情報処理学会論文誌 Vol.64 No.1 214–228 (Jan. 2023)

顧客の閲覧行動分析のための時間窓トピックモデル

伊藤 史世[1,a)] 雲居 玄道[1,b)] 後藤 正幸[1,c)]

受付日 2022年4月28日, 採録日 2022年10月4日

概要：現在，EC サイトをはじめとした様々なサービスがインターネット上で展開されており，膨大な閲覧履歴データを取得することが可能になっている．そのため，マーケティング分析上の観点から，蓄積された閲覧履歴を用いてユーザの嗜好を詳細に分析し，効果的な施策に結び付ける活動も広がっている．ここで，EC サイト上の購買においては，Web ページを閲覧する中で徐々に興味の対象が絞られていき，最終的にある商品を購入するというユーザ行動が想定される．よって，各ユーザの興味の時間変化を分析し，この変化に基づいて適切なタイミングで施策を実施することで，施策の効果を高めることが期待できる．従来，時間とともに変化するユーザの興味をモデル化することが可能な手法として，購買履歴データを対象とした時系列トピックモデルである Topic Tracking Model（以下，TTM）が提案されている．しかし，TTM は比較的長い購買履歴データを想定して提案された手法であり，1 つのセッション内の閲覧系列のように，数十ページ程度のサイトの遷移のなかでユーザの興味が移り変わっていくという状況を想定していない．そのため，TTM を閲覧履歴データに適用した際には，モデルの学習が困難となり，ユーザの興味の変化を検出できない恐れがある．そこで本研究では，モデルのパラメータを段階的に推定する，時間窓トピックモデルを提案する．提案手法により，ユーザの興味が様々に移り変わるという複雑な状況が仮定され，かつ 1 セッションあたりの閲覧ページが数十ページ程度の閲覧履歴データにおいても，安定したパラメータ推定が可能となり，各ユーザの興味の収束度合いに基づいた施策の実施の実現が期待される．最後に，人工データと実データに対して提案手法を適用し，その有効性を検証する．

キーワード：トピックモデル，時系列トピックモデル，閲覧履歴データ，崩壊型ギブスサンプリング

Time Window Topic Model for Analyzing Customer Browsing Behavior

Fumiyo Ito[1,a)] Gendo Kumoi[1,b)] Masayuki Goto[1,c)]

Received: April 28, 2022, Accepted: October 4, 2022

Abstract: Nowadays, various services like EC sites have been expanding on the Internet, and huge amount of browsing history data are being accumulated. It is, therefore, desiblack to take effective marketing action by analyzing users' interest in detail by using accumulated browsing history data. Generally, users narrow down their interest while browsing pages and finally purchase an item on the EC site. Hence, to improve the effectiveness of marketing measures, it is important to analyze the change points of users' interest and to timely conduct it based on those change. Conventionally, Topic Tracking Model (TTM) has been proposed as a method which can model the changes of uses' interest over time by purchase history data. However, TTM assumes the relatively long purchase history data and doesn't consider the situation where users change their interest in short browsing site transition of about several dozen pages. Therefore, TTM cannot estimate the changes of users' interest when applied to the browsing history data. In this research, we propose a new topic model which estimates parameters in step by step, named Time Window Topic Model. The proposed method enables us to estimate parameters and clarify the change of users' interest while browsing. It will be possible to conduct effective marketing measure based on the convergent of users' interest by applying the proposed model. Finally, we apply the proposed method to both of the artificial and real data set and show the usefulness of our proposed method.

Keywords: topic model, time series topic model, browsing history data, collapsed gibbs sampling

1. はじめに

現在，EC サイトをはじめとした様々なサービスがインターネット上で展開されており，各企業は購買履歴に加え，Web ページの閲覧履歴などユーザの行動に関するより詳細なログデータを取得することが可能になっている．そのため，マーケティング分析の観点から，閲覧履歴を活用することでユーザの嗜好をより詳細に分析し，効果的な施策実施に結び付ける取り組みが活発になっている [1], [2], [3], [4], [5]．ここで，一般的な消費者の EC サイト上の購買においては，Web ページを閲覧する中で徐々に購入対象の商品の対象を絞っていき，最終的にある商品を購入するというユーザ行動が想定される．しかし，一般的な EC サイトにおける購買率はたかだか数%であることがほとんどであり，購買に至らないまま，EC サイトを離れてしまうケースが大多数を占める．これらのユーザは，最終的に購入したい商品を絞り込むことができずに，様々な商品を閲覧しただけで離反していってしまうユーザと考えられる．このようなユーザに対しては，EC サイト上の閲覧履歴を用いて，その興味の度合いや商品の絞り込みの程度を分析し，施策実施の最も適切なタイミングを把握することで，その効果を高めることが可能と考えられる．すなわち，時間とともに変化する各ユーザの興味を時系列分析し，商品の絞り込みの度合いに基づいて，適切なタイミングで施策を実施するための手法は実応用上の価値が非常に高い．

一方，顧客の嗜好は多様であり，観測可能な購買や閲覧の背後には，複数の嗜好が異なる潜在的なグループの存在を仮定することが有用である場合が多い．特に，マーケティング分析の領域では，観測可能な状態の背後に潜在的な状態（以下，トピック）を仮定したトピックモデルや潜在クラスマルコフモデルの有効性が示されている [6], [7], [8], [9]．トピックモデルを購買履歴や閲覧履歴に適用することで，ユーザの嗜好を表現するトピック分布と，各トピックにおける商品やサイトの購買または閲覧確率であるアイテム分布を同時に推定し，これらの分布を用いてユーザの興味を分析することが可能である．このうち，時間とともに変化するユーザの嗜好をモデル化することが可能なトピックモデルとして，購買履歴データを対象とした Topic Tracking Model（以下，TTM）[10] が提案されている．TTM は，トピック分布とアイテム分布の両方に時系列性を仮定しており，各時刻において，これらの分布のパラメータを同時に推定する．このモデルは，比較的長い購買履歴データを想定し提案された手法であり 1 つのセッション内の閲覧系列のように，数十ページ程度のサイトの遷移の中でユーザの興味が移り変わっていくという状況を考慮できていない．EC サイト上にアクセスしたユーザの様々なセッションでは，徐々に興味の対象の絞り込みが行われていく少数のユーザ（以下，収束ユーザ）と，興味を絞り込むことができておらず様々な商品を閲覧している状態にとどまる多数のユーザ（以下，未収束ユーザ）が混在していることが想定される．しかし，TTM をそのまま閲覧履歴データに適用した場合，各時刻で観測されるデータ長が十分に存在している場合を除き，トピック分布とアイテム分布を同時に推定することは困難であり，そのためユーザの興味の収束状況を検出できない恐れがある．

そこで，本研究では，ユーザの興味を表現するトピック分布と各トピックの特徴を表現する分布であるアイテム分布を段階的に推定するトピックモデルである，時間窓トピックモデル（Time Window Topic Model：TWTM）を提案する．提案モデルでは，各時刻におけるデータが比較的少数な場合でもパラメータの安定的な推定を可能とするために，はじめに全時刻のデータを用いてアイテム分布の推定を行う．この際，Web サイトの閲覧においては，ユーザは連続する一定期間内の閲覧系列（以下，ウィンドウ）のなかでは類似した興味に従ってページを閲覧するという行動に着目し，ウィンドウごとに潜在トピックを割当てを行う．この後に，推定したアイテム分布のパラメータを固定したもとで，時刻ごとにユーザのトピック分布の推定を行う．提案手法により，収束ユーザと未収束ユーザが混在し，ユーザの興味が様々に移り変わるという複雑な状況が仮定される閲覧履歴データにおいても，安定したパラメータ推定を行うことができ，収束ユーザの興味の変化を明らかにすることが可能となる．これによって，各ユーザの興味の収束度合いに基づいた施策の実施による施策効果の向上が期待される．本研究では，人工データと実データに対して提案手法を適用し，その有効性を検証する．

2. 準備

2.1 関連研究

EC サイトの閲覧履歴データを活用して，様々なマーケティング分析に結び付ける取り組みは広く試みられている．たとえば，武政ら [3] は，顧客の閲覧履歴を利用した商品ランキング生成法を提案している．久松ら [4] は，閲覧行動を考慮した購買予兆の発見モデルを構築している．伊藤ら [5] は，閲覧履歴・購買履歴を活用した顧客セグメンテーションと商品スコアリングにより，購買予測を行うモデルを構築している．

一方，購買履歴や閲覧履歴データから顧客の興味を分析する場合，顧客の嗜好は多様であり，嗜好のまったく異なるユーザグループが混在していることが想定することが多い．そのため，観測可能な購買や閲覧の背後に複数の潜在的な状態（トピック）の存在を仮定したトピックモデル

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan
[a)] fumiyo0607@fuji.waseda.jp
[b)] moto-aries@ruri.waseda.jp
[c)] masagoto@waseda.jp

を起点にして様々な手法が提案されている．代表的なトピックモデルである，Latent Dirichlet Allocation（以下，LDA）[11] はユーザの嗜好や興味をトピック混合割合であるトピック分布として解釈することが可能であり，購買履歴データをはじめとした様々なデータに対して適用と拡張がなされている [8], [9], [12]．さらにこの LDA から発展し，時系列的な特徴をとらえるために文書データを対象として，Dynamic Topic Model [13] に代表される時系列トピックモデル [14], [15], [16] が提案されている．これらの手法は，元来，文書データを対象としており，各時刻において 1 文書ずつ生成されることを想定している．そのため，これらの時系列トピックモデルを購買履歴や閲覧履歴データに適用し，ユーザの嗜好の連続的な変化を分析することはできない．これらの手法に対し，1 人のユーザの嗜好の変化をトピック分布の変化としてモデル化するために提案された手法に TTM がある．TTM を購買履歴に適用することにより，ユーザの嗜好の変化をついて解析することが可能である．

また，購買履歴や閲覧履歴データを対象としてマルコフモデルや隠れマルコフモデル（以下，HMM）を起点とした研究も広く行われている [17]．HMM ではトピック間の遷移のしやすさについて解析することが可能であるが，ユーザの興味の変化を分析するためには，すべての購買や閲覧に対するトピックの変化を分析することが必要となる．これに対し，HMM にトピックモデルの考え方を援用し，トピック間の遷移にマルコフ性を仮定しながら各ユーザのトピック分布を推定可能な手法として，Hidden Topic Markov Model（以下，HTMM）[18] が知られている．Hotoda ら [6] は，この HTMM を閲覧履歴データに適用するために拡張したモデルを提案し，実データに適用して有効性を示している．このモデルでは，トピック間の遷移割合とユーザのトピック分布を推定可能であるが，トピック分布が時刻によって変化することは想定されていない．

さらに，トピックモデルを閲覧履歴に対して適用した最近の研究としては，LDA を活用して閲覧履歴をトピック分布として表現することによってユーザのプライバシを保護することを目的とする手法の研究 [19]，推薦精度の向上を目的として閲覧履歴からトピックを抽出する手法の研究 [20], [21] がある．これらの研究は閲覧履歴に対してトピックモデルを適用しているという点では提案手法と類似しているが，これらの提案はあくまで LDA の活用にとどまっており「ユーザの興味の変化をとらえる」という目的のために適用することができない．

以上のように，ユーザの購買や閲覧行動を分析するための手法は様々提案されているが，閲覧履歴データにおいてユーザの興味の変化に焦点を当てた研究はこれまでなされていない．

表 1 変数の定義
Table 1 Variable definitions.

