていることや、リツイート元の内容を前提とした発言がされていることが多いため、発言間の情報伝搬を展開してコンテキストを解析することは今後の課題と言える。

6. おわりに

本論文では、マイクロブログサービスから過去の発言と位置情報を取得し、行動解析を行うことにより未来の利用者の行動を予測する手法について提案した。提案手法では、人の行動に確率過程モデルを適用し解析した統計情報を蓄積することで、未来の利用者の行動予測を行うものである。また、マイクロブログサービスである Twitter のデータを用いて、提案手法の評価実験を行った。

今後は予測精度の向上のため、発言の分布を基にクラスタリングを行い発言が集中している領域からランドマークを抽出する手法について検討を行う。さらに、移動と発言の関係性について検討を行い、予測精度の向上を目指す。提案手法では発言単語のみを対象としており、発言の意味を考えた予測を行っていない。観光や食事、買い物といった目的別の辞書を与えて学習することにより、より抽象的な意味を反映した予測を行うことが可能となると考えられる。さらに、予測結果を反映させた推薦システムの設計、実装を行う予定である。

謝辞 本研究は、JSPS 科研費 若手研究 (B) 課題番号 24700074 の助成を受けたものである。

参考文献

- [1] Counts, S. and Smith, M.: Where Were We: Communities for Sharing Space-time Trails, *Proc. 15th Annual ACM International Symposium on Advances in Geographic Information Systems (GIS 2007)*, pp.1–8 (2007).
- [2] 平井浩将, 森 傑: 経路探索における GPS ナビゲーションツールの利用とアクション生起との関係—都市空間におけるアクトファインディングに関する研究, *日本都市計画学会都市計画論文集*, No.42-3, pp.541–546 (2007).
- [3] Zheng, Y., Xie, X. and Ma, W.: GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory, *IEEE Data Engineering Bulletin*, Vol.33, No.2, pp.32–39 (2010).
- [4] Hung, C.-C., Chang, C.-W. and Peng, W.-C.: Mining Trajectory Profiles for Discovering User Communities, *Proc. 2009 ACM International Workshop on Location Based Social Networks (LBSN 2009)*, pp.1–8 (2009).
- [5] Kim, J.W., Candan, K.S. and Tatemura, J.: Efficient Overlap and Content Reuse Detection in Blogs and Online News Articles, *Proc. 18th ACM International Conference on World Wide Web (WWW 2009)*, pp.81–90 (2009).
- [6] 田中淳史, 田島敬史: twitter のツイートに関する分類手法の提案, *日本データベース学会他第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010)*, A5-4 (2010).
- [7] Harris, Z.S.: Distributional structure, *Word*, Vol.10, pp.146–162 (1954).
- [8] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, available from (<https://code.google.com/p/mecab/>) (accessed 2013-05-11).
- [9] Hiruta, S., Yonezawa, T., Jurmu, M. and Tokuda, H.: Detection, Classification and Visualization of Place-

triggered Geotagged Tweets, *Proc. 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, pp.956–963 (2012).

- [10] 蛭田慎也, 米澤拓郎, 徳田英幸: 場所誘因型位置情報付き発言の検出と可視化, *情報処理学会論文誌*, Vol.54, No.2, pp.710–720 (2013).
- [11] Lee, R. and Sumiya, K.: Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection, *Proc. 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, LBSN'10*, pp.1–10 (2010).



斉藤 裕樹 (正会員)

1996 年明治大学理工学部情報科学科卒業。2001 年同大学大学院理工学研究科基礎理工学専攻博士課程修了。博士 (工学)。明治大学助手、東京電機大学助手、同大学講師を経て、2013 年より明治大学総合数理学部准教授。オーバレイネットワーク、センシングとネットワークのソフトウェアの研究に従事。日本ソフトウェア科学会、IEEE、ACM 各会員。



高山 翼

1990 年生。2013 年東京電機大学未来科学部情報メディア学科卒業。同年より株式会社富士通システムズ・イースト勤務、現在に至る。データマニングとユビキタスコンピューティングに興味を持つ。



山上 慶

1986 年生。2011 年東京電機大学未来科学部情報メディア学科卒業。2013 年同大学大学院未来科学研究科情報メディア学専攻修士課程修了。同年より株式会社ジェイアール東日本情報システム勤務、現在に至る。SNS へ投稿された発言のコンテキストと位置情報を用いた行動解析について研究を行う。



戸辺 義人 (正会員)

東芝，慶應義塾大学，東京電機大学を経て，2012年青山学院大学理工学部情報テクノロジー学科教授．センサネットワーク，参加型センシングの研究を進めている．IEEE，電子情報通信学会，計測自動制御学会，人間情報学会

各会員．



鉄谷 信二

1980年北海道大学大学院修士課程修了．同年電電公社（現，NTT）入社以来，ファクシミリにおける画像信号処理，電子写真記録および立体表示技術等の研究・実用化に従事．1991年ATR通信システム研究所に出向，臨場

感表示技術に従事．1994年NTTヒューマンインタフェース研究所に復帰，高速ネットワーク用アプリケーションの開発に従事．2000年ATR知能映像通信研究所に出向，コミュニケーション環境生成技術に関する研究，ATRメディア情報科学研究所にて五感メディアの研究に従事．2004年4月，東京電機大学理工学部教授，2008年4月同大学未来科学部教授，マルチメディアに関する研究に従事，現在に至る．工学博士．