Vol.2012-MPS-88 No.8
2012/5/17

推薦システムのための状態遷移確率の構造を未知としたマルコフ決定過程

桑田修平[†1,†2] 前田康成[†3]
松嶋敏泰[†1] 平澤茂一[†4]

推薦問題を扱うためのより一般化されたマルコフ決定過程モデルに対して，ベイズ基準のもとで最適な推薦ルールを履歴データから求める方法を提案する．提案法の特徴は，ある商品を推薦した後に何が買われたのかを考慮していること，さらに，一回の推薦結果だけでなく一定期間内に行った複数の推薦結果を評価している点にある．ここで，従来の推薦手法と大きく異なる点は，推薦ルールを求めるためのプロセスを統計的決定問題として厳密に定式化したことにある．その結果，推薦する目的に対して最適な推薦が行えるようになった．人工データを用いた評価実験により，提案する推薦手法の有効性を示す．

Variable Order Transition Probability Markov Decision Process for the Recommendation System

Shuhei Kuwata,[†1,†2] Yasunari Maeda ,[†3]
Toshiyasu Matsushima [†1] and Shigeichi Hirasawa[†4]

In this paper, we proposed a general markov decision process model for the recommendation system. Furthermore, based on the bayesian decision theory, we derived the optimal recommendation lists from the proposed model using historical data. Our method takes into account not only the purchased items but also the past recommended items within a given period. Here, the unique thing about this paper is that we formulate the process to get the recommendation lists as the statistical decision problem. As a result, we can obtain the most suitable recommendation lists with respect to the purpose of the recommendation. We show the experimental results by using artificial data that our method can obtain more rewards than the conventional method gets.

1. はじめに

これまで，推薦手法に関する多数の研究が行われてきており，特に，Amazon.com 等の EC（Electronic Commerce，電子商取引）サイトにおいては既に実用化が進んでいる．また，最近では，Hadoop を利用することで，大規模なデータに対しても簡単に推薦方式を実装できるようにもなってきている．

従来の推薦手法は，以下のように 3 つのタイプに分けて説明することができる[1)]：

1. メモリーベースアルゴリズム
2. モデルベースアルゴリズム
3. ハイブリッドアルゴリズム

ここで，上記 3 つのタイプに属す殆どの従来法に共通する特徴として，商品を推薦した結果を考慮していない点を挙げることができる．つまり，過去に購入した商品履歴（以降，購入商品履歴と呼ぶ）のみから次に推薦する商品を決める手法が殆どであり，ある商品を推薦した結果どのような商品が購入されてきたかを踏まえて，次に推薦する商品を決めていない．つまり，推薦した商品の履歴（以降，推薦商品履歴と呼ぶ）を考慮していない．

また，別の共通点として，推薦は 1 回のみ行うことを想定している点が挙げられる．しかし，会員制の EC サイト等を考えると，同じユーザに対して，推薦は 1 回限りではなく複数回，継続的に行うことが想定できる場合もある．

†1 早稲田大学
Waseda University
†2 株式会社 NTT データ
NTT DATA CORPORATION
†3 北見工業大学
Kitami Institute of Technology
†4 サイバー大学
Cyber University

そこで，本論文では，上記 2 点を考慮した推薦手法を提案する．すなわち，ある商品を推薦した後に何が買われたのかを考慮し，さらに，一時点の推薦結果だけでなく一定期間内に行った複数の推薦結果を評価する推薦手法を提案する．

ここで，その 2 点を考慮した従来法として文献 3) がある．文献 3) では，マルコフ決定過程をベースにした推薦手法を提案しており，推薦商品履歴や推薦を複数回行うことが考慮されている．具体的には，直近に購入された 3 つの商品からなる順列をマルコフ決定過程モデルの 1 つの “状態” と見なし，次に購入される商品は，1 時点前の状態とその時に推薦された商品によって確率的に定まるものと仮定する．そして，その仮定のもとで，将来に渡って得られる “利得” を最大化する “定常政策” を求めている．ここで，定常政策は推薦ルールに該当し，商品 3 個分の購入商品履歴ごとに推薦する商品が 1 つ定まる．

