Hidden Topic Markov Models に基づく顧客購買行動分析に関する一考察

情報数理応用研究　　5218C033-8　保戸田未桜
指導教員　後藤正幸

A Study on Customer Purchase Behavior Analysis Based on Hidden Topic Markov Models

HOTODA Mio

1. はじめに

近年，EC サイトを通じたオンラインでの商品購買は広く一般消費者に受け入れられるようになり，市場規模が拡大している [1]．EC サイト上では，ユーザの各ページの閲覧行動の詳細なログが取得可能であるため，これらのデータを活用した Web マーケティング技術の重要性が高まっている．一方，EC サイトにおける購買に至る割合（Conversion Rate ; CVR）は通常，高々数%であることが知られている [2]．そのため，多くの EC サイトで CVR を改善するための施策が必要とされている．例えば，ユーザの購買意欲を把握することで効果的なタイミングで施策を打ったり，購買につながりやすいページを把握することでそのページにユーザを誘導したりすることにより，CVR の向上が見込める．

一般に，ユーザは商品を購入する前に EC サイト上の様々なページを閲覧することが多い．したがって，ユーザの思考状態（購買意欲の有無や嗜好，ニーズ等）の変化は EC サイトのページ遷移傾向に反映されていると考えられる．そのため，閲覧履歴データを分析することで，ユーザの購買意欲を抽出できれば，適切な施策の一助になると考えられる．この様な分析においては，ユーザの思考状態に依存して各閲覧ページが生成されたと仮定することは極めて自然であり，観測不可能な思考状態を観測可能な閲覧ページから推測できると考えられる．そこで本研究では，ユーザのページ遷移の背後に潜在トピック（潜在クラス）を仮定するモデルを提案する．

従来の潜在クラスモデルをページ遷移データに適用すると，連続的に閲覧した一連のページ全体に同じ潜在トピックを仮定するか，または各ページに対し毎回異なる潜在トピックを仮定するかのどちらかである．しかし，閲覧中にユーザの思考状態が変化する可能性は十分に考えられるため，連続的に閲覧した一連のページ全体に同じ潜在トピックを仮定することは好ましくない．一方で，ページごとにユーザの思考状態が頻繁に変わることもほとんどないと考えられるため，毎ページ異なる潜在トピックを仮定したトピックモデルの適用も好ましくない．

そこで本研究では， Hidden Topic Markov Models（HTMM）[3] をベースとし，リアルタイムに閲覧履歴データの分析を可能にした購買行動分析モデルを提案する．HTMM は，前後関係を考慮しない Latent Dirichlet Allocation（LDA）[4] に隠れマルコフモデル（Hidden Markov Chain model ; HMM）の考え方を援用したモデルである．HTMM は文書に対して複数の潜在トピックを想定するが，LDA とは異なり，文書中の文単位での潜在トピックを考える．すなわち，各文中の単語は同じ潜在トピックに存し，かつ連続する文は同じ潜在トピックを持つ可能性が高いことを仮定する．このため，文書を同じ統計的特徴を持つ複数の文群に分割することができる．したがって HTMM を閲覧履歴データに援用し，ページ単位での潜在トピックの推定を行うことで，閲覧履歴を同じ特徴を持つ複数の群に分割することができる．そのうえで実際の閲覧履歴データから得られる購買実績を併用することで，各潜在トピックの購買確率を求めることができる．提案モデルにより，ユーザの購買意思をリアルタイムで予測することが可能になると考えられる．加えて，各ユーザの潜在トピックの変化点を検出することが可能になり，ユーザの思考が変化する際にどのようなページ遷移が起こっているかを理解することができる．本研究では，実閲覧履歴データを用いて分析し，提案手法の有効性を検証する．

2. 関連研究

2.1. 隠れマルコフモデル

Hidden Markov Model （HMM）は，観測不可能なマルコフ過程とその各状態に依存して生成されるシンボルの組み合わせによって，シンボルの系列を表現するモデルである．HMM の対象となる系列データは，複数の状態を持ち，それらの状態がある遷移確率により遷移するマルコフモデルと仮定される．これを一次マルコフ性という．

