Vol.2015-MPS-103 No.30
Vol.2015-BIO-42 No.30
2015/6/24

推薦の粒度を学習の進展に応じて切り替えるmulti-armed banditアルゴリズム

宇田川 拓郎[1,a)] 渡邉 卓也[2,b)] 山中 章裕[1,c)] 室井 浩明[2,d)] 東山 昌彦[3,e)] 小田 哲[1,f)]
小田桐 優理[4,g)] 宮井 康宏[4,h)] 志村 誠[4,i)] 本庄 利守[1,j)]

概要：Multi-armed bandit アルゴリズムは、複数の選択肢から一つを選ぶ試行を繰り返しつつ、選択結果に応じて与えられる報酬を元に、選択肢の選び方を最適化する為のアルゴリズムである。同アルゴリズムは、ユーザへの推薦内容とそれに対する応答を元に、推薦内容をオンラインで最適化できることから、リアルタイム性の必要なニュースサイトにおける記事の推薦等に利用されている。その際、嗜好の多様なユーザを対象とするサービスにおいては、ユーザ個々の嗜好に合わせた推薦を行うことが望ましい。しかし、まだ行動履歴が十分に集まっていないユーザに対しては嗜好の学習が十分行われておらず、初期段階においては適切な推薦が行えないことが課題となる。そこで、そうしたユーザについてはまずユーザ集団全体の嗜好に基づいた推薦を行い、ある程度学習が進んだ段階で個別の嗜好に基づく推薦に切り替えるmulti-armed bandit アルゴリズムを提案する。また、ユーザ集団の行動をシミュレートした実験を通じて、提案アルゴリズムの有効性を示す。

A Multi-armed Bandit Algorithm to Switch Recommendation Granularity Based on Learning Progress

Takuro Udagawa[1,a)] Watanabe Takuya[2,b)] Akihiro Yamanaka[1,c)] Hiroaki Muroi[2,d)]
Masahiko Higashiyama[3,e)] Oda Satoshi[1,f)] Yuri Odagiri[4,g)] Yoshihiro Miyai[4,h)]
Makoto Shimura[4,i)] Toshimori Honjo[1,j)]

Abstract: Multi-armed bandit algorithms optimise a way to choose an option from several alternatives through the iteration of choice, based on observed rewards for each choice. They have been used as recommendation algorithms for online services such as web-based news article recommendations in order to optimise recommendations online based on what they recommended to users and their responses. For those services that target users having a wide variety of preferences, it is desirable to produce recommendations according to specific preferences of each user. However, it would be problematic that personalised recommenders tend to produce inappropriate recommendations for those users of whom not enough number of behavioural records have been accumulated. We propose a new multi-armed bandit algorithm that produce recommendations for those new users based on a global preference at first, and then offer personalised recommendations after enough preference of the user is obtained. We also show that the proposed algorithm is effective by conducting a simulation experiment.

[1] NTT ソフトウェアイノベーションセンタ
NTT Software Innovation Center
[2] エヂリウム株式会社
Edirium K.K.
[3] NTT データ先端技術株式会社
NTT DATA INTELLILINK Corporation
[4] 株式会社ドワンゴ
DWANGO Co., Ltd.
[a)] udagawa.takuro@lab.ntt.co.jp
[b)] sodium@edirium.co.jp
[c)] yamanaka.akihiro@lab.ntt.co.jp
[d)] muroi@edirium.co.jp
[e)] higashiyamam@intellilink.co.jp
[f)] oda.sathoshi@lab.ntt.co.jp
[g)] yuri_odagiri@dwango.co.jp
[h)] yoshihiro_miyai@dwango.co.jp
[i)] makoto_shimura@dwango.co.jp
[j)] honjo.toshimori@lab.ntt.co.jp

