# Synapse: 文脈と時間経過に応じて推薦手法の選択を最適化するメタ推薦システム

三宅 悠介[1,2,a)]　峯 恒憲[3]

**概要**：EC サイトで扱う商品種類増大に伴う情報過多問題を解決するため，効果的な推薦手法を選択することが重要である．推薦手法の有効性は様々な要因や時間の経過によって左右されるため，最適な推薦手法の選択には実環境での継続的な評価が不可欠である．しかしながら，実環境での評価では機会損失が課題となる．本研究では，この実環境での評価を，文脈と時間の経過を考慮した多腕バンディット問題とみなして解くことで，機会損失を抑えながら最適な推薦手法を自動かつ継続的に選択するメタ推薦システムを提案する．評価では，実際の EC サイトから取得したデータを用いて最適な推薦手法を選択するシミュレーションを実施した．実験の結果，提案システムが，評価時に生じる機会損失を抑え，文脈と時間の経過を考慮しない場合と比較して累積クリック数を約 9.7%増加させる効果があることを確認した．

## 1. はじめに

EC サイトの市場規模の成長 [1] に伴い，取り扱う商品の種類は増大している．EC サイト利用者の通常の行動では全ての商品を見て回ることは困難であるため，利用者が関心のある商品を効率的に探せるよう，推薦システムは EC サイトにとって不可欠となっている．推薦システムは何らかの方策に基づいて利用者が興味を持つ商品を選定する．そこで採られる方策は推薦手法と呼ばれ，多くの手法が提案されている (e.g.,[2], [3], [4], [5])．そのため，利用者の要求を満たす効果的な推薦手法を選択することが EC サイトの運営者にとって重要となる．

しかしながら，推薦手法の有効性は評価時点での精度や応答速度，扱う商品特性の差異，評価者の状況といった様々な要因に左右されるため，EC サイトにとって，どの推薦手法が最も効果的かを予め知ることは困難である．そのため，各 EC サイトにとっての最適な推薦手法は，各環境での利用者の評価に基づき選定されなければならない．

一方で，実環境における推薦手法の比較評価では，評価の劣る推薦手法を候補として利用し続ける場合や，評価の見切りが早過ぎることで長期的な評価で勝る推薦手法を利用できない場合のような，評価期間中の機会損失が課題となる．これらの機会損失を抑えるためには，ある時点での評価の高い推薦手法を利用しながら，他の候補との評価を並行して行う必要がある．

この利用と評価のトレードオフの最適な解を求める問題は，多腕バンディット問題 [6] として知られており，この問題を解くための解法が提案されている (e.g., [7], [8], [9], [10])．しかしながら，これらの解法は推薦手法の以下の 3 点の特徴の全てを考慮することができないため，機会損失を十分に減らすことができない．第一は，推薦手法の有効性が様々な要因によって変化する点である．推薦手法の相対的な性能の優劣を助長する要因に関する研究が報告されており (e.g., [11], [12], [13]) ，これらの要因を文脈として扱える解法が不可欠である．第二は，推薦手法の有効性が時間の経過に伴って変化する点である．例えば，内容ベース型推薦で用いる内容表現の改善や，協調型推薦で用いる利用者の嗜好情報の蓄積による精度向上が期待できる．加えて，EC サイトの規模変化に伴う推薦システム基盤の処理性能の変化や，EC サイトの負荷増大による間接的な推薦の遅延，推薦システムの不具合による部分的な推薦手法の精度や速度の低下も発生しうる．推薦手法の精度や応答速度はその有効性に影響を与える [4] ことから，推薦システムの状況の変化も考慮することが望ましい．そのため，推薦手法の有効性が非定常であることを前提とした解法が必要となる．第三は，推薦手法間の相対的な有効性が逆転する点である．協調型推薦 [5] は，商品に対する利用者集団の嗜

[1] GMO ペパボ株式会社 ペパボ研究所
Pepabo R&D Institute, GMO Pepabo, Inc., Tenjin, Chuo ku, Fukuoka 810-0001 Japan
[2] 九州大学 大学院システム情報科学府 情報知能工学専攻
Department of Advanced Information Technology, Graduate School of ISEE, Kyushu University
[3] 九州大学 大学院システム情報科学研究院 情報知能工学部門
Faculty of Information Science and Electrical Engineering, Kyushu University
[a)] miyakey@pepabo.com

好情報を用いて商品を推薦する．この推薦手法では嗜好情報が不足する状況で推薦精度が低下する，いわゆるコールドスタート問題 [14] が存在する．このような推薦手法は，嗜好情報の蓄積に伴い精度を次第に向上させ他の推薦手法に比べて有効となっていく可能性があるため，ある時点で劣った推薦手法に対しても継続的に注意を払うような解法が望ましい．

本研究では，多腕バンディットを用いて，EC サイトの文脈と時間の経過に応じて推薦手法の選択を自動的かつ継続的に最適化する推薦システムを提案する．ここで文脈とは，推薦手法の相対的な優劣の差を助長し選択に影響を及ぼす要因の組み合わせによって表現される状態と定義する．加えて，同じ文脈であっても，推薦手法の有効性が変動する状況も考慮する．文脈や時間の経過に応じて推薦方法を選択する解法として，現在の状態の迅速かつ正確な推定が求められる動的環境に適した TVTP（Time-varying Thompson Sampling）[15] に着目した．しかし，推薦手法の有効性が逆転する状況には対処できないため，提案システムでは，この状況に対処可能となるよう TVTP を拡張した Aggressive Exploration TVTP (AE-TVTP) を考案し採用する．

