Vol.2013-GN-87 No.2
2013/3/18

ショップ情報を用いた予測モデル構築法の検討

江崎　健司[1,a)]　石黒　勝彦[2,b)]　倉島　健[1,c)]　高屋　典子[1,d)]　内山　匡[1,e)]

,

概要：本論文では，ネットショッピングにおけるユーザの商品閲覧行動を説明するショップ情報を考慮したユーザモデルを提案する．提案モデルは，ユーザ自身の興味に基づいて，（1）ショップを選択した後，（2）そのショップが取り扱う商品の中から商品を選択するというユーザ行動に関する仮説に基づいている．従来技術が，過去の商品選択履歴のみからユーザの興味を推定するのに対して，提案手法は，商品そのものの選択に加えて，商品を取り扱うショップの選択も考慮することで，ユーザの閲覧商品と興味を高精度に推定可能な特徴がある．

キーワード：予測モデル，商品閲覧，EC サイト，ユーザモデル，ショップ情報

Esaki Kenji[1,a)]　Ishiguro Katsuhiko[2,b)]　Kurashima Takeshi[1,c)]　Takaya Noriko[1,d)]
Uchiyama Tadasu[1,e)]

,

1. はじめに

ネットでのショッピングにおけるユーザの商品閲覧行動のモデル化は，商品推薦や，トレンド解析，戦略立案支援を可能にする．例えば自社の顧客理解を目的として，商品閲覧行動ログから自社にはどんなユーザが来ているのか，ユーザに閲覧され易い商品は何かをモデルから分析することができる．一方で近年トピックモデルを用いた文書や購買ログなどの離散データ解析が注目されている．トピックモデルは階層ベイズモデルの一種であり，あるユーザが閲覧する商品は，ユーザ固有のトピック比率（興味を表す）に従ってあるトピックを選択した後，そのトピック固有の商品閲覧確率分布（流行を表す）に従って生成される，と仮定する．つまり，同じ興味を持つユーザは同じ商品を閲覧しやすいという仮定をおく．ネットでのショッピングにおけるユーザの商品閲覧行動ログにおけるユーザ・商品は，それぞれ，文書データにおける文書・単語に対応する．トピックモデルを用いてデータ解析を行う場合，興味が似ているユーザごとにユーザセグメントを作ることで自社の持つユーザセグメントごとに閲覧され易い商品は何かをモデルから分析し，ユーザセグメント別の推薦や戦略立案を行うことができる．

トピックモデルを用いた研究例は多々あり，トピックモデルを用いたジャンル可視化では視覚的なユーザの興味に基づく商品のセグメント分けを構築する研究 [2] などがある [1][3][4][5].

今回用いる商品閲覧行動ログはネットの発達により実際にサービスを展開しているネットのショッピングであれば蓄積されるものであり，基本的なデータであるといえる．さらに，ネットでのショッピングでは PC やスマホのブラウザの利用データからどのショップで商品を購入したのか

[1] NTT サービスエボリューション研究所
NTT Servise Evolution Laboratories
[2] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
[a)] esaki.kenji@lab.ntt.co.jp
[b)] ishiguro.katsuhiko@lab.ntt.co.jp
[c)] kurashima.takeshi@lab.ntt.co.jp
[d)] takaya.noriko@lab.ntt.co.jp
[e)] uchiyama.tadasu@lab.ntt.co.jp

というショップ情報を獲得することができるようになっている．トピックモデルに代表されるようにこれまで多くの行動モデルが提案されてきているが，商品の選択履歴のみの分析にとどまっている．しかし，実際のネットでのショッピングを考えると次の 2 つの理由から，ショップ情報を用いることの利点を説明できる．第一に，商品の選択に加え，ショップの選択も加味してユーザ固有の興味を推定することができる．ショップの選択にはユーザの興味が色濃く反映されていると考えられるからである．例えば，カジュアルなファッションに興味があるユーザとエレガントで高級なファッションに興味があるユーザでは選択するショップは異なると期待される．第二に，ショップ情報を考慮した行動モデルを構築することで，ショップごとに取り扱っている商品や在庫が異なるため，ショップごとに商品閲覧確率分布を得る事ができる．これによりショップ間の差異を商品閲覧確率分布の観点から明らかにすることができる．また，あるショップを訪れるユーザが他のどのショップに訪れているかといった，ユーザ興味に基づく競合の発見につなげることもできる．