© 2015 Information Processing Society of Japan

Vol.2015-MPS-103 No.30
Vol.2015-BIO-42 No.30
2015/6/24

1. はじめに

今日広く普及しているウェブベースのオンラインサービスにおいては、サービスの内容や提供する商品の推薦内容等について、ユーザの行動を計測することでその良し悪しを評価することが行われている。複数の選択肢についてのユーザ集団の嗜好を調査し推薦の最適化を行いたい場合、ユーザ集団を分割してそれぞれに異なる推薦内容を提供しユーザの反応を測定する、いわゆる A/B テストを実施し、最も結果の良好であった推薦内容を選べばよい。しかし A/B テストを行った場合、もし選択肢の中に期待される水準よりもパフォーマンスの相当程度低いものが含まれていた場合、テスト期間中はその低いパフォーマンスに引きずられ、サービス全体のパフォーマンスが低下することが予想される。

パフォーマンスの低い選択肢の影響は、ユーザの嗜好を測定しつつ、同時にパフォーマンスの良好な選択肢をより多く提示する、オンラインでの最適化を行うことで緩和することができる。Multi-armed bandit アルゴリズムは、この、測定（探索）と良好な選択肢の提示（活用）の両立に対する一定の解を与えるような選択肢の選び方を決定するアルゴリズムである。各選択肢についての測定回数が過少であれば誤ったパフォーマンスの見積もりに基づいた提示を行う危険性があり、しかし良好な選択肢を多く提示できなければサービス全体のパフォーマンスが低下してしまう為、両者のバランスをとった選択肢の選び方をすることが必要である。

Multi-armed bandit アルゴリズムにおける最適化は、選んだ選択肢（腕）と、それに対する報酬に基づいて行われる。最も一般的な multi-armed bandit アルゴリズムの定式化は、regret を最小化する、というものである**?**。μ^* が最良の選択肢の報酬の期待値、$\mu_j(t)$ が t 回目の選択で選んだ選択肢 j の報酬の期待値であるとき、T 回の選択での regret R_T は

$$R_T = T\mu^* - \sum_{t=1}^{T} \mu_j(t) \qquad (1)$$

と定義され、最良の選択肢を選び続けた場合と、実際に選んだ選択肢の報酬の期待値との差である。これをウェブサービスに適用する場合、報酬は、もしそのサービスのパフォーマンス指標がクリック率であれば提示した商品等がクリックされたかどうか、コンバージョン率なのであれば契約や申し込みに至ったかどうかを 1 や 0 等の数値にエンコードすることで与えることができる。

このように、multi-armed bandit アルゴリズムによってユーザに対する推薦のオンラインでの最適化を行うことができるが、対象のサービスのユーザの嗜好が多様な場合には、ユーザ集団全体の嗜好に基づいた推薦を行うよりも、ユーザ個々の嗜好に合わせた推薦を行うことで、より良い結果を得ることが期待できる。このユーザ個々に対する最適化は、それぞれのユーザについて独立に multi-armed bandit アルゴリズムを適用して提示する推薦内容を決定することでひとまず実現することはできる。

しかしユーザ集団全体の嗜好を推定するのに十分な量のデータを得られる場合であっても、ユーザ個々の嗜好を推定するのに十分な量のデータが手に入るとは限らない。一般に、計測されるデータの量はユーザによってばらつきが大きく、その度数分布は裾の長い分布となってデータ量の少ないユーザが多数となることが多い。また新規ユーザについてはそもそもデータがない状態から推定を始めなければならない。この問題は cold-start 問題とよばれ**?**、特に協調フィルタリングでの推薦を行う際に問題とされるが、multi-armed bandit アルゴリズムを個々人の嗜好に合わせた推薦に適用する際にも同様に問題となる。そこで本研究では、まだ学習の十分に進んでいないユーザに対してはユーザ集団全体の嗜好に基づいた推薦を行い、十分に嗜好を学習したユーザに対してはユーザ個々の嗜好に基づいた推薦を行うことでこの問題に対処することを狙ったアルゴリズムを提案する。