変数	概要
U	全ユーザ数
D	全ページ数
$\mathcal{U} = \{a_u\}_{u=1}^{U}$	ユーザ集合
$\mathcal{X} = \{b_d\}_{d=1}^{D}$	ページ集合
K	潜在トピック数
$\mathcal{Z} = \{1, \cdots, K\}$	潜在トピック集合
M_u	ユーザ a_u の閲覧数
$x_m^u \in \mathcal{X}$	ユーザ a_u の m 番目の閲覧
$\boldsymbol{X}_u = x_1^u, x_2^u, \cdots, x_{M_u}^u$	ユーザ a_u の閲覧系列
T	全時刻数
$\mathcal{T} = \{1, \ldots, T\}$	時刻集合
$M_{t,u}$	ユーザ a_u の時刻 t の閲覧数
$x_m^{t,u}$	ユーザ a_u の時刻 t の m 番目の閲覧
$\boldsymbol{X}_{t,u} = x_1^{t,u}, \cdots, x_{M_{t,u}}^{t,u}$	時刻 t のユーザ a_u の閲覧系列（ただし，$\boldsymbol{X}_u = \boldsymbol{X}_{1,u}, \ldots, \boldsymbol{X}_{T,u}$）

2.2 問題設定

EC 市場の拡大にともない，近年では各企業において膨大な閲覧履歴データが蓄積されており，事業効率化のためのデータ利活用が重要な課題となっている．閲覧履歴データからユーザの興味の変化をモデル化することができれば，興味の変化に基づく施策実施が可能となり，より効率的なマーケティング活動の実現が期待される．

そこで本研究では，ユーザの興味が短期間で切り替わることが想定される EC サイトの閲覧履歴においても，ユーザの興味の変化をトピック分布により追跡することを目的とし，これを実現するための手法を提案する．提案手法においては，比較的短い時間単位での分析を想定しており，1 時刻として 1 日などの期間のほかに，30 ページなどの任意の閲覧数を設定することも可能である．ここで，閲覧履歴データに関する変数の定義と各分析手法に共通する変数の定義を**表 1** に示す．なお，各手法で個別に使われる変数は適宜定義を行うものとする．

2.3 従来手法

本研究では，比較的短い時間で変化するユーザの興味の連続的な変化をトピック分布の変化として解析可能な手法の提案を目的としている．すなわち，短い閲覧系列からトピック分布とアイテム分布を推定可能であること，トピック分布の時間にともなう変化をモデル化可能であることが求められる．

従来，短い系列データからトピック分布の推定を可能にした手法として Biterm Topic Model（以下，BTM）[22] が知られている．BTM は SNS などで投稿されるような比較的短い文書を対象に提案されたトピックモデルであり，1 つの文書に含まれる単語の 2 語のペア（バイターム）を定義し，このバイタームごとに 1 つのトピックが割り当

てられる．この手法では，1 つのバイタームに対して，1 つのトピックを定義することによって，1 文書あたりの単語数が少ない場合においても，トピック分布の推定を適切に行いながら，1 文書に複数のトピックが含まれているという状況をモデル化することを可能にしている．いま，潜在トピック $k \in \mathcal{Z}$ が出現する確率を $\theta_k = p(k|\boldsymbol{\theta})$ とし，$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, $\sum_{k=1}^{K} \theta_k = 1$ をトピック分布とする．また，潜在トピック k のもとでページ b_d が出現する確率を $\phi_{k,d} = p(b_d|\boldsymbol{\phi}_k)$, $\sum_{d=1}^{D} \phi_{k,d} = 1$ として，$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$ をアイテム分布とする．トピックモデルにおいては，この 2 つの分布によって観測データの背後にある潜在トピックを表現することができる．このもとで，BTM におけるユーザ a_u の閲覧系列 $\boldsymbol{X}_u$ の出現確率は式 (1) で表現される．

$$
\begin{aligned}
&P(\boldsymbol{X}_u|\boldsymbol{\theta}, \boldsymbol{\Phi}) \\
&= \prod_{\{x_i^u, x_j^u\} \subseteq \boldsymbol{X}_u, i \neq j} \sum_{k=1}^{K} P(x_i^u, x_j^u, z_{u,i,j} = k|\boldsymbol{\theta}, \boldsymbol{\Phi}) \\
&= \prod_{\{x_i^u, x_j^u\} \subseteq \boldsymbol{X}_u, i \neq j} \sum_{k=1}^{K} \Big\{ P(z_{u,i,j} = k|\boldsymbol{\theta}) \\
&\qquad \times P(x_i^u|z_{u,i,j} = k, \boldsymbol{\Phi}) P(x_j^u|z_{u,i,j} = k, \boldsymbol{\Phi}) \Big\} \\
&= \prod_{\{x_i^u, x_j^u\} \subseteq \boldsymbol{X}_u, i \neq j} \sum_{k=1}^{K} \Big\{ \theta_k \phi_{k,x_i^u} \phi_{k,x_j^u} \Big\} \qquad (1)
\end{aligned}
$$

ただし，$z_{u,i,j}$ は，ユーザ u の i 番目と j 番目の組合せ，つまりバイタームに対して割り当てられた潜在トピックであり，B_u は $\boldsymbol{X}_u$ に含まれるバイタームの数である．閲覧履歴データにおいては，ユーザの興味は変化することが考えられるため，バイターム中の 2 つのページに同一潜在トピックを仮定することは適切ではない．そのため，非常に短い閲覧系列内のバイタームのみを作成するなどの工夫が必要であり，閲覧履歴データに直接適用することは困難であると考えられる．

一方で，ユーザの興味の連続的な変化をトピック分布としてモデル化することが可能な手法として Tracking Topic Model (以下, TTM) [10] が提案されている．TTM は, 時間とともに変化するユーザの嗜好とトピック内の流行をモデル化することが可能であり, 時刻ごとにトピック分布とアイテム分布の推定を行う．さらに, ユーザのトピック分布とアイテム分布の両方に対して 1 次のマルコフ性を仮定している．ここで，時刻 t におけるユーザ a_u のトピック分布のパラメータを $\boldsymbol{\theta}_{t,u} = (\theta_{t,u,1}, \ldots, \theta_{t,u,K})$，時刻 t における潜在トピック $k \in \mathcal{Z}$ のアイテム分布 $\phi_{t,k} = (\phi_{t,k,1}, \ldots, \phi_{t,k,D})$ とし, $\boldsymbol{\Phi}_t = (\boldsymbol{\phi}_{t,1}, \ldots, \boldsymbol{\phi}_{t,K})$ としたとき, 時刻 t におけるユーザ u の閲覧系列の出現確率は以下の式 (2) で表現される．

$$
\begin{aligned}
&P(\boldsymbol{X}_{t,u}|\boldsymbol{\Theta}_t, \boldsymbol{\Phi}_t, \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}) \\
&= \prod_{m=1}^{M_{t,u}} \sum_{k=1}^{K} P(x_{t,u,m}, z_{t,u,m} = k|\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Phi}_{t-1}) \\
&= \prod_{m=1}^{M_u} \sum_{k=1}^{K} \Big\{ P(z_{t,u,m} = k|\boldsymbol{\Theta}_t, \boldsymbol{\Theta}_{t-1}) \\
&\times P(x_{t,u,m}|z_{t,u,m} = k, \boldsymbol{\Phi}_t, \boldsymbol{\Phi}_{t-1}) \Big\} \\
&= \prod_{m=1}^{M_u} \sum_{k=1}^{K} \theta_{u,k} \phi_{k,x_{u,n}} \qquad (2)
\end{aligned}
$$

TTM では各ユーザのトピック分布とアイテム分布に対する 1 次のマルコフ性を表現するために，$\boldsymbol{\theta}_{t,u}$，$\boldsymbol{\phi}_{t,k}$ の事前分布を以下の式 (3)，(4) によって定義する．

$$
\begin{aligned}
P(\boldsymbol{\theta}_{t,u}|\boldsymbol{\theta}_{t-1,u}, \alpha_{t,u}) &\propto \prod_{k=1}^{K} \theta_{t,u,k}^{\alpha_{t,u}\theta_{t-1,u,k}-1} \\
&= \mathrm{Dirichlet}(\alpha_{t,u}\boldsymbol{\theta}_{t-1,u}) \qquad (3)
\end{aligned}
$$

$$
\begin{aligned}
P(\boldsymbol{\phi}_{t,k}|\boldsymbol{\phi}_{t-1,k}, \beta_{t,k}) &\propto \prod_{d=1}^{D} \phi_{t,k,d}^{\beta_{t,k}\phi_{t-1,k,d}-1} \\
&= \mathrm{Dirichlet}(\beta_{t,k}\boldsymbol{\phi}_{t-1,k}) \qquad (4)
\end{aligned}
$$

ただし，$\alpha_{t,u}$，$\beta_{t,k}$ は，ディリクレ分布の形状を制御するハイパーパラメータである．

ここで，BTM，TTM，提案手法のグラフィカルモデルの比較を**図 1** に示す．

3. 提案手法

3.1 概要

閲覧履歴データから，ユーザの興味の変化をトピック分布としてモデル化するためには，TTM のような時系列トピックモデルは有効であると考えられる．しかしながら，従来の時系列トピックモデルは比較的長い閲覧系列を前提とし，ユーザの興味もゆっくりと変化することを想定している．一方で，閲覧履歴データにおいてはユーザの興味は数〜数十単位の閲覧の中で切り替わることが想定され，この興味の変化をとらえることが重要である．ここで，TTM では 1 時刻ごとにユーザの閲覧内のページの共起パターンに基づいて全ユーザのトピック分布とアイテム分布を同時に推定する．しかし，ユーザの興味が短期間で切り替わっていく状況では，このページの共起パターンが様々に変化してしまうために，トピック分布とアイテム分布の双方に対して時系列的な変化を許容する TTM では，これらの分布の推定が困難になってしまう．さらに，TTM では 1 時刻前のトピック分布とアイテム分布から各分布を生成するために，全時刻において閲覧のデータが存在している必要がある．分析のために，1 時刻としては 1 日などの期間を設定することが考えられるが，分析対象となるユーザが毎

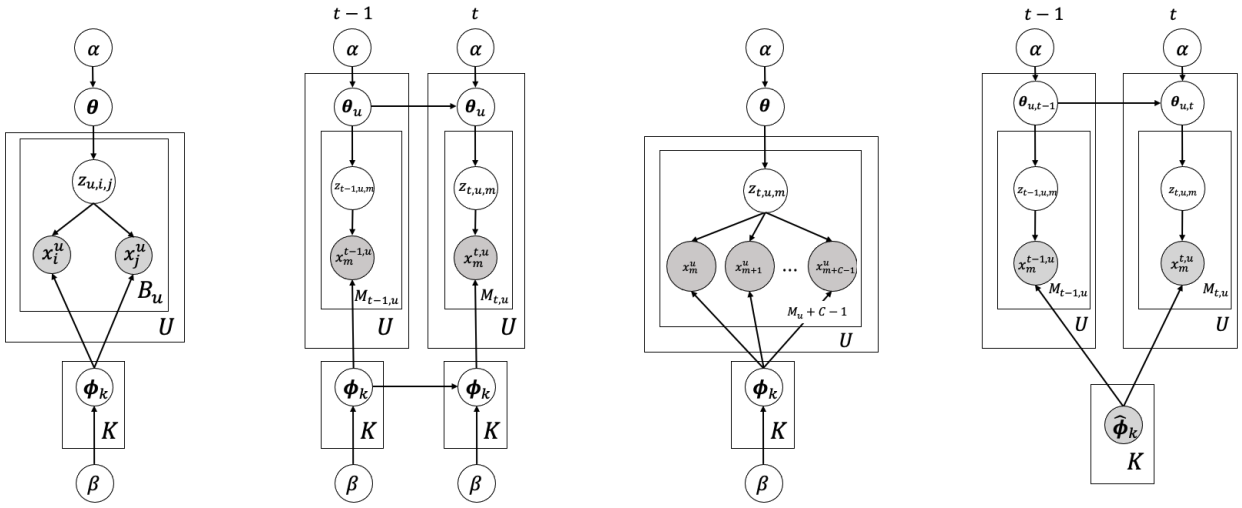

図 1 各手法のグラフィカルモデルの比較
Fig. 1 The comparison of graphical models of each method.