本研究の主な貢献を以下に示す．

(1) EC サイトの文脈と時間の経過を考慮する多腕バンディット解法 AE-TVTP を用いて，自動かつ継続的に推薦手法の選択を最適化する推薦システムを提案する．
(2) 実際の EC サイトから収集したデータを用いて，文脈や時間の経過によって推薦手法の有効性が変化することを示す．
(3) 同データによるシミュレーションの結果，提案システムが評価時に生じる機会損失を抑え，文脈と時間の経過を考慮しない場合と比較して累積クリック数を約 9.7%増加させる効果があることを示す．

本報告の構成を述べる．2 節では関連研究を紹介し，EC サイトにおける推薦手法の選択の課題を述べる．3 節では 2 節で述べた文脈と時間の経過に応じて推薦手法の選択を最適化する課題を解決する提案システムを説明する．4 節では提案システムの有効性を評価し，5 節でまとめる．

## 2. 関連研究

### 2.1 多腕バンディット問題

利用と評価のトレードオフの最適な解を求める問題は，多腕バンディット問題として知られている．この問題は，腕と呼ばれる複数の候補から得られる報酬を最大化する問題である．プレイヤーは各試行で 1 つの腕を選択し，その腕から報酬を得る．各腕はある確率分布に従い報酬を生成するが，プレイヤーは試行の結果からこの確率分布を推測しなければならない．そのため，プレイヤーはある時点の腕ごとの評価に基づき，最も評価の高い腕を用いながらも，真に評価の高い腕の探索を並行して行う．この問題に対する解法では，ある時点で最も評価の高い腕を用いることを活用，各腕の評価を行うことを探索と呼び，これらの活用と探索，報酬による評価の見直しを繰り返し行うことで，短期的には探索による機会損失を，長期的には腕の固定化による機会損失を低減する．同問題に対する基本的な解法には，$\epsilon$-Greedy [16]，UCB1 [17]，Thompson Sampling [18] などが知られている．しかし，これらの解法では，報酬の確率分布が常に同じであるという仮定が置かれている．文脈と時間の経過を考慮して推薦手法を比較評価するためには，報酬の確率分布が変化する問題設定における解法が必要である．多腕バンディット問題では，報酬の確率分布が変化する環境を想定した問題設定について 2 種類の拡張が図られている．

文脈付き多腕バンディット問題では，複数の要因パラメータからなる文脈に応じて腕から得られる報酬の確率分布が決定される [19]．同問題の解法には UCB1 を拡張した LinUCB [19] や Thompson Sampling を同様に拡張した Linear Thompson Sampling (LTS) [20] が提案されている．

非定常な多腕バンディット問題は，同じ文脈においても報酬分布が時間経過によって変化する問題である．同問題の解法では，腕の報酬分布が変化した際に，不利な腕を使い続ける機会損失を抑える必要がある．同問題では，大きく 2 つの課題の解決が求められる．一つ目の課題は，選定する腕の偏りである．定常な問題設定では，試行回数の増加に伴って評価の高い腕を選定する割合を高めるアプローチが採用される．一方，非定常な問題設定では，腕の評価が逆転する可能性があるため，腕の選定に偏りがあると評価の低い腕の変動の検出が遅れてしまう．そのため [21] や [22] では，評価の低い腕に対しても必要な数の試行機会を意図的に設けている．二つ目の課題は，腕の評価の更新である．非定常な問題設定では，腕の報酬分布が変化するため，過去に観測した報酬に捉われずに腕の再評価を迅速に実施しなくてはならない．この課題に対しては，大きく 3 つ，変化検出型 [23]，減衰型 [24]，状態空間モデル型 [25] のアプローチが提案されている．

### 2.2 多腕バンディット解法を用いた推薦手法の選択

実環境での評価をもとに多腕バンディットを用いて最適な推薦方法を選択する推薦システムの研究も行われている．コールドユーザに対するモデル選定に多腕バンディットを用いる推薦 [7] では，予測評価行列をクラスタリングしたものを推薦モデルとしてバンディットで選定するものの，推薦手法の選定において文脈を考慮しない．[8], [9] は，腕となる推薦手法がコンテキストを扱えるが，その選定における多腕バンディットの解法では文脈を考慮できない．腕の選定における文脈考慮の欠如は，文脈によって推薦手法の相対的な優劣の差がある環境において機会損失に

つながる．そこで，文脈に応じた最適な推薦手法を多腕バンディットによって選定する推薦システム [10], [26] が提案されている．これらはロジスティック回帰モデルを組み込んだ多変量の回帰が可能なモデルや文脈付きの解法を用いることで文脈を考慮することができる．しかし，これらを含むいずれの先行研究でも，2.1 項で述べた非定常な問題に対する対策を行なっておらず，時間の経過に伴う推薦手法の有効性の変動への追従性が充分ではない．文脈と時間の経過を考慮した最適な推薦手法を選定するには，文脈付き，かつ，非定常な多腕バンディット問題の解法を用いる必要がある．2.1 項で紹介した解法は，文脈または非定常のいずれか一方のみを扱っている．

両方を考慮した解法として，[27] は利用者の嗜好傾向の変化を扱う多腕バンディットの研究に変化検出アプローチを採用した．この解法では商品特徴量に対する多次元の係数からなる利用者の評価を推定し，その変化を検出する．[28] は，複数の文脈付き多腕バンディットの解法を腕として，各腕の報酬予測の誤差の変化を検出する．これらの変化検出のアプローチでは，検出後に変化前の試行履歴を破棄するため，誤検出であった場合に，不要な探索が増えてしまう．特に要因ごとの観測回数に偏りがある場合にこのリスクが増加する．

Decay LinUCB [29] は LinUCB に割引の概念を導入し，試行回数や平均報酬に減衰パラメータ $\gamma \in (0, 1)$ を乗じた結果を腕の評価に利用する．変化への追従性を高めるには $\gamma$ を小さくする必要があるが，要因ごとの観測回数に偏りがある場合，ある文脈において活用が行われない可能性がある．すべての文脈で活用と探索を適切に行うために $\gamma$ を大きくすると，追従性を高めることができなくなる．

