情報処理学会論文誌　数理モデル化と応用　Vol.6　No.1　20–30 (Mar. 2013)

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,†1,a)]　前田 康成[2]　松嶋 敏泰[1]　平澤 茂一[1]

受付日 2012年4月19日，再受付日 2012年6月7日，
採録日 2012年7月9日

概要：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．推薦問題に関する研究において，これまで，ある商品を推薦した結果どの商品が買われたのか（推薦結果）や，さらには，一定期間内に行った複数の推薦結果が考慮されることはほとんどなかった．これに対して，マルコフ決定過程モデルを用いることで上記 2 点を初めて考慮した手法が提案されている．提案法は，その従来研究のモデルを一般化した点に新規性がある．また，もう 1 つの新規性として，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点がある．従来のモデルを一般化することで，マルコフ決定過程モデルを用いた推薦手法の適用領域が拡大され，かつ，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確認した．

キーワード：推薦問題，マルコフ決定過程，ベイズ決定理論，強化学習

Variable Order Transition Probability Markov Decision Process for the Recommendation System

Shuhei Kuwata[1,†1,a)]　Yasunari Maeda[2]　Toshiyasu Matsushima[1]
Shigeichi Hirasawa[1]

Received: April 19, 2012, Revised: June 7, 2012,
Accepted: July 9, 2012

Abstract: In this paper, we propose a general markov decision process model for the recommendation system. Furthermore, by using historical data, we derive the optimal recommendation lists from the proposed model based on bayesian decision theory. In the recommendation research area, there were little studies that considered both the purchased items and the past recommended items within a given period. In these circumstances, markov decision process based recommend method that can take these two things into account has been proposed. Our method also uses both things as with the previous method. Here, the unique thing about this paper is not only that we generalize the existing model, but also 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 for a wide variety of recommendation scene. By using artificial data, we show the experimental results that our method can obtain more rewards than the conventional method gets.

Keywords: recommendation, Markov decision process, Bayesian decision theory, reinforcement learning

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 [1], [14], [24], [25], [26]，特に，Amazon.com などの

[1] 早稲田大学理工学術院基幹理工学研究科
Department of Computer Science and Engineering, Waseda University, Shinjuku, Tokyo 169–8555, Japan
[2] 北見工業大学
Kitami Institute of Technology, Kitami, Hokkaido 090–8507, Japan
[†1] 現在，株式会社 NTT データ
Presently with NTT DATA CORPORATION
[a)] kuwatas@nttdata.co.jp

EC（Electronic Commerce，電子商取引）サイトにおいて実用化が進んでいる．また，最近では，Hadoop [15] 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 [10]．

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], [20]：

1. メモリベースアルゴリズム：
   利用者間の購買履歴データの類似性をもとに推薦する商品を決定する．協調フィルタリング [8], [11], [12], [16] などが提案されており，購買履歴が類似している利用者の間で人気のある商品を推薦する．
2. モデルベースアルゴリズム：
   購買履歴データに対して確率モデルをあてはめ，得られたモデルをもとに推薦する商品を決定する．顧客セグメントを潜在クラスによって表現した潜在クラスモデル [6], [21], [22] などが提案されており，購入される確率の高い商品を推薦する．
3. ハイブリッドアルゴリズム：
   メモリベースアルゴリズムとモデルベースアルゴリズムを足し合わせた手法 [13]．

ここで，上記 3 つのタイプに属するほとん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という点をあげることができる．すなわち，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場合がほとんどであり，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ふまえて，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）を利用していない．

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あげられる．しかし，会員制の EC サイトなどを考えると，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 [7]．

ここで，それら 2 点を考慮した数少ない従来法として文献 [5] がある．文献 [5] では，マルコフ決定過程 [17], [23] をベースにした推薦手法が提案されており，推薦商品履歴の利用や，推薦を行った結果，これまでどのような商品が購入されてきたかが考慮されている．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とき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，そのモデルのもとで，将来にわたって得られる “利得” を最大化する “定常政策” を求める．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．推薦した結果，次にどの商品が購入されたかを確率モデルとして表現し，かつ，推薦を複数回行った結果得られる利得の合計値に対する最大化が図られており，先に示した 2 点が考慮されているといえる．

ここで，文献 [5] の課題として，つねに商品 3 個分の購入商品履歴ごとに推薦する商品が定められるという点があげられる．適用する対象によっては，たとえば，4 個前の購入商品履歴にも依存するケースも十分考えられる．さらにいえば，何個前までの履歴に依存するかを事前に把握することが困難なケースも考えられる．つまり，依存する履歴数を固定的に取り扱っている点に改善の余地がある．

そこで，本論文では，文献 [5] と同様に，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考慮する推薦手法を提案する．ただし，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，事前に蓄積した履歴データを用いて，最適な推薦ルール（定常政策）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 [5] 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．提案法により，マルコフ決定過程モデルを用いた推薦手法の適用領域が拡大され，かつ，推薦する目的に合わせて，統計的決定の観点で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 章において，本論文で扱う推薦問題を定義し，3 章で，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 章で，マルコフ決定過程モデルを用いた従来法 [5] を説明する．その後 5 章で，一般化したマルコフ決定過程モデルを提案し，さらに，提案したモデルに対して統計的決定理論を用いた最適な推薦ルールの導出法を提案する．6 章で人工データを用いた評価実験を行い，最後に 7 章でまとめる．