2. 関連研究

Multi-armed bandit アルゴリズムには選択肢の選び方の異なる最適化手法が数多く提案され、理論的な解析がなされている。最も単純な手法は ϵ-greedy 法である [3]。この手法は各試行の際に、確率 ϵ ですべての選択肢の中からランダムに選択肢を選び、確率 1 - ϵ でその時点で最も報酬の期待値が高い選択肢を選ぶ。Exp3 法は敵対的 bandit と呼ばれる手法の 1 つで、選択肢から得られる報酬が変動する状況を想定した手法である [4]。この手法では各試行の際に、選択肢が有する重みによってその選ばれる確率が決定される。選ばれる確率が低かった選択肢ほど重みの更新時に大きな報酬が得られるようになっており、報酬の変化に追従しやすい仕組みとなっている。この他にも UCB1 法や softmax 法、Thompson sampling 法などの手法が提案されている [3][7][9]。

Multi-armed bandit の応用はウェブ広告の最適化やニュース記事の推薦等にみられる [1][2]。これらの応用では、ユーザからのフィードバックを得ながらユーザアクセスや CTR (Click Through Rate) を最大化するために multi-armed bandit を利用している。更に、近年では multi-armed bandit をユーザやアイテムの情報をコンテキストとして利用することで拡張した contextual bandit も提案されている。Li らは contextual bandit の LinUCB 法を Yahoo!のフロントページに適用し、ユーザごとに最適化されたニュース記事の推薦による CTR の向上を達成した [6]。

Multi-armed bandit アルゴリズムやその拡張を用いた

© 2015 Information Processing Society of Japan

ユーザごとに最適化した推薦は高い精度が期待される一方で、ユーザ個々の学習では学習データが不十分で適切な推薦が困難になる可能性が議論されている [8]。この問題へのアプローチとして、multi-armed bandit に外部の情報を組み合わせる手法についての研究が報告されている。Caron らはユーザのソーシャル情報を利用し、ユーザの学習の効率化と学習過程の推薦精度の向上を達成している [5]。この手法では、ソーシャルネットワークで近隣のユーザ同士は嗜好が類似するという仮定の下、個々のユーザの学習進度が不十分な場合にはソーシャルネットワーク上で近隣のユーザの学習結果を利用して推薦を行う。

しかしながら、コンテキスト情報や外部の情報は推薦に利用できない場合も考えられる。例えば、コンテキスト情報をユーザが任意に設定することができるようなサービスでは、データの欠損や虚偽の申告が推薦のノイズになりうる。外部の情報については利用可能な情報を当該サービスが有していない状況や、有していても個人情報保護の観点等から利用できない場合が存在する。本研究で提案する手法は推薦結果に対するユーザからのフィードバックのみを用いる手法であり、コンテキスト情報や外部の情報を用いることができない場合にも適用可能な汎用性の高い手法である。

3. 手法

本研究では、一般の multi-armed bandit アルゴリズムを拡張し、ユーザの嗜好の学習の進度に応じて推薦の粒度を切り替える方式を提案する。まず、一般の multi-armed bandit アルゴリズムを一つ選び、これを、基礎とする multi-armed bandit アルゴリズムと呼ぶこととする。また、ユーザ集団全体の嗜好を記録する為の特別なユーザを一つ定め、これをグローバルユーザと呼ぶこととする。各ユーザおよびグローバルユーザについて、基礎とする multi-armed bandit アルゴリズムに従って、期待値ないしは重みのベクトル（以降重みベクトルと総称する）や選択肢の選択回数等を保持する。

あるユーザについて推薦要求があり選択肢を選ぶ際には、そのユーザの嗜好の学習が十分進んでいるか否かをまず判定する。学習が十分に進んでいる場合にはそのユーザの重みベクトルに基づいた選択肢の選出を行い、そうでない場合にはグローバルユーザの重みベクトルに基づいた選出を行う。一方、推薦に対する報酬が得られた際には、個々のユーザとグローバルユーザに対応する重みベクトルや選択回数等の状態を双方同時に更新する。この重みの更新方式は基礎とする multi-armed bandit アルゴリズムに従う。ただし、その際の計算は、個々のユーザとグローバルユーザそれぞれについて、独立に（同量の）報酬が得られたものとして行う。