Time varying Thompson sampling (TVTP) [15] は，状態空間モデルの一種である粒子フィルタを用いて潜在的な状態の変化を捉える．またこのモデルの推定したパラメータを従来の文脈付きの解法である LTS と統合することで文脈付き，かつ，非定常な問題を扱うことができる．これらと本研究で提案する手法は，文脈と時間の経過の考慮に加え，腕の有効性が逆転する環境においても迅速に変化に追従可能な点で異なる．

## 3. 提案手法

本研究では，EC サイトの文脈と時間の経過に応じて推薦手法の選択を自動的かつ継続的に最適化する推薦システム Synapse を提案する．提案システムの処理フローを図 1 に示す．はじめに評価対象となる任意の推薦手法が Method として提案システムである Synapse に登録される．提案システムでは，推薦要求 (Search) に対して登録された推薦手法の中から，AE-TVTP を用いて，文脈と時間の経過を考慮した最適な推薦手法を選択する．次に，提案システムは選択した推薦手法から得られた推薦結果 (Result) を要求元

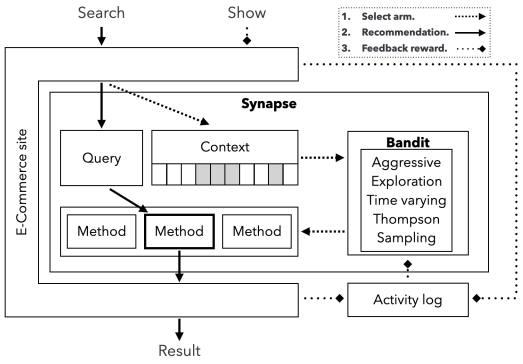


**図 1** 提案システムの処理フロー

に返す．最後に，推薦手法からの推薦結果とこれに対する利用者の行動 (Show) を Activity Log として記録する．この記録は，以降の推薦手法の選択の改善に利用される．提案システムはこれらの一連の処理を定期的に実行することで推薦手法の選択を自動的かつ継続的に最適化する．

### 3.1 推薦手法の登録

提案システムでは，共通のインタフェースを持つモジュールとして実装した推薦手法を登録する．各推薦手法は共通のインタフェースを持つことで，アルゴリズムの詳細に依存せず提案システムにおいて等価的に扱うことができる．提案システムでは，入力として推薦要求を行ったユーザと閲覧中の商品の識別子を，出力として入力に対する推薦結果である商品の識別子一覧を，インタフェースとして定義する．図 1 における Query と Result はこのインタフェースを満たす．提案システムでは，このインタフェースに沿う任意の推薦手法を利用できる．

### 3.2 文脈と時間の経過に応じた推薦手法の選択

TVTP [15] は，現在の状態を素早く正確に推定する必要がある動的な環境に適しているため，文脈や時間の経過に応じた推薦手法の選択に利用できる．他の，変化検出や減衰を用いる解法では，2.2 項で述べたように，一定期間の観測値を用いるため，速度と精度の間にトレードオフが生じてしまう．一方，TVTP も 2.1 項で非定常における最初の課題として述べた「有効性が逆転する状況」への対応が十分ではないことから，これに対処するために TVTP を拡張する．TVTP のアルゴリズムを Algorithm1 に示す．TVTP では腕 $k$ の $t$ 時点における報酬 $y_{k,t}$ を以下のように表す．

$$y_{k,t} \sim \mathcal{N}(x_t^{\mathsf{T}} w_{k,t}, \sigma_k^2) \tag{1}$$

この式において，$x_t$ は $t$ 時点の試行におけるコンテキスト情報，$w_{k,t}$ はそのコンテキストに対応する係数ベクトル，$\sigma_k^2$ は観測誤差の分散，$\mathcal{N}$ はこれらを平均，分散とする正規分布である．ここで，$\sigma_k^2$ の事前分布は，パラメータ $\alpha$ と $\beta$ を

持つ逆ガンマ分布と仮定する．また，$w_{k,t} = c_{w_k} + \theta_k \odot \eta_{k,t}$ である．右辺の初項 $c_{w_k}$ は定常項，第二項は非定常項であり，非定常な項はスケール項 $\theta_k$ と変動項 $\eta_{k,t}$ の要素毎の積である．そして，変動項 $\eta_{k,t}$ は $\eta_{k,t} \sim \mathcal{N}_m(\eta_{k,t-1}, \mathcal{I}_d)$ のようにランダムウォークに従い値が変動すると仮定する．ここで，$\mathcal{N}_m$ は多変量正規分布である．なお，$\mathcal{I}_d$ は $d$ 次元の対角行列であり，その対角成分は $\sigma^2_{v,k}$ である．$\sigma^2_{v,k}$ はカルマンフィルタにおける過程誤差の分散に相当する．TVTP では，この報酬モデルのパラメータ $c_{w_k}, \theta_k, \eta_{k,t}, \sigma^2_k$ の事後分布を粒子フィルタとカルマンフィルタを用いて推定する．なお，$c_{w_k}$ に対する事前分布のパラメータは $\mu_c, \Sigma_c$ であり，同様に $\theta_k$ に対しては $\mu_\theta, \Sigma_\theta$ で，$\eta_{k,t}$ に対しては $\mu_\eta, \Sigma_\eta$ であり，$\sigma^2_k$ に対しては $\alpha, \beta$ である．

TVTP における腕の選定は，LTS と同様に，推定したパラメータの事前分布からのサンプリングによって行う．まず，以下の式に従い係数ベクトル $\bar{w}_{k,t-1}$ を求める．

$$\bar{w}_{k,t-1} \sim \mathcal{N}_m(\bar{\mu}_{w_k}, \bar{\Sigma}_{w_k}) \tag{2}$$