2. 問題設定

本章では，本論文が対象とする推薦問題を定義する．まず，N 人のユーザがいるものとし，各ユーザに対する購入商品履歴と対応する推薦商品履歴が蓄積可能であるものとする．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るものとする．

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

ここで，$a_{t(i)}$（$t = \cdots, -2, -1, 0$）は，時点 t においてユーザ i へ推薦された商品を表し，$x_{t(i)}$ は商品 $a_{t(i)}$ が推薦された後のそのユーザ i の反応（商品を購入する，何も購入し

ないなど）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，

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

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．なお，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．また，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：

$$\mathcal{D} = \{\mathcal{D}_{(1)}, \mathcal{D}_{(2)}, \ldots, , \mathcal{D}_{(N)}\} = \{x^0_{(i)}, a^0_{(i)}\}_{i=1}^N.$$

本論文では，履歴データ $\mathcal{D}$ を蓄積したもとで，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．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．以下に，本論文で想定する推薦の流れを整理する：

1. N 人分の履歴データ $\mathcal{D}$ を蓄積する．
2. 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，何も購入しないなど）．
5. （3. と 4. を繰り返す）．

なお，推薦ルールの更新は考えないものとする[*1]．また，各時点ではつねに 1 個の商品を推薦するものとする（複数個への拡張は自然に行われる）．さらに，推薦対象ユーザによる同一商品の購入は 1 度きりでも複数回でもよいものとする．

以上の設定のもとで，本論文では，一定期間内に行った複数の推薦結果を評価する．すなわち，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本論文の目的とする．

3. 準備：マルコフ決定過程モデル

本章では，従来法 [5]，および，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：$\mathcal{A} = \{1, 2, \ldots, A\}$,

[*1] 本論文では，商品が購入されるたびに得られる履歴データを用いて推薦ルールを逐次更新することは想定しない．実際，推薦ルールの更新は，月に 1 回更新するなど，定期的に行われる場合が多い．

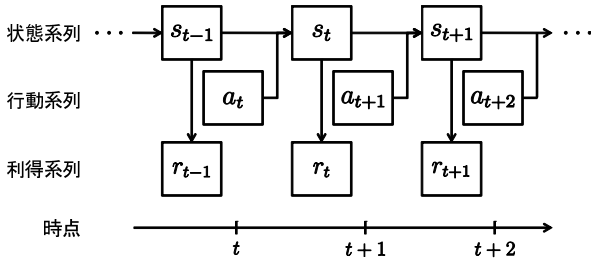

図 1　マルコフ決定過程モデルにおける変数間の関係を表すイメージ図

Fig. 1　The image of the relations between variables on Markov decision process.

- 状態遷移確率：$\{p(s|s', a)|s, s' \in \mathcal{S}, a \in \mathcal{A}\}$,
- 利得集合：$\{r(s, a)|s \in \mathcal{S}, a \in \mathcal{A}\}$.

各構成要素間の関係を**図 1** に示す．図 1 が示すとおり，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 $a_t \in \mathcal{A}$ にのみ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）．ここで，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 $r(s_t, a_t)$ が得られるものとする．

上に示した 4 つの要素すべ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，γ（$0 < \gamma < 1$）は割引率を表す．式 (1) は，一定期間内に得られるすべての利得において，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 [5] 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：

- 推薦対象ユーザが購入する商品 x_t（$t = \ldots, -1, 0, 1, \ldots$）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．式 (2) は，直前に購入された 3 つの購入商品履歴 $(x_{t-3}, x_{t-2}, x_{t-1})$ とそのときに推薦された商品 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 $r(x_t)$ とし（利得関数の引数に行動 a が含まれないことに注意），将来にわたって得られる利得の合計を最大化する推薦ルールを求める．ここで，割引総利得は以下で表される：

$$\sum_{t=1}^{\infty} \gamma^{t-1} r(x_t).$$

- 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，その場合には状態は変化しないものとし，前の時点と同じ状態にいるものとし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 [2]．

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

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 [5] 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. N 人分の履歴データ $\mathcal{D}$ を蓄積する．
2. 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．すなわち，商品 3 個分の順列によって表現される状態を $\{wxy\}$ とおいたとき（$w, x, y \in \mathcal{I}$），期待割引総利得 V から導かれる以下のベルマン方程式，

$$\begin{aligned} & v_{l+1}(\{wxy\}) \\ = & \max_{i \in \mathcal{I}} \sum_{z \in \mathcal{I}} p(x_t = z | s_{t-1} = \{wxy\}, a_t = i; \hat{\theta}) \times \\ & \{r(z) + \gamma v_l(\{xyz\})\}, \end{aligned}$$

