Labeled-LDA に基づくユーザ属性を考慮した Web 閲覧行動分析モデル

1X17C077-7　鈴木凜平
指導教員　後藤正幸

1. 研究背景・目的

多くの企業のマーケティング活動において，年代や職業など複数の属性情報に基づいて施策を検討することは一般的であり，様々な属性を持つ顧客の嗜好を詳細に理解し，顧客属性毎に嗜好の差異を分析することは大変重要である．さらに近年では，膨大な閲覧履歴データやユーザの属性情報が蓄積されており，これの閲覧履歴データと顧客属性の関係性を分析する手法の開発が望まれている．

顧客は多様な嗜好を有することが想定されるため，購買や閲覧などの行動の背後に潜在的な嗜好を仮定することのできる潜在クラスモデルが顧客分析において広く活用されている．これらのなかで代表的な手法である Latent Dirichlet Allocation[1] (以下，LDA) は，近年，閲覧履歴データの分析や推薦システムなどの実データ分析への活用が注目を集めている．しかし，LDA は統計的に推定された潜在的な嗜好（トピック）とユーザ属性との直接的な関係性を扱うことができない．

一方，ユーザ属性と閲覧履歴の関係性分析のため，予め与えられている補助情報をラベルとしてトピックを制限しながら学習を行う Labeled Latent Dirichlet Allocation[2]（以下，L-LDA）の活用が考えられる．L-LDA は，マルチラベルの文書データを対象とした手法であり，文書に複数付与されたラベルがトピックと 1 対1に対応する．文書に付与されているラベルをユーザの属性に，文書中の単語を各閲覧に対応させることで L-LDA を閲覧履歴データに適用することができる．しかし，複数の属性を持つユーザの嗜好を分析するために，各ユーザに対して複数の属性情報を付与したマルチラベルとして学習を行うと，顧客属性の種類が多くなった場合に，その組み合わせ数は膨大になる．そのため，決定されるトピック数も指数的に増加し，1つのトピックに含まれる学習データ数が相対的に少なくなることでトピックの推定精度の低下につながってしまう．

そこで，本研究では，1つの属性を付与したシングルトピックとして，年代や職業などの属性の種類ごとに L-LDA の学習を行い，別個のモデルで得られたトピックの間でクロス分析を行うことによって，複数の属性を持つユーザの閲覧行動の分析を行う手法を提案する．これにより，個々の学習におけるトピック数の増加に伴う推定精度の低下を防ぎ，推定精度を確保しつつ，複数種類の顧客属性の組み合わせに対する購買行動の分析を可能とする．最後に，実際の閲覧履歴データに提案手法を適用し，属性情報の観点から顧客の嗜好を分析することにより提案手法の有効性を示す．

2. 準備

2.1. 変数の定義

本研究では，閲覧履歴データを対象としているため，これを前提として変数の定義を行う．まず，全ユーザ数を U，全サイト数を I とし，ユーザ集合を $\mathcal{U} = \{1, \ldots, U\}$，サイト集合を $\mathcal{X} = \{1, ..., I\}$，ユーザ u が i 番目に閲覧したサイトを $x_{ui} \in \mathcal{X}$ とする．

2.2. Latent Dirichlet Allocation

閲覧履歴データに LDA を適用することで，ユーザにトピックの出現確率分布，そのトピックごとに閲覧サイト出現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．ここで，ユーザ $u \in \mathcal{U}$ のトピック分布を $\boldsymbol{\theta}_u = (\theta_{u1}, ..., \theta_{uK})$ とし，また，潜在トピック集合 $\mathcal{Z} = \{1, ..., K\}$，ユーザ u の i 番目のサイトのトピック割り当てを z_{ui} とする．トピックこのとき，$k \in \mathcal{Z}$ におけるサイト分布を $\boldsymbol{\phi}_k = (\phi_{k1}, ..., \phi_{kI})$ とする．各変数 x_{ui}, z_{ui},$\boldsymbol{\theta}_u$,$\boldsymbol{\phi}_k$ の同時分布は式 (1) で表される．

$$\begin{aligned} &P(x_{ui}, z_{ui}, \boldsymbol{\phi}_k, \boldsymbol{\theta}_u) \\ &= P(x_{ui}|z_{ui}, \boldsymbol{\phi}_k)P(z_{ui}|\boldsymbol{\theta}_u)P(\boldsymbol{\phi}_k)P(\boldsymbol{\theta}_u) \end{aligned} \quad (1)$$

ここで，$P(\boldsymbol{\theta}_u) = \mathrm{Dir}(\boldsymbol{\alpha})$ ，$P(\boldsymbol{\phi}_k) = \mathrm{Dir}(\boldsymbol{\beta}_\mathrm{k})$ である．ただし，$\mathrm{Dir}(\cdot)$ はディリクレ分布，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$，$\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,I})^\top$ は各分布のパラメータである．

