The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

深層強化学習を用いたWebサイト内行動のレコメンド
Deep Reinforcement Learning for Recommender System of Users' Behavior on Website

蓑田 和麻[*1] Kazuma Minoda　阿内 宏武[*1] Hiromu Auchi　川頭 信之[*2] Nobuyuki Kawagashira　石川 信行[*1] Nobuyuki Ishikawa

[*1] 株式会社リクルートテクノロジーズ Recruit Technologies Co.,Ltd.

[*2] 私的著者 Private Author.

For the companies running websites, we need appropriate communication to the visiting users according to their situation such as pages they have visited so far. To achieve that purpose, we should have some strategies to suggest the users to take desirable actions on the website. As one of the strategies, recommendation of users' behavior can be considered. When we recommend some actions not directly related from conversion pages (e.g. reservation pages, purchase pages, etc.) on website such as selecting the searching conditions, it is impossible to recommend the best action that leads to the user's reservation behavior by using the conventional supervised learning methods. In this research, we solve the above problem using deep reinforcement learning method and show its effectiveness by an experiment for actual web access log of users' behavior.

1. 背景・課題

近年，インターネット上に存在する様々な Web サイトや SNS から発信される情報により，Web 上における行動の選択肢が増え，ユーザが求める情報に辿りつくことが困難な傾向がある．Webサイトを運営する企業側は，コンバージョン(旅行予約サイトであれば宿の予約のような,自サイトで目的となる行動，以下 CV)までつながるようなコミュニケーションを Web サイト上で行い，ユーザ体験を向上させることが重要である．つまり「ユーザが求める情報にたどり着くことができること」と「運営するサイト上で目的となる行動を行ってもらうこと」のユーザ側と企業側両者のニーズを満たすことが重要である．

企業側にとって，ユーザを CV に導くためにユーザ体験を向上させ，サイトを訪れるユーザが求めるコンテンツにたどり着けるようなサイトストラクチャ・デザイン・機能を構築することが必要不可欠になっている．しかし，Web サイトの使い方が分からず離脱するユーザや，訪れたものの何をして良いかわからない，あるいは好みのコンテンツが見つからず離脱するユーザが多数存在する．これらの課題については，サイト内で適切なコミュニケーション(例．次に行くべき導線を示す or 最適な検索条件をレコメンドする等)を行うことで解決可能であると考える．一般的なユーザ行動は，サイトを訪問し，検索を行い，アイテムに興味を持ち，興味を持ったアイテムの比較検討を行い，CV するというプロセスで構成される．この中で，検索行動は CV とは直接的な関係を持たないが，CV に至るための重要なプロセスである．

一方で，ユーザを CV に導くことを目的とした従来の施策では，CV と直接的な関係がある「アイテムのレコメンド」や「ソート順の改善」に終始していた．これは，従来の教師あり学習のような手法が，CV との直接的な関係を学習できるためである．しかし，ユーザが次に行くべき導線への誘導や，最適な検索条件のレコメンドの場合は，CV との間接的な関係を学習する必要がある．そのため，報酬(=CV)が遅れて得られる場合のレコメンドは教師あり学習のような手法では不十分であり，実施されていなかった．

本研究では，上記の課題を，深層強化学習を用いて解決し，実際の Web アクセスログデータを用いた実験により，その有用性を示す．

2. 関連研究

近年，強化学習における方策関数や行動価値関数を DNN(Deep Neural Network)を用いて近似する「深層強化学習」に関する研究が盛んに行われている．DNN を用いた強化学習として最も基本的なアルゴリズムである DQN(Deep Q Network)[Mnih 2013]，強化学習における状態・アクションが連続値の場合に適用可能な DDPG(Deep Deterministic Policy Gradient)[Lillicrap 2016]などはよく用いられ，本論文における提案手法の学習アルゴリズムとして利用した．また，深層強化学習をレコメンデーションに適用した研究として，Zheng[Zheng 2018]らによるニュース記事のレコメンド，Zhao[Zhao 2018]らによる page-wise レコメンドの研究などが挙げられる．上記の研究はアイテムのレコメンドに強化学習を適用した例であり，CV との間接的な関係，すなわち報酬が遅れて得られる場合の議論は行われていない．また，Web における機械学習を用いた各種レコメンドは広く研究されているが，Web 上でのアクセスログに対し強化学習を活用し，報酬をサイト上における予約行動とした上での最適な行動のレコメンドについての研究はまだ少ない．