日対象サイトを閲覧することは考えにくく，この制約は実応用の観点からは望ましくない．

そこで，本研究ではユーザの興味が短期間で切り替わる閲覧履歴データにおいても高精度なパラメータ推定を可能とするために，アイテム分布とトピック分布を以下に示す 2 段階で推定することで従来の課題を解決した時間窓トピックモデル（Time Window Topic Model：TWTM）を提案する．提案手法の STEP.1 では，閲覧系列データの特性に基づいて適切なウィンドウサイズ C を決定することで，閲覧系列の局所性を反映した潜在トピックの学習を行う．一方，STEP.2 においては施策の実施頻度など実応用上の観点から，分析者が 1 日や 1 週間など任意の時間単位を決定することが可能である．

STEP.1：アイテム分布の推定：閲覧系列に対して時間窓の概念を導入し，同一ウィンドウ内のページは同じトピックのもとで共起しているという仮定に基づき，トピックモデルを学習する

STEP.2：トピック分布の推定：STEP.1 で求めたアイテム分布のパラメータを固定したもとで，ユーザのトピック分布の推定を行う

提案手法ではアイテム分布における時系列性を排除することによって，各ユーザが全時刻において閲覧データが必要であるという制約も排除している．以下の節では，これらのステップの詳細について述べる．

3.2 STEP.1 アイテム分布の推定

3.2.1 モデル式

連続した短い範囲の閲覧行動の中では，ユーザは同一の興味に従って類似したページを閲覧することが想定される．そこで，この行動をモデル化するため，アイテム分布の推定の際には，Word2vec [23] に代表される多くの自然言語処理の分野で適用されているウィンドウという概念を用い，同一ウィンドウ内のページは同じトピックのもとで共起していると仮定してパラメータの推定を行う．ここで，ユーザ a_u の m 番目のウィンドウ内部分系列 $\boldsymbol{w}_{u,m}$ $(m \in \mathcal{W}_u \coloneqq \{1, 2, \ldots, M_u - C + 1\})$ は，ウィンドウサイズを C として，以下の式 (5) によって定義される．

$$\boldsymbol{w}_{u,m} = x^u_m, x^u_{m+1}, \cdots, x^u_{m+C-1} \tag{5}$$

そして，ユーザ a_u の閲覧系列は式 (6) で表現される．

$$\boldsymbol{X}_u = \boldsymbol{w}_{u,1}, \boldsymbol{w}_{u,2}, \cdots, \boldsymbol{w}_{u,M_u-C+1} \tag{6}$$

このとき，ユーザ a_u の閲覧系列 $\boldsymbol{X}_u$ の出現確率は，以下の式 (7) で与えられる．

$$\begin{aligned} &P(\boldsymbol{X}_u|\boldsymbol{\theta}, \boldsymbol{\Phi}) \\ &= \prod_{m=1}^{M_u-C+1} \sum_{k=1}^{K} P(z_{u,m}=k|\boldsymbol{\theta}) \prod_{i=m}^{m+C-1} P(x^u_i|z_{u,m}=k, \boldsymbol{\phi}_k) \\ &= \prod_{m=1}^{M_u-C+1} \sum_{k=1}^{K} \theta_k \prod_{i=m}^{m+C-1} \phi_{k,x^u_i} \end{aligned} \tag{7}$$

ここで，$z_{u,m}$ はユーザ a_u の m 番目の閲覧の潜在トピックである．また，$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ は，トピック分布のパラメータ，$\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,D})$ は，トピック k のアイテム分布のパラメータであり，$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)^\top \in \mathbb{R}^{K \times D}$ とする．また，トピック分布 $\boldsymbol{\theta}$ とアイテム分布 $\boldsymbol{\phi}_k$ は，以下の式 (8)，(9) のディリクレ分布から生成されるとする．

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \theta_k^{\alpha-1} = \mathrm{Dirichlet}(\boldsymbol{\alpha}) \tag{8}$$

$$p(\boldsymbol{\phi}_k|\boldsymbol{\beta}_k) \propto \prod_{d=1}^{D} \phi_k^{\beta_k-1} = \mathrm{Dirichlet}(\boldsymbol{\beta}_k) \tag{9}$$

アルゴリズム 1 提案手法：STEP.1

for each topic $k = 1, \cdots, K$ **do**
  Draw site distribution $\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}_k)$
for each user $u = 1, 2, \cdots, U$ **do**
  Draw topic distribution $\boldsymbol{\theta}_u \sim \text{Dirichlet}(\boldsymbol{\alpha})$
  for each window $m = 1, \cdots, M_u - C$ **do**
    Draw topic $z_{u,m} \sim \text{Categorical}(\boldsymbol{\theta}_u)$
    for each page $i = m, \cdots, m + C - 1$ **do**
      Draw page $x_i^u \sim \text{Categorical}(\boldsymbol{\phi}_{z_{u,m}})$

ただし，$\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) \in \mathbb{R}^K$，$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$，$\boldsymbol{\beta}_k = (\beta_k, \ldots, \beta_k) \in \mathbb{R}^D$ である．

3.2.2 モデルの学習

提案モデルの学習には，ギブスサンプリングによる潜在トピックの割当てと，不動点反復法による事前分布のパラメータ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ の更新を繰り返す確率的 EM アルゴリズム [24] を用いる．

ここで，全ユーザの閲覧系列集合を $\boldsymbol{X} = \{\boldsymbol{X}_u\}_{u \in \mathcal{U}}$，全ユーザの潜在トピック集合を $\boldsymbol{Z} = \{z_{u,m}\}_{u \in \mathcal{U}, m \in \mathcal{W}_u}$ とする．さらに，$W = \sum_{u=1}^{U} M_u - C + 1$ を全ウィンドウ数，W_k をトピック k に割り当てられたウィンドウの数，$N_{k,d}$ をページ b_d がトピック k に割り当てられた回数とし，$N_k = \sum_{d=1}^{D} N_{k,d}$ とする．ただし，$\boldsymbol{Z}_{\setminus u,m}$ は，$\boldsymbol{Z}$ から $z_{u,m}$ を除いたトピックの割り当てであり，$N_{k \setminus u,m}$ は，$\boldsymbol{w}_{u,m}$ を除いたトピック k に割り当てられたウィンドウの数，$\Gamma(\cdot)$ はガンマ関数である．このとき，ユーザ a_u の m 番目のウィンドウが潜在トピック k に割当てられる確率は，以下の式 (10) によって求められる．

$$
\begin{aligned}
&p(z_{u,m} = k | \boldsymbol{X}, \boldsymbol{Z}_{\setminus u,m}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \int p(z_{u,m} = k | \boldsymbol{Z}_{\setminus u,m}, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \boldsymbol{\alpha}) d\boldsymbol{\theta} \\
&\quad \times \int p(\boldsymbol{w}_{u,m} | \boldsymbol{X}_{\setminus u,m}, z_{u,m} = k, \boldsymbol{Z}_{\setminus u,m}, \boldsymbol{\Phi}) p(\boldsymbol{\Phi} | \boldsymbol{\beta}) d\boldsymbol{\Phi} \\
&\propto (W_{k \setminus u,m} + \alpha) \times \frac{\Gamma(N_{k \setminus u,m} + D\beta_k)}{\Gamma(N_{k \setminus u,m + N_{u,m} + D\beta_k})} \\
&\quad \times \prod_{d=1}^{D} \frac{\Gamma(N_{k,d \setminus u,m} + N_{u,m,d} + \beta_k)}{\Gamma(N_{k,d \setminus u,m} + \beta_k)} \qquad (10)
\end{aligned}
$$

また，トピック分布とアイテム分布の事前分布のパラメータ α と β_k は，不動点反復法を用いて，以下の式 (11)，(12) によって各学習ステップにおいて更新する．

$$
\alpha \leftarrow \alpha \frac{\sum_{k=1}^{K} \Psi(W_k + \alpha) - K\Psi(\alpha)}{K\Psi(W + \alpha K) - K\Psi(\alpha K)} \qquad (11)
$$

$$
\beta_k \leftarrow \beta_k \frac{\sum_{k=1}^{K} \Psi(N_{k,d} + \beta_k) - D\Psi(\beta_k)}{D\Psi(N_k + \beta_k D) - D\Psi(\beta_k D)} \qquad (12)
$$

ここで，$\Psi(x)$ は $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ として定義されるディガンマ関数である．十分な反復を行った後，アイテム分布のパラメータは以下の式 (13) によって求められる．

アルゴリズム 2 提案手法：STEP.2

for each time $t = 1, 2, \cdots, T$ **do**
  for each user $u = 1, 2, \cdots, U$ **do**
    Draw $\boldsymbol{\theta}_{t,u} \sim \text{Dirichlet}(\alpha_{t,u} \hat{\boldsymbol{\theta}}_{t-1,u})$
    for each page $m = 1, 2, \cdots, M_{t,u}$ **do**
      Draw topic
      $z_{t,u,m} \sim \text{Categorical}(\boldsymbol{\theta}_{t,u})$
      Draw page
      $x_m^{t,u} \sim \text{Categorical}(\hat{\boldsymbol{\phi}}_{z_{t,u,m}})$

$$
\hat{\phi}_{k,d} = \frac{N_{k,d} + \beta_k}{N_k + D\beta_k} \qquad (13)
$$

3.3 STEP.2 トピック分布の推定

3.3.1 モデル式

STEP.2 では，式 (13) で求めたアイテム分布のパラメータを固定したもとで，ユーザのトピック分布の推定を行う．ここで，ユーザのトピック分布は時刻 t ごとに変化していることを仮定し，時刻 t（$t \in \mathcal{T} := \{1, \ldots, T\}$）には，1 日や 1 週間などの任意の時間を設定することが可能である．ユーザの興味の連続的な変化をモデル化するため，TTM と同様に，式 (14) によって，1 時刻前のトピック分布に依存していることを表現する．

$$
\begin{aligned}
P(\boldsymbol{\theta}_{t,u} | \hat{\boldsymbol{\theta}}_{t-1,u}, \alpha_{t,u}) &\propto \prod_{k=1}^{K} \theta_{t,u,k}^{\alpha_{t,u} \hat{\theta}_{t-1,u,k} - 1} \\
&= \text{Dirichlet}(\alpha_{t,u} \hat{\boldsymbol{\theta}}_{t-1,u}) \qquad (14)
\end{aligned}
$$