ここで $\bar{\mu}_{w_k}$ と $\bar{\Sigma}_{w_k}$ は各腕に紐づく $p$ 個の粒子から，以下の式で求める．

$$\begin{aligned} \bar{\mu}_{w_k} &= \frac{1}{p}\sum_{i=1}^{p} \mu^{(i)}_{w_k}, \\ \bar{\Sigma}_{w_k} &= \frac{1}{p^2}\sum_{i=1}^{p} \sigma^{2(i)}_k \Sigma^{(i)}_{w_k}. \end{aligned} \tag{3}$$

ここで $p$ は粒子数である．なお，$\mu_{w_k}, \Sigma_{w_k}$ は各粒子で推定した結果から以下の式を用いて求めることができる．

$$\begin{aligned} \mu_{w_k} &= \mu_c + (\Sigma_{\eta_k} + \sigma^2_k \Sigma_\theta)^{-1}(\Sigma_{\eta_k}\mu_\theta + \sigma^2_k \Sigma_\theta \mu_\eta), \\ \Sigma_{w_k} &= \sigma^2_k \Sigma_c + \sigma^2_k \Sigma_\theta \Sigma_{\eta_k}(\Sigma_{\eta_k} + \sigma^2_k \Sigma_\theta)^{-1}. \end{aligned} \tag{4}$$

このようにして求めた係数ベクトル $\bar{w}_{k,t-1}$ とコンテキスト $x_t$ との内積が最も大きかった腕を選定する．これは Algorithm1 の 5 行目と 8 から 11 行目に相当する．

しかし TVTP は，試行回数の増加に伴い，式 (3) の $\bar{\Sigma}_{w_k}$ の値が急激に小さくなることで推薦手法の選定に偏りを生じさせてしまう．この課題の原因の一つは，式 (3) において粒子数 $p$ ではなく $p^2$ による除算がなされている点である．これにより，粒子数の増加に従い急激に $\bar{\Sigma}_{w_k}$ の値が減少する．別の原因は，式 (3) において $\sigma^{2(i)}_k$ で乗算がなされている点である．この操作は式 (4) により各粒子に対して既に施されているため冗長となる．そこで，式 (3) の $\bar{\Sigma}_{w_k}$ の算出を以下のように変更する．

$$\bar{\Sigma}_{w_k} = \frac{1}{p}\sum_{i=1}^{p} \Sigma^{(i)}_{w_k} \tag{5}$$

また，式 (4) の $\mu_{w_k}$ について，算出した値は真の係数ベクトルの値と異なることから，真の係数ベクトルの値となるよう以下のように改修する．

$$\mu_{w_k} = \mu_c + \mu_\theta \odot \mu_\eta \tag{6}$$

ここで，各項は式 (1) の $w_{k,t} = c_{w_k} + \theta_k \odot \eta_{k,t}$ の各値に対する事前分布の平均パラメータである．

我々は，このように TVTP を拡張して積極的な探索を促すことで，推薦手法の選択の偏りを減らし，時間の経過に伴う推薦手法の有効性の変化，特に有効性が逆転する状況での追従性を改善する．また，提案システムでは，様々な情報をコンテキスト情報 $x_t$ として扱うことができる．4 節で，商品カテゴリとして表現される商品特性の差異が，推薦手法を選択する際のコンテキスト情報として有効な要素であることを示す．この時，コンテキスト情報 $x_t$ は，入力インタフェースから得た閲覧中の商品が属する商品カテゴリから所属する商品カテゴリの場合に 1 を，そうでない場合に 0 を値とするダミー変数にエンコードし，商品カテゴリ数を次元数とするベクトルとして表現している．これは図 1 の Context に相当する．

### 3.3 継続的な推薦手法の評価

提案システムでは，利用者からの推薦要求の結果を記録する．記録には，時刻，コンテキスト情報，選択された推薦手法，ならびにその推薦結果である商品一覧と，その推薦結果に対する利用者の反応が含まれる．利用者の反応は，推薦要求の時刻以降で直近の閲覧や購入行動における商品が，推薦結果の商品一覧に含まれたか否かで判断される．システムは，直近の行動の行動種別が，閲覧であればクリック，購入であればコンバージョンとして記録する．そして，これら報酬についての記録を一定期間ごとに集計し，報酬として受け取る．この報酬の反映は，Algorithm1 の 16 から 20 行目に従う．

## 4. 評価

### 4.1 評価データと推薦手法

本研究では，実際の EC サイトから採取した複数の推薦手法の商品カテゴリごとのクリック率の推移実績データを用いて，提案システムの有効性を評価する．対象の EC サイトの推薦システムでは，6 つの推薦手法が利用でき，推薦時にはこのうち一つの推薦手法を，CTR を指標とした $\epsilon$-Greedy [16] を用いて選択している．また，閲覧中の商品の商品詳細ページにおいて EC サイトの推薦システムが選定した関連商品のリスト（最大 12 件）が合わせて表示されている．なお，商品はその特性に応じて EC サイトの提供する 18 個のカテゴリのいずれかに所属している．評価実験では集計期間中の推薦回数が著しく少なかった 2 つを除く以下の 4 つの推薦手法の実績データを用いた．