個々のユーザの嗜好の学習が進んでいるかの判定には様々な手法が考えられる。本研究では個々のユーザの選択肢の選択回数について一定の閾値を予め定めておき、それを超えたかどうかで判定する手法を採用した。

本方式は基礎とする multi-armed bandit アルゴリズムに任意の方法を用いることができるが、かなり学習が進んだ段階にあっても各選択肢をある程度均等に選ぶようなアルゴリズムがより好ましい。これは、個々のユーザの嗜好の学習の初期段階においては、その学習を進める為にはある程度均等に各選択肢を選出する必要があるところ、既にユーザ集団全体の嗜好の学習が進行していて、選出が特定の選択肢に極端に偏っている場合、個々のユーザの嗜好の学習が進みにくくなる為である。そこで本研究では、一定の比率で各選択肢に均等な選出を行うのが特徴である Exp3 法を、基礎とする multi-armed bandit アルゴリズムとして用いた。

Exp3 法では、t 時点における選択肢 i を引く確率 $p_i(t)$ は

$$p_i(t) = (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{K} w_j(t)} + \frac{\gamma}{K} \quad i = 1, \ldots, K \quad (2)$$

により決定する。ここで、$w_i(t)$ は t 時点での選択肢 i の重みであり、Exp3 法ではこの各選択肢の重みを記録・更新していく。$\gamma \in [0, 1]$ は各選択肢の重みに応じた配分と各選択肢に均等な配分との比率を調整するパラメータで、ランダムな探索を行う割合を制御する役割をもつ。

Exp3 法にて t 時点で選択肢 j を選出し、報酬 $x_j(t) \in [0, 1]$ が得られたとき、$t+1$ 時点での j の重み $w_j(t+1)$ は

$$\hat{x}_j(t) = x_j(t)/p_j(t) \quad (3)$$

$$w_j(t+1) = w_j(t)\mathsf{exp}(\gamma \hat{x}_j(t)/K) \quad (4)$$

により算出する。その他の選択肢の重みは変化させない。また、重みの初期値は 1 である。式 (4) にて示されている通り、Exp3 法では各選択肢の重みは指数関数的に増大する為、学習が進行すると特定の選択肢の重みが支配的になり易い。

なお、計算機で広く用いられている IEEE 754 フォーマットでの浮動小数点表現では、倍精度でも指数部分が 11 bit であり、10^{308} 程度までしか扱うことができない。従って、Exp3 法で重みの更新を繰り返せばいずれオーバフローする為、その対策が実装にあたっては必要となる。任意精度の浮動小数点表現を用いればこの問題は発生しないが、既存のハードウェアを利用できないことで速度面での問題の発生するおそれがあった為、本研究では単精度または倍精度の浮動小数点表現を用いつつ、オーバフロー発生時に重みの範囲を圧縮する方式を採用した。

オーバフロー発生時、選択肢 i の重み $w_i(t)$ は

$$w'_i(t) = \mathsf{max}(w_i(t)/\mathsf{min}(\mathbf{w}(t))/\delta, 1) \quad i = 1, \ldots, K \quad (5)$$

© 2015 Information Processing Society of Japan

Vol.2015-MPS-103 No.30
Vol.2015-BIO-42 No.30
2015/6/24

により圧縮する。$\mathbf{w}(t)$ は t 時点での重みベクトル、δ は圧縮幅を指定するパラメータである。$\min(\mathbf{w}(t))$ により各選択肢の重みが比較的均等に伸びた場合に効果的に圧縮し、δ により重みの伸びていない選択肢があった場合にも圧縮する。[*1]