2.2. Hidden Topic Markov Model

Hidden Topic Markov Model（HTMM）は，LDA に HMM の考え方を援用した，文書生成モデルである．図 1 に HTMM のグラフィカルモデルを示す．

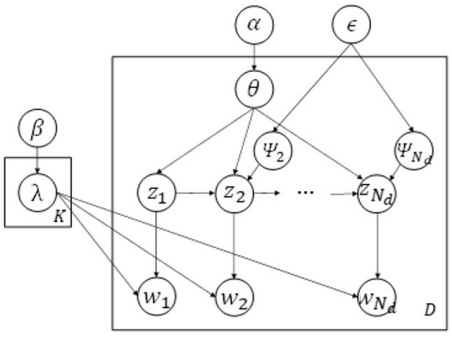

図 1: HTMM のグラフィカルモデル

K はトピック数，D は文書数，N_d は文書 d 内の単語数を表わしている．HTMM は LDA 同様，各文書 d はト

ピック分布 θ_d を持ち，文書内の全単語について，トピック分布 θ_d に従ってまずトピック $z_{d,n}$ が選ばれ，そのトピック $z_{d,n}$ に対応する単語分布 λ に従って単語 w が生成される．トピック分布 θ_d は文書ごとに生成され，単語分布 λ はトピックごとに生成される．また，α と β はハイパーパラメータであり，それぞれトピック分布 θ が従うディリクレ分布のパラメータ，単語分布 λ が従うディリクレ分布のパラメータを示す．LDA では，各文書 d におけるトピック分布 θ_d から生成された単語トピック z は互いに独立であるが，HTMM では一次マルコフ性を仮定する．図 1 に示すように，HTMM では各トピック $z_{d,n}$ がトピック分布 θ_d とトピック遷移変数 $\psi_{d,n}$ に依存する遷移確率を持つマルコフ連鎖を形成している．トピック遷移変数 $\psi_{d,n}$ は式 (1) で表現され，各単語が 1 つ前の単語トピックを引き継ぐか否かを表すパラメータである．文頭の単語に対してはハイパーパラメータ ϵ に従う二項分布により $\psi_{d,n}$ が決定され，文内の単語に対しては $\psi_{d,n} = 0$ となる．

$$\begin{cases} z_{d,n} = z_{d,n-1} & (\psi_{d,n} = 0) \\ z_{d,n} \sim multinomial(\theta_d) & (\psi_{d,n} = 1) \end{cases} \tag{1}$$

HTMM のパラメータのうち，実際に観側される変数は文書上に現れている単語 w のみである．そのためその他の潜在変数やパラメータについては Expectation-Maximization Algorithm (EM) を使用して推定する．EM は現在の変数を用いて事後分布を計算する操作（E-step）と事後分布を最大化するように変数を更新する操作（M-step）を繰り返すことにより観測できない変数を探索的に推定する手法である． HTMM の場合，求めるべき変数は θ_d，λ と ϵ である．θ と ϵ が推定された後，それに応じて遷移行列が更新される．HTMM にこおける EM アルゴリズムの各ステップの更新式は以下の式 (2)–(7) で計算される．

E-step :

$$Pr(z_n, \psi_n \mid d, w_1 \cdots w_{N_d} \ ; \ \theta_d, \lambda, \epsilon) \tag{2}$$

ただし，式 (2) は M-step から出力された θ_d，λ と ϵ に対し Forward-backward Algorithm を用いて算出する．

M-step :

$$\theta_{d,z} \propto E(C_{d,z}) + \alpha - 1 \tag{3}$$

$$\lambda_{z,w} \propto E(C^{z,w}) + \beta - 1 \tag{4}$$

$$\epsilon = \frac{\sum_d \sum_{n=2}^{N_d} Pr(\psi_{d,n} = 1 \mid w_1 \cdots w_{N_d})}{\sum_d (N_d^s - 1)} \tag{5}$$