ただし，$\alpha_{t,u}$ は，ディリクレ分布の形状を制御するハイパーパラメータであり，$\hat{\boldsymbol{\theta}}_{t-1,u}$ は，$t-1$ 時点において推定されたトピック分布のパラメータである．このとき，ある時刻 t におけるユーザ a_u の閲覧系列 $\boldsymbol{X}_{t,u}$ の出現確率は以下の式 (15) で与えられる．ここで，$z_{t,u,m}$ はユーザ a_u の時刻 t における m 番目の閲覧の潜在トピックである．

$$
\begin{aligned}
&P(\boldsymbol{X}_{t,u} | \boldsymbol{\theta}_{t,u}, \hat{\boldsymbol{\Phi}}) \\
&= \prod_{m=1}^{M_{t,u}} \sum_{k=1}^{K} P(z_{t,u,m} = k | \boldsymbol{\theta}_{t,u}) P(x_m^{t,u} | z_{t,u,m} = k, \hat{\boldsymbol{\phi}}_k) \\
&= \prod_{m=1}^{M_{t,u}} \sum_{k=1}^{K} \theta_{t,u,k} \hat{\phi}_{k,x_m^{t,u}} \qquad (15)
\end{aligned}
$$

ただし，$\hat{\boldsymbol{\Phi}} = (\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_K)$ であり，$\hat{\boldsymbol{\phi}}_k = (\hat{\phi}_{k,1}, \ldots, \hat{\phi}_{k,D})$ は式 (13) によって推定されたアイテム分布のパラメータである．

3.3.2 モデルの学習

STEP.1 と同様に，STEP.2 においても確率的 EM アルゴリズムを用いた学習を行う．

ここで，$\boldsymbol{Z}_t = \{z_{t,u,m}\}_{u \in \mathcal{U}, m \in \mathcal{M}_{t,u}}$（$\mathcal{M}_{t,u} = \{1, 2, \ldots, M_{t,u}\}$），$N_{t,u,k}$ をユーザ a_u が時点 t においてトピック k に割り当てられた回数とし，$N_{t,u} = \sum_{k=1}^{K} N_{t,u,k}$，$\hat{\boldsymbol{\Theta}}_{t-1} = \{\hat{\boldsymbol{\theta}}_{t-1,u}\}_{u \in \mathcal{U}}$ であるとする．ただし，$\boldsymbol{Z}_{t \setminus t,u,m}$ は $\boldsymbol{Z}_t$ から

$z_{t,u,m}$ を除いたトピックの割当てであり，$N_{t,u,k\backslash t,u,m}$ は $x_m^{u,t}$ を $\boldsymbol{X}_t$ から除いた際に，トピック k に割り当てられたページ数である．このとき，ユーザ a_u の m 番目の閲覧ページに対してトピック k が割り当てられる確率は，以下の式 (16) によって与えられる．ただし，$\boldsymbol{\alpha}_t = (\alpha_{t,u}, \ldots, \alpha_{t,U})$ である．

$$
\begin{aligned}
&p(z_{t,u,m} = k | \boldsymbol{X}_t, \boldsymbol{Z}_{t\backslash t,u,m}, \hat{\boldsymbol{\Theta}}_t, \boldsymbol{\alpha}_t) \\
&\propto p(z_{t,u,m} = k | \boldsymbol{Z}_{t\backslash t,u,m}, \hat{\boldsymbol{\Theta}}_t, \boldsymbol{\alpha}_t) p(x_m^{t,u} | \boldsymbol{X}_{t\backslash t,u,m}, \hat{\boldsymbol{\Phi}}) \\
&= \int p(z_{t,u,m} = k | \boldsymbol{Z}_{t\backslash t,u,m}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} | \hat{\boldsymbol{\Theta}}_t, \boldsymbol{\alpha}_t) d\boldsymbol{\Theta} \times \hat{\phi}_{k, x_m^{t,u}} \\
&\propto \frac{\Gamma(N_{t,u,k\backslash t,u,m} + \alpha_{t,u} \hat{\theta}_{t-1,u,k})}{\Gamma(N_{t,u\backslash t,u,m} + \alpha_{t,u})} \times \hat{\phi}_{k, x_m^{t,u}} \qquad (16)
\end{aligned}
$$

さらに，ハイパーパラメータ $\alpha_{t,u}$ は，不動点反復法によって式 (17) によって更新され，更新後のパラメータは時点 $t+1$ の更新に用いられる．

$$
\alpha_{t,u} \leftarrow \alpha_{t,u} \frac{\sum_{k=1}^{K} \hat{\theta}_{t,u,k} A_{t,u,k} \alpha_{t,u}}{\Psi(N_{t,u} + \alpha_{t,u}) - \Psi(\alpha_{t,u})} \qquad (17)
$$

ただし，$A_{t,u,k} = \Psi(N_{t,u,k} + \alpha_{t,u} \hat{\theta}_{t-1,u,k}) - \Psi(\alpha_{t,u} \hat{\theta}_{t-1,u,k})$ である．学習の反復後，時点 t におけるユーザ a_u のトピック分布は以下の式 (18) によって推定される．

$$
\hat{\theta}_{t,u,k} = \frac{N_{t,u,k} + \alpha_{t,u} \hat{\theta}_{t-1,u,k}}{N_{t,u} + \alpha_{t,u}} \qquad (18)
$$

提案手法の STEP.2 において閲覧データが存在しない時刻が存在した場合には，その時刻の閲覧は取り扱わず，次に閲覧が存在した時間において，以前に閲覧が存在した時刻を 1 時刻前のトピック分布として扱い分布の推定を行う．一方，TTM においては，時刻ごとにアイテム時刻ごとにアイテム分布を推定するため各トピックが時刻によって変化することを許容している．そのため，異なる時刻の閲覧が混在した状態である時刻のアイテム分布が推定される構造となっていることから同様の方法をとることができず，TTM ではすべての時刻においてユーザの閲覧が存在しているという制約を満たす必要がある．

4. 人工データによる評価実験

ここでは，提案手法の有用性を検証するため，ユーザの興味の切替え頻度と未収束ユーザの割合を変化させた場合についてモデルの性能の変化の評価，さらにモデルの処理時間の評価を行う．

4.1 人工データの生成

本研究においては，閲覧履歴データにおいて以下の 3 つの特性を仮定して，人工データを生成する．

(1) 閲覧の中でユーザの興味が変化する．
(2) 閲覧系列内の連続した短い範囲の閲覧行動では，ユーザは類似したページを閲覧する．

ここでは，閲覧履歴データにおける仮定を再現するために，興味トピックとステージという概念を導入する．興味トピックとは，ユーザが特に興味を向けているトピックであり，ステージとは，ユーザの興味が同一である期間のこととする．すなわち，同一のステージにおいては，あるユーザの興味トピックは一定である．ユーザ a_u のステージ s における興味トピックを $\mathcal{Z}_{s,u}^*$ とすると，ユーザ a_u のステージ s におけるトピック分布は式 (19) によって定義される．

アルゴリズム 3 人工データの生成

```
for each stage s = 1, 2, ··· , S do
    for each user u = 1, 2, ··· , U do
        for each window m = 1, 2, ··· , M do
            Draw topic
            z_{s,u,m} ~ Categorical(θ*_{s,u})
            for each page n = 1, ··· , N_m do
                Draw page
                x^{s,u}_{m,n} ~ Categorical(φ*_{z_{s,u,m}})
```

$$
\theta_{s,u,k}^* = \begin{cases} \frac{1 - \left(K - n\left(\mathcal{Z}_{s,u}^*\right)\right) \times \theta_{\text{noise}}}{n\left(\mathcal{Z}_{s,u}^*\right)}, & \text{if } k \in \mathcal{Z}_{s,u}^* \\ \theta_{\text{noise}}, & \text{otherwise} \end{cases} \qquad (19)
$$

また，アイテム分布は全時刻で共通とし，以下の式 (20) を各値に持つ多項分布として定義する．

$$
\phi_{k,x}^* = \frac{\int_{x-1}^{x} \mathcal{N}(x | \mu_k, \sigma^2) dx}{\int_0^D \mathcal{N}(x | \mu_k, \sigma^2) dx} \qquad (20)
$$

$$
\mu_k = \left(k - \frac{1}{2}\right) \times \frac{D}{K} \qquad (21)
$$

最終的に，閲覧履歴データは生成アルゴリズム 3 によって作成される．

4.2 ユーザの興味の切替え頻度による性能の変化

本提案手法は，ユーザの興味が頻繁に切り替わる状況に対しても適用可能な手法であることを目指している．そこで，本節において興味の切替わり頻度が短くなった場合にモデルの性能が低下が認められないかどうかを確認する．

4.2.1 実験条件

全ユーザ数 $U = 1000$，全 Web ページ数 $D = 1600$，潜在トピック数 $K = 8$ とし，$\theta_{\text{noise}} = \frac{1}{K} \times 0.05$，$\sigma = \frac{D}{K \times 6}$，$M = 10$，$N_m = 5$ とした．さらに，ウィンドウサイズは $C = 3$，ハイパーパラメータの初期値は，STEP.1 においては $\alpha = 0.1$，$\beta_k = 0.1$（$k = 1, \ldots, K$），STEP.2 においては $\alpha_{1,u} = 50.0$（$u = 1, \ldots, U$）とした．ここでは，ユーザの興味の切り替わる期間が変化した場合の推定性能を比較するために，**表 2** のように S とそれに対応する M，N_s を変化させ実験を行った．ただし，N_s は 1 ステージあたりの閲覧回数であり，ステージ数 S は全閲覧においてユーザのトピック分布が切り替わる回数を意味する．ここで，すべての S において総閲覧数は一定で 300 であり，S が大きいほどユーザのトピック分布が頻繁に切り替わっている

表 2 実験条件
Table 2 The experiment setting.

