

少数の正解ラベルを用いた移動履歴からの移動手段判定

Detecting Transportation Mode From GPS Trajectories with A Small Quantity of Labeled Data

数原 良彦*1 鈴木 雅大*2 戸田 浩之*1 鷺崎 誠司*1
Yoshihiko Suhara Masahiro Suzuki Hiroyuki Toda Seiji Susaki

*1 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, NTT Corporation

*2 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

This paper tackles the task of transportation mode detection from raw GPS trajectories. Zheng et al. propose a two-step approach that formalizes the task as a multiclass classification problem. Their method first split GPS trajectories into segments based on detecting change-points. The segments are considered as instances after being labeled based on annotations, following which their features are extracted. Our work verifies that this method requires a large quantity of annotating in order to achieve a particular level of accuracy. Collecting many annotations is difficult due to expensive cost and privacy concerns. Hence, we introduce a semi-supervised learning algorithm to use unlabeled data so as to improve classification accuracy with a small quantity of labeled data.

1. はじめに

スマートフォンやタブレット端末の普及により、ユーザの位置情報取得が容易となり、近年では位置情報を利用したサービスが普及し、利用するシーンが増えている。たとえば、Google Now*1 や i コンシェル*2 などのサービスでは、位置情報を元に検索結果や提示する情報を変化させることにより、ユーザ状況に合わせた情報提示を行っている。このような背景から、端末が利用可能なセンサ情報を利用してユーザ状況を推定する技術の必要性が一層高まっている。本稿では、世の中に普及しているスマートフォン、タブレット端末での利用を想定して、GPS などを用いて得られる移動履歴情報を用いたユーザ状況推定技術の開発に取り組む。

具体的には、GPS や WiFi 測位に基づいて得られた位置情報と時刻情報を持つ測位点系列が与えられた際に、当該ユーザの状態推定を行う課題に取り組む。本稿ではこの測位点系列を測位方法によらず GPS 軌跡と呼ぶ。GPS 軌跡を用いたユーザ状態推定の研究としては、ユーザが訪問した地域を推定する滞在地推定 [西田 13] や、どのような移動手段を用いたかという移動手段判定 [Zheng 08, Zheng 10] の研究が行われてきた。本稿ではこの中で、移動手段判定手法の開発に取り組む。

GPS 軌跡を用いた移動手段判定の従来手法 [Zheng 08, Zheng 10] においては、GPS 軌跡とそれに対する移動手段アノテーションを用いて教師あり学習の枠組みで移動手段予測モデルを構築し、移動手段が未知の GPS 軌跡に対して移動手段を付与する方法を用いている。約 4 年半にわたる 182 名の被験者の GPS 軌跡と、一部の GPS 軌跡に対するアノテーションデータを用いて評価を行っている。しかしながら、実サービスを提供する上では、そもそもユーザが自らの GPS 軌跡に対して移動手段のアノテーションを行う必要があり、コストの観点から大量のラベル付きデータが利用できないという問題がある。また、プライバシーの問題からユーザが GPS 軌跡に対し

て全てのアノテーションを行うわけではない。

このような背景から、我々は GPS 軌跡とそれに対するアノテーションを用いた移動手段判定タスクにおいては、利用可能なラベルが少量であると想定するのが妥当であると考えた。既存研究においてはデータセットに含まれるラベル付きデータをすべて利用した教師あり学習の手法を用いた方法と結果が報告されており、利用するラベル付きデータの量に応じて精度がどのように変化するかという知見は報告されていない。

我々は利用可能なラベル付きデータとラベルなしデータを利用し、半教師あり学習の枠組みで予測モデルを構築することにより、少量のラベル付きデータのみを利用する既存手法に比べて高精度な移動手段判定が可能な方法を提案する。前処理として Zheng らの方法を利用し、教師あり学習を用いた予測モデル構築に関して本問題設定に合わせた改良手法を提案する。本稿では、予備実験として既存手法の問題設定において、利用可能なラベル付きデータの増加によって予測精度がどのように変化するか学習曲線を描くことで確認し、教師あり学習による移動手段判定においては大量のラベル付きデータが必要であることを示す。また、本稿で提案する半教師あり学習の枠組みでラベルなしデータを利用する提案手法を述べ、評価実験を通じてその有効性を検証する。