3. 問題設定

リクルートが運営する Web サイトは主に，企業がアイテム(「じゃらん https://www.jalan.net/」であれば宿，「SUUMO https://suumo.jp/」であれば不動産物件 など)を掲載し，ユーザは掲載されたアイテムに対して検索を行い CV する，といったユーザ行動により成り立っている．Web サイトを運営する企業側は日々，「デザインの変更」や「新規機能の追加」など Web サイト改善のために様々な施策の検討・開発を行っているが，ユーザとクライアントのマッチングを増やす(= CV を増やす)という目的は共通している．上記の目的のために Web サイト上でレコメンド等の施策を展開する際，用いるデータとしてはアイテム情報とユーザ情報(アクセスログ情報)の 2 種類に大別される．

本論文では Web 上でリアルタイムにユーザにレコメンドを行うことを想定しているため，リアルタイムに取得可能な情報を強化学習における状態に用いる必要がある．深層強化学習モデルの学習には，すでに蓄積されているアイテム情報とユーザのアクセスログ情報を用いる．また，本論文でレコメンド対象とするサ

イト内行動は「検索条件の指定」とした．一章で述べた通り，検索条件の指定はCVとは間接的な行動だが，CVにたどり着くために重要な絞り込みの機能である．

上記の問題設定のもと，深層強化学習を用いたWebサイト内行動のレコメンドアルゴリズムを提案し，その妥当性について検証する．

4. 強化学習によるサイト内アクションのレコメンド

本章では，Web サイトにおけるサイト内アクションのレコメンドを実現するための提案手法について述べる．

4.1 強化学習の構成定義

(1) エージェントとの定義

強化学習では，ある状態における価値を最大化するようにエージェントを学習させる．Web サイトにおけるエージェントとは，レコメンドを行うシステムのことを指す．本システムは Web サイトを訪問したユーザの各状態（4.1-(2)）において，将来的に得られる報酬（4.1-(4)）が最大になるような方策を学習する．

(2) 状態の定義

将棋や囲碁に代表されるボードゲームにおいて，状態はマス目に置かれた駒の情報であり，状態の表現に必要な情報は現在の盤面の状況のみで表現可能である．しかしながらWebサイトの場合は，ユーザ一人ひとりの状態を表現するための情報は，ユーザが「閲覧したアイテム情報(“じゃらん”の場合は温泉付きや禁煙などの宿に関する付帯情報)」，「ユーザが過去に指定した検索条件」，「アクセスした日時」，「流入元サイト」，「アイテムページ閲覧数」など，様々な変数により表現可能である．本研究では事前に解析を行い，ユーザの状態表現に有効と思われる変数を抽出した．

また，ユーザの状態を表現する際，過去に遷移してきたページも重要な情報である．例えばユーザのページ遷移を状態として扱う際は，各ページをone-hot-vectorで表現し，ユーザが過去に遷移したページの論理和で表現する．ユーザが過去に遷移したページ数を *step_size*，*t* 回目の遷移ページの one-hot-vector を $\boldsymbol{p}_t$ とすると，状態 $\boldsymbol{x}_t$ は以下の式(1)で定義できる．

$$\boldsymbol{x}_t = \begin{cases} \boldsymbol{p}_t \mid \boldsymbol{x}_{t-1} & (1 \le t \le step\_size) \\ \boldsymbol{p}_0 & (t = 0) \end{cases} \quad (1)$$