S	1	2	3	6
M	60	30	20	10
N_s	300	150	100	50

アルゴリズム 4 トピックの対応関係の判定

$\mathcal{Z}_{\text{true}} = \{1, 2, \cdots, K\}$
$\mathcal{Z}_{\text{pblack}} = \{1, 2, \cdots, K\}$
for each association $i = 1, 2, \cdots, K$ **do**
  for each true topic $k \in \mathcal{Z}_{\text{true}}$ **do**
    for each pblackicted topic $l \in \mathcal{Z}_{\text{pblack}}$ **do**
      calculate KL divergence
      $D_{k,l} = \sum_{d=1}^{D} \phi^*_{k,d} \log \frac{\phi^*_{k,d}}{\hat{\phi}_{k,d}}$
  $k^*, l^* = \underset{\mathcal{Z}_{\text{true}}, \mathcal{Z}_{\text{pblack}}}{\arg\min} (D_{k,l})$
  associate topic $k^* := l^*$
  update $\mathcal{Z}_{\text{true}} = \{\mathcal{Z}_{\text{true}} \setminus \{k^*\}\}$
  update $\mathcal{Z}_{\text{pblack}} = \{\mathcal{Z}_{\text{pblack}} \setminus \{l^*\}\}$

ことを意味する．また，ここでは以下の式 (22) によって，ユーザの興味トピックを定義する．

$$\mathcal{Z}^*_{s,u} = \text{sample}\left(\mathcal{Z}^* \setminus \mathcal{Z}^*_{s-1,u}, 2\right) \tag{22}$$

ただし，$\mathcal{Z}^* := \{1, 2, \cdots, K\}$，sample$(\mathcal{A}, n)$ は，集合 $\mathcal{A}$ からランダムに選択された n 個の要素からなる集合を返す関数であり，$\mathcal{Z}^*_{0,u} = \mathcal{Z}^*$ であるとする．なお，最後のステージにおいては 10 閲覧多く生成し，これをテストデータとして使用した．

4.2.2 アイテム分布の推定性能の評価

アイテム分布の推定性能を明らかにするため，以下の式 (23) によって定義されるトピック間距離を用い，真の構造と推定された構造の差異によって評価を行う．ただし，$\hat{\boldsymbol{\Phi}} = (\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_K)$ は各手法によって推定されたアイテム分布である．

$$\text{KL}(\boldsymbol{\Phi}^*, \hat{\boldsymbol{\Phi}}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{d=1}^{D} \phi^*_{k,d} \log \frac{\phi^*_{k,d}}{\hat{\phi}_{k,d}} \tag{23}$$

なお，真のアイテム分布と推定されたアイテム分布のトピックはアルゴリズム 4 によって対応づけを行った．ここで，トピック間距離はトピック間の類似性を評価する指標であり，この値が小さいほど真の分布との距離が小さく，真のアイテム分布を正確に推定できていることを意味する．

いま，ステージ数 S を変化させた場合の真のアイテム分布と各手法で推定されたアイテム分布のトピック間距離を**図 2** に示す．図 2 より，TTM や LDA はステージ数 S が増加するにつれて，真のトピック分布との距離が大きくなりアイテム分布の推定性能が低下していることが分かる．一方で，提案手法における真の分布とのトピック間距離は，ステージ数 S によらず低い値を示している．このことから，提案手法ではユーザの興味の変化が頻繁に発生する場合においても適切なアイテム分布の推定が可能であるといえる．

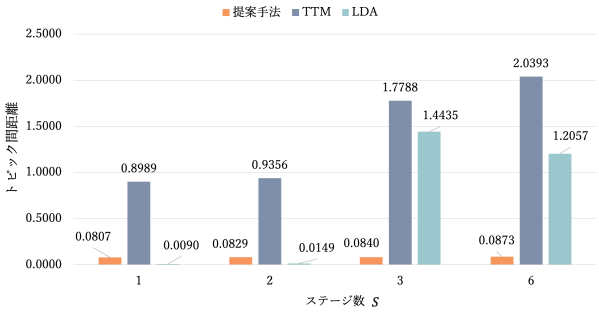

図 2 ステージ数 S を変化させたときのトピック間距離の比較
Fig. 2 Comparison of the distance between topics for varying S.

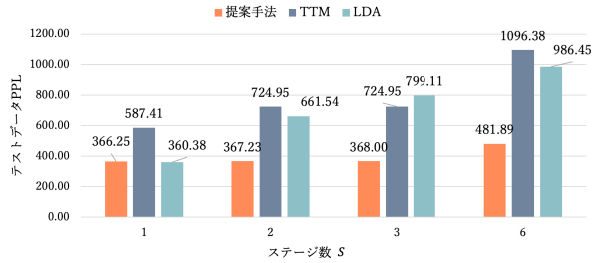

図 3 ステージ数 S を変化させたときのテストデータ PPL の比較
Fig. 3 Comparison of test data PPL for varying S.

4.2.3 閲覧の予測性能の比較

ユーザの興味が頻繁に切り替わる状況においてもモデルの性能が低下しないかを評価するために，ここでは表 2 に示した S ごとに式 (24) によって定義されるテストデータ Perplexity（以下，PPL）の比較を行う．PPL とはトピックモデルの性能を評価する指標であり，小さいほどモデルの性能が高いことを意味する．ここで，PPL は式 (24) によって求められる．

$$\text{PPL} = \left\{ \prod_t \prod_u P(\boldsymbol{X}_{t,u} | \boldsymbol{\theta}_{t,u}, \boldsymbol{\Phi}_t) \right\}^{-\frac{1}{N}} \tag{24}$$

ただし，$N = \sum_t \sum_u N_{t,u}$ である．

いま，ユーザのトピック分布の切替わり回数を意味するステージ数 S を変化させたときの各手法のテストデータ PPL の値を**図 3** に示す．図 3 より，TTM や LDA では，S が大きくなる，すなわちユーザの興味が短期間で切り替わるようになるにつれてテストデータ PPL が増加し，性能が低下しているのに対して，提案手法は S が増加してもテストデータ PPL は大きく変化しないことが分かる．よって，提案手法は従来手法の TTM と比較してユーザの興味が短期間で切り替わる状況においても，ユーザの興味をトピック分布としてとらえることが可能であるといえる．

4.3 未収束ユーザの割合による性能の変化

本提案手法は収束ユーザと未収束ユーザが混在してい

る状況において収束ユーザを適切に検出するというユースケースを想定している．ここで，LDA などの従来のトピックモデルにおいてはユーザの興味がランダムに切り替わる未収束ユーザが多く存在している場合，トピック間での閲覧の偏りが存在しなくなってしまうためにパラメータの推定が困難となる．一方，実応用において未収束ユーザが多いという状況は，特定の EC サイトにおいては購買に対してきわめて慎重で，アイテムの探索に近い意味で閲覧を行っているユーザが大多数を占めるというケースであり，多くの EC サイトでも生じている状況であると考えられる．よって，このような場合において施策実施の効果が見込まれる収束ユーザを適切に検知することは，実応用上重要な課題である．以上のことから，ここでは未収束ユーザの割合を増加させ，この割合によってモデルの性能がどのように変化するのか評価する．

4.3.1 実験条件

全ユーザ数 $U=1000$, 全 Web ページ数 $D=1600$, ステージ数 $S=4$, 潜在トピック数 $K=8$ とし, $\theta_{\text{noise}}=\frac{1}{K}\times 0.05$, $\sigma=\frac{D}{K\times 6}$, $M=10$, $N_m=5$ とした．また，ここでは未収束ユーザと収束ユーザが混在していることを仮定し，収束ユーザ集合を $\mathcal{U}_{\text{conv}}$，未収束ユーザ集合を $\mathcal{U}_{\text{random}}$ とする．このとき，ユーザ a_u のステージ s における興味トピック集合は以下の式 (25) のように定義される．

$$\mathcal{Z}^*_{s,u}=\begin{cases}\text{sample}\left(\mathcal{Z}^*_{s-1,u},\frac{K}{2^{(s-1)}}\right), & u\in\mathcal{U}_{\text{conv}}\\ \text{sample}\left(\mathcal{Z}^*\setminus\bigcup_{s'=1}^{s}\mathcal{Z}^*_{s'-1,u},\frac{K}{S}\right), & u\in\mathcal{U}_{\text{random}}\end{cases} \tag{25}$$

ただし，$\mathcal{Z}^*_{0,u}$ は以下の式 (26) によって定義される．

$$\mathcal{Z}^*_{0,u}=\begin{cases}\mathcal{Z}^*, & u\in\mathcal{U}_{\text{conv}}\\ \varnothing, & u\in\mathcal{U}_{\text{random}}\end{cases} \tag{26}$$

加えて，実問題においては，ユーザの興味が切り替わる時刻，すなわちステージが変化するタイミングは未知であるため，トピック分布を推定する際には，1 時点あたりの閲覧ページ数 N_t を設定する．本実験においては，N_t は $M\times N_s$ の約数であり，ステージ s と時点 t は 1 対 1 に対応するものとする．ここで，ステージ s に対応する時点 t を s_t と定義し，ここでは $N_t=25$，全時点数 $T=8$ とした．このとき，1 ユーザあたりの閲覧ページ数は $N_t\times T=25\times 8=200$，各ユーザの 1 ステージあたりの閲覧ページ数は $M\times N_s=10\times 5=50$ であり，$s_2=1, s_3=2$ となる．また，各ステージにおける興味トピック数は以下のようである．

$$\{n(\mathcal{Z}^*_{s,u\in\mathcal{U}_{\text{conv}}})|s=1,\cdots,4\}=\{8,4,2,1\}$$
$$\{n(\mathcal{Z}^*_{s,u\in\mathcal{U}_{\text{random}}})|s=1,\cdots,4\}=\{2,2,2,2\}$$

さらに，ウィンドウサイズは $C=3$，ハイパーパラメータの初期値は，STEP.1 においては $\alpha=0.1$, $\beta_k=0.1$ $(k=1,\ldots,K)$，STEP.2 においては $\alpha_{1,u}=10.0$ $(u=1,\ldots,U)$ とした．ここでは，全ユーザに占める未収束ユーザの割合 r_{random} を $r_{\text{random}}=0.6,0.7,0,8,0.9$ と変化させて実験を行った．

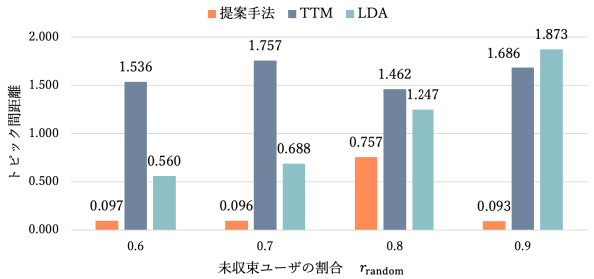

図 4 r_{random} を変化させたときのトピック間距離の比較

Fig. 4 Comparison of the distance between topics for varying r_{random}.

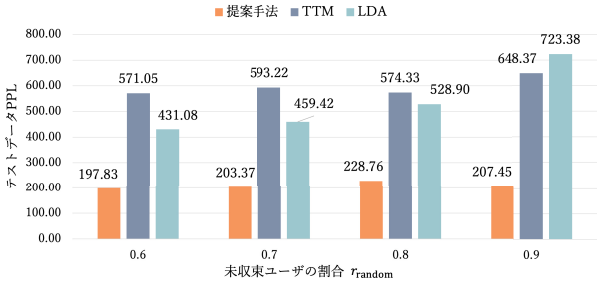

図 5 r_{random} を変化させたときの収束ユーザのテストデータ PPL の比較

Fig. 5 Comparison of test data PPL for varying r_{random} of $\mathcal{U}_{\text{conv}}$.