これに対して，本論文では，商品購入履歴と推薦商品履歴を考慮するための，より一般化されたマルコフ決定過程モデルを提案する．さらに，提案するモデルに対して，事前に得られている履歴データを用いて，最適な定常政策（推薦ルール）を求める方法を提案する．ここで，マルコフ決定過程モデルベースの従来法[3] を含め，従来の推薦手法と大きく異なる点は，推薦ルールを求めるためのプロセスを統計的決定問題として厳密に定式化したことにある．本論文では特に，ベイズ決定理論に基づいて最適な推薦ルールを求める方法を提案する．提案法を用いることにより，推薦する目的に合わせて，統計的決定の観点で常に最適な推薦が行えるようになる．

本論文の構成は次のとおりである：まず，2 節において，本論文で扱う推薦問題を定義し，3 節で，提案法がベースとして用いるマルコフ決定過程モデルの概要を説明する．続いて 4 節で，マルコフ決定過程モデルをベースにした従来法[3] を説明した後，5 節で一般化したマルコフ決定過程モデルを提案し，さらに，統計的決定理論に基づいて最適な推薦ルールを導出する方法を提案する．6 節で人工データを用いた評価を行い，最後に 7 節でまとめる．

2. 問題設定

本節では，本論文が対象とする推薦問題を定義する．まず，ユーザの購入商品履歴とそのユーザに対する推薦商品履歴の 2 種類の履歴データが既に N 人分あるものとする．ただし，両履歴においては，ユーザ i ごとに以下に示すような順番が分かっているものとする．

$$a_{n_i-1(i)}, x_{n_i-1(i)}, \ldots, a_{-2(i)}, x_{-2(i)},$$
$$a_{-1(i)}, x_{-1(i)}, a_{0(i)}, x_{0(i)}, \quad i = 1, 2, \ldots, N.$$

ここで，$a_{t(i)}(t = \cdots, -2, -1, 0)$ は，時点 t においてユーザ i に対して推薦された商品を表し，x_t は商品 a_t が推薦された後のそのユーザ i の反応（推薦されたその商品を購入する等）を表すものとする．n_i はユーザ i の履歴数を表す．また，購入商品と推薦商品はいずれも同じ商品集合 $\mathcal{I}$ に含まれるものとし，

$$x_{t(i)}, a_{t(i)} \in \mathcal{I} = \{1, 2, \ldots, I\},$$
$$t = \ldots, -2, -1, 0, 1, 2, \ldots, \quad i = 1, 2, \ldots, N,$$

時点 t までのユーザ i の購入商品履歴，および，推薦商品履歴をそれぞれ $x^t_{(i)}, a^t_{(i)}$ と表す．ただし，購入商品履歴 $x_{t(i)}$ として，“何も購入しない” も含むものとする．さらに，両履歴をまとめた履歴データを $\mathcal{D}_{(i)}$ で表し，ユーザ N 人分をまとめて $\mathcal{D}$ で表すものとする：$\mathcal{D} = \{\mathcal{D}_{(1)}, \mathcal{D}_{(2)}, \ldots, , \mathcal{D}_{(N)}\} = \{x^0_{(i)}, a^0_{(i)}\}^N_{i=1}$.

本論文では，履歴データ $\mathcal{D}$ が与えられたもとで，履歴 (x^0, y^0) を持つ推薦対象ユーザに対する推薦商品を自動で決めるためのルール（推薦ルール）を求める問題を考える[⋆1]．ここで，推薦対象ユーザは，履歴データ $\mathcal{D}$ に含まれる N 人のユーザとは異なるユーザであるものとする．また，履歴データ $\mathcal{D}$ から推薦ルールを求めた後は，その推薦ルールを更新することは考えず，かつ，ユーザに対して複数回の推薦を行うことを想定する．以下に，本論文で想定する推薦の流れを整理する：

1. 履歴データ $\mathcal{D}$ を蓄積する．
2. N 人分の履歴データ $\mathcal{D}$ から推薦ルールを求める．
3. 2. で求めた推薦ルールを用いて，推薦対象ユーザに対して商品を 1 個推薦する．
4. 推薦対象ユーザが反応を示す（商品を購入する，何も購入しない等）．
5. （3. と 4. を繰り返す）．