本論文では，ネットでのショッピングにおけるユーザの商品閲覧行動を説明するショップ情報を考慮したユーザモデルを提案し，ユーザの興味や次の閲覧商品を高精度に推定可能なことを示す．提案モデルは，ユーザ固有の興味に基づいて（1）ショップを選択した後，（2）そのショップが取り扱う商品の中から商品を閲覧するというユーザ行動に関する仮説に基づいている。これにより，商品そのものの選択に加えて、商品を取り扱うショップの選択も考慮すること，ショップごとに商品閲覧確率分布（流行）が異なることも考慮することが可能になる．実験では実際の EC サイト閲覧データを用いて，ユーザが訪れたショップ情報がわかっている時に次の閲覧商品を高精度に推定可能なことを示し，ユーザの興味を高精度に推定可能な事を示す．

2. 準備

本論文の目的は，ユーザの商品閲覧行動モデルを表す，ショップ s においてユーザ u が商品 i を閲覧する確率 $P(i|u,s)$ を推定することとする．ここで$\boldsymbol{U} = \{u\}_{u=1}^{U}$ はユーザ集合，$\boldsymbol{I} = \{i\}_{i=1}^{I}$ は閲覧商品集合，U はユーザ数，I は商品数を表す．$\boldsymbol{S} = \{s\}_{s=1}^{S}$ はショップ集合，S はショップ数を表す．ショップ s においてユーザ u が商品 i を閲覧する確率 $P(i|u,s)$ 全てを推定すると膨大なパラメータ数になるが，トピックモデルを用いて潜在トピックを導入することにより，商品閲覧行動のモデルにおける推定すべきパラメータ数を大幅に削減することができる．トピックモデルでは，ショップ s においてユーザ u が商品 i を閲覧する確率を，ショップ s においてユーザ u があるトピック k に興味を持つ確率とそのトピック k でショップ s において商品 i が閲覧される確率の二つの要素に行列分解して計算する．具体的には，トピックモデルでは，潜在トピック k が与えられたときユーザ u と商品 i は条件付独立であると仮定し，確率 $P(i|u,s)$ を以下の式で表す．

$$P(i|u,s) = \sum_{k=1}^{K} \theta_{s,u,k}\phi_{s,k,i} \tag{1}$$

ここで K は潜在トピック数である．$\theta_{s,u,k} = P(k|u,s)$ はショップ s においてユーザ u がトピック k に興味がある確率を表し，ユーザの興味に相当する $(\theta_{s,u,k} \geq 0, \sum_k \theta_{s,u,k} = 1)$．また，$\phi_{s,,k,i} = P(i|k,s)$ はショップ s におけるトピック k の中で商品 i が閲覧される確率を表し，商品の流行に相当する $(\phi_{s,k,i} \geq 0, \sum_k \phi_{s,k,i} = 1)$．

3. 従来法

本章ではトピックモデルを従来法として紹介する．トピックモデルを用いたデータ解析を行うことで，ネットショッピングを行っているユーザのセグメント分けやある興味を持っているユーザはどのような商品を閲覧しやすいのかといった分析が可能になり、ユーザセグメント別の商品推薦や戦略立案支援が可能となる．トピックモデルは商品閲覧行動をモデル化しているため，そのモデル化の仮説も合わせて紹介する．