4.3.2 アイテム分布の推定性能の評価

アイテム分布の推定性能を明らかにするため，式 (23) によって定義されるトピック間距離を用い，真の構造と推定された構造の差異によって評価を行う．ここで，図 4 に未収束ユーザの割合 r_{random} を変化させたときの LDA, TTM，および提案手法のトピック間距離の推移を示す．図 4 より，いずれの r_{random} においても従来手法と比較して提案手法は高い値を示していることが分かる．LDA が未収束ユーザ割合 r_{random} が大きくなるにつれてトピック間距離が小さくなっているのに対し，提案手法では，あまり変化していない．これは，LDA ではユーザごとにトピック分布を推定するのに対して，提案手法では，全ユーザに単一のトピック分布を仮定し，ウィンドウごとにトピックの割当てを行っているために，局所的なトピックの偏りをとらえることができるためと考えられる．そのため，提案手法は LDA と比較して，未収束ユーザの影響受けることなく適切にアイテム分布の推定を行うことが可能であるといえる．

4.3.3 閲覧の予測性能の比較

提案モデルの性能を評価するために，LDA，TTM と式

表 3 オフラインの処理時間
Table 3 Off-line processing time.

手法	学習時間（秒/イテレーション）
提案手法（STEP.1）	419
提案手法（STEP.2）	35
TTM	58
LDA	34

表 4 オンラインの処理時間
Table 4 On-line processing time.

手法	z のサンプリング（秒/回）	$\boldsymbol{\theta}_{t,u}$ の更新（秒/t・人）	合計処理時間（秒/30 閲覧）
提案手法	1.16×10^{-4}	4.70×10^{-5}	3.53×10^{-3}
TTM	1.88×10^{-4}	4.63×10^{-5}	5.68×10^{-3}
LDA	1.26×10^{-4}	5.13×10^{-5}	3.84×10^{-3}

(24) によって定義されるテストデータ PPL の比較を行った．ここでは，最後の 1 時刻（$t = 8$）をテストデータとして用いた．ここで，未収束ユーザの割合を変化させたときの各手法のテストデータ PPL の値を**図 5** に示す．図 5 の結果より，提案手法はいずれの条件においても従来手法と比較して低い PPL を示しており，提案手法による段階的なパラメータの推定とユーザのトピック分布への時系列性の仮定が有効であったといえる．

4.4 モデルの処理時間

提案手法の実用化のためには，オンラインの処理時間が実行上問題ないとならない範囲に収まっていることが非常に重要である．そのため，本節において提案手法における処理をオンラインとオフラインに分け，それぞれについて従来手法と処理時間の比較を行う．ここで，オフラインの処理には過去に蓄積されたすべてのデータを使用した，全時刻における全ユーザのトピック分布およびアイテム分布の推定が含まれる．一方で，オンラインの処理時間はある時刻において新たな閲覧系列が取得されたときの更新処理であり，特定時刻，特定ユーザのトピック分布の更新が含まれいるものと定義する．

4.4.1 実行条件

オンライン，オフラインの処理時間それぞれについて，Intel(R) Xeon(R) CPU @ 2.20 GHz を 2 つ搭載した 1 台のマシンで計測した．ただし，ユーザ数は 1,000，ユーザ 1 人あたりの総閲覧回数は 300 回とし，オンライン処理においては 1 人のユーザについて 30 閲覧を新規に取得した場合を想定した．その他のモデルの学習条件については，4.2.1 項と同様であるとし，ステージ数は $S = 6$ とした．

4.4.2 実行結果

オフライン，オンラインの処理時間をそれぞれ**表 3**，**表 4** に示す．さらに，オフライン処理時の対数尤度の変化を

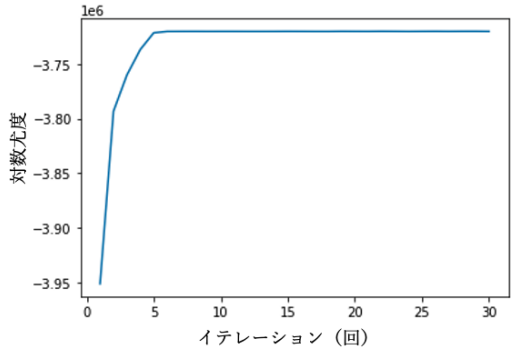

図 6 提案手法（STEP.1）のオフライン処理時の対数尤度の変化
Fig. 6 The change in log-likelihood of the proposed method (STEP.1) during offline processing.

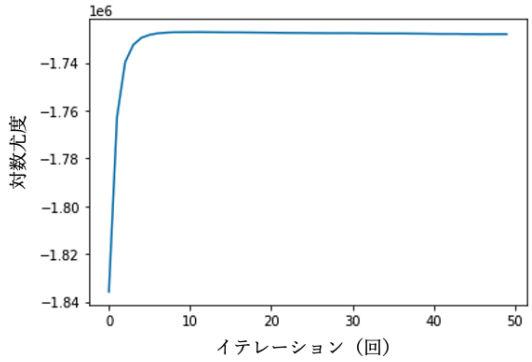

図 7 提案手法（STEP.2）のオフライン処理時の対数尤度の変化
Fig. 7 The change in log-likelihood of the proposed method (STEP.2) during offline processing

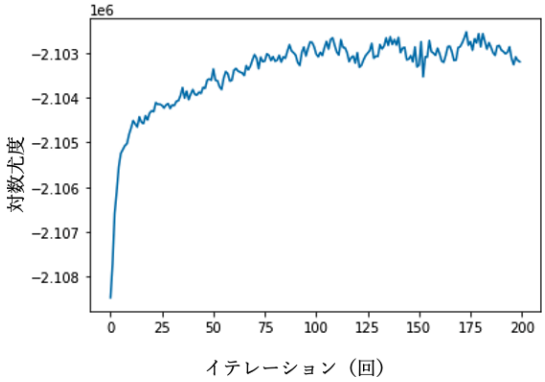

図 8 TTM のオフライン処理時の対数尤度の変化
Fig. 8 The change in log-likelihood of TTM during offline processing.

図 7，**図 8**，**図 9** に示す．まず，オンライン処理においては，従来手法と比較して，提案手法の STEP.1 における 1 イテレーションあたりの処理時間が長くなっているものの，図 7–図 9 に示すように収束までに必要となるイテレーション回数が少なくなるために，合計の処理時間としては従来手法と比較して短くなっている．また，オフラインの処理時間については表 4 より，1 秒未満で更新が完了しており，実用において十分耐えうる処理時間であると考えられる．

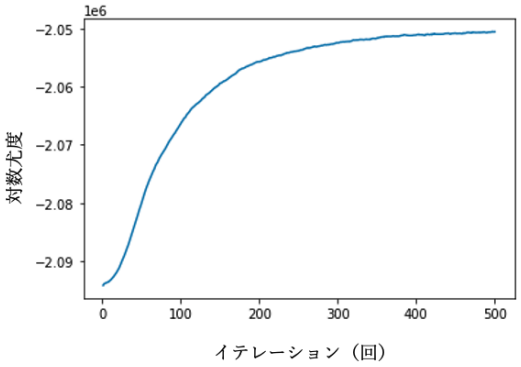

図 9 LDA のオフライン処理時の対数尤度の変化

Fig. 9 The change in log-likelihood of LDA during offline processing.

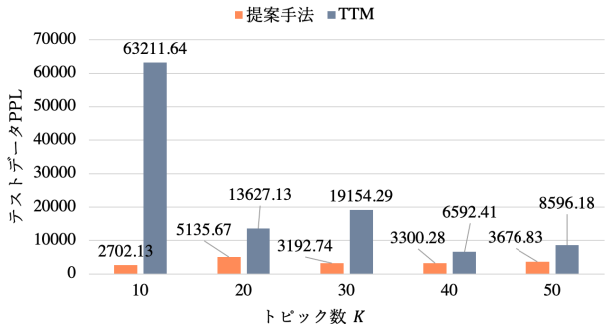

図 10 実データにおけるテストデータ PPL の比較

Fig. 10 The comparison of the test data PPL of real data.

5. 実データ分析

本研究では，Web サイトの閲覧履歴を対象値として実ビジネスへの適用を目指した手法の提案を目的としている．そのため，本章において提案手法を実データに適用した場合のモデルの性能の評価を行う．さらに，モデルの学習によって得られる推定結果の具体例を示し，その結果をふまえてどのようなマーケティング上のアクションをとることができるのかを検討する．

5.1 分析条件

ここでは，株式会社ヴァリューズ提供の楽天市場の閲覧履歴データを用いる．対象データのデータ取得期間は 2019 年 2 月 1 日–4 月 31 日であり，5.2 節においては，1 期間を 1 週間として設定[*1]し，全期間において閲覧が存在したユーザを抽出した．その結果抽出されたユーザ数は $U = 47$ であり，全 Web ページ数 $D = 13,527$，総閲覧回数は 29,888 であった．このとき，ウィンドウサイズは $C = 5$ として設定[*2]した．また，5.3，5.4 節においては，1 期間を 1 日として設定し，1 期間あたりの閲覧回数が 10 回以上存在する期間が 30 以上あるユーザを抽出した．その結果，全ユーザ数 $U = 178$，全 Web ページ数 $D = 5,670$，総閲覧回数は 301,916 であり，ウィンドウサイズは $C = 10$ とした[†]．これらの閲覧のうち各ユーザについて最後の 1 閲覧をテストデータとした．

さらに，ハイパーパラメータの初期値は共通して，STEP.1 では $\alpha = 0.1$，$\beta_k = 0.1$（$k = 1, \ldots, K$），STEP.2 では $\alpha_{1,u} = 10.0$（$u = 1, \ldots, U$）とした．

5.2 閲覧の予測性能の比較

提案モデルの性能を評価するために，TTM とテストデータ PPL の比較を行った．ここで，トピック数 K を 10 から 50 まで変化させたときの提案手法と TTM のテストデータ PPL の値を**図 10** に示す．図 10 の結果より，いずれのトピック数においても提案手法は TTM よりも低い PPL を示しており，実データにおいても提案手法による段階的なパラメータの推定が有効であったといえる．

5.3 提案手法によって推定したトピックの分析

ここで，提案手法により推定されたアイテム分布 ϕ_k（$k = 2, 3, 4, 6$）の出現確率上位 5 ページを**表 6**，**表 7**，**表 8**，**表 9** に示す．ただし，衣類においては非常に多様なページが存在したため，各ページを実際に確認し，括弧書きとして服飾の系統を記載した．また，ページ名は各ページ URL の店舗を意味する階層部分を抽出して記載している．

まず，表 7 に示すトピック 3 においては，ファッション系のページが多く含まれており，ファッション系のトピックであると解釈することができる．さらに，このトピックでは，ファッション系というカテゴリが同じだけでなく，非常に類似性の高いページが帰属していた．これは，潜在トピックの割当てをウィンドウごとに行うことで，ウィンドウ内の閲覧ページの内容の類似性を仮定したためと考えられる．一方で，表 6 より，トピック 2 においては時計や衣類，天然アクセサリーパーツなどファッションに関する多様なカテゴリのページが出現しており，ファッション雑貨のトピックであると考えられる．また，表 9 よりトピック 6 においても園芸やゴルフ用品などアウトドアに関する複数のカテゴリのページが出現しており，園芸やアウトドアに関するトピックであることがうかがえる．ここで，これらと同様の解釈をすべてのトピックに対して行った結果を**表 5** に示す．また，トピック 2，6 のように多様なカテゴリのページが含まれるトピックが得られた理由として，日用品や園芸，アウトドア関連のページはユーザが連続して類似したページを見ることは比較的少なく，ウィンドウ内においても様々なページを閲覧していることがあげられる．このことが正しければ，トピックの内容により一貫性を持たせるためには，ウィンドウサイズ C の設定を比較的

