LL-006

輸送車両の先見的運行経路設定

Proactive Route Planning for Transport Vehicles

向 直人　　渡邉 豊英
Naoto Mukai　Toyohide Watanabe

１．はじめに

近年，位置情報システムの精度向上や携帯端末の小型化に伴い，オン・デマンドな輸送システムが注目されている．例えば，デマンド・バスは，顧客の要求に応じて発車時間や運行経路を柔軟に変化させるシステムである．従来，これらのシステムは反射的な方法で運行経路を設定してきた．つまり，要求発生後に，顧客の利便性・運行会社の採算性を考慮した上で運行経路を最適化する．また，その最適化の多くは人手によるものである．本稿では，強化学習手法（Q-学習）に基づいた，先見的な運行経路設定手法を提案する．本手法において，輸送車両は学習エージェントである．エージェントはサービス・エリアの顧客発生傾向を学習することによって，報酬を最大にするように経路を選択する．最後に，学習前・後のエージェントを比較し，その有用性を評価する．

２．形式化

本問題を以下のように形式化する．輸送システムのサービス・エリアを式（１）で与える．サービス・エリアはグラフ構造で表される．ノードnは顧客の乗降位置であり，エッジeはノード間の経路である．一方通行や車線数等の制限は考慮しない．

$$\begin{cases} G=(N,E) \\ N=\{n_1,n_2,\cdots\} \\ E=\{e(n,n')\,|\,n,n'\in N\} \end{cases} \quad (1)$$

台数kのエージェント（輸送車両）を式（２）で与える．エージェントは一定速度でグラフ上を移動する．また，本稿では車両の最大乗員数を考慮しない．つまり，車両タイプ（バスやタクシー等）に依る輸送コスト（ガソリン代等）を同一とみなす．

$$V=\{v_1,v_2,\cdots,v_k\} \quad (2)$$

一般に，顧客の発生分布は一様ではない．出勤時間であれば，住宅街からビジネス街へといった傾向が存在すると考えられる．そこで，顧客を式（３），サービス・エリア内に存在する顧客の発生傾向（フロー）を式（４）で与える．顧客cはフローfに従って発生する．ここで，フローfは，乗車位置n_r，降車位置n_d，発生確率ηによって特徴付けられる．

$$C=\{c_1,c_2,\cdots\} \quad (3)$$

$$F=\{f_1,f_2,\cdots,f_m\} \qquad f=(n_r,n_d,\eta) \quad (4)$$

名古屋大学大学院　情報科学研究科　社会システム情報学専攻　〒464-8603　名古屋市千種区不老町

本問題の目的関数を式（５）で与える．顧客cの待機時間を$T_w(c)$，乗車時間を$T_r(c)$で表す．よって，式（５）は顧客の平均輸送時間の最小化を表している．通常，車両台数が制限されたとき，バス・システムであれば，待機時間が短く，乗車時間が長くなる．逆に，タクシー・システムであれば，待機時間が長く，乗車時間が短くなる．よって，サービス・エリア内の発生傾向に応じて，エージェントが自身の振舞を変化させることで，輸送システムの効率向上が期待できる．

$$\min\left(\sum_{c\in C}\frac{T_w(c)+T_r(c)}{|C|}\right) \quad (5)$$

３．学習

提案手法において，輸送車両は学習エージェント[1]である．本節では，強化学習の代表的手法であるQ-学習について述べ，さらに，Q-学習に基づいた輸送車両の先見的経路の獲得手法について述べる．

3.1 Q-学習

エージェントの学習過程は，マルコフ決定過程（MDP）によって表現される．エージェントvは，状態s_tにあるとき，行動a_tを選択することにより，状態s_{t+1}に遷移する．このときの遷移確率は，式（６）に示す条件確率Pで与えられる．

$$P=P\{s_{t+1}\,|\,s_t,a_t\} \quad (6)$$

エージェントはある状態に到達すると報酬R_tを得る．ここで，状態sにおいて，行動aを選択する確率を政策$\pi(s,a)$と表す．マルコフ決定過程における，エージェントの目的は，式（７）に示す割引報酬和を最大化する政策$\pi^*(s,a)$を獲得することにある．パラメータγは割引率と呼ばれ，値が小さければ，即時報酬を重視し，逆に，値が大きければ，将来獲得するであろう報酬を重視する．

$$V_t^\pi=R_t+\gamma\cdot R_{t+1}+\gamma^2\cdot R_{t+2}\cdots \quad (7)$$

一般的に，割引報酬和V_t^πは既知ではない．よって，その推定値$Q(s_t,a_t)$をエージェントの経験に従って更新することで，割引報酬和V_t^πの値に近付ける．その代表的手法であるQ-学習は式（８）に従って推定値$Q(s_t,a_t)$を更新する．パラメータαは学習率と呼ばれ，学習の進行速度を調整する．学習率αを適切な値に設定することで，推定値$Q(s_t,a_t)$が収束することが証明されている．

$$Q(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha\left[R_t+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1})\right] \quad (8)$$

3.2 先見的経路の獲得

前節で述べたQ-学習を，我々の輸送システム問題に適用するために，“状態-行動”，“報酬”，“刑罰”を定義しなくてはならない．本節では，それらの定義を順に述

べ，次に，得られた推定値$Q(s_t,a_t)$に基づき，どのように経路を選択するかを述べる．

本稿では，エージェントはノード間を時間T_eで遷移すると仮定する．つまり，エッジ長を全て同一とみなす．よって，時間T_eはマルコフ決定過程（ＭＤＰ）における単位時間となる．