本研究の貢献は以下の通りである：

- 大量の GPS 軌跡と少量の正解ラベルのみが利用可能な状況において、ラベルなしデータを利用して移動手段判定を行う方法を提案する。
- 実データを用いた実験において、少量のラベルでは十分な精度が達成できないことを検証し、そのような状況において提案手法を用いることで精度向上が可能であることを示す。

2. 関連研究

GPS ログなどの移動履歴を用いた研究として、ユーザ活動状態推定 [Cho 11, Liao 07]、移動手段判定 [Zheng 08, Zheng 10]、地域分析 [Xiao 10, Yuan 12] などが挙げられる。Zheng ら [Zheng 08, Zheng 11] は移動手段が変化する際には、

連絡先: suhara.yoshihiko@lab.ntt.co.jp

*1 <http://www.google.com/landing/now/>

*2 <https://www.nttdocomo.co.jp/service/customize/iconcier/>

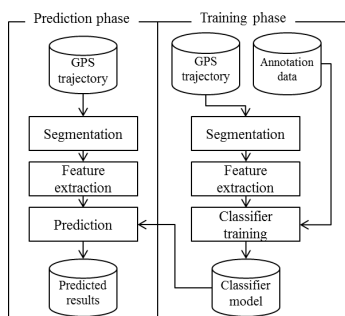


図 1: システム構成

徒歩状態が挟まれる仮定に基づき、測位点から得られる速度と加速度を用いた変化点抽出手法を提案し、これによって分割された複数のセグメントに対してマルチクラス分類問題として移動手段を推定する方法を提案している。しかしながら、彼らの研究ではデータセットに含まれる全ユーザのラベル付きデータは全て利用可能な状況を想定している。本稿では基本的な問題設定は Zheng らの研究を踏襲し、利用可能なラベル付きデータが少量の場合の既存手法の評価結果、およびラベルなしデータ利用の提案とその検証を報告する。

3. 移動手段判定

移動手段判定のシステム構成を図 1 に示す。移動手段判定は予測モデル構築フェーズ (Training phase) と、予測フェーズ (Prediction phase) の 2 つに分けられる。まず、予測モデル構築を行う訓練フェーズについて述べる。与えられた各ユーザの GPS 軌跡情報について、測位点系列に基づき、複数のセッションに分割する。この際、日付をまたぐ、または 20 分以上の間隔が空いた場合に GPS 軌跡を異なるセッションとして分割する。これにより得られた複数のセッションそれぞれについて、後述するセグメント分割手法を用いてセグメント抽出を行う。ここで得られたセグメントを学習や予測における 1 事例として扱う。各事例について特徴抽出を行い、訓練フェーズにおいては、セグメントと GPS 軌跡に対するアノテーションを照らし合わせることでラベルを付与し、ラベル付きデータを生成する。得られたラベル付きデータを用いて、教師あり学習の枠組みでマルチクラス分類器を生成し、これを予測モデルとして用いる。予測フェーズにおいては、アノテーションに基づいてラベル付与する点を除いて訓練フェーズと同様の処理を行い、得られた特徴ベクトルと生成された予測モデルを用いて移動手段判定結果を出力する。

本稿においては、比較のため Zheng ら [Zheng 08, Zheng 10] による変化点検出に基づくセグメント分割の方法をそのまま利用する。本稿の貢献は予測器の学習部分に関するものであり、セグメント分割の方法に依存しないことに注意されたい。

3.1 変化点抽出に基づくセグメント抽出

本稿では Zheng ら [Zheng 08, Zheng 10] による変化点抽出に基づくセグメント分割を用いて事例の生成を行う。Zheng らの方法では、大きく 4 つのステップで変化点抽出を行う。

1. 各測位点に対して位置情報をもとに速度、加速度を計算し、速度閾値 θ_v と加速度閾値 θ_a のいずれかを越えた場合には測位点タイプを non-walk とし、そうでない場合には測位点タイプを walk と判定する。
2. 同じ測位点タイプが連続する測位点系列をセグメントとみなす。セグメントに含まれる測位点系列の総移動距離

が θ_L 以下、またはセグメントに含まれる測位点系列の総経過時間が θ_T 以下の場合には、セグメントに含まれる測位点系列の測位点タイプをひとつ前のセグメントと同じものに変更し、セグメントを併合する。