[*1] TTM ではすべての期間において閲覧履歴が存在しないと学習を行うことができないため，この条件を満たすように 1 週間を 1 期間とした．

[*2] ウィンドウサイズ C は，$C = 2, 3, 4, 5, 7, 10$ と変化させたときに最もテストデータ PPL が低かったものを採用した．

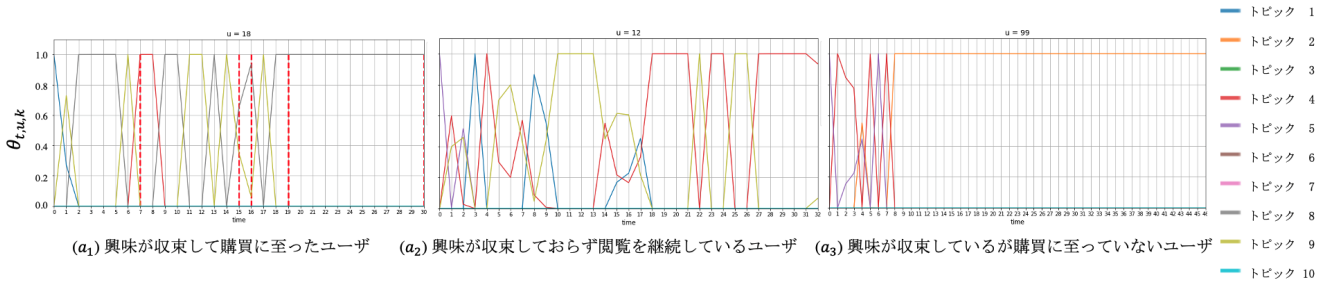

図 11 各ユーザのトピック分布の変化
Fig. 11 The change in topic distribution for each user.

表 5 得られたトピックの解釈
Table 5 The interpretation of estimated Topics.

k	解釈	k	解釈
1	ファッション・手芸	6	園芸・アウトドア
2	ファッション雑貨	7	中古ファッション
3	女性用衣類	8	生活雑貨
4	日用品	9	スポーツ系ファッション
5	女性用衣類	10	ファッション・手芸

表 6 トピック 2 の出現確率 $P(x_i|z)$ 上位 5 ページ
Table 6 Top 5 pages for $P(x_i|z)$ of Topic 2.

	ページ名	ページカテゴリ	$P(x_i|z)$
1	kinkodo	時計・眼鏡用品	0.0425
2	paty	衣類（カジュアル）	0.0409
3	sockkobe	女性用下着	0.0404
4	furusato	特産品	0.0308
5	lifestone	天然石アクセサリーパーツ	0.0265

表 7 トピック 3 の出現確率 $P(x_i|z)$ 上位 5 ページ
Table 7 Top 5 pages for $P(x_i|z)$ of Topic 3.

	ページ名	ページカテゴリ	$P(x_i|z)$
1	auc-ecoloco	女性用衣類（カジュアル）	0.0380
2	woody-h	衣類（カジュアル）	0.0335
3	onepi-c	女性用衣類（きれいめ）	0.0298
4	auc-tresor	女性用衣類（カジュアル）	0.0262
5	aranciato	女性用衣類（カジュアル）	0.0222

表 8 トピック 4 の出現確率 $P(x_i|z)$ 上位 5 ページ
Table 8 Top 5 pages for $P(x_i|z)$ of Topic 4.

	ページ名	ページカテゴリ	$P(x_i|z)$
1	rakuten24	日用品	0.0795
2	soukai	日用品	0.0512
3	kenkocom	日用品・健康食品	0.0368
4	yunica	開運アイテム	0.0227
5	chanet	ペット用品	0.0205

小さくすればよいと考えられる．

5.4 ユーザの興味の変化に対する分析

ここでは，ユーザの嗜好の変化に関する分析と分析結果を用いたマーケティング施策について具体的な考察を行うために，事例として以下の 3 タイプのユーザを抽出した．

- a_1：興味が収束して購買に至ったユーザ
- a_2：興味が収束しておらず閲覧を継続しているユーザ
- a_3：興味が収束しているが購買に至っていないユーザ

ここで，a_1，a_2，a_3 のユーザのトピック分布の変化を**図 11** に示す．ただし，図の縦軸は，各時刻における所属確率 $\theta_{t,u,k}$ を表し，赤色の波線は購買の発生を意味する．

まず，ユーザ a_1 は，時刻 $t = 18$ 以降，トピック 8 への所属確率が大きくなり，時刻 $t = 19$ において購買を行っている．さらに，購買後も同一のトピックに対する所属確率が高い状態が継続していることから，このユーザはトピック 8 への興味が収束しており，このトピックへの所属確率が高い店舗のクーポンの発行やメルマガの送付により追加購買を促すことが可能であると考えらえる．

一方で，ユーザ a_2 は時刻ごとのトピック分布の変動が大きく，ユーザの興味の対象が十分に絞り込めていないと想定できる．このようなユーザが購買に至る可能性は低く，クーポンの発行など施策の効果は十分に見込めないことが考えられるため，コストのかかる施策実施の優先度が低いユーザ群である．しかし，ユーザ a_2 については，時刻 $t = 18$ 以降にトピック 4 とトピック 8 の間で興味が移り変わるようになってきている．このことから，ユーザ a_2 に対してこれらのトピックに関する商品をトップページに表示させるなど，コストが比較的かからない推薦などを実施することにより興味を収束させ，購買につなげられる可能性がある．

また，ユーザ a_3 は時刻 $t = 8$ 以降にトピック 2 への興味の高まりと収束が確認されるが，購買には至っていないユーザである．このようなユーザについては，十分に興味

表 9 トピック 6 の出現確率 $P(x_i|z)$ 上位 5 ページ
Table 9 Top 5 pages for $P(x_i|z)$ of Topic 6.

	ページ名	ページカテゴリ	$P(x_i|z)$
1	locondo	靴	0.0715
2	engei2	園芸用品	0.0657
3	hanaipn	鉢植え	0.0493
4	golfpartner	ゴルフ用品	0.0412
5	fan-annex	インポート布地	0.0281

の対象が絞れていることから，直近の閲覧店舗のクーポンの発行や商品の追加情報の提示などによって，購買を誘発できる可能性がある．

以上のように，提案手法を用いることによってユーザの興味の変化を明らかにし，興味の収束度合いに基づくマーケティング施策の立案が可能であるといえる．

6. 考察

6.1 提案手法の実応用範囲

多くのECサイトにおいては，ユーザの購買率を向上させるために推薦システムを実装していることが一般的となっている．推薦システムとは，ユーザの過去の閲覧や購買情報に基づいて算出されたユーザの予測購買確率が高い，つまりユーザの嗜好と合致すると推定されたアイテムを各ユーザに表示するなどの仕組みを指す．日用品などの比較的低価格なアイテムに関しては，この推薦システムによってユーザの購買意欲を高めることが可能になると考えられるが，衣類や家電，家具などの比較的高価格帯のアイテムに関しては推薦システムが提供する情報のみではユーザは購買に踏み切ることができない可能性がある．さらに，これらの高価格帯の商品を取り扱うECサイトにおいては，1商品あたりの利益も大きく，購買を検討している顧客に対して確実にアイテムを購買してもらうことが利益を向上させるために非常に重要である．そのため，このような高価格帯の商品を扱うECサイトでは，購買意欲が高まっているユーザに対して，クーポンの発行や追加情報の提示などのアプローチをとる必要があり，多くのサービスにおいてこれらの仕組みが実装されつつある．しかしながら，これらの施策は過去の購買回数や金額などの一定のルールに基づいて実施されることが多く十分に最適化がなされているとは言い難い．このような場合に提案手法による分析結果を活用することで，ユーザの興味の収束度合いに基づいた施策の実施が可能となり，施策の費用対効果の向上が期待される．

6.2 ユーザの興味の収束判定

提案手法を用いてユーザのトピック分布を推定することにより，この分布を用いてユーザの興味の変化を解釈することが可能になる．しかし，興味が収束したユーザ群などを厳密に定義する場合には，トピック分布 $\boldsymbol{\theta}_{t,u}$ $(t = 1, \ldots, T)$ を用いた収束判定の方法論の確立が必要になる．ここで，この収束判定のためには以下の2つの事項を考慮する必要があると考えられる．

(1) 少数のトピックに興味が偏っていること
(2) ユーザのトピック分布の変化が少ないこと

まず，(1) を判定するためには判定時点の時刻 t におけるユーザのトピック分布のうち，$\theta_{t,u,k}$ が最大となる k においてその値が一定以上であるかどうかによって判定することができる．この閾値 δ としては，$\delta = 0.5$ や $\delta = (K-1)/K$ などを用いることが考えられる．また，(2) を判定するためには，式 (27) によって定義される判定時点の時刻 t から S 時刻前までのトピック分布の変化率の平均などを基準にすることが考えられる．

$$c(u,t,k,S) = \frac{1}{S}\sum_{s=1}^{S}\frac{|\theta_{t-s+1,u,k} - \theta_{t-s,u,k}|}{\theta_{t-s,u,k}} \tag{27}$$

以上より，アルゴリズム5に示す収束判定アルゴリズムが考えられる．ただし，ここでは (1) の判定基準における閾値を δ，(2) の判定基準における閾値を ϵ としている．

アルゴリズム 5 収束判定アルゴリズム

```
for ユーザ u = 1, ··· , U do
    所属確率が最大の潜在トピック l を判定 l = arg max_k θ_{u,t,k}
    if θ_{t,u,l} > δ then
        if c(u, t, l, S) < ε then
            収束ユーザである a_u ∈ U_conv
        else
            収束ユーザでない a_u ∉ U_conv
    else
        収束ユーザでない a_u ∉ U_conv
```

この収束判定アルゴリズムにおけるパラメータ δ，ϵ については，値を変化させたときの収束ユーザの割合とトピック分布の変化などを確認することにより，実験的に調整する必要があると考えられる．収束の判定を厳しく設定するためには δ をより大きく，ϵ をより小さく設定すればよい．

6.3 ハイパーパラメータの設定

提案手法を用いて施策に関する意思決定を行う際には，意思決定の透明性を保つため，解釈性の高いトピックを形成可能なトピック数とウィンドウサイズの決定が重要である．これらの決定のためには，定量的・定性的，2つの観点からの評価が必要であると考えられる．まず，定量的な評価としてはPPLやCoherence [24] などのトピックモデルの性能を評価する指標を用いることが考えられる．PPLについては，一般にトピック数を増加させると減少していくという傾向があるため，Coherenceと併用してトピック数を決定する必要がある．一方で，定性的な評価としては，トピック数やウィンドウサイズを変化させて得られたトピックに対し，知識を持った分析者が解釈を与えるなどの方法があげられる．この解釈の観点としては，形成されたトピックが解釈可能であり施策立案の際の情報として有用であるか，形成されたトピックは今までの事前知識と照らし合わせて大きく乖離がないか，などが考えられる．