エージェントの状態を，現在ノードn_tと，そこに至る長さδのノード履歴$(n_{t-\delta},\cdots,n_{t-1})$の列で与える．よって，状態$s_t$は式（９）で表される．

$$s_t=(n_{t-\delta},\cdots,n_{t-1},n_t) \qquad (9)$$

ここで，ノードnに隣接するノードの集合を$L(n)$と表す．状態s_tにおいて選択可能な行動は，式（１０）で示すように，隣接ノードで与えられる．

$$a_t=n_{t+1}\in L(n_t) \qquad (10)$$

これらから，履歴長δが１のとき，状態と行動の組は式（１１）で与えられる．これは，エージェントの経路選択は，現在ノードのみでなく，どの経路を辿ったかに依存することを表している．図１に具体例を示す．エージェントv_1，v_2が共にノードn_cに位置している．しかし，エージェントv_1はノードn_aを経由，一方，エージェントv_2はノードn_bを経由して，ノードn_cに到達している．このとき，エージェントv_1とv_2の状態は異なり，行動n_d，n_eを選択する確率も異なる．

$$(s_t,a_t)=(n_{t-1},n_t,n_{t+1}) \qquad (11)$$

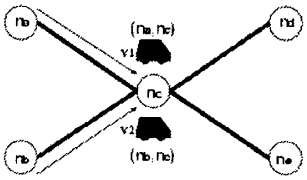

図１．エージェントの状態と行動

エージェントがノードnに到達したときに得られる報酬を式（１２）で与える．ここで，$|C_r(n)|$はノードnで乗車する顧客の人数であり，ω_rはその重み係数である．また，$|C_d(n)|$はノードnで降車する顧客の人数であり，ω_dはその重み係数である．ここで，重み係数のバランスは，車両の輸送タイプに影響する．例えば，乗車係数ω_rを大きく，降車係数ω_dを小さくすると，エージェントはバスのように多くの顧客を乗車させることを好む．逆に，乗車係数ω_rを小さく，降車係数ω_dを大きくすると，エージェントはタクシーのように顧客を素早く降車させることを好む．

$$R(n)=\omega_r\cdot|C_r(n)|+\omega_d\cdot|C_d(n)| \qquad (12)$$

エージェントがノードnに到達すると，ノードnで乗降車する顧客の要求は満足される．よって，到達直後のノードnで得られる報酬は０である．時間が経過することによって，顧客がある確率に従って発生し，本来の報酬期待値に近付くと考える．そこで，本手法では，ノードnのアイドル時間$idle(n)$に従って，推定値$Q(s_t,a_t)$を低く見積もる．ここで，ノードnのアイドル時間$idle(n)$とは，エージェントがノードnに到達してから経過した時間を表す．刑罰関数$P(a_t)$は，式（１３）で与えられ，推定値$Q(s_t,a_t)$の重み係数として用いられる．ζは刑罰を与える最大回数を表している．刑罰関数$P(a_t)$は，時間経過と共に，０から１に増加する．最大回数ζを超えると常に１となる（$P(a_t)=1$は本来の報酬期待値を表す）．

$$P(a_t)=\log\left(\frac{(idle(n_{t+1})-T_e)(e-1)}{\zeta\cdot T_e}\right) \qquad (13)$$

ここまで定義した，“状態-行動”，“報酬”，“刑罰”によって，推定値$Q(s_t,a_t)$が定まる．エージェントの行動選択はε-ルーレット手法（$0<\varepsilon<1$）で決まる．すなわち，確率εで，エージェントはランダムに行動を選択する．また，確率$1-\varepsilon$で，エージェントは，式（１４）で示される，ルーレット手法で行動を選択する．

$$\Pr(s_t,a_t)=\frac{P(a_t)\cdot Q(s_t,a_t)}{\sum_{a'_t\in L(n_t)}P(a'_t)\cdot Q(s_t,a'_t)} \qquad (14)$$

４．実験

提案手法の有効性を評価する．実験環境を以下に述べる．車両は１台とし，ノード数$|N|$=１０，エッジ数$|E|$=１５で構成されるグラフGをサービス・エリアとする．また，１０パターンの発生傾向（$F_1,\cdots,F_{10}$）をランダムに生成した．各パターンは異なる発生分布から構成されるが，発生する顧客の総数は同数である．他のパラメータ設定を表１に示す．

表１．パラメータ設定

学習率α	0.1	乗車係数ω_r	1
割引率γ	0.5	降車係数ω_d	1
履歴長δ	1	刑罰数ζ	9

本実験では，２種類のエージェントを比較する．固定経路エージェント（Fixed Route Agent）は，全ノードを最短距離で巡回する．この経路は，顧客発生が一様分布であるときの，最適戦略である．学習経路エージェント（Learned Route Agent）は，提案手法によって獲得した，政策に従って行動する．図２は，顧客の輸送時間（待機時間＋乗車時間）を示している．８パターンにおいて，学習経路エージェントが優位性を示した．発生分布の偏りが顕著である程，提案手法によって獲得された政策が有効となった．

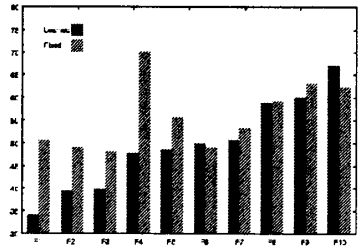

図２．輸送時間（単位時間T_e=5）

５．まとめ

本稿では，輸送システムを対象とし，従来の反射的手法とは異なる先見的手法に注目した．提案手法は，強化学習に基づき，先見的な運行経路の獲得を可能にした．

参考文献

[1] Santana, H.; Ramalho, G.; Corruble, V.: and Ratitch, B. 2004. Multi-Agent Patrolling with Reinforcement Learning. In Proceedings of International Conference on Autonomous Agents and Multi-Agents Systems, 1120-1127.