トピックの階層性を考慮した購買行動分析モデルに関する研究

1X17C117-5　松岡佑以
指導教員　後藤正幸

1. 研究背景と目的

近年 EC サイトの普及に伴い，蓄積された購買履歴データに含まれているユーザの嗜好を分析し，マーケティング施策に活用することは企業にとって重要な課題となっている．購買履歴データからユーザの嗜好の分析を行う手法としてトピックモデルが知られており，代表的な手法に Latent Dirichlet Allocation(LDA)[1] がある．LDA はユーザの購買行動の背後に潜在クラス（トピック）を仮定しており，得られたトピックからユーザの嗜好が解釈可能である．

一方，ユーザの嗜好には階層性が存在していると考えられ，LDA の拡張手法としてトピックに階層性を仮定した hierarchical Latent Dirichlet Allocation (hLDA)[2] や Pachinko Allocation Model(PAM)[3] が提案されている．hLDA はトピックの階層性が木構造で表現される分岐構造に限定されるため，ある下位トピックを基準として複数の上位トピックとの関係性を分析することができない．一方で PAM は，全ての上位トピックと下位トピックがネットワークで結合されており，複数のトピック間の関係性が分析可能な表現能力の高いモデルである．しかし PAM は初期値に影響を受けやすく，安定したトピックを得ることが難しいため，トピックの解釈が不安定になる．このような不安定な結果の解釈からビジネス上の意思決定を行うことは危険である．

そこで本研究では，ユーザの購買行動に基づいて推定されるトピックの階層性ではなく，EC サイト側がアイテムの管理のために付与している「カテゴリ」という階層性を有した情報を活用することで，施策の検討に有用であり，かつ階層的なトピック分析が可能な手法を提案する．具体的には，初期値に影響を受けにくい LDA に 2 つの粒度の異なるカテゴリの購買履歴データを適用し，得られたトピック間の関係性を表すトピック関連度スコアを算出することで，ユーザの嗜好の階層性を表現したモデルを示す．最後に，実際の EC サイトの評価履歴データに提案手法を適用してユーザの嗜好を分析し，有用性を示す．

2. Latent Dirichlet Allocation(LDA)

購買履歴データに LDA を適用することで，ユーザに潜在クラス（トピック）の出現確率分布，そのトピックごとにアイテムの生起確率分布が仮定され，トピックからユーザの嗜好を分析することが可能である．

ユーザ集合を $\mathcal{U} = \{1, \ldots, U\}$，アイテム集合を $\mathcal{X} = \{1, \ldots, X\}$，$K$ 個の潜在トピック集合 $\mathcal{K} = \{1, \ldots, K\}$ とする．また，ユーザ $u \in \mathcal{U}$ のトピック分布を $\boldsymbol{\theta}_u = (\theta_{u,1}, \ldots, \theta_{u,K})$ とし，ユーザ u の i 番目のアイテムのトピック割り当てを $z_{u,i}$ とする．トピック $k \in \mathcal{K}$ におけるアイテム分布を $\boldsymbol{\phi}_k = (\phi_{k,1}, \ldots, \phi_{k,X})$ とする．また，ユーザ u が i 番目に購買したアイテムを $x_{u,i} \in \mathcal{X}$ とする．各変数 $x_{u,i}$，$z_{u,i}$，$\boldsymbol{\phi}_k$，$\boldsymbol{\theta}_u$ の同時分布は式 (1) で表される．

$$
\begin{aligned}
&p(x_{u,i}, z_{u,i}, \boldsymbol{\phi}_k, \boldsymbol{\theta}_u) \\
&= p\left(x_{u,i} \mid z_{u,i}, \boldsymbol{\phi}_k\right) p\left(z_{u,i} \mid \boldsymbol{\theta}_u\right) p\left(\boldsymbol{\phi}_k\right) p\left(\boldsymbol{\theta}_u\right)
\end{aligned} \tag{1}
$$

3. 提案手法

3.1. 概要

hLDA ではトピックの階層構造が木構造であり，全ての上位トピック，下位トピック間の関係性が表現できない．また PAM は全ての上位トピック，下位トピックがネットワークで結合されているが，推定されるモデルのパラメータが初期値に影響を受けやすく，トピックにばらつきが生じ，安定した解釈結果を得ることが困難である．そこで，初期値に対する安定性が高い LDA を用いてユーザの嗜好の階層性を表現することを考える．しかし LDA はトピックの階層が 1 つであるため，そのままでは階層性を表現することができない．一方，EC サイト上の全アイテムは付与されている大カテゴリ，小カテゴリといった階層構造を有したアイテムマスタによって管理されている．

