生花 EC サイトの閲覧履歴データを学習する改良型 Latent LSTM Allocation モデルの提案

情報数理応用研究　　5219C024-4　張志穎
指導教員　後藤正幸

A Proposal of Improved Latent LSTM Allocation Model to Learn Browsing History Data of Flower EC Site

ZHANG Zhiying

1. 研究背景と目的

近年，インターネットを通じた購買活動は広く消費者に受け入れられるようになり，EC（Electronic Commerce）サイトの市場規模は引き続き増大傾向となっている．このような背景のもと，蓄積された膨大なユーザの購買行動データを活用し，効果的なマーケティング施策に結び付けることの重要性が高まっている．そのため，ユーザの行動履歴から各商品に対する購買意欲を分析し，購買に直結するユーザ行動などの情報を把握する必要がある．そのような中，ユーザの閲覧履歴は，購買履歴よりもデータ量が膨大であり，かつユーザの嗜好や各商品に対する購買意欲の差異が閲覧行動に現れると考えられるため，購買に至る割合を向上させるためには，この閲覧行動をどのように扱うかがポイントとなる．

従来，顧客の購買履歴データの分析には，トピックモデルが有用であることが，様々な事例で報告されている．しかし，単純なトピックモデルは，時系列で与えられる個々の閲覧が独立に生成されるという仮定を置いたモデルであるため，EC サイト上のページ遷移ではこの仮定は相応しくないと考えられる．そのため，Zaheer らは，トピックモデルである Latent Dirichlet Allocation（略称:LDA）[1] と時系列データの長期依存を学習可能な長・短期記憶（Long short-term memory，略称:LSTM）[2] モデルを組み合わせた新たなモデル Latent LSTM Allocation（以下，LLA）[3] を提案した．しかし，LLA モデルは時系列性を持つ閲覧データの特徴分析は可能である一方，その後の購買有無の要素を考慮していない．また，各ページの閲覧時間はユーザの迷いや熟考の程度を示していると考えられるため，ページ遷移の時間間隔を考慮することも望まれる．

そこで，本研究では，ユーザの閲覧行動と購買の有無の関係を分析するためのモデルとして，購買の有無と行動の時間間隔を表現した LLA の拡張モデルを提案する．これにより，購買行動と商品選択するための閲覧行動の双方の観点から類似したグループを表現したモデルが構築され，ユーザの閲覧行動と購買行動を統合的に分析することが可能となる．本研究では，提案モデルを実際の利用履歴データ分析に適用し，得られる知見について考察を行う．

2. 問題設定

本研究で分析対象とするデータは，生花商品専用の EC サイトを運営する企業 A 社が提供する注文履歴データと閲覧履歴データである．本 EC サイトでは，誕生日や母の日などのイベントに紐付いて購入される生花商品を取り扱っているため，各顧客の年間平均購買回数は生活必需品などに比べて非常に少ないという特徴がある．そのため，個々の顧客に対してある程度のボリュームの購買履歴が利用できることを前提とした分析手法は，そのまま適用することができない．1 人あたりの年間購買回数の統計量を表 1 で示す．

表 1: 年間購買回数の統計量

年間購買回数の平均値	2.31
年間購買回数の標準偏差	212
年間購買回数 5 回以下ユーザの割合	61.34%

また，EC サイトにおける顧客の購買行動には，様々なパターンがある．特に，ギフト用商品がメインとなる生花業界では，複数の商品を比較し時間をかけて吟味するユーザや，購買する商品がアクセスする前から決まっているユーザなど，顧客による行動パターンの差異が大きい．そのような差異により，EC サイト上での商品の探索行動も大きく異なると考えられる．本研究ではこのような特徴を考慮した上で，潜在的特徴が異なるユーザのクラスタリングモデルを構築する．これにより，購買に至りやすいグループの特徴と購買に至りにくいグループの特徴の差異を分かるようになり，企業側が購買のキーポイントとなるページを宣伝するなどの施策に通じて，販売の促進に結びつくと考えられる．

3. 準備

3.1. LDA[1]

代表的トピックモデルである LDA モデルは，自然言語処理分野だけでなく，マーケティングデータにも応用がなされている．一般に，ページ遷移データは，嗜好が全く異なるユーザによる行動履歴の集まりであると仮定することが自然であり，統計的性質の異なるグループから成り立っていると考えられる．これらのデータをグルーピングすることで，購買に至りやすいグループの性質と購買に至りにくいグループの性質の差異が分析可能となる．その結果を，購買のキーポイントとなるページを宣伝するなどの施策に結び付けることで，販売の促進につながると期待される．LDA のグラフィカルモデルを図 1 に示す．

ここで，z'_k は k 番目のトピックを表しており，アイテム分布 ϕ'_k と潜在トピック z'_k が与えられたもとで，1 つの購買アイテム v'_n が生成される．また，θ'_m，ϕ'_k に対して，それぞれ α'，β' をパラメータとするディリクレ事前分布を仮定する．