**Algorithm 1:** Time varying Thompson sampling (TVTP) [15].

**procedure** MAIN($p$): ▷ main entry
  Initialize arms with $p$ particles.
  **for** $t \leftarrow 1, T$ **do**
    Get $x_t$.
    $a^{(k)} = \arg\max_{j=1,K} EVAL(a^{(j)}, x_t)$
    Receive $r_{k,t}$ by pulling arm $a^{(k)}$.
    $UPDATE(x_t, a^{(k)}, r_{k,t})$.
**procedure** EVAL($a^{(k)}, x_t$): ▷ get a score for $a^{(k)}$, given $x_t$.
  Learn the parameters based on all particles' inferences of $a^{(k)}$ by Equation (2).
  Compute a score based on the parameters learnt.
  **return** the score.
**procedure** UPDATE($x_t, a^{(k)}, r_{k,t}$): ▷ update the inference.
  **for** $i \leftarrow 1, p$ **do** ▷ Compute weights for each particle.
    Compute weight $\rho^{(i)}$ of particle $\mathcal{P}_k^{(i)}$.
  Re-sample $\mathcal{P}'_k$ from $\mathcal{P}$ according to the weights $\rho^{(i)}$s.
  **for** $i \leftarrow 1, p$ **do** ▷ Update statistics for each particle.
    Update the sufficient statistics $(\mu_{\eta_k}, \Sigma_{\eta_k})$ for $\eta_{k,t}$.
    Sample $\eta_{k,t} \sim \mathcal{N}_m(\mu_{\eta_k}, \Sigma_{\eta_k})$.
    Update the sufficient statistics $(\alpha, \beta, \mu_c, \Sigma_c, \mu_\theta, \Sigma_\theta)$ for $\sigma_k^2, c_{w_k}, \theta_k$.
    Sample $\sigma_k^2 \sim \mathcal{IG}(\alpha, \beta)$.
    Sample
    $$(c_{w_k}^\top, \theta_k^\top)^\top \sim \mathcal{N}_m((\mu_c^\top, \mu_\theta^\top)^\top, \begin{bmatrix} \Sigma_c & 0 \\ 0 & \Sigma_\theta \end{bmatrix}).$$

(1) 閲覧導線による推薦 (browsing_path)
(2) 人口統計学的属性による推薦 (demographic)
(3) お気に入り履歴によるアイテムベース協調型推薦 (llr)
(4) 画像による内容ベース型推薦 (similar_image)

評価実験のため，2019 年 6 月 20 日から 8 月 4 日までの各推薦手法による推薦回数と推薦結果の閲覧回数から，各商品カテゴリに対する一時間ごとの各推薦手法のクリック率ならびに推薦試行回数を集計した．総推薦試行回数は 2,252,236 回であった．各推薦手法のクリック率には，各時間から 3 日前までの推薦回数に対する閲覧回数の指数移動平均を用いる．なお，クリック率の平均を安定させるため，初期の 7 日間の集計期間については平均を算出するためだけに用いて以降の評価では利用しない．

利用する推薦手法の期間中のクリック率の推移を図 2 に示す．全ての商品カテゴリをまとめた場合，期間中に最もクリック率の高い推薦手法が similar_image から browsing_path へ切り替わった．

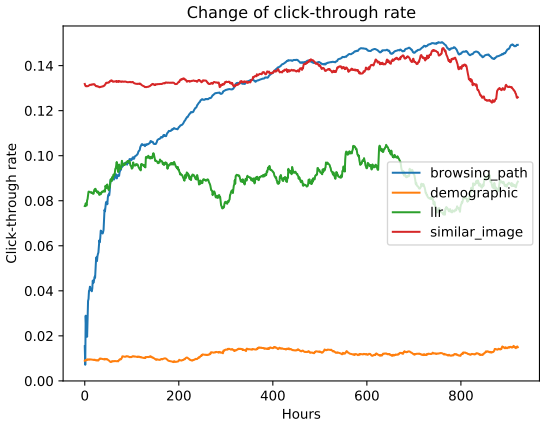


**図 2** 推薦手法ごとのクリック率の推移

### 4.2 評価方法

評価には 4.1 項で得た，各商品カテゴリに対する一時間ごとのクリック率と推薦試行回数の推移の実績データを用いる．評価は，この実績データの期間中に選択した推薦手法によって得られるクリック数のシミュレーションによって行う．ここで，推薦手法は多腕バンディットの解法によって選択され，各推薦手法は設定したクリック率のベルヌーイ分布に従い推薦結果がクリックされるものとする．乱数を用いた確率の計算結果を平均化するために，異なる乱数シードを用いてシミュレーションを 50 回行い，この平均を結果として用いた．また，各解法のパラメータ調整のため上述とは異なる乱数シードでの予備評価を 10 回実施した．なお，選択した推薦手法とその推薦手法から得られた報酬は，都度受け取ることはできず，一時間ごとの試行回数を経過した後に各シミュレーションで用いる多腕バンディットの解法へフィードバックされる．

提案システムでは，AE-TVTP を用いて文脈と時間の経過に応じて推薦手法を選択する．そこで，文脈と時間経過のそれぞれを考慮する場合としない場合を組み合わせた 4 つのシミュレーショングループ A，B，C，D を設定する．

- **A グループ**は文脈も時間の経過も考慮しないグループである．これは期首の時点で全ての商品カテゴリを通して最もクリック率の高かった推薦手法を全期間で一貫して用いる方針に相当する．また，ベースラインとして，全ての推薦手法を均等に用いる A/B テストのシミュレーションもこのグループに含める．
- **B グループ**は文脈を考慮するグループである．A グループと同様のシミュレーションを行うが，商品カテゴリごとにクリック率を判断する．これは期首の時点で各商品カテゴリで最もクリック率の高かった推薦手法を全期間で一貫して用いる方針に相当する．
- **C グループ**は時間の経過を考慮するグループである．

時間の経過に伴う推薦手法のクリック率の変化を多腕バンディット問題の解法によって探索する方針に相当する．各解法には常に一種類の商品カテゴリであるようコンテキスト情報を指定する．

- **D グループ**は文脈と時間の経過の両方を考慮するグループである．C グループと同様のシミュレーションを行うが，各時間に加え，商品カテゴリごとにもクリック率を判断する．