以上の設定のもとで，本論文では，時点 $t=1$ 以降に購入された商品 $x_1, x_2, \ldots$ がもたらす利益を最大にする推薦ルールを求めることを目的とする．すなわち，推薦する目的は，将来に渡って得られる利益を最大化することであり，かつ，将来に渡って得られる利益を最大化するような推薦ルールを求めることを本論文の目的とする．

3. 準備：マルコフ決定過程モデル

本節では，従来法[3]，及び，提案法がベースとして用いるマルコフ決定過程モデルの概要を説明する．ここで，(有限）マルコフ決定過程モデルは，以下の 4 つの要素で構成される確率過程である：

- 有限状態集合：$\mathcal{S} = \{1, 2, \ldots, S\}$,
- 有限行動集合：$\mathcal{A} = \{1, 2, \ldots, A\}$,
- 状態遷移確率：$\{p(s|s', a)|s, s' \in \mathcal{S}, a \in \mathcal{A}\}$,
- 利得集合：$\{r(s, a)|s \in \mathcal{S}, a \in \mathcal{A}\}$.

⋆1 本論文では，履歴データに関するユーザと推薦対象ユーザの嗜好（データの傾向）は同一であると見なす．嗜好が異なるユーザが含まれる場合には，ユーザを事前にクラスタリングすることで類似した嗜好を持つユーザ群に分割することなどが考えられる．

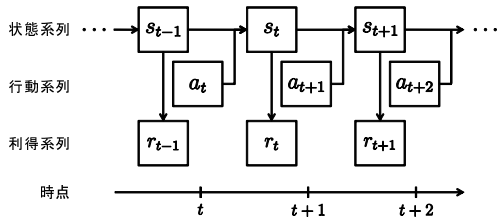

図 1 マルコフ決定過程モデルにおける変数間の関係を表すイメージ図
Fig. 1 The image of the relations between variables on Markov decision process

各構成要素間の関係を**図 1** に示す．図 1 が示す通り，時点 t の状態 $s_t \in \mathcal{S}$ は，1 つ前の時点の状態 $s_{t-1} \in \mathcal{S}$ と時点 t での行動 $a_t \in \mathcal{A}$ のみに依存して確率的に定められる．つまり，時点 t の状態 s_t は，条件付確率 $p(s_t|s_{t-1}, a_t)$ に従って定まる（この条件付き確率は状態遷移確率と呼ばれる）．ただし，時点 t における行動 a_t は，時点 t での状態 s_t に基づいて決定される．このとき，状態に基づいて次の行動を定めるルールを政策 $d(s_t)$ と呼ぶ．さらに，行動 a_t を選択したもとで状態 s_t に遷移した場合には，利得 $r(s_t, a_t)$ が得られる．

上に示した 4 つの要素全ての値が既知であるもとで，最適な政策 $d(s_t)$ を求める種々の方法が提案されている（価値反復法，動的計画法など[2)]）．ここで，最適な政策とは，以下の式で表される割引総利得（一定期間の間に得られる利得の総和），

$$\sum_{t=1}^{T} \gamma^{t-1} r(s_t, a_t), \tag{1}$$

を最大化する政策であることを意味する．ただし，現在の時点を $t=0$ とし，$\gamma (0<\gamma<1)$ は割引率を表す．式 (1) は，一定期間内に得られる全ての利得において，直近に得られる利得ほど重視することを意味している．