を用いて，状態ごとの評価値 $v(\{wxy\})$ を求める．ここで，添え字 l，$l+1$ は，計算上のループ回数を表すインデックスを表し，評価値 v の値が収束するまで繰り返し v_l の値を v_{l+1} の値で更新する．そして，最終的に得られた評価値 v^* をもとに，推薦ルール（状態ごとに推薦する商品 a を定めたルール）を求める：

$$\begin{aligned} a = & \ d(\{wxy\}), \\ = & \ \arg\max_{i \in \mathcal{I}} \sum_{z \in \mathcal{I}} p(x_t = z | s_{t-1} = \{wxy\}, a_t = i; \hat{\theta}) \\ & \times \{r(z) + \gamma v^*(\{xyz\})\}. \end{aligned}$$

なお，文献 [5] では，初期の履歴データ $\mathcal{D}$ には推薦商品履歴は含まれないものとし，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，という仮説をおいたもとで，混合多項分布をあてはめることで状態遷移確率を求める．ただし，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が更新されるため，本論文では，前述した近似的な算出法については考えないものとする．

文献 [5] では，マルコフ決定過程モデルを用いた推薦ルールの方が，従来の推薦方式（モデルベースアルゴリズム）よりも，より多くの割引総利得が得られることを実験的に示している．すなわち，1 人のユーザに対して推薦を複数回行うような場合には，商品を推薦した後のユーザの反応，および，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示されている．しかし，文献 [5] では，既存のマルコフ決定過程モデルをそのまま推薦問題へ適用するにとどまっており，より多くの推薦問題への適用を考えた場合，より表現能力の高いマルコフ決定過程モデルを用いる必要がある．また，その有効性は実験的に示されただけである．

5. 提案法

本章では，文献 [5] 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，履歴データ $\mathcal{D}$ から，統計的決定理論に基づき最適な定常政策 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 [5] で用いられている状態遷移確率（式 (2)）を，以下のように一般化する．

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

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，購入商品履歴と推薦商品履歴から状態を一意に定める関数を表す．たとえば，式 (4) を用いると，

$$\begin{aligned}
&\cdot\ p(x_t|\sigma_1(x^{t-1},a^{t-1}),a_t;\theta_1,1) = p(x_t|x_{t-1},a_t;\theta_1,1),\\
&\cdot\ p(x_t|\sigma_2(x^{t-1},a^{t-1}),a_t;\theta_2,2)\\
&\quad = p(x_t|x_{t-2},x_{t-1},a_t;\theta_2,2),\\
&\cdot\ p(x_t|\sigma_3(x^{t-1},a^{t-1}),a_t;\theta_3,3)\\
&\quad = p(x_t|x_{t-3},x_{t-2},x_{t-1},a_{t-2},a_{t-1},a_t;\theta_3,3),
\end{aligned}$$

といった様々なパターンのマルコフ連鎖による状態遷移確率の表現が可能となる．文献 [5] 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．文献 [5] のように，すべての状態が商品 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依存する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可能となる．その結果，マルコフ決定過程モデルを用いた推薦手法の適用可能領域が拡大される．

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：

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

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

ただし，s_0 は推薦対象ユーザの初期状態を表す（$s_0 = \sigma_m(x^0,a^0)$）．なお，パラメータ m，θ_m が既知であるもとでは，文献 [5] 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　履歴データを用いた最適な推薦ルールの学習

文献 [5] は，モデル m が既知ではあるが，当該モデルのパラメータ θ_m が未知という設定をおいたものととらえることができる．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を考える．すなわち，状態遷移確率の構造自体も未知である場合に，N 人分の履歴データ $\mathcal{D}$ から推薦ルールを学習する方法を提案する．ここで，統計的決定理論，特にベイズ決定理論 [3], [4], [19], [27]，に基づいて最適な推薦ルールを導出する方法を示す．

ただし，本論文では，以下の 2 つの仮定をおく．

- 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 $\mathcal{M}$，Θ_m は既知であるものとする：
$$m \in \mathcal{M},\quad \theta_m \in \Theta_m.$$
- 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．

5.2.1　ベイズ決定理論に基づく推薦ルールの導出法

まず，決定関数 d を以下のように表現する：

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

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持つ関数である．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，

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

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 $\mathrm{E}_{\mathcal{D}}$ を最大化すること考える．

$$\begin{aligned}
&\mathrm{E}_{\mathcal{D}}\left[V(s_0,d(\mathcal{D}),\theta_m,m)\right]\\
&= \sum_{\mathcal{D}} V(s_0,d(\mathcal{D}),\theta_m,m)p(\mathcal{D};\theta_m,m)\\
&= \sum_{x^0_{(1)}}\sum_{a^0_{(1)}}\cdots\sum_{x^0_{(N)}}\sum_{a^0_{(N)}} V(s_0,d(\mathcal{D}),\theta_m,m)\times\\
&\quad p(x^0_{(1)},a^0_{(1)};\theta_m,m)\cdots p(x^0_{(N)},a^0_{(N)};\theta_m,m).
\end{aligned}$$