時間の経過を考慮する C と D のグループでは，以下の 3 つの解法を用いて評価する．

(1) **LTS**

(2) **TVTP**: LTS に対し，時間の経過に対する迅速な評価の更新の効果を比較する．係数ベクトルの算出には式 (6) を用いる．

(3) **AE-TVTP**: TVTP に対し，腕の選定の偏りの改善効果を比較する．

なお，TVTP と AE-TVTP には，調整用パラメータとして過程誤差 $\sigma_v^2$ が存在するが，予備評価において候補の値 $\sigma_v \in \{0.01, 0.001, 0.0001\}$ のうち最も累積報酬が多かった値を結果として採用する．粒子数 $p$ は $p \in \{1, 5, 15, 48\}$ に対し，予備評価では 5 以上であっても累積報酬の改善が確認されなかったため，実行時間の最も短い $p = 5$ を全てのシミュレーションで用いる．また，LTS，TVTP，AE-TVTP の観測誤差の分散 $\sigma_k^2$ は各解法で推定される値を用いる [15]．

評価指標には，利用者の推薦要求の満足度を最大化する性能を測るため，各シミュレーションによって得られた報酬の合計である累積報酬ならびに累積リグレットを用いる．累積リグレットは推薦手法のうち最大の期待値と選択した推薦手法の期待値の差を期間までに合計したものである．加えて，各解法の特性を調べるため各解法による推定クリック率も計測する．

### 4.3 文脈と時間経過の考慮に対する効果

4.2 項の評価方法でシミュレーションを行った累積報酬の結果を図 3 に示す．また，図 4 に各シミュレーションの最終累積報酬について，A グループの推薦手法を一貫して用いるシミュレーションとの差を示した．図 5 には，商品カテゴリごとの推薦手法の有効性の変動の大きさと，累積報酬の改善率との相関を示した．推薦手法の有効性の変動の大きさには，期首に最もクリック率が高かった推薦手法に対して，各時点で最大のクリック率との差を期間までの合計を用いた．また，累積報酬の改善率には，D グループの各解法の累積報酬と，B グループのうち推薦手法を一貫して用いた結果に対する比を用いた．

#### 4.3.1 文脈の考慮

図 3 の左上は文脈も時間の経過も考慮しない A グループの結果である．全ての推薦手法を均等に用いるシミュレーション (random) と，期首に最もクリック率の高かった推薦手法を一貫して用いるシミュレーション (best_at_first_time) を行った．それぞれの累積報酬は 200053.1 と 285203.6 であった．best_at_first_time はクリック率の低い推薦手法の利用を排除することで random と比べ累積報酬が増加した．

図 3 の左下は B グループの結果である．A グループと同様のシミュレーションを行ったが，商品カテゴリごとにクリック率を判断する点が異なる．それぞれの累積報酬は

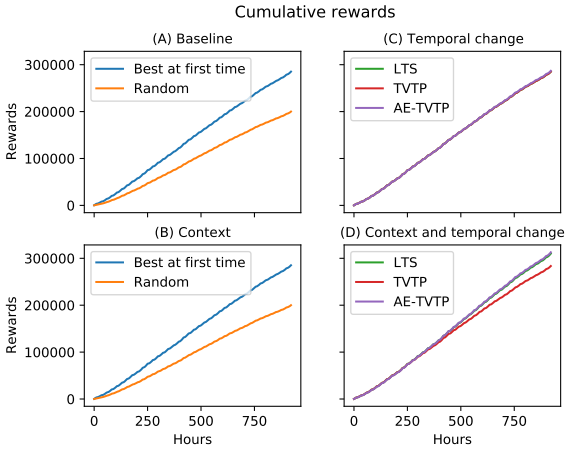


図 3 累積報酬のシミュレーション間比較

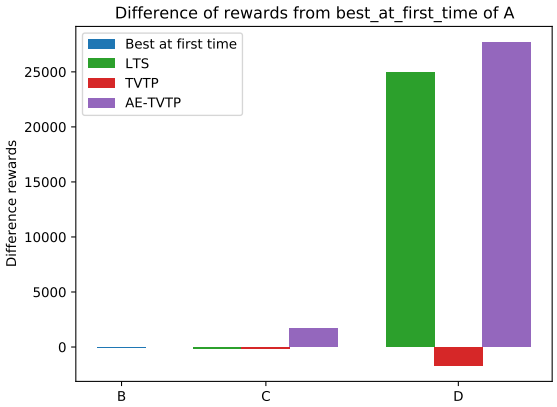


図 4 A グループを基準とした累積報酬の差の比較

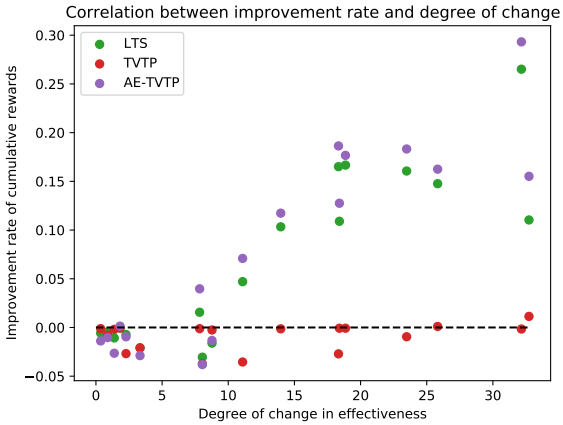


図 5 推薦手法の有効性の変動度合いと累積報酬の改善率の相関

200053.1 と 285174.7 であった．商品カテゴリごとに期首に最もクリック率が高い推薦手法が 1 つの商品カテゴリを除いて同じであったことから，A グループとほぼ同じ累積報酬となった．

#### 4.3.2 時間経過の考慮

図 3 の右上は時間の経過を考慮する一方，文脈は考慮しない C グループの結果である．LTS，TVTP，AE-TVTP の累積報酬は順に 285027.7，285029.8，286906.4 であった．なお，TVTP と AE-TVTP における調整用のパラメタ値 $\sigma_{v,k}$ はそれぞれ 0.01, 0.001 である．図 4 の C グループの結果からも分かるように AE-TVTP の累積報酬は A グループと比較して改善した．一方で，LTS と TVTP は A グループと比較して累積報酬は下がった．これは，AE-TVTP と比較して追従が遅れたことで，結果的に探索のコストを回収できなかったことに起因する．