3. 2. の処理後のセグメントの総移動距離が θ_C 以上の場合にはセグメントタイプを certain と判定し、そうでない場合には uncertain とする。セグメントタイプが uncertain と判定されたセグメントが T_U 以上連続して存在する場合、当該セグメントに含まれる測位点系列の測位点タイプを non-walk とする。
4. セグメントを構成する測位点の測位点タイプが walk から non-walk またはその逆に変化する点を変化点として抽出する。

本稿では変化点抽出の各閾値について Zheng らが文献 [Zheng 11] の実験で用いられた $\theta_v = 2.5[\text{m/s}]$, $\theta_a = 1.5[\text{m/s}^2]$, $\theta_L = 20[\text{m}]$, $\theta_T = 10[\text{s}]$, $\theta_C = 200[\text{m}]$, $T_U = 3$ を用いる。

3.2 教師あり学習を用いた移動手段判定

3.1 の処理によって得られた各セグメントを事例とみなし、各セグメントに対してあらかじめ用意された複数の移動手段候補の中から移動手段を推定する問題設定と捉えることで、移動手段判定をマルチクラス分類の枠組みで定式化できる。Zheng ら [Zheng 08, Zheng 11] は各セグメントについて、セグメントに属する測位点系列によって得られた情報を元に特徴抽出を行い、ラベル付きデータを用いて移動手段予測器の学習を行っている。

本稿では、既存研究 [Zheng 08, Zheng 11] と同様に以下の情報を特徴ベクトルとして用いる。

- セグメント長 $L = \sum_{i=0}^{n-1} L_i$ ここで $L_i = \text{Dist}(P_i, P_{i+1})$
- 平均速度 $\bar{V} = L / (P_n.T - P_0.T)$
- 速度の期待値 $\mathbb{E}[V] = (\sum_{i=0}^{n-1} V_i) / (n - 1)$
- 速度の分散 $\sum_{i=0}^{n-1} (V_i - \mathbb{E}[V])^2 / (n - 1)$
- 当該セグメントにおける上位 3 件の速度
- 当該セグメントにおける上位 3 件の加速度

ここで、セグメントに含まれる測位点系列を (P_0, P_1, \dots, P_n) とし、2 つの測位点の距離を返す関数を $\text{Dist}(\cdot, \cdot)$ とする。なお速度 V_i については上記定義に従うと $i = 0, \dots, n - 1$ までしか計算できないため、 n 個の測位点系列によって構成されるセグメントの場合には $n - 1$ 個の速度が計算されることに注意されたい。

移動手段予測器の学習には、マルチクラス分類を行うことが可能な機械学習手法を用いることができる。既存研究 [Zheng 08, Zheng 11] においては学習手法として決定木、SVM, Bayesian Net, CRF を比較し、決定木が最も高い精度を示したと報告している。本稿では Zheng らの結果を踏まえて決定木と、決定木の性質を引き継ぐアンサンブル学習である Random Forest [Breiman 01] の評価も行うことで、既存研究で提案された枠組みにおいても新たな知見を与えることを目指す。

3.3 半教師あり学習を用いた移動手段判定

本稿では任意の分類器に適用可能なラッパ型の半教師あり学習である self-training [Yarowsky 95][Zhu 09] を用いたラベルなしデータ利用を提案する。ラベルありデータを利用して学習した分類器の予測結果に基づいて疑似的に正解ラベルを付与し、疑似的に生成されたラベルありデータを加えた訓練データを用いて分類器を学習する。

Algorithm Self-training with confidence-based filtering

Input: $D_L = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^L$, $D_U = \{\mathbf{x}^{(i)}\}_{i=L+1}^N$, C

Output: $f: \mathcal{X} \rightarrow \mathcal{Y}$

- 1: Build classifier f_L with D_L .
- 2: $D^* \leftarrow D_L$
- 3: **FOR** i in $L+1$ to N
- 4: $\hat{y} \leftarrow f_L(\mathbf{x}^{(i)})$
- 5: **IF** $\text{Conf}(\hat{y}) > C$ **THEN**
- 6: $D^* \leftarrow D^* \cup (\mathbf{x}^{(i)}, \hat{y})$
- 7: **ENDIF**
- 8: Build classifier f^* with D^* .
- 9: **RETURN** f^*