4. 従 来 法

本論文で設定した問題に対する従来法として文献 3) がある．具体的には，以下のような対応付けを行うことで，マルコフ決定過程モデルを推薦問題に適用している：

- ユーザが購入する商品 $x_t (t=\ldots,-2,-1,0,1,2,\ldots)$ は，以下に示す状態遷移確率に従うものとする：
$$x_t \sim p(x_t|x_{t-3}, x_{t-2}, x_{t-1}, a_t; \theta). \tag{2}$$
ここで，θ は，状態遷移確率を規定する未知のパラメータである．また，式 (2) は，直前に購入された 3 つの購入商品履歴 $(x_{t-3}, x_{t-2}, x_{t-1})$ とその時に推薦された商品 a_t に依存して，次の商品 x_t が選択されることを表している．
- 商品 x_t が購入されることで得られる利益を，時点 t における利得 $r(x_t)$ とする．
- 将来に渡って得られる利益を，以下の割引総利得として表現する：$\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t)$.
- 履歴 (x^0, a^0) を持つ推薦対象ユーザに対する推薦商品を定める推薦ルールを，定常政策 d として表現する：$a_t = d(x_{t-3}, x_{t-2}, x_{t-1})$．ここで，定常政策とは，時点に依存せずに当該時点の状態のみによって選択すべき行動が定まる政策である．つまり，直前に購入された 3 つの購入商品履歴のみから，その時点での推薦商品が定まる．
- 推薦した結果，何も購入されなかった場合には利得を 0 とする．さらに，状態は変化しないものとする．つまり，前の時点と同じ状態に基づいて次に推薦する商品を定める．

上記の対応付けは，商品 3 個分の購入商品履歴 (x_{t-2}, x_{t-1}, x_t) を 1 つの状態 s_t と見なしたマルコフ決定過程モデルとして解釈できる（図 1 参照）．

ここで，定常政策 d は，以下に示す期待割引総利得 V を最大化することで求められる．

$$\begin{aligned} & V((x_{-2}, x_{-1}, x_0), d, \theta) \\ &= \sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) p(x_t|x_{t-3}, x_{t-2}, x_{t-1}, \\ & \quad a_t = d(x_{t-3}, x_{t-2}, x_{t-1}); \theta). \end{aligned} \tag{3}$$

ただし，(x_{-2}, x_{-1}, x_0) は推薦対象ユーザの初期状態 s_0 を表す．

文献 3) で提案されている，推薦ルールの導出手順を以下に示す：

1. 履歴データ $\mathcal{D}$ を蓄積する．
2. N 人分の履歴データ $\mathcal{D}$ からパラメータ θ の最尤推定量 $\hat{\theta}$ を求める．
3. 2. で求めたパラメータの推定値を式 (2) に埋め込んだもとで，価値反復法を適用する．

なお，文献 3) では，初期の履歴データには推薦商品履歴が含まれないものとしている．そのため，購入商品履歴のみから近似的に状態遷移確率を求める手法を提案している．具体的には，推薦された商品は，推薦されなかった場合と比べて購入される確率が上がる，というような仮説を置いたもとで，混合多項分布を当てはめることで状態遷移確率を求める．そして，実際に推薦を行っていくことで推薦商品履歴が蓄積された後は，上記に示した手順のとおり，最尤推定により状態遷移確率を定期的に更新する．

また，文献 3) では，これまでの推薦方式（モデルベースアルゴリズム）よりも，より多くの利益が得られることを実験的に示している．つまり，1 人のユーザに対して推薦を複数回行うような場面を想定した場合には，商品を推薦した後のユーザの反応，及び，一定期間内に行った複数の推薦結果を考慮することの有用性が認められている．

5. 提 案 法

本節では，文献 3) におけるマルコフ決定過程モデルを，より一般化したマルコフ決定過程モデル（一般状態マルコフ決定過程モデルと呼ぶことにする）を提案する．さらに，提案する一般状態マルコフ決定過程モデルに対して，事前に得られている N 人分の履歴データ $\mathcal{D}$ から，最適な定常政策 d を求める方法を提案する．

5.1 マルコフ決定過程モデルの一般化

文献 3) で用いられている状態遷移確率（式 (2)）を，以下にように一般化する．

$$x_t \sim p(x_t|\sigma_m(x^{t-1}, a^{t-1}), a_t; \theta_m, m). \quad (4)$$

ここで，m はモデルのインデックスを表し，θ_m はモデル m における状態遷移確率を規定するパラメータを表す．また，$\sigma_m(\cdot)$ は，購入商品系列と推薦商品系列から状態を一意に定める関数を表す．これは，式 (4) によって，例えば，

$$\begin{aligned} &p(x_t|\sigma_1(x^{t-1}, a^{t-1}), a_t; \theta_1, 1) \\ =\ &p(x_t|x_{t-1}, a_t; \theta_1, 1), \\ &p(x_t|\sigma_2(x^{t-1}, a^{t-1}), a_t; \theta_2, 2) \\ =\ &p(x_t|x_{t-2}, x_{t-1}, a_t; \theta_2, 2), \end{aligned}$$