上記の状態に，サイト独自の変数を追加し，ユーザ一人ひとりの状態を表現する変数を定義した．

(3) アクションの定義

本研究では，アクションをサイトにおける検索条件に設定した．ユーザはサイト内において検索条件を設定し検索行動を行う．この検索条件の設定により表示されるアイテムが変わり，ユーザがコンバージョンページにたどり着けるかどうかが大きく変わる．そこで，強化学習を用い，ユーザ一人ひとりに対して 4.1-(2)で定義した各状態において最適な検索条件(= アクション)をレコメンドするシステムを構築する．4.1-(1)で定義したエージェントは，ある状態において各アクションに対するQ値を算出し，それらの最大値をとるアクションをレコメンドする．

(4) 報酬の定義

ユーザが Web サイトで CV した際にプラスの報酬，ユーザが CV せずに離脱した際にはマイナスの報酬を与える．本研究では前述の報酬を定義したが，状態の定義と同様，サイトによりゴールとする指標は異なるが，今回実施する検索条件のレコメンドは CV 数増加を目的とした施策のため，上記の定義とした．

4.2 深層強化学習を用いた学習方法

(1) DDPG を用いた学習

本論文では，Web サイト内アクションのレコメンドを実現するために，DDPG を採用した．一般的な強化学習アルゴリズムとしてよく用いられる Q-Learning や DQN は，アクションを離散値で与えて学習を行う．今回の検索条件をアクションとしたタスクにおいては，複数の検索条件を指定する，すなわち複数のアクションが同時に起こるという問題を扱う必要がある．この問題を DQN で実現するためには，全てのアクションの組み合わせを別のアクションとして定義する必要がある．今回対象とした検索条件は 42 種類のため，アクションは(2^{42}-1)通りになり，計算量的に不可能である．

DDPG は actor-critic アルゴリズムを用い，各状態におけるアクションを決定する actor と，状態と actor で出力したアクションを入力として方策を評価する critic により安定した学習を行う．この actor 関数は各状態におけるアクションを連続値として算出するため，DQN では実現できなかった複数アクションの表現が可能であるため，本研究に適している．

(2) エピソードの定義

本研究では，あるユーザの一定期間のアクセスログを 1 エピソードとして強化学習を行う．ユーザは 1 セッション(サイトへの訪問から離脱までのアクセス)で予約を行うとは限らず，複数セッションに渡って検索行動を行いコンバージョンまたは離脱に至る．今回はドメインの特性上 7 日間を一区切りと設定し，一人のユーザが 7 日間の間隔を空けてアクセスした場合は別のエピソードとして学習を行う．

(3) アクションのレコメンド

ある状態において，4.2-(1)で述べた actor 関数が出力するアクションベクトルの中で最も高い値をとる次元のアクションをレコメンドする．状態 s のユーザにレコメンドするアクションを A(s)とする．$\mu(\cdot|s)$を学習済みの actor ネットワークとすると，A(s)は以下のように表せる．

$$\boldsymbol{A}(\boldsymbol{s}) = \underset{\boldsymbol{a}}{\boldsymbol{argmax}}\ \boldsymbol{\mu}(\cdot|\boldsymbol{s}) \quad (2)$$

5. 実験

本章では，提案手法を用いたサイト内アクションのレコメンドによるCVへの貢献度についてオフラインで検証を行う．実験パターンと比較手法は複数用意し，提案手法の有効性を証明する．以下で評価方法と比較手法について述べる．

5.1 評価方法

実際のユーザの状態を入力し，そのユーザが CV したか否かを判定する，教師あり学習を用いた予測モデルを作成する．手法は SVM（Support Vector Machine）を用いた．このモデルは状態ベクトルを入力し，CV 確率を予測値として出力する．このモデルを活用し，提案手法によりユーザのアクションを変更させた場合の状態をインプットにすることで，出力値である CV 確率がどの程度向上するかを観察する．この上昇値が本来のユーザの行動に比べて大きいほど CV する可能性が高まり，良いレコメンドであると言える．