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．

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

$$= \sum_{m\in\mathcal{M}} \int_{\Theta_m} \mathrm{E}_{\mathcal{D}}\left[V(s_0, d(\mathcal{D}), \theta_m, m)\right] p(\theta_m|m)p(m)d\Theta_m. \tag{9}$$

以降，式 (9) をベイズ期待効用関数と呼ぶことにする．

まとめると，本論文で提案した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 が未知である場合には，N 人分の履歴データ $\mathcal{D}$ を用いて，ベイズ期待効用関数を最大化する決定関数 d を求めることを考える．その結果得られる決定関数 d は，ベイズ基準のもとで最適な定常政策（推薦ルール）となる．

5.2.2　定常政策の具体的な導出手順

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：

$$\sum_{m\in\mathcal{M}} \int_{\Theta_m} \mathrm{E}_{\mathcal{D}}\left[V(s_0, d(\mathcal{D}), \theta_m, m)\right] p(\theta_m|m)p(m)d\Theta_m$$

$$= \sum_{m\in\mathcal{M}} \int_{\Theta_m} \sum_{\mathcal{D}} V(s_0, d(\mathcal{D}), \theta_m, m)p(\mathcal{D}|\theta_m, m) \times p(\theta_m|m)p(m)d\Theta_m \tag{10}$$

$$= \sum_{\mathcal{D}} p(\mathcal{D}) \underline{\sum_{m\in\mathcal{M}} \int_{\Theta_m} V(s_0, d(\mathcal{D}), \theta_m, m)p(\theta_m|\mathcal{D}, m)\times p(m|\mathcal{D})d\Theta_m}. \tag{11}$$

ここで，式 (10) から式 (11) への展開にはベイズの定理を用いた．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よいことが分かる．そこで，式 (11) の下線部のみに着目して，さらに，

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

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

$$= \sum_{t=1}^{\infty} \gamma^{t-1} r(x_t) \underline{\sum_{m\in\mathcal{M}} p(m|\mathcal{D})\times \underline{\int_{\Theta_m} p(x_t|\sigma_m(x^{t-1}, a^{t-1}), a_t = d(\sigma_m(x^{t-1}, a^{t-1}), \mathcal{D}), \theta_m, m) \times p(\theta_m|\mathcal{D}, m)d\Theta_m}} \tag{12}$$

と変形する．すると，式 (12) の下線部の計算結果を，

$$q(x_t|x^{t-1}, a^{t-1}, a_t = d(x^{t-1}, a^{t-1}, \mathcal{D})), \tag{13}$$

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 (13) 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 (13) 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 $\mathcal{M}$ に含まれるすべ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．複数のモデルを仮定する提案法と比較した場合，文献 [5] は，モデルを 1 つに固定した場合の提案法と見なすこともできる[*2]．

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 (13) 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．ゆえに，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. N 人分の履歴データ $\mathcal{D}$ を蓄積する．
2. 履歴データ $\mathcal{D}$ から，重み付き状態遷移確率 q（式 (13)）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．

5.3　実装上での検討事項

推薦ルールの導出手順を実装するにあたって，特に検討しておくべき事項について触れておく．

5.3.1　履歴データの蓄積

前述のとおり，文献 [5] 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が提案されている．これに対して提案法では，文献 [18] 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せず，準備期間においては，たとえば，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履歴データ $\mathcal{D}$ を蓄積するものとする．そして，推薦ルールの運用期間に入る直前に，蓄積した履歴データ $\mathcal{D}$ から推薦ルールを導出する．ここで，準備期間における仮の推薦ルールとして，ランダムに推薦商品を選択する以外にも，これまでに推薦したことがないような商品を積極的に推薦する方法など，後に控える運用期間を考慮した様々な仮の推薦ルールを考えることができる．蓄積期間における仮の推薦ルールの設定法の検討については今後の課題とし，本論文ではランダムに推薦する商品を選択するものとした．

5.3.2　履歴データが疎である場合の対処

多くの場合，各ユーザが購入している商品は，提供され

[*2] モデルを 1 つに固定した場合においても，パラメータに関する事後分布で周辺化するため，厳密には文献 [5] の手法とは異なる．

ている全商品数と比べてごく少数である．そのような“疎”な履歴データを用いた場合，たとえば，状態遷移確率を精度良く求めることができず，その結果，“使えない”推薦ルールが得られてしまうことが想定される．実際，ユーザを“行”，商品を“列”で表現した購買履歴行列において，購買履歴行列の第 (i,j) 要素は，ユーザ i が商品 j を購入している場合 1，購入していない場合 0 として表現されているものとすると，1（購入）の占める割合は，全要素の10%にも満たない場合がほとんどである．

履歴データが持つこのような性質への対処法として，たとえば，文献 [21] 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．すなわち，履歴データのクラスタリングを事前に行い，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，“密”な履歴データを用意することができる．このとき，推薦ルールは，ユーザクラスごとに導出される．