4. 実験

4.1 実験設定

4.1.1 シミュレーション概要

提案手法の有効性を確認するために本研究では推薦システムを模したシミュレーションで評価を行った。この推薦システムはユーザアクセスに対してアイテムを一つ推薦する。アイテムは特定のジャンルに属するものとし、ユーザがジャンルごとの嗜好を有するものと想定する。推薦結果をユーザが選択（クリック）したか否かに応じて、1/0 の報酬を受け取り、各ユーザのジャンルごとの嗜好をシステムが学習していく。

シミュレーションは以下の処理を 1 ラウンドとし、このラウンドを規定回数繰り返す。

(1) 全ユーザについて、そのラウンドにおいて推薦システムにユーザがアクセスするか否かを決定する
(2) 推薦システムにアクセスするユーザについて、推薦システムがアイテムを推薦する
(3) ユーザがそのアイテムを選択するか否かを決定する
(4) ユーザの選択結果を推薦システムが記録し、モデルの更新を行う

4.1.2 ユーザ行動モデル

シミュレーションではユーザおよびユーザ群の行動モデルを以下のように制御する。

ユーザアクセス頻度

実際のサービスでは利用頻度の高いユーザや低いユーザなど様々なアクセス頻度のユーザが存在することが想定される。多くの場合その分布は裾の長い分布になる。本研究では指数分布を一般化したガンマ分布を用いてユーザのアクセス頻度の分布を定義した。

推薦結果のクリック率

本研究ではユーザがジャンルごとに嗜好を有すると想定してクリック率を設定した。ユーザの嗜好は好き、普通、嫌いの 3 段階で設定する。好きなジャンルについては平均 15%, 普通のジャンルについては平均 10%, 嫌いなジャンルについては平均 1%程度のクリック率を有するように設定した。

ユーザの嗜好の設定は図 1 のように行った。まず、それぞれジャンルに対する嗜好の異なるユーザセグメントを作成する。セグメントは好き、普通、嫌いの組み合わせの数だけ作成する。また、全ジャンルのうち 2 ジャンル（図 1 では右端の 2 要素）については全セグメント共通で普通、嫌いとした。これは、どんなユーザからもある程度関心を持たれるジャンルと、逆にほとんど関心を持たれないジャンルがあることを想定している。本実験の設定ではジャンル数を 8 と設定したため、28 種類の嗜好の組み合わせを作成した。

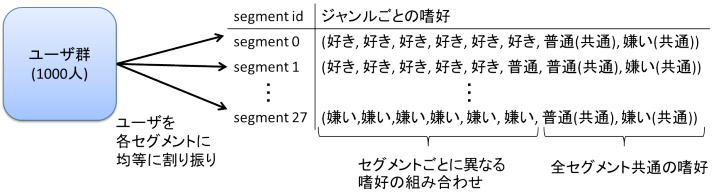

図 1 ユーザの嗜好の設定方法

表 1 ユーザパラメータの設定表

ユーザパラメータ	分布	分布のパラメータ
アクセス頻度	ガンマ分布[*2]	$\alpha = 0.1, \beta = 0.7$
ジャンルの嗜好	ベータ分布[*3]	好き ($\alpha = 1.5, \beta = 10$)
		普通 ($\alpha = 1.0, \beta = 10$)
		嫌い ($\alpha = 1.0, \beta = 100$)
嗜好の変動	正規分布	平均 0, 分散 1

次に各ユーザに一つのセグメントを割り当てる。これは一様分布に基づき各セグメント均等にユーザが割り振られるように行った。そしてユーザは割り当てられたセグメントのジャンルの嗜好に基づき、ベータ分布からサンプリングした各ジャンルのクリック率を設定する。