そのため，このアイテムの階層的カテゴリ情報の粒度を変えて，上位階層 LDA と下位階層 LDA を構成し，これら 2 つの LDA を組み合わせてトピックの階層性を表現可能なモデルを考える．そのためにまず，大カテゴリ，小カテゴリごとの購買履歴データに LDA を適用し，トピックを生成する．次に，大カテゴリと小カテゴリのトピック間の関係性を表すトピック関連度スコアを算出し，その値を用いることで安定したトピックの階層構造を構築する．これにより，カテゴリを活用し安定したトピックを得ることができ，かつユーザの嗜好の階層性の分析が可能となる．

3.2. LDA の学習

アイテムの大カテゴリの購買履歴データに LDA を適用し得られたトピックを B-topic，小カテゴリの購買履歴データから得られたトピックを S-topic とする．また，それらのトピック数を K，M とする．提案手法ではまず各トピック $Z = \{z_k \mid 1 \leq k \leq K\}$，$V = \{v_m \mid 1 \leq m \leq M\}$ を推定し，z_k，v_m におけるアイテム分布 $\boldsymbol{\phi}_k$，$\boldsymbol{\phi}_m$ を推定する．

3.3. トピック関連度スコアの算出

トピック関連度スコアは，B-topic と S-topic のトピック間の関係性を表現した値である．はじめに，トピック関連度スコアを算出するために，購買履歴データの小カテゴリを，それらが所属する大カテゴリに変換する．そしてトピック間の関連度を定義するために，EC サイト側が保有するアイテムカテゴリの階層性を活用する．

具体的には，ある B-topic と，その B-topic に含まれるアイテムの下位階層の小カテゴリに属するアイテムが含まれる S-topic は関連度が高いと仮定する．例えば，大カテゴリで得られた食品のトピックと，小カテゴリで得られたスイーツのトピックは関連度が高いとみなすことができる．大カテ

ゴリ数を Q とし，q 番目の大カテゴリを G_q，q 番目の大カテゴリに属する小カテゴリ数を N_q，q 番目の大カテゴリに属する n 番目の小カテゴリを g_{qn} とする．このとき，トピック関連度スコアは式 (2) によって算出される．

$$\gamma_{k,m} = \sum_{q=1}^{Q} p\left(G_q \mid z_k\right) \sum_{n=1}^{N_q} p\left(g_{qn} \mid v_m\right) \tag{2}$$

ここで $\gamma_{k,m}$ を z_k と v_m のトピック関連度スコアと定義する．$p\left(G_q \mid z_k\right)$ は z_k の下で q 番目の大カテゴリ G_q が出現する確率，$p\left(g_{qn} \mid v_m\right)$ は v_m の下で，q 番目の大カテゴリに属する小カテゴリ g_{qn} が出現する確率を表す．式 (3)，(4) で，B-topic， S-topic ごとに，各トピック関連度スコアの和が 1 になるように正規化を行う．B-topic における S-topic の所属確率を式 (3)，S-topic のトピックにおける B-topic のトピックの所属確率を式 (4) で定義する．

$$\widetilde{p}\left(v_m \mid z_k\right) = \frac{\gamma_{k,m}}{\sum_j \gamma_{k,j}} \tag{3}$$

$$\widetilde{p}\left(z_k \mid v_m\right) = \frac{\gamma_{k,m}}{\sum_l \gamma_{l,m}} \tag{4}$$

以上の所属確率より，大カテゴリ，小カテゴリの双方を基準とした各トピック間の関係性を表現でき，ユーザの嗜好の階層性を分析することが可能である．

4. 実データ分析

提案手法の有用性を示すため，2019 年の楽天市場の評価履歴データ [4] を購買履歴とみなして提案手法を適用し，得られた結果に対する考察を行う．

4.1. 分析条件

実験データの大カテゴリ数 Q，小カテゴリ数はそれぞれ 39，437 である．B-topic 数，S-topic 数は事前実験の結果と解釈性の観点から K=6，M=18，ハイパーパラメータ $\boldsymbol{\alpha}$ の要素は全て 0.1，β=0.1 とし，モデルの学習を行った．