ただし，$\theta_{d,z}$ は正規化されており，θ_d は分布，N_d^s は文書 d 内の文数である．また，$C_{d,z}$ は文書 d の θ_d に従ってトピック z が選ばれた回数，$C^{z,w}$ は，単語 w が $\lambda_{z,w}$ に従ってトピック z から選ばれた回数を示す．

$$E(C_{d,z}) = \sum_{n=1}^{N} Pr(z_{d,n} = z, \psi_{d,n} = 1 \mid w_1 \cdots w_{N_d}) \tag{6}$$

$$E(C^{z,w}) = \sum_{d=1}^{D} \sum_{n=1}^{N} Pr(z_{d,n} = z, w_{d,n} = w \mid w_1 \cdots w_{N_d}) \tag{7}$$

3. 提案モデル

3.1. 着想・モデル化

従来，閲覧履歴データに潜在クラスモデルを適用する際は，1 つのセッションまたは 1 人のユーザに対し 1 つの潜在トピックを仮定していた [5]．これは，ユーザはセッション中，嗜好や心情が変化しないということを意味する．しかし，閲覧中にユーザの嗜好や心情が変化することは一般的に起こり得る．このような思考状態の変化をトピックの変化で表現するために，潜在トピックに一次マルコフ性を仮定し，前後関係を考慮することを考える．この場合，現在の潜在トピック z_t は一時刻前の潜在トピック z_{t-1} と遷移確率によって毎回決定される．しかし，閲覧中にユーザの嗜好や心情が毎回変化するとは考えにくく，遷移確率とは独立に潜在トピックが遷移しない場合を考慮する必要がある，そこで，潜在トピックの遷移に制約を設けた HTMM を EC サイトの閲覧履歴データに適用することを考える．

従来 HTMM は文書データに適用されるモデルであるが，文書データと閲覧履歴データには類似した特徴があると考えられるため適当である．例えば，文書内の各単語はその文書の趣旨や話題を反映していると考えられる．同様に一連の閲覧履歴内の各ページは閲覧ユーザの思考を反映していると考えられる．HTMM は「状態変化が起こるか否か」を表すパラメータを導入することで，連続する文は同じトピックを持つ可能性が高いことを仮定しているが，必ずしも状態変化が起こるわけではなく，前後関係を考慮した閲覧履歴データの分析に適していると考えられる．

一方，HTMM では同一文内の単語は同一トピックに属することを仮定している．すなわち，文の先頭単語以外では $\psi_n = 0$ という制約がある．文書データと比較し，閲覧履歴データでは「文書」に相当するユーザという単位,「単語」に相当するページという単位は存在するものの,「文」という単位は存在しない．また文書データの場合,「文書」内の「文」ごとに潜在トピックを求めることで「文」単位での潜在トピックの遷移を把握することができる．対して閲覧履歴データに関しては，EC サイト側ではできるだけリアルタイムにユーザの購買確率を知りたいというニーズがあり，分析単位をできるだけ細かくし，正確な潜在トピックを把握することが求められているため，ページ単位で潜在トピックを推定するモデルは適当だと考えられる．そのため，先述の HTMM の ψ_n に関する「文の先頭単語以外では $\psi_n = 0$」という制約を取り除き，ページ単位で潜在トピックを推定するモデルを提案する．すなわち本提案手法では，LDA の潜在トピックに一次マルコフ性を仮定し，ページ毎にパラメータ ϵ を用いて ψ_n が決定されることでページごとの潜在トピックの推定に対し「1 つ前の潜在トピック z_{n-1} と同じ潜在トピックになりやすい」という制約を設ける．この制約のため，毎ページごとに潜在トピックの遷移が生起しないモデル化が可能になる．