5.3.3　式 (13) の計算例

最後に，式 (13) の計算例として，各モデルのパラメータ θ_m が，既知のハイパーパラメータ α_i（$i=1,2,\ldots,I$）を持つディリクレ分布に従うものとした場合の計算結果を示す：

$$p(\theta_{m,\sigma_m,a};\{\alpha_i\}_{i=1}^I)=\frac{\Gamma(\sum_{i=1}^I\alpha_i)}{\prod_{i=1}^I\Gamma(\alpha_i)}\prod_{i=1}^I\{\theta_{m,s_m,a,i}\}^{\alpha_i-1},\quad a\in\mathcal{I},\ \sigma_m\in\mathcal{S}_m,\ m\in\mathcal{M}. \tag{14}$$

ここで，$\theta_{m,\sigma_m,a}=\{\theta_{m,\sigma_m,a,1},\ldots,\theta_{m,\sigma_m,a,I}\}$，$\theta_m=\{\{\theta_{m,\sigma_m,a}\}_{s_m\in\mathcal{M},a\in\mathcal{I}}\}$ は，モデル m において状態 σ_m のもとで商品 a を推薦した際の商品購入確率を表す（$\sum_i\theta_{m,\sigma_m,a,i}=1$）．また，$\mathcal{S}_m$ はモデル m における状態集合を表し，$\theta_{m,\sigma_m,a,i}$ は，状態 σ_m のもとで商品 a を推薦した際，商品 i が購入される確率を表す．さらに，$\Gamma(\cdot)$ はガンマ関数を表すものとし，モデル m に関しては一様分布を仮定する（$p(m)=1/|\mathcal{M}|$）．

このとき，モデルパラメータ θ_m の事後確率 $p(\theta_m|\mathcal{D},m)$ は以下のように表される．

$$\begin{aligned}&p(\theta_m|\mathcal{D},m)\\ &\propto\prod_{\sigma_m\in\mathcal{S}_m}\prod_{a\in\mathcal{I}}p(\theta_{m,\sigma_m,a}|m)\prod_{i=1}^N p(x^0_{(i)},a^0_{(i)}|\theta_{m,\sigma_m,a},m)\\ &\propto\prod_{\sigma_m\in\mathcal{S}_m}\prod_{a\in\mathcal{I}}\prod_{i=1}^I\{\theta_{m,\sigma_m,a,i}\}^{\alpha_i+h_{\sigma_m,a,i}-1}.\end{aligned}\tag{15}$$

ただし，$h_{\sigma_m,a,i}$ は，履歴データ $\mathcal{D}$ において，モデル m 上の状態 σ_m のもとで商品 a を推薦した際に実際に商品 i が購入された総回数を表す．式 (15) は，事後分布がハイパーパラメータ $\alpha_i+h_{\sigma_m,a,i}$ を持つディリクレ分布によって表現されることを意味する．したがって，式 (12) の下線部中の積分計算は以下のように解析的に解ける．

$$\begin{aligned}&\int_{\Theta_m}p(x_t=i|s_{t-1}=\sigma_m,a_t=a,\theta_m,m)p(\theta_m|\mathcal{D},m)d\Theta_m\\ &=\frac{\alpha_i+h_{\sigma_m,a,i}}{\sum_{i=1}^I\alpha_i+h_{\sigma_m,a,i}}.\end{aligned}\tag{16}$$

次に，モデル m の事後確率 $p(m|\mathcal{D})$ は，以下のように表される．

$$\begin{aligned}&p(m|\mathcal{D})\\ &\propto\int_{\Theta_m}p(\mathcal{D}|\theta_m,m)p(\theta_m|m)p(m)d\Theta_m\\ &\propto\int_{\Theta_m}p(\mathcal{D}|\theta_m,m)p(\theta_m|m)d\Theta_m\\ &\propto\prod_{\sigma_m\in\mathcal{S}_m}\prod_{a\in\mathcal{I}}\int_{\Theta_m}\frac{\Gamma(\sum_{i=1}^I\alpha_i)}{\prod_{i=1}^I\Gamma(\alpha_i)}\times\prod_{i=1}^I\{\theta_{m,\sigma_m,a,i}\}^{\alpha_i+h_{\sigma_m,a,i}-1}d\Theta_m\\ &\propto\prod_{\sigma_m\in\mathcal{S}_m}\prod_{a\in\mathcal{I}}\frac{\Gamma(\sum_{i=1}^I\alpha_i)}{\prod_{i=1}^I\Gamma(\alpha_i)}\frac{\prod_{i=1}^I\Gamma(\alpha_i+h_{\sigma_m,a,i})}{\Gamma(\sum_{i=1}^I\alpha_i+h_{\sigma_m,a,i})}.\end{aligned}\tag{17}$$

したがって，先に示した事前分布（ディリクレ分布および一様分布）を仮定した場合，モデル m ごとに，$h_{\sigma_m,a,i}$ の値を履歴データ $\mathcal{D}$ からカウントするだけで，式 (13) が計算できることが分かる．つまり．モデル m ごとに，状態 $\sigma_m(\sigma_m(=x^{t-1},a^{t-1}))$ における商品購入確率を式 (16) によって計算しておき，続いて，式 (17) によって計算される値で，すべてのモデルに関する重み付けを状態ごとに行えばよい．このとき，重み付けられた状態遷移確率 q 上での状態は，モデル集合 $\mathcal{M}$ に含まれるすべてのモデルによって表現される状態の和集合と一致する．