ユーザの嗜好の変動

サービスを利用している間にユーザの嗜好が変動することが考えられる。本研究ではそういったユーザの嗜好の変動に追従しレコメンドの精度を維持できるかも確認する。具体的には事前に設定されたセグメントの嗜好パラメータに対して、一定のラウンドごとに正規分布に従うノイズを乗せることで各ジャンルのクリック率を変動させる。ただし、パラメータを変動させる際には全体のクリック率が大幅に変動することを避けるため、期待値が一定となるようにパラメータを変動させた。本研究では 2000 ラウンドに一度セグメントの嗜好を変動させるようにして実験を行った。

各設定に用いた分布とパラメータを表 1 にまとめる。

4.1.3 評価方法

手法の効果の測定は CTR で行った。すなわち推薦したアイテムがユーザにクリックされた割合である。本研究では二つの観点による結果の比較を行う。

実験 1: グローバルな推薦から個別の推薦への切り替えの有無の影響

この実験では提案手法である、学習の進行度に応じてグローバルな推薦から個別の推薦に切り替える手法の有効性

[*1] 除算を 2 回実行すべきであることに注意されたい。除数同士を掛け合わせてはならない。

[*2] 平均:$\alpha\beta$ 分散:$\alpha\beta^2$

[*3] 平均:$\frac{\alpha}{\alpha+\beta}$ 分散:$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

© 2015 Information Processing Society of Japan

を確認する。効果の比較はランダムに推薦を行う random、常にグローバルなモデルに基づき推薦を行う global、初めからユーザ個別の推薦を行う personal、そして提案手法であり学習の進行度に応じて推薦手法を切り替える switch で行った。グローバルな推薦から個別の推薦への切り替えはそのユーザのアクセス数が 200 回に達した際に行う。

実験 2: 推薦の切り替えの閾値の影響

二つ目の実験ではグローバルな推薦から個別の推薦に切り替えるアクセス数を変化させて CTR を計測した。比較ではアクセス数 50 回、100 回、200 回、300 回で切り替える場合についてシミュレーションを行い、その過程での CTR を測定した。

実験 1、2 ともに基礎とするアルゴリズムには Exp3 法を選び，パラメータ γ は 0.1 とした。また、シミュレーションはユーザ数 1000 人, ジャンル数 8, ラウンド数を 8000 として 3 回実施し、その平均の CTR を用いて評価を行った。

4.2 実験結果

4.2.1 実験 1 の結果

実験 1 の結果を図 2 に示す。図は横軸がラウンド数、左側の縦軸が対応するラウンド時における CTR の値で実線で示されている結果に対応する軸である。右側の縦軸はグローバルな推薦からユーザ個別の推薦に切り替えたユーザの割合を表す軸であり、グラフ中で破線で表されている結果に対応する。

グローバルな推薦からユーザ個別の推薦に切り替える提案手法は、シミュレーションの開始から一定ラウンド経過後最も高い CTR を達成している。シミュレーションの開始直後はグローバルに学習・推薦を行う手法が安定した CTR を達成しているが、9%〜10%のあたりで CTR の上昇が鈍くなり、ラウンドを経過するにつれユーザ個別に推薦をする手法の方が高い CTR を示すようになる。ユーザ個別の推薦を初めから行う手法ではシミュレーション開始直後の CTR が低い。この主な原因は、ユーザ個別に学習を行っているため十分な学習データを得るまでに CTR の低い選択肢を選んでしまっていることと考えられる。ユーザ個別に推薦を行う手法は 2500 ラウンド経過したあたりから CTR が急激に向上し提案手法と同程度の CTR を達成している。提案手法と個別に推薦する手法は最終的にはどちらも全ユーザ個別に学習・推薦を行うため、ラウンドの経過につれて CTR の差が小さくなるのは期待通りの振る舞いと言える。

ユーザ個別の推薦を伴う手法では、定期的に CTR が一時減少する事象が見られた。CTR の減少が生じるのはラウンドが 2000 の倍数の時であることから、ユーザの嗜好の変動によって一時的に CTR の減少が起きていると考えられる。しかし、CTR の減少は一時的なものでありそこから更に一定ラウンド経過後には CTR が回復しており、