2.3. Labeled Latent Dirichlet Allocation

L-LDA では，各ユーザに付与されたラベルとトピックが1対1に対応するようにトピックを制御する．すなわち，ラベルに性別が含まれている場合には，男性のトピック・女性のトピックなどが生成される．ここで，全ラベル数を M，ユーザ u に付与されているラベル数を M_u とし，ユーザ u のラベルベクトルを $\boldsymbol{\Lambda}^{(u)} = (\Lambda_1^{(u)}, \ldots, \Lambda_M^{(u)})$ $(\Lambda_m^{(u)} \in \{0, 1\})$ とする．そして，このラベルベクトル $\boldsymbol{\Lambda}^{(u)}$ に基づき，ラベル指示ベクトル $\boldsymbol{\lambda}^{(\boldsymbol{u})}$ を $\boldsymbol{\lambda}^{(u)} = \{m|\Lambda_m^{(u)} = 1\}$ $(m = 1, \ldots, M)$ と定義する．さらに，ラベル射撮行列 $\boldsymbol{L}^{(u)} \in \mathbb{R}^{M_u \times M}$ の m 行 n 列目の要素 $L_{m,n}^{(u)}$ を式 (2) のように定義する．

$$L_{m,n}^{(u)} = \begin{cases} 1 & (\lambda_m^{(u)} = n) \\ 0 & (\lambda_m^{(u)} \neq n) \end{cases} \quad (2)$$

そして，ユーザ u のトピック分布の事前分布であるディリクレ分布のパラメータ $\boldsymbol{\alpha}^{(u)}$ を以下の式 (3) によって求める．

$$\boldsymbol{\alpha}^{(u)} = \boldsymbol{L}^{(u)}\boldsymbol{\alpha} = (\alpha_{\lambda_1^{(u)}}, \ldots, \alpha_{\lambda_{M_u}^{(u)}})^\top \quad (3)$$

ただし，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^\top$ である．式 (2),(3) により，L-LDA では，事前に付与されたラベルによりトピックを制御しながらトピック分布，サイト出現分布の推定を行うことが可能となる．

3. 提案手法

3.1. 概要

本研究では，ユーザが持つ複数の属性をラベルに対し，ラベルの種類ごとに独立で L-LDA を学習し，各モデルで得られ

たパラメータをもとにクロス分析を行う手法を提案する．その結果，属性の種類毎に学習を行った上でクロス分析を行うことにより，推定精度を保ちつつ，複数の属性種を持つユーザの嗜好を分析することが可能となる．

3.2. 提案手法を用いた分析手順

本研究では，性別年代・職業・世帯年収の3種のラベルを用いる．属性種について職業を a，年代性別を b，世帯年収を c とする．さらに，各属性種の属性集合を $\mathcal{Z}_a = \{z_{(a,1)}, \ldots, z_{(a,N_a)}\}$，$\mathcal{Z}_b = \{z_{(b,1)}, \ldots, z_{(b,N_b)}\}$，$\mathcal{Z}_c = \{z_{(c,1)}, \ldots, z_{(c,N_c)}\}$ とする．ここで，提案手法による分析手順を以下に示す．

Step.1) 属性種ごとの L-LDA の学習

まず属性の種類ごとに独立に L-LDA の学習を行い，属性ごとの出現確率 $P(x_i|z)(z \in \{\mathcal{Z}_a \cup \mathcal{Z}_b \cup \mathcal{Z}_c\})$ を求める．その後，サイトごとに各属性での出現確率を用いて，各属性の各サイトへの所属確率 $P(z|x_i)$ の算出を行う．

Step.2) 属性種間のクロス分析の実施

属性種間でクロス分析を行うために，2つの属性が同時に各サイトへ所属する確率を算出する．具体的には，サイト x_i における $z_{(a,s)} \in \mathcal{Z}_a$ かつ $z_{(b,t)} \in \mathcal{Z}_b$ の同時確率を以下の式 (4) のように定義する．

$$P(z_{(a,s)}, z_{(b,t)}|x_i) = P(z_{(a,s)}|x_i) \times P(z_{(b,t)}|x_i) \quad (4)$$

従来，クロス分析を活用した分析手法として劉らの研究 [3] がある．この手法では，異なるデータで学習した2つの潜在クラスモデルで推定された各潜在トピックが同時に各ユーザに所属する確率を算出しており，学習データの違いによって独立性を仮定している．一方，本手法では，クロス分析に用いるモデルの学習データが共通なため，属性種ごとの独立した学習とサイト x_i による条件付き確率とすることよって独立性を仮定する．そして，式 (4) の計算をサイトごとに行い，この結果を用いて複数の属性を持つユーザの嗜好に関するクロス分析を行う．