4.2. 分析結果と考察

B-topic と S-topic の解釈の違いを確認するために，表 1 に B-topic への所属確率が高い上位 3 大カテゴリの結果の一部を，表 2 に S-topic への所属確率が高い上位 3 小カテゴリの結果の一部を示す．

表 1 より，z_1 は出産を機に生活に変化が生まれたユーザ，z_2 は美容意識が高いユーザ，z_2 は食品を購買する傾向にあるユーザのトピックであることがわかる．

表 1: B-topic の上位 3 大カテゴリ（一部抜粋）

z_1	z_2	z_3
キッズ・マタニティ	美容・コスメ	食品
日用品	ダイエット	スイーツ・お菓子
家具	日用品	水・ソフトドリンク

表 2 より，v_2 はダイエット食品，v_4 は食品，v_{11} は子供用品のトピックであり，表 1 と比較すると，より詳細なトピックが得られていることが確認できた．

表 2: S-topic の上位 3 小カテゴリ（一部抜粋）

v_2	v_4	v_{11}
サプリメント	水産加工品	ベビー
ナッツ	惣菜	おもちゃ
ダイエット	フルーツ	マタニティ・ママ

さらに，ユーザの嗜好の階層性を確認するために，z_1 における各 S-topic の所属確率を表 3，v_2 における各 B-topic の所属確率を表 4 に示す．

表 3: z_1 における $\widetilde{p}\left(v_m \mid z_1\right)$

	v_0	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
z_1	0.0551	0.0056	0.0047	**0.1320**	0.0040	0.0672	**0.1247**	0.0076	0.0106
	v_9	v_{10}	v_{11}	v_{12}	v_{13}	v_{14}	v_{15}	v_{16}	v_{17}
z_1	**0.1327**	0.0057	**0.1318**	0.0682	0.0472	**0.1505**	0.2087	0.0231	0.0084

表 4: v_2 における $\widetilde{p}\left(z_k \mid v_2\right)$

	z_0	z_1	z_2	z_3	z_4	z_5
v_2	0.0139	0.0221	**0.6154**	**0.3451**	0.0004	0.0032

表 3 より，z_1 において S-topic v_3,v_6,v_9,v_{11},v_{14} の所属確率が高いことがわかる．これらの S-topic は家具，ペット用品，子供用品，マタニティ用品，手芸アイテムのトピックであり，z_1 の「出産を機に生活に変化が生まれ，子供用品や日用品を購買する傾向のあるグループ」の中に，「子供用品だけでなく，ペット用品も購買するグループ」が存在することがわかる．

また表 4 より，v_2 のダイエット食品のトピックは美意識が高いユーザの B-topic z_2，食品を購買する傾向が高いユーザの B-topic z_3 への所属確率が高くなっている．したがって，同じダイエット食品を購買する傾向のグループの中にも，美容に興味があるユーザ嗜好のトピックと，ダイエット食品に限らず様々な食品を購買するユーザ嗜好のトピックという 2 つの側面が存在することがわかる．以上より，提案手法によりユーザの嗜好の階層性を考慮した分析が可能であるといえる．

5. まとめと今後の課題

本研究では，アイテムが持つカテゴリという階層性を用いることで，従来の階層性を考慮したモデルと比較して安定したユーザの嗜好を分析可能な手法を提案した．さらに，実際の EC サイトの評価履歴データを用いることで提案手法の有用性を示した．今後の課題として，ユーザの属性や店舗情報などの補助情報を考慮したモデルの構築が挙げられる．

謝辞

本研究では，国立情報学研究所の IDR データセット提供サービスにより楽天株式会社から提供頂いた「楽天データセット」を用いた．貴重なデータの提供に深く感謝致します．

参考文献

[1] Blei, D.M., Ng, A.Y., Jordan, M.I., “Latent Dirichlet Allocation,” *Machine learning*, vol.45, No.1, pp.5–32, 2001.

[2] Griffiths, T .L., et al.,”Hierarchical topic models and the nested Chinese restaurant process,” *Advances in neural information processing systems*,pp.17–24, 2004.

[3] Li, W., McCallum, A., “Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations,” *Proceedings of the 23rd international conference on Machine learning*, pp. 577–584, 2006.

[4] 楽天株式会社, “楽天市場データ,” 国立情報学研究所情報学研究データリポジトリ．(データセット), https://doi.org/10.32130/idr.2.1, 2014.