3.2. 購買分析方法

潜在トピックを推定した後，ユーザの購買完了ページの閲覧データを用いて，各潜在トピックにおける購買率を求める．

式 (8) に購買率の計算方法を示す．S_z は潜在トピック z のページ数，C_z は潜在トピック z において閲覧後 10 ページ以内に購買の起こったページ数を示す．これにより，購買に至りやすい潜在トピックを発見することができ，閲覧途中でも潜在トピックから，ユーザが購買に至りやすいか否か予測することができると考えられる．

$$\text{購買率} = \frac{C_z}{S_z} \tag{8}$$

加えて，潜在トピックの変化点や潜在トピック間の遷移割合を知ることで，ページ遷移とユーザの思考の変化の関係性を理解することができ，ユーザを購買率の高い潜在トピックに誘導することができる可能性がある．

4. 実データ分析

提案手法の有効性を検証するために，実際の閲覧履歴データを使用した分析を行う．これにより，抽出された特徴を明らかにし考察を行う．

本分析は，ユーザのインターネット上での行動ログ分析を行う企業である株式会社ヴァリューズが保有する Web ページの閲覧履歴データを対象とした．データ収集期間は 2017 年 8 月 1 日から 2017 年 10 月 31 日の 3ヵ月間であり，この期間に楽天市場のサイトを閲覧したユーザ 766 人，総セッション数 35,958 件の楽天市場に関する閲覧履歴データを分析対象とする．各潜在トピックにおける購買率は式 (8) を用いて計算する．また，ハイパーパラメータは $\alpha = 1.001$，$\beta = 1.0001$，トピック数は $K = 8$ とした．

4.1. 購買率について

各潜在トピックにおける購買率を図 2 に示す．ただし，潜在トピックは購買率の大きい順に示している．図中の線は閲覧履歴データ全体の平均購買率を表す．

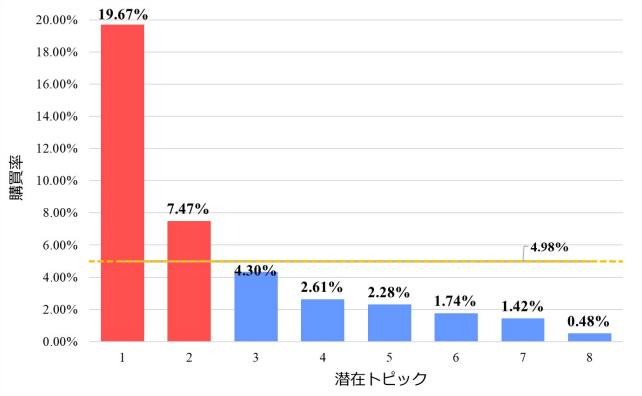

図 2: 各潜在トピックにおける購買率

図 2 より，潜在トピックによって購買率に大きな差があることがわかった．すなわち，提案手法で閲覧中のユーザの潜在トピックを知ることによって，購買に至りやすいユーザか否かを予測することができると考えられる．例えば，あるユーザがトピック 1 と推定された場合，このユーザは他の潜在トピックのユーザに比べて閲覧 10 ページ以内に購買に至る可能性が高いが，必ず購買に至るわけではないため CVR 向上施策の対象になると考えられる．逆に，トピック 8 と推定された場合は，このユーザを施策対象にすることは非効率であると考えられる．

また，購買率の高い潜在トピック 1,2 と購買率の低い潜在トピック 8 に含まれるページ種別のうち，全体のページ種別割合と比較して各潜在トピックに含まれている割合が大きいるページ種別上位 10 件について表 1 に示す．

表 1: 各潜在トピックに含まれているページ種別上位 10 件

（i）潜在トピック 1

ページ	all	topic1	差分
カートページ	3.71%	15.45%	11.74%
注文操作ページ	2.51%	12.82%	10.31%
注文完了ページ	0.68%	3.84%	3.16%
商品詳細ページ	33.10%	34.77%	1.67%
商品画像ページ	0.19%	0.77%	0.58%
レビューページ	2.24%	2.79%	0.55%
検索ページ	1.94%	2.32%	0.38%
価格比較ページ	0.93%	1.10%	0.16%
イベントページ	0.27%	0.40%	0.13%
商品サムネイルページ	0.03%	0.04%	0.01%