6.　評価実験

前章において，ベイズ期待効用関数を最大化する推薦ルール，すなわち，ベイズ基準のもとで最適な推薦ルールを導出する方法を示した．これにより，理論的に最適な推薦ルールの導出が可能となった．本章では，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．ここで，本論文と文献 [5] では問題設定が厳密には一致しないため，1 つのモデル（状態遷移確率）のみを用いて推薦ルールを導出する手法を従来法と見なし，間接的に従来法と比較を行うものとする．具体的には，複数のモデルを重み付ける提案法が，1 つのモデルを固定的に利用する従来法と比べて，より多くの割引総利得が得られることを確認する．また，モデルベースアルゴリズムとの比較も行い，購入商品履歴だけでなく推薦商品履歴も利用することの有用性についてもあわせて確認する．

6.1　実験手順

本論文では，次に示すモデル集合 $\mathcal{M}$ を用いて評価実験を行った：

表 1　実験結果：モデルを 1 つに固定した場合と提案法の割引総利得（真のモデルを含む場合）
Table 1　Results: The discounted sum of rewards with respect to each true model ($m^* \in \mathcal{M}$).

	モデルを 1 つに固定した場合						
真のモデル	1	2	3	4	5	6	提案法
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**

表 2　実験結果：モデルを 1 つに固定した場合と提案法の割引総利得（真のモデルを含まない場合）
Table 2　Results: The discounted sum of rewards with respect to each true model ($m^* \notin \mathcal{M}$).

	モデルを 1 つに固定した場合						
真のモデル	1	2	3	4	5	6	提案法
モデル 1	-	125.47	**129.66**	123.79	**127.53**	117.50	121.83
モデル 2	115.51	-	124.85	115.88	126.61	**129.27**	**128.06**
モデル 3	116.32	116.33	-	107.48	113.25	**123.94**	**117.22**
モデル 4	113.90	111.93	112.57	-	**132.37**	125.38	**130.00**
モデル 5	104.74	113.11	115.52	116.45	-	**129.16**	**118.20**
モデル 6	115.61	**115.89**	108.18	107.71	111.79	-	**116.77**

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

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. 式 (14) で表されるディリクレ分布に従って，真のパラメータ θ_{m^*} を生成する．
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 $\mathcal{D}$ を生成/蓄積する．
4. 履歴データ $\mathcal{D}$ 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. 提案法/従来法によって得られた推薦ルールを評価する．

ここで，履歴データの準備期間に該当する 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，推薦する商品を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，推薦ルールの運用期間に該当する 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，割引総利得を算出する．具体的な計算には，式 (16)，(17) を用いた．

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（何も購入しない場合は利得 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**（真のモデルを含む場合），**表 2**（真のモデルを含まない場合）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した場合，および，複数のモデルを重み付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，値が大きいほど平均的により多くの割引総利得が得られていることを示し，各行において上位 2 つの割引総利得を太字で表示している．また，表 1 の最下行は，列ごとの平均値を表す．

表 1 より，モデルを 1 つに固定した結果において，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．また，真のモデルと一致していない場合には，真のモデル

表 3　実験結果：同じ商品を 2 回購入しない場合（真のモデルを含む場合）
Table 3　Results: In the case that the same items are not purchased twice ($m^* \in \mathcal{M}$).

	モデルを 1 つに固定した場合						
真のモデル	1	2	3	4	5	6	提案法
モデル m_1	43.40	**43.77**	**44.52**	42.53	41.94	41.81	42.60
モデル m_2	38.91	43.25	**43.58**	39.49	42.14	41.38	**43.73**
モデル m_3	38.36	38.31	**42.81**	38.44	37.85	41.86	**43.33**
モデル m_4	39.92	38.01	40.01	**43.74**	43.36	42.06	**44.18**
モデル m_5	38.30	38.04	38.85	39.53	**42.52**	41.20	**42.52**
モデル m_6	38.01	37.76	38.00	38.11	37.42	**41.66**	**39.17**
平均	39.48	39.86	41.29	40.31	40.87	41.66	**42.59**

表 4　実験結果：モデルベースアルゴリズムとの比較結果
Table 4　Results: The comparison with the model-based approach.