3.1 LDA

Latent Dirichlet Allocation（LDA）[1] は代表的なトピックモデルであり，図 1 にグラフィカルモデルを示す．ここで $x_{u,n}$ はユーザ u が n 番目に閲覧した商品であり，$\boldsymbol{I} = \{i\}_{i=1}^{I}$ の閲覧商品集合の中から選択される．また N はユーザが閲覧した商品数でありユーザ毎に異なる．LDA では興味ベクトル $\theta_u = \{\theta_{u,k}\}_{k=1}^{K}$，流行ベクトル $\theta_k = \{\phi_{k,i}\}_{i=1}^{I}$ は以下のディリクレ事前分布から生成されると仮定する．

$$P(\theta_u|\alpha) \propto \prod_{k=1}^{K} \theta_{u,k}^{\alpha_k - 1} \tag{2}$$

$$P(\phi_k|\beta) \propto \prod_{k=1}^{K} \phi_{k,i}^{\beta_i - 1} \tag{3}$$

ここで，α はトピック k のトピック選択確率の事前分布を表すパラメータ，β は商品 i の商品閲覧確率の事前分布を表すパラメータである．目的であるユーザの商品閲覧行動モデルを表す，ショップ s においてユーザ u が商品 i を閲覧する確率 $P(i|u,s)$ は下記式で推定される．

$$P(i|u,s) = \sum_{k=1}^{K} \theta_{u,k}\phi_{k,i} \tag{4}$$

LDA のモデルで表現される商品閲覧行動は，ユーザごとにユーザ固有の興味をもっており，その興味にあった商品を選択する，というものである．

Vol.2013-GN-87 No.2
2013/3/18

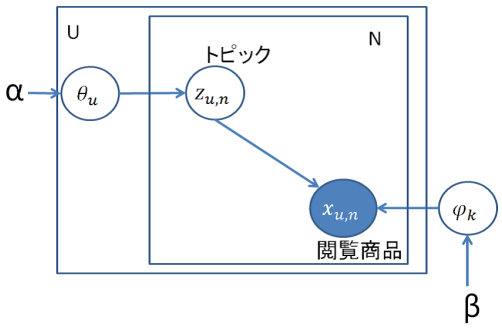

図 1 LDA のグラフィカルモデル

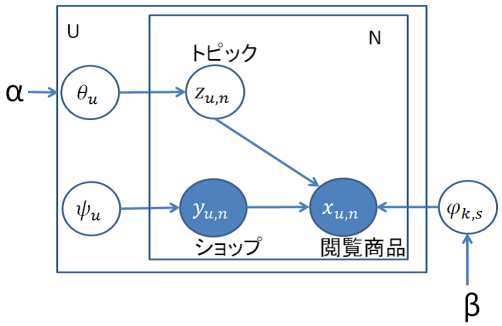

図 2 LDAwithShop のグラフィカルモデル

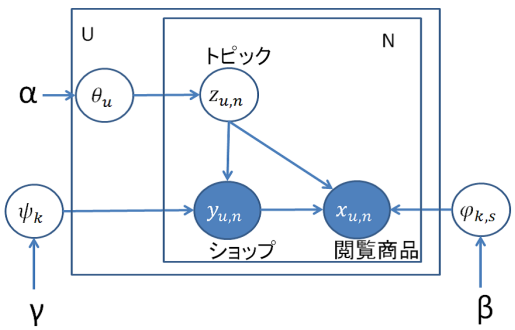

図 3 shopLDA のグラフィカルモデル

4. 提案法

本章では EC サイトにおけるショップ情報を用いたユーザモデルを 2 つ提案する．

4.1 LDAwithShop

LDAwithShop は，LDA をショップごとに学習するトピックモデルであり，図 2 にグラフィカルモデルを示す．ここで $y_{u,n}$ はユーザ u が n 番目に閲覧した商品を閲覧したショップであり，$\boldsymbol{S} = \{i\}_{i=1}^{S}$ のショップ集合の中から選択される．興味ベクトル $\theta_u = \{\theta_{u,k}\}_{k=1}^{K}$，ショップごとの流行ベクトル $\phi_{k,s} = \{\phi_{s,k,i}\}_{i=1}^{I}$ は以下のディリクレ事前分布から生成されると仮定する．

$$P(\theta_u|\alpha) \propto \prod_{k=1}^{K} \theta_{u,k}^{\alpha_k - 1} \tag{5}$$

$$P(\phi_{k,s}|\beta) \propto \prod_{k=1}^{K} \phi_{s,k,i}^{\beta_{i,s} - 1} \tag{6}$$