のような様々なパターンのマルコフ連鎖によって状態遷移確率が表現されうることを意味している．従来法との違いは，購入商品履歴と推薦商品履歴の個別の組合せごとに，1 つの状態を定義できることにある．つまり，一般状態マルコフ決定過程モデルにおいては．従来法のように，全ての状態が商品 3 個分の商品購入履歴によって表現されるモデルや，状態ごとに購入商品や推薦商品の履歴数が変わるモデルなど，状態を柔軟に表現することができる．

状態遷移確率を式 (4) によって表現したことから，定常政策 d と期待割引総利得 V はそれぞれ以下の式で表される：

$$a_t = d(\sigma_m(x^{t-1}, a^{t-1})), \quad (5)$$

$$\begin{aligned} &V(s_0, d, \theta_m, m) \\ =\ &\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) p(x_t|\sigma_m(x^{t-1}, a^{t-1}), \\ &a_t = d(\sigma_m(x^{t-1}, a^{t-1})); \theta_m, m). \end{aligned} \quad (6)$$

ただし，s_0 は推薦対象ユーザの初期状態を表す（$s_0 = \sigma_m(x^0, a^0)$）．ここで，定常政策 d は，パラメータ m, θ_m が既知であるもとでは，従来法と同様に価値反復法を用いることで，式 (6) を最大化する定常政策 $d(\sigma_m(x^{t-1}, a^{t-1}))$ を求めることができる．

5.2 モデルと状態遷移確率が未知である場合の推薦ルールの導出法

従来法は，モデル m が既知のもとで，当該モデルのパラメータ θ_m が未知という設定を置いているものとして捉えることができる．これに対して，本節では，一般状態マルコフ決定過程モデルにおいて，パラメータ m, θ_m がともに未知である場合，つまり，状態遷移確率の構造自体が未知である場合を考える．このとき，既に得られている N 人分の履歴データ $\mathcal{D}$ を利用して推薦ルールを学習することが考えられる．そこで，本論文では，パラメータ m, θ_m が未知である場合に，履歴データ $\mathcal{D}$ から推薦ルールを求める方法を提案する．ここで，統計的決定理論，特にベイズ決定理論[6]，に基づいて推薦ルールを導出する．

ただし，本論文では，以下の 2 つの仮定を置く．

- パラメータ m, θ_m は未知ではあるが，それぞれのパラメータが属す集合 $\mathcal{M}, \Theta_m$ は既知であるものとする：$m \in \mathcal{M}, \theta_m \in \Theta_m$．
- 真のパラメータ $m^*, \theta^*_{m^*}$ が存在し，かつ，モデル集合 $\mathcal{M}$ とモデルパラメータ集合 Θ_m にそれぞれ含まれるものとする．

また，これまでの議論と区別するために，以降では統計的決定理論の言葉で推薦ルールの導出法を説明する．まず，決定関数 d を以下のように表現する：

$$a_t = d(\sigma_m(x^{t-1}, a^{t-1}), \mathcal{D}). \quad (7)$$

ここで，決定関数の引数に履歴データ $\mathcal{D}$ が入っていることに注意．式 (5) で表される定常政策は，提案するマルコフ決定過程モデルにおける “状態” を表す変数のみを引数にもつ関数である．これに対して，式 (7) の決定関数は，決定関数を履歴データ $\mathcal{D}$ から学習することを考慮した関数であり，式 (5) とは本質的に異なる関数である．すると，履歴データ $\mathcal{D}$ とモデルに関するパラメータ m, θ_m から定まる決定関数 d を評価する関数（評価関数）V は，