	真のモデルを含む場合		真のモデルを含まない場合	
真のモデル	提案法	モデルベース	提案法	モデルベース
モデル 1	**121.83**	109.78	**121.83**	120.11
モデル 2	**131.10**	114.56	**128.06**	118.33
モデル 3	**126.67**	110.69	117.22	**117.77**
モデル 4	**126.75**	111.00	**130.00**	119.10
モデル 5	**129.10**	112.87	118.20	**120.90**
モデル 6	**116.77**	114.67	116.77	**118.26**
平均	**125.37**	112.26	**122.02**	119.08

と一致したモデルを用いた結果と比べて，おおむねより少ない割引総利得が得られていることも分かる．その他，全体的な傾向として，状態遷移確率が高次のマルコフ連鎖によって表現されているほど，より多くの割引総利得が得られている．

これに対して，提案法は，モデルを 1 つに固定した場合と比べ，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．

次に，表 2 より，真のモデルを含まない場合には，おおむね，モデルを 1 つに固定した場合の中で最も良い結果が得られているモデルに次ぐ割引総利得が得られていることが分かる．これは，仮定するモデル集合の中で，真のモデルに似たモデルの重みは大きく，あまり似ていないモデルの重みは小さく設定されたと解釈できる．モデル集合に含まれるマルコフ連鎖よりも，高次のマルコフ連鎖によって真の状態遷移確率が表現される場合（真のモデルが 5 の場合）においても，提案法が有用であることが確認できる．つまり，真のモデルが含まれない場合においても，複数のモデルを重み付けることの有効性を示唆している．

表 1，表 2 の実験結果は，過去に購入したことのある商品を再度購入する可能性がある場合の結果である．これに対して**表 3** は，1 度購入された商品は 2 度と購入されない状態遷移確率を設定値として用いた場合の結果である．表の見方は表 1 と同様である．ここで，同一商品の購入は 1 回きりであるため，推薦対象ユーザへの推薦回数は 10 とした．表 3 より，同一商品の購入が 1 回きりの場合であっても，先の結果と同様に，提案法の方が平均的により多くの割引総利得が得られていることが確認できる．ユーザの購買傾向が状態遷移確率として正しく反映され，その結果，有用な推薦ルールが導出できているといえる．

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．

最後に，モデルベールアルゴリズムとの比較結果を**表 4** に示す．ここで，今回の評価実験においては，多項分布に基づく手法をモデルベースアルゴリズムとして用いた．具体的には，N 人分の購入商品履歴のみを用いて多項分布のパラメータ（各商品の購入確率）を最尤推定した後，最も購入される確率の高い順に商品を推薦していくものとした．表 4 より，真のモデルを含む場合においては，つねに提案法がより多くの割引総利得を得ていることが分かる．また，真のモデルを含まない場合には，モデルベースアルゴリズムの方が良い結果を得ている場合があるものの，平均的には提案法の方が良い結果を得ていることが分かる．以上の結果から，長期的な視点に立ち，より多くの利益を得ようとする推薦目的のもとでは，購入される確率が高い商品を推薦するだけでは不十分であることが分かる．

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．

提案法の特徴は，表現能力のより高いモデルのもとで，推薦商品履歴を考慮していること，および，複数の推薦結果を考慮しているという点にある．また，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．提案した枠組みを用いることで，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．提案法の有効性を確認するため人工データを用いた評価実験を行い，1 つの固定されたモデルを利用する場合と比べて，複数のモデルを重み付ける提案法は平均的により多くの利得が得られることを示した．

今後の取組みとして，購入商品履歴以外の変数をマルコフ決定過程モデルの状態に組み込むことが考えられる．たとえば，ユーザのデモグラフィック情報や商品購入時の時間帯など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．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，さら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．

ただし，提案法は複数のモデルを用いるため，1 つのモデルを固定した場合と比べて計算量がかかるという課題がある．そこで今後は，たとえば文献 [9] で提案されている計算アルゴリズムなどを参考にすることで，複数のモデルの重み付け計算を効率良く行う方法についても検討していきたい．

謝辞　本研究を行う機会を与えてくださった，株式会社 NTT データ技術開発本部木谷強本部長，ならびに，同技術開発本部サービスイノベーションセンタ上島康司センタ長，中川慶一郎部長に深く感謝いたします．