図 2: アルゴリズム

アルゴリズムを図 2 に示す．本稿では，self-training で疑似ラベルとして採用する際に，確信度に対して閾値を設けて正解ラベルとして採用するかフィルタリングを行うヒューリスティクスを導入した方法 [Yarowsky 95] を用いる．確信度に基づくフィルタリングを用いた self-training では，ラベル付きデータ D_L ，ラベルなしデータ D_U ，確信度閾値 C を入力として受け取る．ラベル付きデータを用いて分類器を生成し，生成された学習器を用いてラベルなしデータの各事例について予測を行い，予測結果 \hat{y} と確信度 $\text{Conf}(\hat{y})$ を得る．得られた確信度が C 以上の場合には，予測結果をラベルとして訓練データに追加し，そうでない場合には追加データとして利用しない．全てのラベルなしデータについて，上述の処理を終えたのちに，最終的に得られた訓練データ集合を用いて新たな分類器を生成し，これを最終結果として返す．

4. 評価

評価実験を通じて以下の 2 点を検証する．(1) 訓練データ数が限定された状況下において，既存手法の教師あり学習の枠組みでどの程度移動手段判定精度が達成可能であるか知見を得る (実験 1)．(2) 少量の正解ラベルが与えられた状況において，提案手法を用いることで，教師あり学習に比べて高精度な予測が達成可能であることを検証する (実験 2)．データセットについて述べたのちに，実験 1 と実験 2 について述べる．

4.1 データセット

本実験では，Microsoft Research によって公開されている GeoLife データセット^{*3} を利用した．本データセットには，2007 年 4 月から 2012 年 8 月にかけて 182 ユーザの GPS 軌跡が含まれており，そのうち 69 ユーザが一部の GPS 軌跡に対して 11 種類の移動手段のアノテーションを行っている．本研究では，Zheng ら [Zheng 08, Zheng 11] と同様に walk, bus, car, bike の 4 種類の移動手段を評価対象とした．この際，Zheng らと同様に taxi ラベルも car ラベルと同じものとして扱う．3.1 で述べたセグメント抽出を行い，抽出されたセグメントに対して特徴抽出と移動手段のアノテーションに基づいて上記 4 クラスのラベル付与を行った結果，合計 37,717 件のラベル付きデータを抽出した．

4.2 実験 1: 学習曲線

学習に利用する訓練データ数の増加に伴い，予測精度がどのように変化するか実験を行った．学習アルゴリズムとしては，決定木 (DT) とアンサンブル学習である Random Forest (RF) を用いた．学習アルゴリズムの実装は scikit-learn 0.14.1^{*4} を用いた．ラベル付けと特徴抽出された全事例について 10-fold nested cross validation (nCV) を用いてパラメータ選択と評

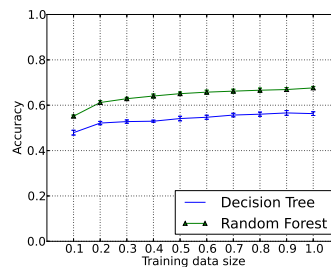


図 3: 10-fold nCV におけるテストデータに対する学習曲線

価を行った．10-fold nCV においては，データセットを 10 分割し，9 個のブロックを訓練データ，残りのブロックをテストデータとして評価を行う試行を 10 回行う．その際，各試行において訓練データについて交差検定を用いて評価指標が最大となるようにパラメータを選択する方法である．本実験においては，評価指標として正解率を利用し，パラメータ探索についてはグリッド探索を用いた．決定木のパラメータ探索については，枝分割の方法 (Gini 係数またはエントロピー) と分類に利用する最大特徴数 (0.1, 0.2, ..., 1.0) の中から正解率が最大になるものを選択した．Random Forest のパラメータ探索については，生成する木の数 (10, 30, 50)，最大特徴数 (0.1, 0.2, ..., 1.0) の中から正解率が最大になるものを選択した．

結果を図 3 に示す． x 軸はモデル生成に利用した訓練データの割合を表し， y 軸は正解率を表す．グラフは各手法について，訓練データサイズ毎の正解率の平均値を表し，エラーバーは標準偏差を表す．図より，訓練データのサイズによらず RF が安定して高い正解率を示すことがわかった．Zheng らの研究では DT が SVM, Bayesian Net, CRF に比べて高い正解率を示しており，この DT の結果を上回ることから，本タスクにおいて RF が有効である新たな知見が得られた．

図 3 より，訓練データの割合が 0.2 になるまで正解率が大きく増加し，それ以降なだらかに増加している．全ての訓練データを用いた場合においても精度がまだ向上する余地があることが伺える．なお，各 fold において訓練データは最大 33,945 事例利用しているため，学習曲線において 0.1 増加するたびに約 3,400 事例多く利用していることを表す．この結果より，安定した精度を得るまで約 7,000 事例のアノテーションが必要であることがわかる．