$$V(s_0, d(\mathcal{D}), \theta_m, m), \quad (8)$$

と表現できる．上記の評価関数は式 (6) で表される期待割引総利得に相当し，本論文では式 (8) を効用関数と呼ぶことにする．ここで，学習に用いる履歴データ $\mathcal{D}$ が変わると，そこから求められる決定関数も変わる可能性があるため，式 (7) と同じく効用関数の引数に履歴データ $\mathcal{D}$ が含まれていることに注意．そのため，従来法のように式 (3) を直接最大化するのではなく，履歴データ $\mathcal{D}$ の出現確率で平均化した期待効用関数 $\mathrm{E}_{\mathcal{D}}$ を最大化する．

$$\mathrm{E}_{\mathcal{D}}\left[V(s_0, d(\mathcal{D}), \theta_m, m)\right].$$

ここで，パラメータ m, θ_m が未知である場合，期待効用関数 $\mathrm{E}_{\mathcal{D}}$ を最大化する決定関数を求めることは困難である．なぜならば，履歴データ $\mathcal{D}$ に依存して最適なパラメータ m, θ_m の値が変動するためであり，一般に，m, θ_m が変動する全範囲にわたって期待効用関数 $\mathrm{E}_{\mathcal{D}}$ を最大化する決定関数を求めることは難しい．

そこで，ベイズ決定理論の枠組みでは，未知のパラメータ m, θ_m が確率分布に従うものとして，パラメータ m, θ_m が従う確率分布（パラメータ m, θ_m に対する事前分布）$p(m), p(\theta_m)$ で期待効用関数 $\mathrm{E}_{\mathcal{D}}$ をさら

に平均化した以下の値の最大化を図る．

$$\mathrm{E}_{\mathcal{M}}\left[\mathrm{E}_{\Theta_m}\left[\mathrm{E}_{\mathcal{D}}\left[V(s_0, d(\mathcal{D}), \theta_m, m)\right]\right]\right] \quad (9)$$

以降，式 (9) をベイズ期待効用関数と呼ぶことにする．

5.3 定常政策の導出手順

ベイズ基準のもとで最適な定常政策を求める具体的な手順を示す．まず，式 (9) は以下のように展開できる：

$$\sum_{\mathcal{D}} p(\mathcal{D}) \underline{\sum_{m\in\mathcal{M}} \int_{\Theta_m} V(s_0, d(\mathcal{D}), \theta_m, m) p(\theta_m|\mathcal{D}, m) p(m|\mathcal{D}) d\Theta_m}. \quad (10)$$

すると，$p(\mathcal{D})$ の項は定常政策を決める際には考慮しなくて良いことが分かる．そこで，式 (10) の下線部のみに着目して，さらに，

$$\underline{\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) \sum_{m\in\mathcal{M}} p(m|\mathcal{D}) \int_{\Theta_m} p(x_t|\sigma_m(x^{t-1}, a^{t-1}), a_t = d(\sigma_m(x^{t-1}, a^{t-1}), \mathcal{D}), \theta_m, m) p(\theta_m|\mathcal{D}, m) d\Theta_m} \quad (11)$$

と変形すると，式 (11) の下線部の計算結果を，

$$q(x_t|x^{t-1}, a^{t-1}, a_t = d(x^{t-1}, a^{t-1}, \mathcal{D})), \quad (12)$$

のように 1 つの状態遷移確率 q として見ることができるようになる．ここで，式 (12) は，モデル $m\in\mathcal{M}$，および，モデルごとのパラメータ $\theta_m\in\Theta$ に関して周辺化した状態遷移確率である．また，式 (12) は，モデルパラメータ θ_m の事後確率 $p(\theta_m|\mathcal{D}, m)$ で重み付けた状態遷移確率を，さらに，集合 $\mathcal{M}$ に含まれる全てのモデルに関して，その事後確率 $p(m|\mathcal{D})$ で重み付けた状態遷移確率と解釈できる．なお，複数のモデルを仮定する提案法と比較した場合，従来法[3] は，モデルを 1 つに固定した場合の提案法と見なすことができる．

以上の結果から，ベイズ期待効用関数の最大化は，次に示す関数の最大化に帰着される．

$$\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) q(x_t|x^{t-1}, a^{t-1}, a_t = d(x^{t-1}, a^{t-1}, \mathcal{D})).$$