上記の評価を下記の手法ごとに比較して，提案手法の有効性を示す．

1. No_recommend
   レコメンドなし（実際のユーザの状態）
2. Random
   ランダムで検索条件をレコメンド
3. Fixed
   集計結果から CV 確率が高い，検索軸 1 件を全員にレコメンド
4. Supervised
   CV したユーザのログから状態をインプット，検索軸をアウトプットにした教師あり学習モデルによる出力値を活用したレコメンド
5. DQN
   学習方法に DQN を用いたレコメンド
6. DDPG_1 ~ DDPG_5
   提案手法によるレコメンド．パラメータを 5 種類用意

4.2 章(1)でアクションの連続性に対応できないという DQN のデメリットについて言及したが，実験で用いた DQN では複数アクションがある場合，ランダムで 1 つのアクションを選択し学習することで次元数を削減した．DDPG_1～DDPG_4 の Actor 及び Critic のニューラルネットワーク構成は同じであるが，DDPG_5 は，Actor のニューラルネットワーク構成において第 0 層からの次元圧縮と中間第2層の次元数が異なっている．DQN のニューラルネットワーク構成は，DDPG_1 のポリシーネットワークと同じ構成である．DQN，DDPG_1～DDPG_5 の割引率 γ は，それぞれ 0.80，0.80，0.85，0.90，0.98，0.85 であり，ターゲットネットワークの更新間隔は，200，200，1000，1000，1000，200 となっている．

また，DQN，DDPG_1～DDPG_5 の学習には，232637 エピソード，2937642 ステップの Web アクセスログを用いた．検証用の SVM の学習にはステップ数が 20 以上のエピソードのみを用いたため，5272 エピソードで学習を行い，1336 エピソードでテストを行った．

5.2 結果

Random(ランダムレコメンド)，Fixed(固定レコメンド)，Supervised(教師あり学習レコメンド)，DQN を用いたレコメンドの場合は，No_recommend(レコメンドなし)と比較して平均 CV 確率が低下した．提案手法の DDPG によるレコメンドでは，平均 CV 確率が上昇した．DDPG_1～DDPG_4 が同様に高く，中でも DDPG_4 がもっとも高い平均 CV 確率を示した．

表 1．予測 CV 確率の平均と標準偏差の比較結果

手法	平均 CV 確率	CV 確率標準偏差
No_recommend	0.39019	0.11512
Random	0.38978	0.12347
Fixed	0.38745	0.11386
Supervised	0.38994	0.11384
DQN	0.38917	0.11539
DDPG_1	**0.40189**	0.12393
DDPG_2	**0.40195**	0.12390
DDPG_3	**0.40197**	0.12405
DDPG_4	**0.40205**	0.12414
DDPG_5	**0.39438**	0.11364

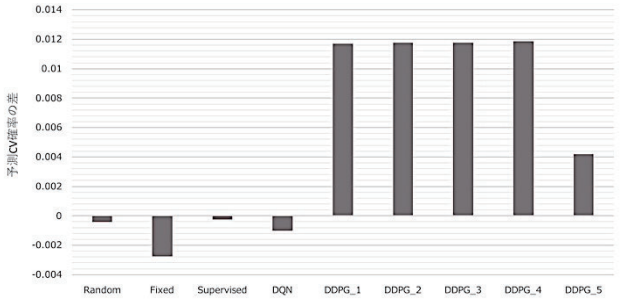

図 1．No_recommend と平均 CV 確率の差

6. 考察

DDPG によるレコメンドではいずれの条件でも平均 CV 確率を上昇させることが分かった．DDPG_4 が最大の平均 CV 確率となるのは，割引率 γ = 0.98 が高いため，CV の影響がかなり以前のステップ状態にまで波及しているからと考えられる．一方 DQN の平均 CV 確率が下降したのは，DQN は複数のアクションから 1 つをランダムに選択する処理を行う必要があったために精度が落ちたと考えられる．一方，DDPG は Actor-Critic モデルを採用しているためアクション定義の自由度が高いので，Web サイトへの強化学習の応用に適していると考えられる．ただし，DDPG の場合，深層ニューラルネットワークの構成の違いによって精度が大幅に変わってくるので，適切なニューラルネットワークの設計が重要である．