ここで，β はショップ s における商品 i の商品閲覧確率の事前分布を表すパラメータである．また，ユーザが訪れるショップの選択確率は $\psi_u = \{\psi_{u,s}\}_{s=1}^{S}$ とする．$\psi_{u,s} = P(s|u)$ はユーザ u がショップ s を選択する確率を表す $(\psi_{u,s} \geq 0, \sum_s \psi_{u,s} = 1)$．目的であるユーザの商品閲覧行動モデルを表す，ショップ s においてユーザ u が商品 i を閲覧する確率 $P(i|u,s)$ は下記式で推定される．

$$P(i|u,s) = \psi_{u,s} * \sum_{k=1}^{K} \theta_{u,k}\phi_{s,k,i} \tag{7}$$

LDAwithShop のモデルで表現される商品閲覧行動は，ユーザはユーザ固有の興味 θ_u とショップを選択する確率確率 ψ_u を持っており，ショップが取り扱う商品の中からその興味にあった商品を $\phi_{k,s}$ に従って選択する，というものである．LDAwithShop では在庫や取扱い商品が異なるためにショップごとに異なる流行を持つ事を正しく推定することができる．これにより，ユーザが訪れたショップごとに異なる閲覧しやすい商品を得る事ができ，それに基づく推薦や戦略立案が可能になる．

4.2 ShopLDA

ShopLDA は，ユーザ固有の興味に基づいて（1）ショップを選択した後，（2）そのショップが取り扱う商品の中から商品を選択するというユーザ行動に関する仮説に基づいている．これにより，商品そのものの選択に加えて、商品を取り扱うショップの選択も考慮すること，ショップごとに商品閲覧確率分布（流行）が異なることも考慮することが可能になる．図 3 にグラフィカルモデルを示す．興味ベクトル $\theta_u = \{\theta_{u,k}\}_{k=1}^{K}$，ショップごとの流行ベクトル $\phi_{k,s} = \{\phi_{s,k,i}\}_{i=1}^{I}$ は以下のディリクレ事前分布から生成さ

れると仮定する．

$$P(\theta_u|\alpha) \propto \prod_{k=1}^{K} \theta_{u,k}^{\alpha_k - 1} \tag{8}$$

$$P(\psi_k|\gamma) \propto \prod_{k=1}^{K} \psi_{s,k}^{\gamma_s - 1} \tag{9}$$

$$P(\phi_{k,s}|\beta) \propto \prod_{k=1}^{K} \phi_{s,k,i}^{\beta_{i,s} - 1} \tag{10}$$

ここで，γ はショップ s のショップ選択確率の事前分布を表すパラメータである．ここで ShopLDA では，トピック k の中で訪れるショップ s の選択確率は $\psi_k = \{\psi_{k,s}\}_{s=1}^{S}$ とする．$\psi_{k,s} = P(s|k)$ はトピック k でショップ s が選択される確率を表す $(\psi_{k,s} \geq 0, \sum_s \psi_{k,s} = 1)$．目的であるユーザの商品閲覧行動モデルを表す，ショップ s においてユーザ u が商品 i を閲覧する確率 $P(i|u,s)$ は下記式で推定される．

$$P(i|u,s) = \sum_{k=1}^{K} \theta_{u,k}\psi_{k,s}\phi_{s,k,i} \tag{11}$$

ShopLDA のモデルで表現される商品閲覧行動は，ユーザ固有の興味に基づいて（1）ショップを選択した後，（2）そのショップが取り扱う商品の中から商品を選択する，というものである．ShopLDA では，ユーザ固有の興味ベクトル θ_u に基づいてあるショップ s をショップ選択確率 ψ_k で選択したということを考慮することで，そのショップが選択され易いトピック k をユーザは持っていたと推定することができる．LDAwithShop では，ユーザ u がショップ s を選択する確率はトピック k に依存しないため，閲覧した商品 i のみからトピック k を推定しなければならない．閲覧した商品 i に加え，選択したショップ s も用いる ShopLDA の方がユーザ u の興味を高精度に予測することが期待される．また，shopLDA では ψ_k に基づいてユーザの興味に基づくショップのセグメント分けをすることで，同じ興味を持つユーザが訪問する競合ショップの分析が可能になる．さらに，shopLDA ではユーザが訪問していないショップに対してもユーザの興味に基づき選択確率を算出できることから，ユーザ u が訪問していないショップ s で閲覧する商品 i を予測することが可能になる．