4.3 実験 2: ラベルなしデータ利用の効果

提案手法の有効性を確認するため，ラベルなしデータ利用の効果を検証した．学習器は実験 1 において高い精度を示した RF のみを用いた．訓練データを実験 1 における最小値である全体の 0.1 に固定し，残りの訓練データをラベルなしデータとみなす．実験 1 と同様に 10 nCV を実行し，学習器のパラメータ調整も同様に行った．self-training で利用対象とするラベルなしデータについては，ラベルありデータとして利用されない訓練データのラベルを無視することで，ラベルなしデータとして扱う．実験 2 では，ラベルなしデータの利用サイズを 0% から 100% 利用するという 10 段階の結果を評価した．また，self-training における閾値は $C = \{0, 0.6, 0.7, 0.8, 0.9, 0.95\}$ を用いた．

$C = \{0, 0.6, 0.7, 0.8\}$ について，ラベルなしデータ増加に伴う正解率の変化を表すグラフを図 4 に示す．図中，エラーバーなしの赤字で表示されている線がラベルありデータのみを使った分類器の結果を表し，エラーバー付きの青線が半教師あり学習の結果を表す．各グラフのエラーバーは標準偏差を表す． y

*3 <http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/>

*4 <http://scikit-learn.org/>

軸は正解率を表し、 x 軸は利用したラベルなしデータの量を表す。ラベルなしデータを利用しない教師あり学習に基づく RF の結果 (図中赤線) はラベルなしデータ増加によっても精度が変化せず、これが提案手法に対するベースラインとなる。

図 4 (a),(b),(c) より, self-training を利用することで, 正解率が向上していることが確認できる。特に図 4 (a),(b),(c) においては, ラベルなしデータ利用が 0.7 を境にゆるやかに正解率が減少しており, ラベルなしデータの追加が必ずしも正解率の単調増加に貢献するわけではないことが確認できた。図 4 (d) より, self-training における確信度に対する閾値 C を高く設定すると, ラベルなしデータ追加の効果を得られないことが確認できた。この場合は, 利用するラベルなしデータサイズを増加しても, 正解率が増減しており, 半教師あり学習の効果を得られていない。グラフに示しているものの $C = 0.9, 0.95$ においても $C = 0.6$ と同様の結果が得られた。これらの結果から, 本タスクに self-training を適用する場合には, 閾値設定によるフィルタリングをすることで効果が得られなくなることを, ラベルなしデータの数を一定数以上に増やしても半教師あり学習の効果がそれ以上得られなくなることを確認した。この結果は, 高い確信度のラベルなしデータを採用することが望ましいとされる self-training の一般的な知見 [Cheng 13][Yarowsky 95] とは異なる。

self-training によって半教師あり学習の効果が得られる仕組みを (1) 正しいか, (2) 学習データとして有益であるか, の 2 つに分けて考察する。前提として予測ラベルが真のラベルではない場合には, 学習データとして必ず不利益を与えるとする。確信度が高い予測については (1) に関しては必然的に高いと見なせる。今回のケースでは (2) について問題があるため, 学習効果が得られなかったと考える。RF においては, 複数の決定木の予測の投票に基づいて予測と確信度を決定するため, 確信度が高いということは, すなわち RF を構成する多くの決定木において正しく分類されていることを意味する。そのため, 確信度が高いデータを新たな訓練データとして追加したところで, RF 全体としての予測結果が変わるような学習効果が得られない。それどころか, 仮に確信度が高い予測が誤っている場合には, そのような予測に基づくラベルなしデータの利用によって, モデルの性能低下を引き起こすおそれがある。これが図 4 (d) においてラベルなしデータの効果が得られなかった理由と考えている。

5. おわりに

本稿では, GPS 軌跡とそれに対するアノテーションデータを用いた移動手段判定タスクにおいて, 少量の正解ラベルのみが利用可能な状況における既存手法の適用を検討し, その場合の既存手法の性能評価と, 半教師あり学習に基づくラベルなしデータを利用する方法を提案した。評価実験を通じて以下の 3 点を確認した: (1) 訓練データ数の増加に伴うテストデータに対する正解率の学習曲線を描くことで, 一定の正解率を達成するためには, 多数のアノテーションが必要になること。(2) マルチクラス分類の学習器として Random Forest を利用することで, 既存研究において最大性能を示した決定木を上回る正解率を達成した。(3) 少量の正解ラベルのみ利用可能な状況において, ラベルなしデータを用いて self-training に基づく半教師あり学習を用いることで, 精度向上が可能なことを検証した。