（ii）潜在トピック 2

ページ	all	topic2	差分
検索ページ	14.51%	23.71%	9.20%
商品詳細ページ	33.10%	41.43%	8.34%
メルマガ関連ページ	0.41%	6.31%	5.90%
メルマガ関連ページ	0.22%	3.31%	3.09%
レビューページ	0.16%	2.45%	2.29%
キャンペーンページ	0.13%	1.79%	1.67%
検索ページ	1.94%	3.10%	1.15%
レビューページ	2.24%	2.76%	0.52%
商品画像ページ	0.05%	0.36%	0.31%
カテゴリーページ	0.42%	0.67%	0.25%

（iii）潜在トピック 8

ページ	all	topic8	差分
商品詳細ページ	33.10%	56.75%	23.66%
検索ページ	14.51%	18.21%	3.70%
ショップページ	2.55%	6.01%	3.46%
検索ページ	0.39%	3.52%	3.13%
レビューページ	2.24%	3.79%	1.55%
ショップページ	2.25%	2.87%	0.62%
カード関連ページ	0.11%	0.30%	0.19%
商品画像ページ	0.05%	0.16%	0.11%
イベントページ	0.27%	0.36%	0.09%
FAQ ページ	0.12%	0.20%	0.08%

表 1（i）は最も購買率の高い潜在トピック 1 について示しているが，やはり購買の際に必ず通るカートページ（購入の際に選んだ商品をまとめておくページ）や注文操作ページ（支払方法や配送先などを入力するページ）の割合が高かった．あわせて，商品画像ページやレビューページ，価格比較ページが多いことも特徴として挙げられる．ここから，商品に関して詳しく調べていることが伺え，この段階に入ると直後に購買に至る場合が多いことがわかる．表 1（ii）は次いで購買率の高い潜在トピック 2 について示しており，潜在トピック 1 に比べてメールマガジンに関するページの割合が増えることがわかる．ここから，メールマガジンを購読するようなユーザは優良ユーザだと考えられ，他のユーザより購買しやすい傾向があると考えられる．表 1（iii）は最も購買

率の低い潜在トピック 8 について示しており，潜在トピック 1 と同様に商品を探している様子（商品を検索，ショップページや商品詳細ページを閲覧している様子）が伺える．ここからユーザは商品を探している状態が長く続くため，商品を探しているからと言ってすぐに購買につながるわけではないことがわかる．潜在トピック 1 の結果からもわかるように商品画像等，商品そのものに対する深い興味が無いと購買につながらないと考えられる．また，他の潜在トピックには見られないポイントカードや FAQ に関するページの割合も高いことがわかる．これらのページは直接購買に関係しないものであるため，ショッピング以外の目的で EC サイトを利用している可能性もあると考えられる．

次に現在の潜在トピックが前の潜在トピックを引き継いだものである場合（継承）とそうでない場合（選択）の各潜在トピックにおける購買率について考察する．結果より“選択”でも“継承”でも購買率の高い潜在トピックは同じであることがわかった．しかし，“選択”よりも“継承”での購買率が高いトピックやその逆のトピックも存在した．すなわち，潜在トピックによって施策を行うべきタイミングが異なると考えられる．

4.2. 潜在トピックの遷移について

期間中，終始同一の潜在トピックであったユーザもいたが，データ収集期間中に閲覧されたページ数は平均で約 249 ページと多いため，潜在トピックの遷移がよく見られた．潜在トピックが変化する場合のトピック遷移割合を表 2 に示す．各潜在トピックの混合率より大きな遷移割合を太字で示す．

表 2: 潜在トピックが変化する場合のトピック遷移割合 (%)

		To							
		1	2	3	4	5	6	7	8
From	1		**8.96**	**37.17**	**10.98**	7.80	**9.19**	**11.91**	13.99
	2	**17.33**		**29.07**	**8.60**	**8.60**	8.37	6.98	**21.05**
	3	**22.56**	**11.74**		**8.57**	**10.93**	**10.93**	**8.90**	**26.37**
	4	**24.08**	7.63	24.55		5.48	7.75	4.77	**25.74**
	5	13.28	**9.69**	**32.06**	7.06		8.85	5.86	**23.21**
	6	**16.84**	6.63	**25.79**	7.79	**8.21**		5.58	**29.16**
	7	**28.49**	5.43	**30.53**	5.29	6.11	5.70		18.45
	8	15.06	7.55	**40.99**	**9.61**	**8.51**	**12.52**	5.76	
混合比		16.37	8.19	24.76	8.05	7.91	8.95	7.18	18.57