5. 定量評価実験

ブラウザの利用履歴をもとに EC サイトにおける商品閲覧行動ログを抽出した．本研究で用いる EC サイトにおける商品閲覧行動ログは約 800 名に関するものであり，対象コマース数は 150 サイトである．この抽出した閲覧行動ログを用いて提案モデルの有効性を検証した．具体的には，各ユーザの 2011 年 9 月から 10 月までに観測された商品閲覧行動をもとに，各ユーザが 2011 年 11 月に一番最初に閲覧した商品 $x_u = i$ を予測し，その予測精度に基づいてモデルの妥当性を評価する．学習データとして用いたのは 2011 年 9 月から 10 月までに観測された全ユーザの閲覧行動ログである．ただし，今回の予測実験においては，正解商品（2011 年 11 月に一番最初に閲覧した商品）のブランドと，モデルが予測した商品のブランドが一致した場合に，予測が成功したと判断した．比較した手法は以下の 4 つである．

表 1 各手法ごとの予測精度（ショップ情報が未知の場合）
Table 1 Prediction accuracy of product viewing.

手法	予測精度
個人の過去履歴	0.08
LDA	0.04
AuthorTopic	0.01
LDAwithShop	0.04
shopLDA	0.03

- 個人の過去履歴に基づく予測モデル: ユーザが過去に閲覧した回数の多い商品から順に提示する手法
- LDA
- LDAwithShop（提案法 1）
- shopLDA（提案法 2）

評価尺度として N ベスト正解率を用いた．N ベスト正解率は，閲覧確率 $P(x_u = i|u, \theta_u, \phi_{s,k})$ の値のトップ N 個の商品の中に，実際にユーザが選択した商品が含まれた場合に 1 を，含まれない場合に 0 を出力する指標である．最終的に，全ユーザの中で 1 が出力された割合を算出した．なお，本実験では，予測対象の商品が一つであるためN＝1 を用いている．表 1 に各手法のNベスト正解率を示す．結果として，個人の過去履歴を用いたものが最も精度がよかったことがわかる．次に精度がよかったのは LDA と LDAwithShop であった．今回の実験でのブランドの異なり数は 2887 ブランドであるが，個人の平均的閲覧ブランド数は 5.7 と少なくまた再訪問率も 4 割と高かったことが原因であると考える．

提案手法の重要な特徴は，ショップ情報を有効に用いてユーザの行動を予測可能なことである．ユーザは，実際の商品閲覧，あるいは購買行動を起こす前に，まず，ショップを訪問する．提案法が，ショップを訪れたそのタイミングで，高精度にユーザの行動予測が可能なことを示すために，x_u を閲覧したショップ $y_u = s$ は与えられているという条件のもとでの予測評価を行った．次式のようにショップ s でユーザ u が商品 i を選択する確率を計算する．

$$P(x_u = i|u, y_u = s, \theta_u, \phi_{k,s}) = \sum_{k=1}^{K} \phi_{u,k}\theta_{s,k,i} \tag{12}$$

今回も評価尺度として N ベスト正解率を用いた．今回の N ベスト正解率は，閲覧確率 $P(x_u = i|u, y_u = s, \theta_{u,s}, \phi_{k,s})$

表 2 各手法ごとの予測精度
Table 2 Prediction accuracy of product viewing given consideration of shop information.

手法	予測精度
個人の過去履歴	0.17
全体の人気	0.17
ショップの人気	0.18
LDA	0.17
AuthorTopic	0.18
LDAwithShop	0.21
shopLDA	0.19