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 Trans. Knowledge and Data Engineering*, Vol.17, No.6 (2005).
[2] Bellman, R.: *Dynamic Programming* (*Princeton Landmarks in Mathematics*), Princeton University Press (2010).
[3] Bernardo, J.M. and Smith, A.F.M.: *Bayesian Theory*, Wiley (2000).
[4] Box, G.E.P. and Tiao, G.C.: *Bayesian Inference in Statistical Analysis*, Wiley (1992).
[5] G. Shani, D. Heckerman and Brafman, R.I.: An MDP-Based Recommender System, *Journal of Machine Learning Research*, Vol.6, pp.1265–1295 (2005).
[6] Hofmann, T.: Latent semantic models for collaborative filtering, *ACM Trans. Information Systems*, Vol.22, No.1, pp.89–115 (2004).
[7] Iwata, T., Saito, K. and Yamada, T.: Recommendation Method for Improving Customer Lifetime Value, *IEEE Trans. Knowledge and Data Engineering*, Vol.20, No.9, pp.1254–1263 (2008).
[8] Linden, G., Smith, B. and York, J.: Amazon.com recommendations: item-to-item collaborative filtering, *IEEE Internet Computing*, Vol.7, No.1, pp.76–80 (2003).
[9] Matsushima, T. and Hirasawa, S.: Bayes universal coding algorithm for side information context tree models, *IEEE International Symposium of Information Theory* (*ISIT2005*), pp.2345–2348 (2005).
[10] Owen, S., Anil, R., Dunning, T. and Friedman, E.: *Mahout in Action*, Manning Publications Co. (2011).
[11] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proc. ACM CSCW1994*, pp.175–186 (1994).
[12] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.: Item-Based Collaborative Filtering Recommendation Algorithms, *Proc. ACM WWW2001*, pp.285–295 (2001).
[13] Si, L. and Jin, R.: Unified Filtering by Combining Collaborative Filtering and Content-Based Filtering via Mixture Model and Exponential Model, *Proc. ACM CIKM2004*, Washington D.C., US, pp.156–157 (2004).
[14] Su, X. and Khoshgoftaar, T.M.: A Survey of Collaborative Filtering Techniques, *Advances in Artificial Intelligence* (2009).
[15] White, T.: *Hadoop: The Definitive Guide*, O'Reilly (2010).
[16] Xue, G.-R., Lin, C., Yang, Q., Xi, W., Zeng, H.-J., Yu, Y. and Chen, Z.: Scalable Collaborative Filtering Using Cluster-based Smoothing, *Proc. ACM SIGIR2005*, Salvador, Brazil (2005).
[17] 森村英典，高橋幸雄：マルコフ解析，日科技連 (1979).
[18] 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).
[19] 繁桝算男：ベイズ統計入門，東京大学出版会 (1985).
[20] 桑田修平，上田修功：一括予測型協調フィルタリング，情報処理学会論文誌（数理モデル化と応用），Vol.48, No.SIG 15 (TOM 18), pp.153–162 (2007).
[21] 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).
[22] 小野智弘，本村陽一，麻生英樹：ベイジアンネットによる映画コンテンツ推薦方式の検討，電子情報通信学会技術研究報告（NC），Vol.104, No.348, pp.55–60 (2004).
[23] 金子哲夫：マルコフ決定理論入門，槇書店 (1973).
[24] 神嶌敏弘：推薦システムのアルゴリズム（1），人工知能学会誌，Vol.22, No.6, pp.826–837 (2007).
[25] 神嶌敏弘：推薦システムのアルゴリズム（2），人工知能学会誌，Vol.23, No.1, pp.89–103 (2008).
[26] 神嶌敏弘：推薦システムのアルゴリズム（3），人工知能学会誌，Vol.23, No.2, pp.248–263 (2008).

[27]　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).

桑田 修平　（正会員）

平成 13 年早稲田大学理工学部卒業．平成 15 年同大学院理工学研究科修士課程修了．（株）NTT データ，日本電信電話（株）NTT コミュニケーション科学基礎研究所を経て，平成 20 年（株）NTT データ技術開発本部，現在に至る．博士（工学）．統計的学習の応用に関する研究に従事．平成 19 年電子情報通信学会 PRMU 研究会研究奨励賞受賞．電子情報通信学会，人工知能学会，OR 学会等各会員．

前田 康成　（正会員）

平成 7 年早稲田大学理工学部卒業．平成 9 年同大学院理工学研究科修士課程修了．日本電信電話（株），東日本電信電話（株），北見工業大学助手，助教を経て，平成 22 年同大学准教授，現在に至る．博士（工学）．統計的決定理論の学習問題への応用に関する研究に従事．電子情報通信学会等会員．

松嶋 敏泰　（正会員）

昭和 53 年早稲田大学理工学部工業経営学科卒業．昭和 55 年同大学院修士課程修了．同年日本電気（株）入社．昭和 61 年早稲田大学大学院理工学研究科博士後期課程入学．平成元年横浜商科大学講師．平成 3 年同大学助教授．平成 4 早稲田大学理工学部工業経営学科助教授．平成 9 年同大学教授．平成 19 年早稲田大学基幹理工学部応用数学科教授，現在に至る．知識情報処理および情報理論とその応用に関する研究に従事．博士（工学）．平成 13 年ハワイ大客員研究員．平成 23 年カリフォルニア州立大学バークレイ校客員教授．IEEE，人工知能学会，OR 学会，日本経営工学会等各会員．

平澤 茂一　（正会員）

昭和 36 年早稲田大学理工学部数学科卒業．昭和 38 年同大学同学部電気通信学科卒業．同年三菱電機（株）入社．昭和 56 年早稲田大学理工学部工業経営学科（現在経営システム工学科）教授．平成 21 年早稲田大学名誉教授，早稲田大学理工総研名誉研究員，サイバー大学 IT 総合学部客員教授，現在に至る．情報理論とその応用，ならびに，計算機応用システムの開発，情報検索システム等の研究に従事．工学博士．昭和 54 年 UCLA 計算機科学科客員研究員，昭和 60 年ハンガリー科学アカデミー，昭和 61 年伊トリエステ大学客員研究員，平成 14 年 UCLA 計算機科学科訪問教員．平成 5 年電子情報通信学会小林記念特別賞，業績賞受賞．IEEE Life Fellow，電子情報通信学会，経営情報学会等各会員．