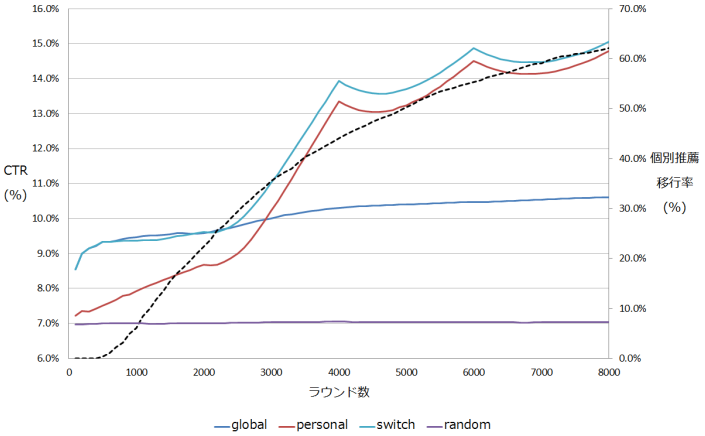

図 2 Exp3 法のシミュレーション結果

嗜好の変化に追従する学習ができている結果が得られている。追従可能な変動の間隔や変動後 CTR の回復に必要な学習の量については今後の課題とする。

4.2.2 実験 2 の結果

個別の推薦に切り替える学習の進行度（本実験の場合はユーザごとのアクセス数の閾値）を変化させた場合の実験結果を図 3 に示す。

実験 1 同様横軸がシミュレーションのラウンド数、縦軸がそのラウンドでの CTR を示す。シミュレーション開始当初最も CTR の高い手法は 300 回のアクセスで推薦方法を切り替える手法である。この手法は開始後の CTR の立ち上がりが良く、グローバルな推薦を行う手法とほぼ同等な CTR を達成しながら推移している。逆に 50 回や 100 回のアクセスで切り替える手法は開始直後の CTR が低めになっている。2500 ラウンドを過ぎたあたりからはすべての手法の CTR が急激に向上している。

シミュレーションが経過するにつれて CTR の高い手法の逆転が見られる。3000 ラウンドを過ぎたあたりから、100 回や 200 回のアクセスで推薦方法を切り替える手法が上位に来るようになり、最終的には 200 回で切り替える手法が最も良い CTR を達成している。一方で、8000 ラウンド経過後には 300 回で切り替える手法が他の個別に推薦を行う手法に比べて最も低い CTR の結果となっている。

これらのことから、推薦方法の切り替えまでの学習期間を大きくとると、推薦開始直後の CTR を高めやすい傾向があると言える。一方で学習期間を大きく取り過ぎると後期に CTR が伸び悩むことがある。この手法を用いる際には、サービスの特性を考慮し最も CTR を高めたいタイミングをどこに定めるかを狙って切り替えに必要な学習期間を設定する必要があると考えられる。