また，提案手法において，ウィンドウサイズの設定はモデルの性能に大きく影響するため適切な設定を行うことが重要である．適切なウィンドウサイズ C は，学習対象問題の統計的性質に依存し，局所的なユーザの興味の移り変わりが激しいようなデータに対しては短く，局所的な興味の

移り変わりがあまりみられない場合には比較的長めに設定する必要がある．そのため，ウィンドウサイズ C は，データの特性などを加味して試行結果を確認しながら探索的に調節する必要がある．

6.4 提案手法の有効性と拡張性

提案手法の最も大きな特徴はアイテム分布とトピック分布のパラメータの段階的な推定であり，ユーザごとのトピック分布は，アイテム分布が既知の状態で推定を行う．そのため，各ユーザの各時刻におけるトピック分布 $\boldsymbol{\theta}_{t,u}$ の推定の際に必要となるデータは，ユーザの閲覧系列 $\boldsymbol{X}_{t,u}$ のみである．これにより，提案手法では TTM や LDA と比較して少ない計算量でユーザのトピック分布を推定することが可能であり，実用上のメリットは大きい．さらに，事前に推定するアイテム分布については任意の分布を設定することも可能である．そのため，ユーザの属性情報などの補助情報を活用して潜在トピックの推定を行うことが可能なトピックモデル [25], [26], [27] や，多くの人の閲覧される人気のサイトを特定のトピックに集約することの可能なトピックモデル [28] で推定したアイテム分布を用いるなど，分析観点によって柔軟に潜在トピックを指定することができる．

7. まとめと今後の課題

本研究では，Web サイトの閲覧履歴データを対象とし，ユーザの興味の収束度合いを明らかにするために，ユーザの興味が短期間において切り替わるという状況においても安定してモデルのパラメータを推定可能な手法を提案した．人工データセットを用いた評価実験において，提案手法は未収束ユーザが多い状況においても収束ユーザの興味の収束をとらえることが可能であることを示した．さらに，実データセットを用いた分析により，提案手法は実データに対しても従来手法と比較して高精度なパラメータ推定が可能であり，マーケティング上の施策の施策の立案に対しても有用であることを示した．

本研究の成果により，閲覧履歴データからユーザの興味の変化をトピック分布としてモデル化し，この分布の変化を解析することにより，興味の収束度合いについて明らかにすることが可能になった．これにより，実ビジネスにおける様々なマーケティング施策の設計により新たな視座を提供することが可能になり，実務上の貢献も期待できる．

今後の課題として，閲覧行動の多様性の考慮したモデル化があげられる．トピックモデルで推定された潜在トピックは一般的に各トピックにおける出現確率が高い単語やアイテムによって特徴づけられ，きわめて大局的な傾向を表している [29]．一方，閲覧履歴データにおいては様々な閲覧パターンが想定され，少数のユーザにみられる特徴的な傾向を検出できることは，新たなニーズの発見など有用な情報となりうる．また，提案手法による収束判定に基づき施策を実施した場合の施策効果の実証的な検証も重要な課題の 1 つである．

謝辞 本研究にあたり，貴重な実データをご提供頂いた株式会社ヴァリューズの皆様に深く感謝致します．また，本研究の一部は科研費 21H04600 の助成を受けたものである．

参考文献

[1] Bucklin, R.E. and Sismeiro, C.: A model of web site browsing behavior estimated on clickstream data, *Journal of Marketing Research*, Vol.40, No.3, pp.249–267 (2003).
[2] Johnson, E.J., Bellman, S. and Lohse, G.L.: Cognitive lock-in and the power law of practice, *Journal of Marketing*, Vol.67, No.2, pp.62–75 (2003).
[3] 武政孝師，後藤順哉：EC サイトにおける顧客の閲覧履歴を利用した商品ランキング生成法，オペレーションズ・リサーチ：経営の科学，Vol.59, No.8, pp.465–471 (2014).
[4] 久松俊道，外川隆司，朝日弓未，生田目崇：EC サイトにおける購買予兆発見モデルの提案，オペレーションズ・リサーチ：経営の科学，Vol.58, No.2, pp. 93–100 (2013).
[5] 伊藤孝太朗，澤邊剛，保坂桂佑，松下亮祐，雪島正敏：顧客のセグメンテーションと商品のスコアリングによる購買予測，オペレーションズ・リサーチ：経営の科学，Vol.60, No.2, pp. 75–80 (2015).
[6] Hotoda, M., Kumoi, G. and Goto, M.: A Study on Customer Purchase Behavior Analysis Based on Hidden Topic Markov Models, *Industrial Engineering & Management Systems*, Vol.20, No.1, pp. 48–60 (2021).
[7] 松嵜祐樹, 三川健太, 後藤正幸：マルコフ潜在クラスモデルに基づく EC サイトにおける施策実施効果分析に関する一考察，情報処理学会論文誌，Vol. 58, No.12, pp.2034–2045 (2017).
[8] Giri, R., Choi, H., Hoo, K.S. and Rao, B.D.: User behavior modeling in a cellular network using latent dirichlet allocation, *International Conference on Intelligent Data Engineering and Automated Learning*, pp.36–44, Springer (2014).
[9] Iwata, T. and Sawada, H.: Topic model for analyzing purchase data with price information, *Data Mining and Knowledge Discovery*, Vol.26, No.3, pp.559–573 (2013).
[10] Iwata, T., Watanabe, S., Yamada, T. and Ueda, N.: Topic tracking model for analyzing consumer purchase behavior, *21st International Joint Conference on Artificial Intelligence* (2009).
[11] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation, *The Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
[12] Zhu, Q., Shyu, M.-L. and Wang, H.: Videotopic: Content-based video recommendation using a topic model, *2013 IEEE International Symposium on Multimedia*, pp.219–222, IEEE (2013).
[13] Blei, D.M. and Lafferty, J.D.: Dynamic topic models, *Proc. 23rd International Conference on Machine Learning*, pp.113–120 (2006).
[14] Wang, X. and McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends, *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.424–433 (2006).
[15] Iwata, T., Yamada, T., Sakurai, Y. and Ueda, N.:

Sequential modeling of topic dynamics with multiple timescales, *ACM Trans. Knowledge Discovery from Data* (TKDD), Vol.5, No.4, pp.1–27 (2012).
[16] Pruteanu-Malinici, I., Ren, L., Paisley, J., Wang, E. and Carin, L.: Hierarchical Bayesian modeling of topics in time-stamped documents, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.32, No.6, pp.996–1011 (2009).
[17] Zhang, H., Ni, W., Li, X. and Yang, Y.: Modeling the heterogeneous duration of user interest in time-dependent recommendation: A hidden semi-Markov approach, *IEEE Trans. Systems, Man, and Cybernetics: Systems*, Vol.48, No.2, pp.177–194 (2016).
[18] Gruber, A., Weiss, Y. and Rosen-Zvi, M.: Hidden topic markov models, *Artificial Intelligence and Statistics*, *PMLR*, pp.163–170 (2007).
[19] Beigi, G., Guo, R., Nou, A., Zhang, Y. and Liu, H.: Protecting user privacy: An approach for untraceable web browsing history and unambiguous user profiles, *Proc. 12th ACM International Conference on Web Search and Data Mining*, pp.213–221 (2019).
[20] Zhang, J., Li, Y., Chen, M. and You, L.: An implicit feedback integrated LDA-based topic model for IPTV program recommendation, *2016 16th International Symposium on Communications and Information Technologies* (ISCIT), IEEE, pp.216–220 (2016).
[21] Rajendran, D.P.D. and Sundarraj, R.P.: Using topic models with browsing history in hybrid collaborative filtering recommender system: Experiments with user ratings, *International Journal of Information Management Data Insights*, Vol.1, No.2, p.100027 (2021).
[22] Cheng, X., Yan, X., Lan, Y. and Guo, J.: Btm: Topic modeling over short texts, *IEEE Trans. Knowledge and Data Engineering*, Vol.26, No.12, pp.2928–2941 (2014).
[23] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, pp.3111–3119 (2013).
[24] Wallach, H.M.: Topic modeling: Beyond bag-of-words, *Proc. 23rd International Conference on Machine Learning*, pp.977–984 (2006).
[25] Yao, L., Zhang, Y., Wei, B., Jin, Z., Zhang, R., Zhang, Y. and Chen, Q.: Incorporating knowledge graph embeddings into topic modeling, *31st AAAI Conference on Artificial Intelligence* (2017).
[26] Ramage, D., Hall, D., Nallapati, R. and Manning, C.D.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, *Proc. 2009 Conference on Empirical Methods in Natural Language Processing*, pp.248–256 (2009).
[27] Xue, F., Hong, R., He, X., Wang, J., Qian, S. and Xu, C.: Knowledge-Based Topic Model for Multi-Modal Social Event Analysis, *IEEE Trans. Multimedia*, Vol.22, No.8, pp.2098–2110 (2019).
[28] Dieng, A.B., Ruiz, F.J. and Blei, D.M.: Topic modeling in embedding spaces, *Trans. Association for Computational Linguistics*, Vol.8, pp.439–453 (2020).
[29] Wu, X., Li, C. and Miao, Y.: Discovering Topics in Long-tailed Corpora with Causal Intervention, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp.175–185 (2021).

伊藤 史世

1997 年生．2022 年早稲田大学大学院創造理工学研究科経営システム工学専攻修了．2022 年同大学院創造理工学研究科経営システム工学専攻修了．2022 年 PwC コンサルティング合同会社入社．機械学習を用いたデータ分析によるビジネス課題の解決に従事．

雲居 玄道

2008 年早稲田大学理工学部経営システム工学科卒業，2020 年同大学大学院創造理工学研究科博士後期課程退学．2008 年早稲田大学理工学術院総合研究所嘱託研究員．2015 年浄土真宗本願寺派総合研究所研究助手．2019 年早稲田大学創造理工学部経営システム工学科助手．2022 年早稲田大学データ科学センター講師，現在に至る．博士（工学）．機械学習・データマイニングの研究に従事．IEEE，経営情報学会，日本気象学会等各会員．

後藤 正幸 （正会員）

1994 年武蔵工業大学大学院修士課程修了．2000 年早稲田大学大学院理工学研究科博士課程修了．博士（工学）．1997 年同大学理工学部助手．2000 年東京大学大学院工学研究科助手．2002 年武蔵工業大学環境情報学部助教授．2008 年早稲田大学創造理工学部経営システム工学科准教授．2011 年同大学教授．情報数理応用とデータサイエンス，ならびにビジネスアナリティクスの研究に従事．著書に，『入門パターン認識と機械学習』コロナ社（2014），『ビジネス統計—統計基礎とエクセル分析』オデッセイコミュニケーションズ（2015）等．IEEE，INFORMS，電子情報通信学会，人工知能学会，日本経営工学会，日本オペレーションズリサーチ学会，経営情報学会等各会員．