上式は，式 (12) が求まった後は，状態遷移確率が既知であるマルコフ決定過程モデルとして扱えることを意味している．故に，本論文で提案する定常政策（推薦ルール）の導出手順は次のように書ける．

1. 仮の推薦ルールを用いて（後述），履歴データ $\mathcal{D}$ を蓄積する．
2. N 人分の履歴データ $\mathcal{D}$ から，重み付き状態遷移確率 q（式 (12)）を計算する．
3. 2. で求めた重み付き状態遷移確率 q を用いて価値反復法を適用し，定常政策 d を求める．

5.4 履歴データの蓄積期間と推薦ルールの運用期間

前述のとおり，従来法においては，当初の履歴データには推薦商品履歴が存在しないものとし，初期時点では状態遷移確率を近似的に求める対処案を提案していた．これに対して，提案法では，文献 4) で提案されているように，履歴データを蓄積するための期間（準備期間）と，履歴データから導出した推薦ルールを運用する期間（運用期間）とを明確に分けて考えるアプローチをとる．つまり，準備期間において得られる利得については最大化の対象とはせず，準備期間においては，例えば，推薦する商品をランダムに選択するなど，仮の推薦ルールに基づいて推薦商品履歴を蓄積するものとする．

5.5 クラスタリングの事前適用

通常，各ユーザが購入している商品は，提供されている全商品数と比べてごく少数である．このような場合には，前処理として，例えば，文献 5) で提案されているような，ユーザと商品の共クラスタリングが有用である．つまり，クラスタリングを行うことで，類似した購買履歴を持つユーザ群，および，当該ユーザ群に購入されやすい商品群のみを抽出することで，履歴データが密なユーザと商品のみを対象に，提案法を適用することができる．

6. 評 価 実 験

本節では，人工データを用いた評価実験を行うことにより，提案法の有効性を確認する．

6.1 実 験 手 順

次に示すモデル集合 $\mathcal{M}$ を用いて評価実験を行った：

$$\mathcal{M} = \{p(x_t|a_t, \theta_1, 1), p(x_t|x_{t-1}, a_t, \theta_2, 2), p(x_t|x_{t-2}, x_{t-1}, a_t, \theta_3, 3), p(x_t|a_{t-1}, a_t, \theta_4, 4), p(x_t|x_{t-1}, a_{t-1}, a_t, \theta_5, 5), p(x_t|x_{t-2}, x_{t-1}, a_{t-1}, a_t, \theta_6, 6)\}.$$

上記のモデル集合は，状態と行動に関して，最大 2 次のマルコフ連鎖を仮定したモデルで構成されている．評価手順を以下に示す：

1. モデル集合から，真のモデル m^* を 1 つ選択する．
2. 既知のハイパーパラメータを持つディリクレ分布に従って，真のパラメータ θ_{m^*} を生成．
3. 真のモデル m^* とパラメータ θ_{m^*} を用いて，N 人分の履歴データを生成する．
4. N 人分の履歴データに対して提案法/従来法を適用し，推薦ルールを導出する．
5. 得られた推薦ルールを評価する．

ここで，3. では，真のモデル m^* とパラメータ θ_{m^*} によって表現される真の状態遷移確率に対して，推薦する商品（行動）を逐次ランダムに選択することで N 人分の履歴データを生成する．また，5. では，真のモデル m^* とパラメータ θ_{m^*} によって表現される真の状態遷移確率と，提案法/従来法により導出した推薦

表 1 実験結果：真のモデルを変化させた場合の割引総利得.
Table 1 Results: The discounted sum of rewards with respect to each true model($m^* \in \mathcal{M}$).