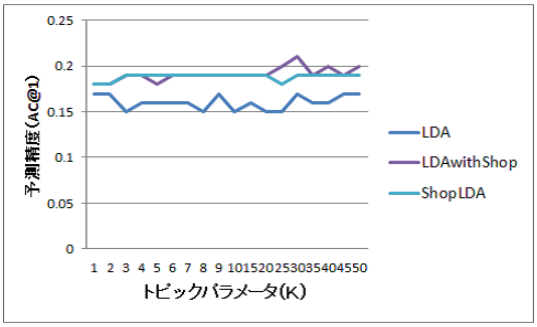

図 4 各手法の予測精度（トピック数 1-50）

の値のトップ N 個の商品の中に，実際に閲覧された商品が含まれている割合を表す．前の実験と同じ 4 手法に加え，全ショップを通じて人気商品から順に提示する手法（全体の人気）と，与えられたショップ内での人気商品から順に提示する手法（ショップ内での人気）も比較手法として追加した．なお，各手法の出力に与えられたショップ s で取り扱わない商品が含まれた場合があるが，それを予測値として出力しないようにしている．

表 2 に各手法の N ベスト正答率を示す．LDA，LDAwithShop，ShopLDA のパラメータであるトピック数を 1 から 50 まで変化させ，その中で予測精度が最大の値を記載している．また，LDA，LDAwithShop，ShopLDA のトピックパラメータと精度との関係性を図 4 に示す．表 2 に示すように，提案法である LDAwithShop と shopLDA が比較手法を上回る予測精度を達成した．つまり，ショップごとの商品閲覧確率分布を推定することにより，ショップが与えられた条件のもとで提案法が高精度に行動予測できることを示せた．LDAwithShop の方が ShopLDA よりも予測精度が高かった理由として，ShopLDA が推定すべきパラメータ（トピック依存のショップ選択確率）が増えたため，うまく学習出来なかったことが挙げられる．今後，学習データを増やすなどして，検討を進める余地がある．

閲覧ユーザ数が多かった上位 5 件のショップについて，ショップ依存の潜在トピック分布を表 3 に示す．今回は ShopLDA の K=5 の場合を分析した．また，閲覧ユーザ数の多いショップから順に A，B，C，D，E としている．これを見ると，トピック 2 に所属するショップが多いもの

表 3 メジャー EC サイトの所属トピック（shopLDA@K=5）
Table 3 topic distribution of top 7 shops (shopLDA@K=5).

EC サイト	所属トピック
ショップ A	1,2
ショップ B	1
ショップ C	5
ショップ D	2
ショップ E	2
ショップ F	1
ショップ G	2

の，よく閲覧されるショップはそれぞれバラバラのトピックに所属することがわかる．ショップ D について競合分析すると閲覧数が多いショップの中ではショップ E，G が競合であり，他の閲覧数が上位のショップとは競合していないことがわかる．商品が異なるため分布が異なることもあるが，在庫数も異なるため閲覧商品確率分布も異なっていた．これにより同一のブランドでもショップが異なると流行が異なることがモデル化できるため表 2 のようにショップごとに商品閲覧確率分布が異なる方が予測精度がでたと考えられる．

6. むすび

本論文では，EC サイトにおけるユーザの商品閲覧行動を説明するショップ情報を考慮したユーザモデルを提案し，ユーザの興味や次の閲覧商品を高精度に推定可能なことを示した．提案モデルは，ユーザ固有の興味に基づいて（1）ショップを選択した後，（2）そのショップが取り扱う商品の中からて商品を選択するというユーザ行動に関する仮説に基づいている。これにより，商品そのものの選択に加えて、商品を取り扱うショップの選択も考慮すること，ショップごとに商品閲覧確率分布（流行）が異なることも考慮することが可能になった．実際の EC サイト閲覧データを用いて次の閲覧商品を高精度に推定可能なことを示し，ユーザの興味を高精度に推定可能な事を示した．

参考文献

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan.: *Latent dirichlet allocation*, the Journal of machine Learning research (2003).
[2] 岩田具治，山田武士，上田修功: トピックモデルに基づく文書群の可視化, 情報処理学会論文誌 Vol.50 No.6 (2009).
[3] T.Hofmann: *Probabilistic latent semantic analysis*, UAI (1999).
[4] T.Hofmann: *Collaborative filtering via gaussian probabilistic latent semantic analysis*, SIGIR (2003).
[5] Mark Steyvers,Padhraic Smyth,Thomas Grifths.: *Probabilistic Author-Topic Models for Information Discovery*, KDD (2004).