また，入力と出力が直接的な関係を持っている問題に対しては教師あり学習は有効であるが，Web サイトのユーザ行動のような入力(状態)と出力（CV）が間接的な関係を持っている問題に対しては，強化学習の方が，状態・アクションと CV との関係性を学習することができたと考えられる．これにより，ユーザ一人ひとりの状態に応じて，ユーザを CV に導くための適切なアクションをレコメンドすることが可能になった．

7. まとめ

Web サイトにおいて，検索軸のレコメンドのような CV からは遠い施策においては，従来の教師あり学習では CV との直接的な関係性を学習するため，限界があった．しかし，本論文で提案した CV を報酬として設計した強化学習モデルにより，有効なレコメンドが可能になった．さらに CV 確率を求めるモデルを作成し，その出力を評価指標にしたところ，提案手法である DDPG によるレコメンドが最も効果が高くなることが示された．

現在，上記のようなオフラインの検証ではなく，実際の Web サイトに導入し，A/B テストによる効果計測を始めている．今後，ユーザの状態をより詳細に定義する変数の作成や，強化学習のメリットの 1 つであるオンラインでの「探索」の仕組みについても検討したい．

参考文献

[Lillicrap 2016] Timothy P. Lillicrap, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, Daan Wierstra: Continuous control with deep reinforcement learning, International Conference on Learning Representations (2016)

[Mnih 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller: Playing Atari with Deep Reinforcement Learning, NIPS Deep Learning Workshop (2013)

[Zheng 2018] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. In Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 167–176 (2018)

[Zhao 2018] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang.2018 . Deep Reinforcement Learning for Page-wise Recommendations. In Twelfth ACM Conference on Recommender Systems (RecSys '18) (2018)

Appendix

CV ユーザと非 CV ユーザでの, ステップごとの平均 Q 値について比較した.

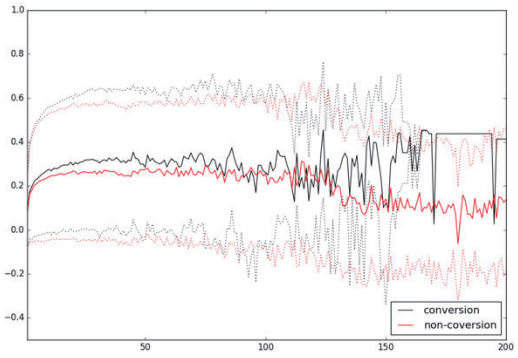

図 2.【 DDPG_4】ステップ毎の平均 Q 値と標準偏差(横軸:ステップ数, 縦軸:Q 値)

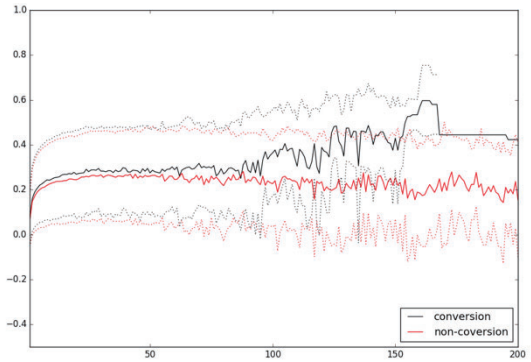

図 3.【 DDPG_5】ステップ毎の平均 Q 値と標準偏差(横軸:ステップ数, 縦軸:Q 値)