4. 実データ分析

4.1. 分析条件

本研究では，株式会社ヴァリューズから提供を受けた閲覧履歴データを使用する．全ユーザ数は $U = 6,068$，全サイト数は $I = 479$ であり，総閲覧回数は $7,443,579$ であった．さらに，ユーザの属性のうち，「職業 $(\mathcal{Z}_a)$」「年代+性別 $(\mathcal{Z}_b)$」「世帯年収 $(\mathcal{Z}_c)$」を属性ラベルとして使用し，総ラベル数は 44 であった．また，ハイパーパラメータは $\alpha = 0.1$，$\beta = 0.1$ とした．

4.2. 属性の種類数とトピックの推定精度に関する分析

属性種の数 Q を変化させたときの出現確率の上位との比較を行うために，「会社勤務（$z_{a,p} \in \mathcal{Z}_a$）」の属性を持つユーザのサイト閲覧回数上位 5 件を表 1 に示す．次に，3種類の属性 $(\mathcal{Z}_a \cup \mathcal{Z}_b \cup \mathcal{Z}_c)$ を付与して L-LDA で学習を行うことによって得られた「会社勤務」の出現確率上位 5 件を表 2 に示す．さらに，2 種類の属性 $(\mathcal{Z}_a \cup \mathcal{Z}_b)$ を付与した場合の結果を表 3 に，最後に 1 種類のラベル $(\mathcal{Z}_a)$ を付与した場合の結果を表 4 に示す．

表 1 から表 4 より，属性種が増えるたびに出現回数上位と出現確率上位の共通するサイトが減少し，またそれらの確率が小さくなっていることから推定精度が低下していることが考えられる．そのため，ラベル種ごとに独立して学習を行うことが有効であるといえる．

表 1: 会社勤務 $(z_{a,p})$ のサイト閲覧回数上位 5 件

サイト名	被閲覧回数
アダルトサイト	131774
Yahoo! ニュース	122434
Yahoo!検索	121999
Google（日本）	108322
国税庁 確定申告書等作成コーナー	77026

表 2: $Q = 3$ の場合の $P(x_i|z_{a,p})$ 上位 5 件

サイト名	$P(x_i\|z_{a,p})$
Twitter	0.0531
永久不滅プラス	0.0381
Facebook	0.0362
Gmail	0.0240
SBI 証券	0.0217

表 3: $Q = 2$ の場合の $P(x_i|z_{a,p})$ 上位 5 件

サイト名	$P(x_i\|z_{a,p})$
楽天ウェブ検索	0.0491
永久不滅プラス	0.0440
Twitter	0.0411
ネットアンケート	0.0402
Facebook	0.0294

表 4: $Q = 1$ の場合の $P(x_i|z_{a,p})$ 上位 5 件

サイト名	$P(x_i\|z_{a,p})$
Google（日本）	0.0852
アダルトサイト	0.0588
Twitter	0.0516
DMM.ADULT アニメ通販	0.0447
楽天ウェブ検索	0.0354

4.3. 提案手法を用いた顧客の嗜好に関する分析

クロス分析結果を活用した分析の一例として，「会社勤務 $(z_{a,p})$」かつ「50 代男性 $(z_{b,q} \in \mathcal{Z}_b)$」の属性を持つユーザに対して同時確率を算出した結果を表 5 として示す．

表 5: $P(z_{a,p}, z_{b,q}|x_i)$ 上位 10 件

サイト名	$P(z_{a,p}, z_{b,q}\|x_i)$
Response.	0.0742
Google+	0.0558
Yahoo! オークション	0.0500
日経 xTECH（クロステック）	0.0420
Vector ソフトライブラリ& PC ショップ	0.0418
車・自動車 SNS みんカラ	0.0418
Google Sites	0.0372
Engadget 日本版	0.0369
Yahoo!かんたん決済	0.0355
トヨタ自動車	0.0314

表 5 より，「会社勤務」かつ「50 代男性」は，自動車関連サイトとニュースサイトの所属確率が高く，関心が高いことが分かる．そのため，これらの属性を持つユーザに対しては，自動車関連サイトとニュースサイトを中心にマーケティング施策を行うことが効果的であると考えられる．

5. まとめと今後の課題

本研究では属性の種類ごとに L-LDA の学習を行い，クロス分析を行うことで複数属性の組み合わせ毎の特徴を分析可能な手法を提案した．さらに，実際の閲覧履歴データに対して提案手法を適用し分析を行い，提案手法の有効性を示した．今後の課題としては，購買履歴などの異なる実データへの応用や，サイトの多様性を考慮した分析手法の改良が挙げられる．

参考文献

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

[2] D.Ramage, D.Hall, R.Nallapati, C.D.Manning ”Manning:Labeled LDA: A supervised topic model for credit attribution in multi-label corpora”, *EMNLP2009*, pp.248-256, 2009.

[3] 劉 佩潔, 山下 遥, 岩永二郎, 樽石将人, 後藤正幸 ”グルメサービスにおけるレストラン推薦投稿へのリアクション数増加を目的とした潜在クラスモデル分析”, 情報処理学会論文誌, Vol.59, No.1, pp.211-226, 2018.