表 2 のトピック 7 と 8 に着目すると，両方とも購買率が低いトピックであるが，トピック 7 では購買率の高いトピック 1 に遷移する確率が高いのに対して，トピック 8 では購買率が中程度のトピックへの遷移が大半であることがわかる．そのため，潜在トピック間の遷移確率に着目することで，現状のユーザの購買意欲だけでなく，今後のユーザの思考状態の変化をある程度推定することも可能であると考えられる．

また，どのようなページで潜在トピックの遷移が起きることが多いのか調べた．その結果，次の特徴が見られた．

・EC サイト自体や各ショップのトップページ
　→ユーザが思考をリセットした
・商品詳細ページ
　→商品を探す段階から商品を絞る段階に移った
・検索ページ
　→ショッピング以外の利用（ポイント機能など）からショッピング利用に替わった

これらのページは閲覧履歴データ内に多く見られるページであるが，あわせてユーザの思考状態の変化を反映したページだと考えられる．加えてこれらのページの詳細や前後のページ，また潜在トピックとの関係を分析することによってユーザのページ遷移傾向を理解できる可能性がある．

5. 考察

提案モデルを EC サイトの施策対象ユーザ決定の支援に用いる際には，リアルタイムに閲覧ユーザに対して施策を行うことが効果的か否かを判断する必要がある．本提案モデルは，モデルの学習によってあらかじめ潜在トピックごとに購買率を算出し，購買に至りやすい潜在トピックであるか否かを定めておく．これにより，ユーザが閲覧したページからユーザがどの潜在トピックに所属するのかをリアルタイムに推定することで，施策対象ユーザを特定するための一助になると考えられる．実際，実データを用いた分析から，潜在トピックごとに購買率に大きな差があり，潜在トピックを推定することの有用性を示した．さらに，潜在トピックごとに含まれるページ種別が異なること，また潜在トピック間の遷移には偏りがあること，潜在トピックの遷移が起きる際に閲覧されやすいページを発見した．これにより，有効な施策内容決定の一助になると考えられる．例えば，表 1（ii）からメールマガジンは有効な施策であることがわかる．

6. まとめと今後の課題

本研究の目的は，EC サイトにおいて購買に至りやすいユーザを発見することで，EC サイトでの CVR 向上のための施策検討を支援することである．そのため，HTMM をユーザの閲覧履歴データに適用し購買実績と併用することで，ユーザの購買意欲を予測する手法を提案した．ユーザの潜在トピックと各潜在トピックの購買率から施策のターゲットユーザを特定したり，潜在トピックの遷移のしやすさを分析したりすることは，CVR を改善するための具体的な施策に役立つと考えられる．今後の課題として，最適な潜在トピック数決定などによる提案手法の改善が挙げられる．

参考文献

[1] 経済産業省 商務情報政策局 情報経済課，“平成 30 年度我が国におけるデータ駆動型社会に係る基盤整備（電子商取引に関する市場調査），” 2019.

[2] カイロスマーケティング株式会社，“コンバージョン率の目安と業界別平均値，” https://blog.kairosmarketing.net/contentmarketing/conversion-average-140320/, 2014 年 3 月 20 日, 最終閲覧日 2019 年 9 月 19 日.

[3] Gruber Amit, Yair Weiss, and Michal Rosen-Zvi, “Hidden topic markov models,” *Artificial intelligence and statistics*, pp.163–170, 2007.

[4] Blei David M., Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, pp.993–1022, 2003.

[5] 松嵜祐樹，三川健太，後藤正幸，“マルコフ潜在トピックモデルに基づく 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.