#### 4.3.3 文脈と時間経過の考慮

図 3 の右下は文脈と時間の経過を考慮する D グループの結果である．LTS，TVTP，AE-TVTP の累積報酬は順に 310104.4，283546.7，312865.8 であった．なお，TVTP と AE-TVTP のパラメタ値 $\sigma_{v,k}$ はそれぞれ 0.001, 0.0001 である．図 5 から，推薦手法の有効性の変動が小さい商品カテゴリでは，探索のコストを回収できず，全ての解法において累積報酬は改善しないことが分かる．反対に，推薦手法の有効性の変動が大きく，有効な推薦手法を切り替えないことによる損失が大きくなる商品カテゴリでは，LTS と AE-TVTP が大きく改善する．一方で，TVTP では累積報酬が伸びなかった．これは，探索が少ない課題から，機首に最もクリック率が高いと判断した推薦手法を期間中継続して利用したためである．これに対し，AE-TVTP では積極的な探索によって推薦手法の選定の偏りを解消した．

以上の結果，文脈と時間の経過を考慮した多腕バンディット問題の解法を利用した D グループの AE-TVTP でのシミュレーションにより，文脈と時間の経過を考慮せずにある時点で商品カテゴリを通して最もクリック率が高い推薦手法を一貫して利用し続けた A グループと比較して累積クリック数を約 9.7%増加させる効果があることを確認した．なお，各シミュレーションの累積報酬に対する $\chi^2$ 検定は有意水準 1%において有意である．

### 4.4 考察

図 6 に，先ほどの D グループのシミュレーションのうち二つの商品カテゴリ Clothing，Dolls の結果を示す．1 段目は対象の商品カテゴリにおける推薦手法ごとのクリック率，2 段目から 4 段目はそれぞれ LTS，TVTP，AE-TVTP による推定クリック率の推移である．5 段目は累積リグレットの推移である．なお，解法ごとのクリック率の推定値と累積リグレットの推移は，各解法の挙動の特性を明確にするため，50 回のシミュレーションの平均ではなく，各解法

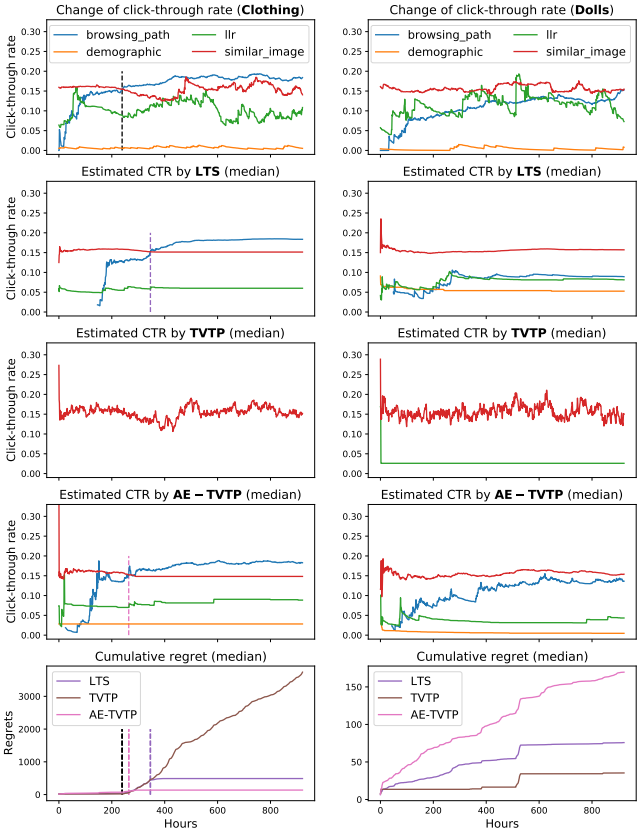


**図 6** D グループの各シミュレーションにおけるクリック率，推定クリック率，累積リグレットの商品カテゴリ間比較．縦線はクリック率または推定クリック率が逆転した時点を示す．

の累積報酬が中央値となった回の結果を示している．

図 6 の列左の商品カテゴリ Clothing では，ある時期から browsing_path が最もクリック率の高い推薦手法となった．2 段目と 4 段目から，LTS と AE-TVTP では時間の経過に伴いこの変化への追従が行えたことが分かる．特に提案手法で採用した AE-TVTP は LTS と比較して，変化後の browsing_path のクリック率を素早く推定できたことで，リグレットの増加を抑えることができている．一方で，期首から最もクリック率の高い推薦手法に変化がない期間では，LTS と比較して AE-TVTP のリグレットがわずかに増加した．また 3 段目では，TVTP が早急に推薦手法の評価を確定し，期首の評価に基づき選定を判断したことでリグレットが一層増加したことがわかる．列右の商品カテゴリ Dolls では，期間中，最もクリック率が高い推薦手法にほぼ変更がなかったため，全期間にわたり AE-TVTP のリグレットが増加した．同様の現象が他の 5 つの商品カテゴリ（Interior，Knitting，Houseware，Toys，Aroma）でも見られた．これらは AE-TVTP の積極的な探索機能に起因する．本評価では，総合的には累積報酬や累積リグレットの改善につながったが，今後は最もクリック率の高い推薦手法が変化しない期間においても機会損失を低減する適応的な探索手法の研究を進めたい．