真のモデル	従来法（モデル固定）						提案法
	1	2	3	4	5	6	
モデル 1	125.07	125.47	**129.66**	123.79	**127.53**	117.50	121.83
モデル 2	115.51	**133.15**	124.85	115.88	126.61	129.27	**131.10**
モデル 3	116.32	116.33	**131.41**	107.48	113.25	123.94	**126.67**
モデル 4	113.90	111.93	112.57	**134.80**	**132.37**	125.38	126.75
モデル 5	104.74	113.11	115.52	116.45	126.29	**129.16**	**129.10**
モデル 6	115.61	115.89	108.18	107.71	111.79	**118.49**	**116.77**
平均	115.19	119.31	120.37	117.69	122.97	123.96	**125.37**

ルールを用いて，推薦対象ユーザ分の評価用データを生成することで，割引総利得を算出した．

評価実験に用いたパラメータ値は次のとおりである：ユーザ数 $N=10000$，ユーザごとの履歴数は 5〜100 個の中からランダムに選択（最初の 2 時点分は初期状態用として使用），商品数 $I=20$，推薦対象ユーザの履歴数 100（初期状態用 2 時点，評価用 98 時点），ディリクレ分布のハイパーパラメータ $\alpha_i=10^{-3}(i=1,2,\ldots,I)$，割引率 $\gamma=0.95$．また，利得は $(1,10)$ の間でランダムに生成し（何も購入されない場合は利得 0），価値反復法の収束判定は，価値関数の差分が 10^{-6} を下回ったとき収束したと判断した．

6.2 実験結果

先に示した実験手順を 10 回繰り返して得られた割引総利得の平均値を**表 1** に示す．表の各行は，第 1 列に示したモデルを真のモデルとしたもとで，1 つのモデルを固定的に使用する従来法，及び，複数のモデルを重みづける提案法によって得られた割引総利得をそれぞれ表す．ここで，表 1 の最下行は，列ごとの平均値を表す．

表 1 より，従来法において，真のモデルと一致している場合には，他のモデルと比べて比較的より多くの割引総利得が得られていることが分かる．これに対して，提案法は，モデルを 1 つに固定する従来法と比べ，平均的により多くの割引総利得を得られていることが分かる．これは，真のモデルに合わせてモデルやモデルパラメータに関する事後確率が適切に計算され，その結果，複数のモデルで重み付けた状態遷移確率が真のモデルに近いものとなっていることを意味している．通常，真のモデルを知ることは困難である．そのため，真のモデルが自明でない場合には，提案法のように，複数のモデルを事前に用意しておき，得られた履歴データからモデルの重みを適切に調節するアプローチが有効であると言える．

7. まとめ

本論文では，推薦問題を扱うための，より一般化されたマルコフ決定過程モデル（一般状態マルコフ決定過程モデル）を提案した．さらに，提案したモデルに対して，統計的決定理論に基づく推薦ルールの導出法を提案した．

今後の取組としては，本論文では，購入商品履歴のみによって状態を表現したが，購入商品履歴以外の変数を状態に組み込むことが考えられる．例えば，ユーザのデモグラフィック情報や商品購入時の時間帯等がユーザの購入行動に影響することが分かっている場合には，状態を表現する変数群にそれらの変数を追加することで（状態遷移確率の条件部にそれらの変数を追加することで），ユーザの購入行動をより反映したマルコフ決定過程モデルを定義することができる．このように，適用する問題や事前に得られている知識をモデルに組み込むことで，更に現実に即した推薦が行えるようになる．モデルに変数を追加した場合においても，提案法を用いることにより，ベイズ基準のもとで最適な推薦が実現される．

参考文献

1) Adomavicius, G. and Tuzhilin, A.: Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6 (2005).
2) Bellman, R.: *Dynamic Programming (Princeton Landmarks in Mathematics)*, Princeton University Press (2010).
3) G.Shani, D.Heckerman and Brafman, R.I.: An MDP-Based Recommender System, *Journal of Machine Learning Research*, Vol. 6, pp. 1265–1295 (2005).
4) 前田康成，浮田善文，松嶋敏泰，平澤茂一：学習期間と制御期間に分割された強化学習問題における最適アルゴリズムの提案，情報処理学会論文誌，Vol.39, No.4, pp.1116–1126 (1998).
5) 桑田修平，山田武士，上田修功：ディリクレ過程混合モデルに基づく離散データの共クラスタリング，情報処理学会論文誌（数理モデル化と応用），Vol.1, No.1, pp.60–73 (2008).
6) 松嶋敏泰：帰納・演繹推論と予測-決定理論による学習モデル-，情報論的学習理論ワークショップ (IBIS'98)，pp.1–8 (1998).