ただし, (3) においては疑似的な正解ラベルとして採用するための基準の閾値を高く設定するとラベルなしデータを利用する効果が得られないことも確認した。本稿では初期検討とし

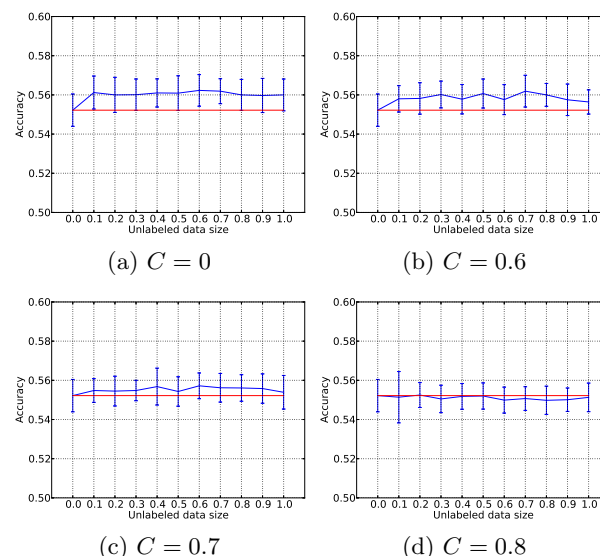


図 4: 実験 2 の結果

て, 任意の学習手法と組み合わせの容易な self-training を採用し, ラベルなしデータを利用することで精度向上可能なことを示した。ラベルなしデータ利用方法として, さらに高度な半教師あり学習の開発が今後の課題である。また, 行動パターンはユーザに依存するところが大きいと考えており, 個人に特化したパターンと全体ユーザ共通のパターンを考慮した予測モデル構築を検討する予定である。

参考文献

- [Breiman 01] Breiman, L.: Random Forests, *Mach. Learn.*, Vol. 45, pp. 5–32 (2001)
- [Cheng 13] Cheng, S., Huang, Q., Liu, J., and Tang, X.: CoSTra: Confidence-based Self-training, *Journal of Computational Information Systems*, Vol. 9, No. 24, pp. 9761–9769 (2013)
- [Cho 11] Cho, E., Myers, S. A., and Leskovec, J.: Friendship and mobility: user movement in location-based social networks, in *Proc. KDD '11*, pp. 1082–1090 (2011)
- [Liao 07] Liao, L., Fox, D., and Kautz, H.: Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields, *Int. J. Rob. Res.*, Vol. 26, No. 1, pp. 119–134 (2007)
- [Xiao 10] Xiao, X., Zheng, Y., Luo, Q., and Xie, X.: Finding similar users using category-based location history, in *Proc. SIGSPATIAL '10*, pp. 442–445 (2010)
- [Yarowsky 95] Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, in *Proc. ACL '95*, pp. 189–196 (1995)
- [Yuan 12] Yuan, J., Zheng, Y., and Xie, X.: Discovering regions of different functions in a city using human mobility and POIs, in *Proc. KDD '12*, pp. 186–194 (2012)
- [Zheng 08] Zheng, Y., Liu, L., Wang, L., and Xie, X.: Learning transportation mode from raw GPS data for geographic applications on the web, in *Proc. WWW '08*, pp. 247–256 (2008)
- [Zheng 10] Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y.: Understanding transportation modes based on GPS data for web applications, *ACM Trans. Web*, Vol. 4, No. 1, pp. 1:1–1:36 (2010)
- [Zheng 11] Zheng, Y. and Zhou, X. eds.: *Computing with Spatial Trajectories*, Springer (2011)
- [Zhu 09] Zhu, X., Goldberg, A. B., Brachman, R., and Dietterich, T.: *Introduction to Semi-Supervised Learning*, Morgan and Claypool Publishers (2009)
- [西田 13] 西田京介, 戸田浩之, 倉島健, 内山匡: 確率的訪問 POI 分析: 時空間行動軌跡からのユーザモデリング, マルチメディア, 分散, 協調とモバイル (DICOMO2013) シンポジウム, pp. 334–345 (2013)