強化学習に基づく人の流れシミュレーションフレームワークの構築

Yanbo Pang・樫山武浩・関本義秀

A reinforcement learning based modeling framework for people flow simulation

Yanbo Pang, Takehiro KASHIYAMA and Yoshihide SEKIMOTO

Abstract: Understanding individual and crowds dynamics in urban environments is critical for numerous applications, such as urban planning, traffic forecasting and location-based services. However, developing such models needs travel survey data that are high cost and updated in-frequently. Thanks to the emerging mobility data collection techniques such as Call Detailed Record (CDR), smartphone applications and social network check-ins, which enables researchers to observe and analysis people flow directly. However, because of the privacy policy, personal information are not available so that such data sources are hardly to be used for existing travel demand modeling approach. In this study, we develop a reinforcement learning based approach that is capable of reproducing individual's daily travel behavior from anonymous locational data and reconstruct people flow on citywide level.

Keywords: 交通需要（travel demand），人の流れ（people flow），位置情報（location data），強化学習（reinforcement learning）

1. はじめに

都市計画，交通需要予測，モビリティサービスなどを提供するため，さまざまなシナリオにおいて人の流れを理解・把握することは重要である(Sekimoto et al. 2011). しかし、既存の交通需要モデルを構築するには、行動者の個人属性や意識調査が不可欠である．こうした調査はコストが高く、更新が遅いため，日々に変動している人の流れを反映することが難しい．一方，近年の様々な技術発展で CDR，GPS，check-in など様々な位置情報データセットが使えるようになっている．しかし，個人データの取り扱いには課題がある.個人情報を隠匿する上で，個人を特定できないように匿名化がなされる．こうしたデータをエージェントモデリングおよびシミュレーションにどのように活かすことできるのかが課題である．

そこで，本研究では，位置情報を活用することで，より現実の交通行動意思決定を反映したモデルを構築し，シミュレーション結果の精度を向上するフレームワークを作成することを目標とする．具体的には，機械学習の一つであり，逐次的な意思決定にも用いられる強化学習を利用したエージェントの行動制御に反映した end-to-end フレームワークを構築する．その中では、逆強化学習を用いてデモンストレーション軌跡から人間の行動要因についてのパラメーターを推定し、それをエージェントモデルに組み込むことで，より現実な人の流れシミュレーションを行う．

2. 手法の概略

図 1 には位置情報に基づく強化学習エージェントシミュレーションのフレームワークを示す.

Yanbo Pang
東京大学生産技術研究所　関本研究室
pybdtc@iis.u-tokyo.ac.jp

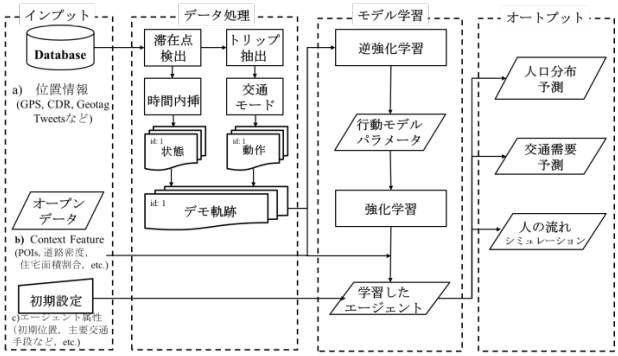

図-1 本研究のフレームワーク

2.1 エージェントモデル

(1) 強化学習による交通行動モデルの定式化

強化学習エージェントは，現在の状態を観測し，行動を選択することで環境から報酬を獲得し，一連の行動を通じて報酬が最も多く得られるような方策(policy)を学習する． エージェントモデルは，マルコフ決定過程(Markov Decision Process, MDP)として定式化され，**状態 S** は，各時刻におけるエージェントの位置を表す. **行動 A** はエージェントが環境に対して行う働きかけの種類を表し，一つのトリップ（起点・終点・交通手段から構成）がこれに相当する. **遷移関数**$P_{sa}(\cdot)$は状態 s において行動 a を取った時の次の状態への状態遷移確率を表す． **割引因子**は現在の報酬と未来の報酬との間における重要度の差異を表す. **報酬 R** はその行動の即時的な良さを表す. エージェントは環境から現在の状態と報酬を受け取り，行動集合の中から行動を決定し，これを環境に引き渡す.

一方，強化学習問題を解くことは，最適な方策を求めることである. 方策の評価指標として，状態価値関数

$$V^{\pi}(s) = R^{\pi}(s) + \gamma \sum_{a \in A} \sum_{s\prime \in S} \pi(s,a) Pr(s,a,s')\, V^{\pi}(s')$$

を表す.これは，方策 π のもとで状態 s からエージェントが方策 π に基づいて行動を決定した場合の期待値である.すると，最も良い方策は次式で更新する:

$$\pi(s) \leftarrow \underset{a \in A_s}{argmax} \sum_{s\prime \in S} T(s,a,s') \big(R(s,a,s') + \gamma V^{\pi}(s') \big)$$

これらの操作πがすべての状態に対し変化しなくなるまで繰り返すことで，最適な政策を得る.