## 5. おわりに

本研究では，多腕バンディットを用いて，文脈と時間の経過に応じて推薦手法の選択を自動的かつ継続的に最適化する推薦システムを提案し，その有効性を示した．実験から，推薦手法の選択に影響を及ぼす文脈を適切に選定すること，ならびに時間の経過を考慮した推薦手法の選定の最適化によって，考慮しない場合と比較して累積クリック数の向上に繋がることがわかった．また，推薦手法の優劣に変化のない期間では結果が逆転する可能性が確認されたことで，機会損失を低減する解法が重要であることも示唆された．今後の課題として，推薦手法の選択に影響を及ぼす効果の高い文脈の発見，ならびに特定の文脈において評価の高い推薦手法の確立による文脈ごとの効果向上の実現が挙げられる．

### 参考文献

[1] The Ministry of Economy, T. and (METI), I.: FY2019 Global Economy Survey for Formulating an Integrated Domestic and External Economic Growth Strategy (E-Commerce Market Survey) (2020).
[2] Burke, R.: Hybrid recommender systems: Survey and experiments, *User Modeling and User-Adapted Interaction*, Vol. 12, No. 4, pp. 331–370 (2002).
[3] Bobadilla, J., Ortega, F., Hernando, A. and Gutiérrez, A.: Recommender systems survey, *Knowledge-based Systems*, Vol. 46, pp. 109–132 (2013).
[4] Lops, P., De Gemmis, M. and Semeraro, G.: Content-based recommender systems: State of the art and trends, *Recommender Systems Handbook*, Springer, pp. 73–105 (2011).
[5] 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).
[6] Katehakis, M. N. and Veinott Jr, A. F.: The multi-armed bandit problem: decomposition and computation, *Mathematics of Operations Research*, Vol. 12, No. 2, pp. 262–268 (1987).
[7] Felício, C. Z., Paixão, K. V., Barcelos, C. A. and Preux, P.: A multi-armed bandit model selection for cold-start user recommendation, *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 32–40 (2017).
[8] Brodén, B., Hammar, M., Nilsson, B. J. and Paraschakis, D.: Ensemble recommendations via thompson sampling: an experimental study within e-commerce, *23rd international conference on intelligent user interfaces*, pp. 19–29 (2018).
[9] Cañamares, R., Redondo, M. and Castells, P.: Multi-armed recommender system bandit ensembles, *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 432–436 (2019).
[10] Santana, M. R., Melo, L. C., Camargo, F. H., Brandão, B., Soares, A., Oliveira, R. M. and Caetano, S.: Contextual Meta-Bandit for Recommender Systems Selection, *Fourteenth ACM Conference on Recommender Systems*, pp. 444–449 (2020).
[11] Ekstrand, M. and Riedl, J.: When recommenders fail: predicting recommender failure for algorithm selection and combination, *Proceedings of the sixth ACM conference on Recommender systems*, pp. 233–236 (2012).
[12] Braunhofer, M., Codina, V. and Ricci, F.: Switching hybrid for cold-starting context-aware recommender systems, *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 349–352 (2014).
[13] Anderson, A., Maystre, L., Anderson, I., Mehrotra, R. and Lalmas, M.: Algorithmic effects on the diversity of consumption on spotify, *Proceedings of The Web Conference 2020*, pp. 2155–2165 (2020).
[14] Ahn, H. J.: A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, *Information Sciences*, Vol. 178, No. 1, pp. 37–51 (2008).
[15] Zeng, C., Wang, Q., Mokhtari, S. and Li, T.: Online context-aware recommendation with time varying multi-armed bandit, *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 2025–2034 (2016).
[16] Sutton, R. S. and Barto, A. G.: *Reinforcement learning: An introduction*, MIT press (2018).
[17] Auer, P., Cesa-Bianchi, N. and Fischer, P.: Finite-time analysis of the multiarmed bandit problem, *Machine Learning*, Vol. 47, No. 2-3, pp. 235–256 (2002).
[18] Thompson, W. R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, Vol. 25, No. 3/4, pp. 285–294 (1933).
[19] Li, L., Chu, W., Langford, J. and Schapire, R. E.: A contextual-bandit approach to personalized news article recommendation, *Proceedings of the 19th international conference on World wide web*, pp. 661–670 (2010).
[20] Agrawal, S. and Goyal, N.: Thompson sampling for contextual bandits with linear payoffs, *International Conference on Machine Learning*, pp. 127–135 (2013).
[21] Liu, F., Lee, J. and Shroff, N.: A change-detection based framework for piecewise-stationary multi-armed bandit problem, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1 (2018).
[22] Cao, Y., Wen, Z., Kveton, B. and Xie, Y.: Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit, *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 418–427 (2019).
[23] Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O. and Sebag, M.: Multi-armed bandit, dynamic environments and meta-bandits, 2006, *NIPS-2006 workshop, Online trading between exploration and exploitation, Whistler, Canada* (2006).
[24] Gupta, N., Granmo, O.-C. and Agrawala, A.: Thompson sampling for dynamic multi-armed bandits, *2011 10th International Conference on Machine Learning and Applications and Workshops*, Vol. 1, IEEE, pp. 484–489 (2011).
[25] Granmo, O.-C. and Berg, S.: Solving non-stationary bandit problems by random sampling from sibling kalman filters, *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, pp. 199–208 (2010).
[26] 三宅悠介，峯恒憲: Synapse: 文脈に応じて継続的に推薦手法の選択を最適化する推薦システム, **電子情報通信学会論文誌** *D*, Vol. 103, No. 11, pp. 764–775 (2020).
[27] Hariri, N., Mobasher, B. and Burke, R.: Adapting to user preference changes in interactive recommendation, *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).
[28] Wu, Q., Wang, H., Li, Y. and Wang, H.: Dynamic ensemble of contextual bandits to satisfy users' changing interests, *The World Wide Web Conference*, pp. 2080–2090 (2019).
[29] Hassan, M. A.: Non-Stationary Contextual Multi-Armed Bandit with Application in Online Recommendations., PhD Thesis, University of Virginia (2015).