5. まとめと今後の課題

本研究では、ユーザ個々の嗜好についての学習の進度に応じて、ユーザ集団全体の嗜好に基づいた推薦から、ユー

© 2015 Information Processing Society of Japan

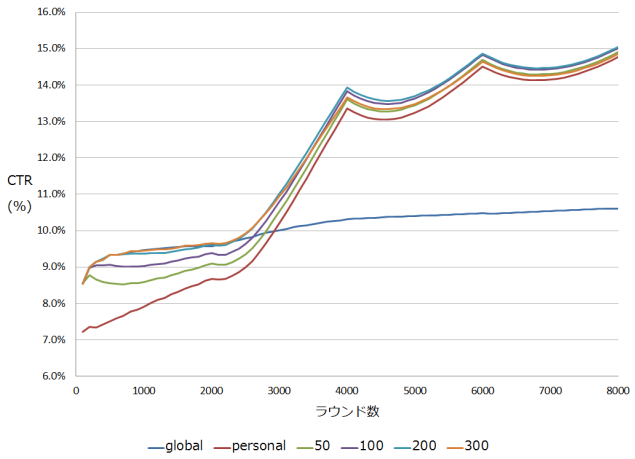

図 3　個別推薦に移行するアクセス数による比較

ザ個々の嗜好に基づいた推薦に切り替える multi-armed bandit 手法を提案した。また、提案手法の有効性を確認する為のシミュレーションによる実験を行った。Exp3 法を基礎となるアルゴリズムとして利用した場合に、提案手法の CTR が、全体の嗜好だけもしくはユーザ個々の嗜好だけに基づく推薦を行った場合に比べ、上回る場合の多いことを示した。

今後の課題としては以下の様な事項が挙げられる。

学習の進行度の測定方法

本研究ではユーザのアクセス数に基づいて学習の進行度を測定したが、その他には以下のような手法が考えられる。

- （0 以外の）報酬の登録回数が閾値を超えたかどうかで判定する手法
- 重みベクトル中のいずれかの重み、あるいは重みの合計が閾値を超えたかどうかで判定する手法
- 重みベクトル中の重みの偏りで判定する手法

また学習の進行度の測定方法に加え、閾値の設定の仕方も今後検証を要する事項である。

パラメータの影響

本研究では 1 種類のパラメータの組み合わせのシミュレーションにより、提案手法の有効性を示した。しかし、提案手法の適用可能範囲を明確にするため、ユーザ数、ジャンル数、各ジャンルのクリック率や嗜好の変動など各種パラメータを変化させた際の効果について検証を行うことが必要である。

他の multi-armed bandit アルゴリズムへの応用

本研究ではユーザの嗜好の変動を考慮して、乱択的な要素の強い Exp3 法を提案手法に用いて評価を行った。一方で前述の通り multi-armed bandit アルゴリズムにはこれまで複数の手法が提案されてきている。それらの手法についても提案手法を応用し、効果の測定を行い、本手法が効果を発揮する条件についての検討が必要である。

参考文献

[1] Abe, N. and Nakamura, A.: Learning to optimally schedule internet banner advertisements, *ICML*, Vol. 99, pp. 12–21 (1999).
[2] Agarwal, D., Chen, B.-C. and Elango, P.: Explore/exploit schemes for web content optimization, *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, IEEE, pp. 1–10 (2009).
[3] 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).
[4] Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E.: The nonstochastic multiarmed bandit problem, *SIAM Journal on Computing*, Vol. 32, No. 1, pp. 48–77 (2002).
[5] Caron, S. and Bhagat, S.: Mixing bandits: A recipe for improved cold-start recommendations in a social network, *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, ACM, p. 11 (2013).
[6] 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*, ACM, pp. 661–670 (2010).
[7] Luce, D. R.: Individual Choice Behavior (1959).
[8] Madani, O. and DeCoste, D.: Contextual recommender problems [extended abstract], *Proceedings of the 1st international workshop on Utility-based data mining*, ACM, pp. 86–89 (2005).
[9] Thompson, W. R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, pp. 285–294 (1933).

© 2015 Information Processing Society of Japan

【正誤表】

p.2 左列 下から 16 行目

誤) regret を最小化する、というものである?

正) regret を最小化する、というものである[10]

p.2 右列 上から 12 行目

誤) cold-start 問題とよばれ?

正) cold-start 問題とよばれ[11]

p.6 右列最終行

以下を追加

[10] Kuleshov, V. and Precup, D.: Algorithms for multi-armed bandit problems, CoRR, Vol. abs/1402.6028 (online), available from (http://arxiv.org/abs/1402.6028) (2014).

[11] Schein, A. I., Popescul, A., Ungar, L. H. and Pennock, D. M.: Methods and Metrics for Cold-start Recommendations, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02, ACM, pp. 253-260 (2002).