(2) 逆強化学習による報酬関数パラメータの推定

しかし，決定過程で最適な方策を計算方法を示したが，報酬関数の構造は自明ではない.言い換えると，現実の世界では，報酬，つまり行動の「良さ」を定量的に定義することは難しい問題である. 例えば,「いい行動」をどう評価するか，人によって同じ目的でも，場所・交通手段・移動距離・料金に対するが趣味嗜好が異なるため，一意に行動にともなう報酬を決定できない. これらの行動要因とそれらのパラメータを意識調査から推定できるが，位置情報から推定する方法は存在しない.

本研究では，Andrew Ng et al.(2000) が提唱した逆強化学習(Inverse Reinforcement Learning: IRL)という手法を用い，エキスパートの行動軌跡に基づいて報酬関数を推定する. 具体的には，Ziebart et al. (2008) の最大エントロピー逆強化学習 (Maximum Entropy Inverse Reinforcement Learning) アルゴリズムを用いて位置情報から抽出した移動軌跡によって報酬関数のパラメータを学習させる. この手法は、行動をフィーチャー$f_a \in R$が特徴づけ，重みベクトルθを用いてこれらフィーチャをパラメータ化した線形関数に従う.軌跡のフィーチャf_ζ は，軌跡ζに含まれる動作のフィーチャの合計である.したがって，フィーチャに適用される軌跡の報酬の線形和は次のように定義する：

$$reward(f_\zeta) = \theta^T f_\zeta = \sum_{s_i \in \zeta} \theta^T f_{sj}$$

行動軌跡の分布を次のように定義する：

$$P(\zeta|\theta) = \frac{exp\left(\sum_{s_i \in \zeta} \theta^T f_{sj}\right)}{Z(\theta)}$$

ここでは，ζが経路，s_iは i 番目の状態，$Z(\theta)$は分配関数である. したがって，行動フィーチャの制

約を受ける経路の分布に対し，エントロピーを最大化することによって，フィーチャの重みベクトルfを機械学習させる.また，最大エントロピー分布において観測されたデータの尤度を，次のように最大化することを意味する：

$$\theta = \underset{\theta}{argmax} \sum_{examples} logP$$

パラメータ更新の勾配は，期待される実証的フィーチャ数と学習者フィーチャ数の差であり，次のように表す：

$$\nabla L(\theta) = \tilde{f} - \sum_{\zeta} P(\zeta|\theta,T) f_{\zeta} = \tilde{f} - \sum_{si} D_{s_i} f_{s_i}$$

D_{s_i}は期待される行動の頻度であり，全てのパスを数え上げ，各パスにおいてその状況に行きつく回数を確率的に数えることによって計算する.

2.3　シミュレーション空間と環境モデル

シミュレーション空間では，離散化した都市空間，道路網，場所属性を記述する数値情報，交通シミュレーターによる構成される．本研究では標準地域メッシュにて空間を離散化する.

3.　適用例

3.1 使用データ

実験は東京都市圏を対象として行う.本研究では移動者の意思決定を表す位置情報，環境を表現するデータとして，表 1 に示すものを使用した.

表-1　使用データ詳細

内容	対象データ	詳細
訓練用データ	人の流れデータセット	サンプル数：1000
検証用データ	人の流れデータセット	サンプル数：10000
特徴量	国勢調査	夜間人口
	経済センサス	従業員数
	経済センサス	事業所数
	国土数値情報	道路密度・延長
	国道数値情報	鉄道乗降客数
	国道数値情報	公共施設数

東京都市圏の人の流れの全体を実際に計測することは難しいため，東京都の人の流れを再現しても検証することが難しい．本研究では提案する手法の検証を行うために，PT 調査データの内，1 万人分の人の流れを真値とするような仮想都市圏を想定して行う.

3.2　シミュレーション結果の可視化

10000 エージェントの一日分の行動シミュレーションを行い，その結果を可視化する．図 2 では，エージェントの行動を道路・鉄道ネットワーク上で最短経路ベースで経路探索を行い，1 分ごとの位置情報を内挿した結果である.ここでは，6:00, 8:00, 10:00, 18:00, 20:00, 22:00 のタイムスライスにおける，エージェントの分布を示す. 各点はエージェントとし，当時点の交通手段により色分けしている（青：滞在；黄：徒歩；緑：車；赤：電車）.

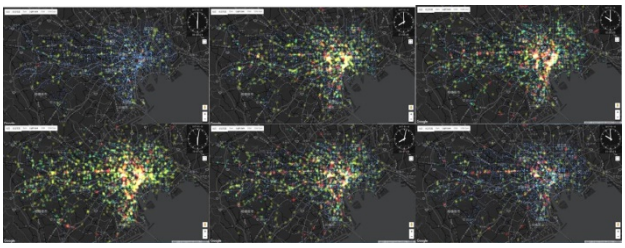

図-2　シミュレーション結果の可視化

3.3　人口分布状況の再現度

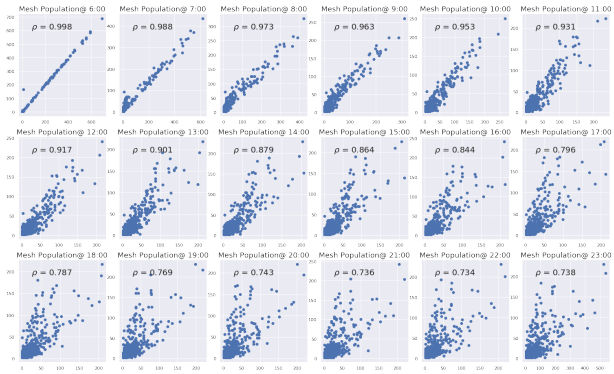

図-3　1Km メッシュベースの人口分布

シミュレーション結果の検証について，まず人口分布の推定精度を検証した. 人口分布については，仮想都市圏の個々メッシュ内の人口数を真値とし，エージェントの分布を比較する. 図 3 では時間ごとのシミュレーション結果と推定人口の

分布を散布図で示す. 図の中の数字は真値と推定値との相関係数である.

推定値と真値との平均相関は 0.8 以上の高精度を示している. 朝 6 時には真値と完全一致しているが，その後通勤時間帯（7 : 00-9 : 00）には分布の広がりは見られるものの，相関がやや低下していく傾向がみられる. また，日中から夕方時間帯にも相関は次第に低下しているが，これは行動範囲の広がりが原因だと考えられる. さらに，夜の帰宅時間帯（17 : 00-19 : 00）における相関の低下は，主に帰宅行動するエージェントは間違った自宅に辿り着くことが原因であると考えられる.

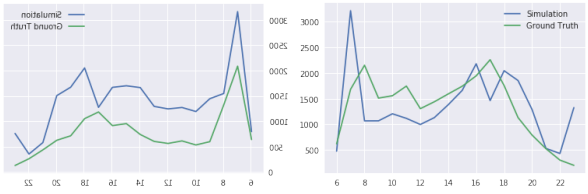

図-4　時間帯別鉄道（左）と車（右）利用者数

時間帯別の鉄道利用者数変化は図 4（左）に示す結果となった． 本手法を用いたことにより通勤時間帯の交通混雑状況を再現できていることがわかる. しかしながら，全体的に鉄道利用者数の推定値は真値より，過大評価されており，特に日中の時間帯は移動者数が少ない様子がうまく反映できなかった. これは「電車」という交通手段のコストが過少評価とされ，エージェントが「電車」で移動すると現実より多い報酬が獲得すると考えられる. 一方，図 4（右）に示している車利用者数は，鉄道利用者数より誤差が少なく，鉄道利用者に比べ，推定精度が高い.

まとめると，実際の人間の軌跡を教師データとした強化学習エージェントは，逐次的に行動意思決定を行い，結果としての模擬行動データセットは対象地域の人口動態を精度よく再現できることが分かった.一方で, エージェントの鉄道や自動車などのモビリティ利用率が正解値より，多く推定される傾向があった.これは，エージェントは「滞在」より，移動すれば多い報酬をもらえることに起因する.また, 学習された報酬関数の特徴量は，必ずしもリアルな人間と一致するわけではない.人間の行動選択は複雑な意思決定プロセスであるが，交通利便性，商業発達，時間帯以外の要素の検討が欠けていることは，エージェントの行動選択がリアルな人間と異なることの要因となっている.これを踏まえ, 報酬関数の設計に関してさらなる検証が必要だと考えられる.

4. おわりに

本研究では，移動軌跡から，データ処理，行動モデル構築，シミュレーションのパイプラインを組み立て，人手を多くかけずにエージェントの行動モデルを構築するフレームワークを提案した. さらに，人の流れデータセットに適用させ，その新たなモデルの逐次的な意思決定枠組みを用いて，人の交通行動シミュレーションを行い，生成した軌跡データの精度検証を行った. 今回の提案手法に対して，様々な位置情報ビッグデータへの応用が可能となる. 今後は，強化学習エージェントモデルのもとで，時系列データの分析・モデリングを考慮することで，より人間の振舞いに近い行動モデルの構築を目指す.

参考文献

Zheng, Y., Lizhu Zhang, Xing Xie, WeiYing Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web.pp. 791-800.

Sekimoto, Y., R.Shibasaki, H.Kanasugi, T.Usui, & Y.Shimazaki. 2011. PFLOW : Reconstruction of people flow recycling large-scale social survey data, IEEE Pervasive Computing, Vol.10, No.4, pp.27-35.

大野夏海，関本義秀，中村敏和，Horanont Teerayut，柴崎亮介，東京都市圏における長期の GPS データを用いた移動経路の推定に関する研究，第 21 回地理情報システム学会講演論文集, Vol.21, CD-ROM, 2012

Ng, Andrew Y., and Stuart J. Russell. "Algorithms for inverse reinforcement learning." Icml. Vol. 1. 2000.

Ziebart, Brian D., et al. "Maximum entropy inverse reinforcement learning." Aaai. Vol. 8. 2008.