3.2. LSTM[2]

LSTM モデルは，深層学習の分野において用いられる再帰型ニューラルネットワーク（RNN）アーキテクチャである．一般的な LSTM のユニットは，RNN とは異なり，図

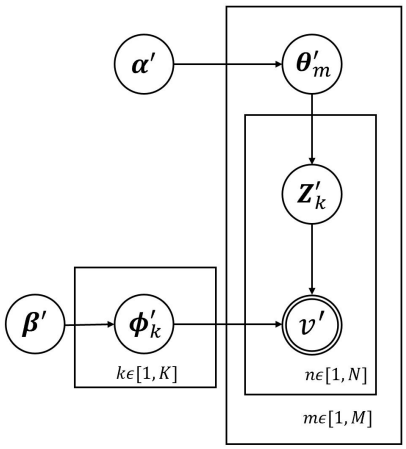

図 1: LDA のグラフィカルモデル

2 で示したように，セル，入力ゲート，出力ゲート，および忘却ゲートから構成される．セルは任意の時間間隔にわたって値を記憶し，3 つの「ゲート」はセルを出入りする情報の流れを制御する．そのような構造によって，メモリーのコントロールが可能となり，時系列データの長期的な統計的特徴を学習できるようになる．LSTM は，従来の RNN を訓練する際に遭遇しうる勾配爆発と勾配消失の問題に対処するために開発された．ギャップの長さに対する相対的な鈍感さが，多数の応用における RNN や隠れマルコフモデル，その他のモデルに対する LSTM の優位性となっている．

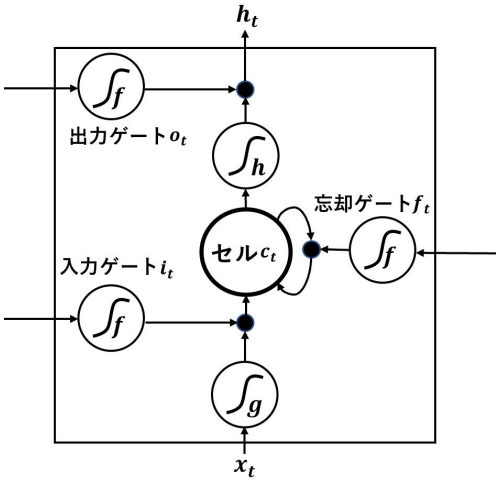

図 2: LSTM ユニットのイメージ図

4. 従来手法

LSTM モデルは時系列性を持つ離散データを処理するための強力なツールとなっているが，数多くのパラメータの学習を必要とし，そのパラメータを解釈することも難しくなっている．この点は，多様なユーザ行動を分析するためのモデリング手法としては望ましいとは言えない．一方，LDA などのトピックモデルは解釈性を持つが，時系列的特徴を学習することはできない．そこで，Zaheer らは LSTM モデルと LDA モデルを組み合わせた Latent LSTM Allocation モデルを提案した [2]．このモデルにより，各ユーザ行動のトピックを抽出し，行動ではなくトピックの時系列性を考慮し，学習を行うことが可能となる．

ユーザがサイトにアクセスしてから購買や離脱までの一連の閲覧行動の単位をセッションと呼ぶ．セッション数を D，トピック数を K，ユーザ閲覧履歴データの長を N_d とそれぞれ仮定する．また，ユーザの下でのトピック分布を θ として，トピック z_k の下でのユーザ分布を ϕ_k と表記する．ϕ_k に対し，β をパラメータとするディリクレ事前分布を仮定する．また，セッション d の t 時刻における LSTM の状態値を $s_{d,t}$ とし，その状態のもとで生成するトピックを $z_{d,t}$，閲覧ページを $w_{d,t}$ と表記する．θ を生成するために用いる重みを W_p とし，バイアスを b_p とする．LLA の生成プロセスとグラフィカルモデルは以下のようになる．

生成アルゴリズム

Step1 k は 1 から K まで

(a) ϕ_k を決める: $\phi_k \sim Dir(\beta)$

Step2 D におけるすべての d に対して

(a) 初期化：LSTM の $s_{d,0}=0$ とする

(b) t は 1 から N_d まで

i. $s_{d,t}$ を更新する：

$$s_{d,t} = \mathbf{LSTM}\left(z_{d,t-1}, s_{d,t-1}\right)$$

ii. θ を生成する：

$$\theta = softmax_K\left(W_p s_{d,t} + b_p\right)$$

iii. トピックを生成する：

$$z_{d,t} \sim Categorical\left(\theta\right)$$

iv. ページを生成する：

$$w_{d,t} \sim Categorical\left(\phi_{z_{d,t}}\right)$$

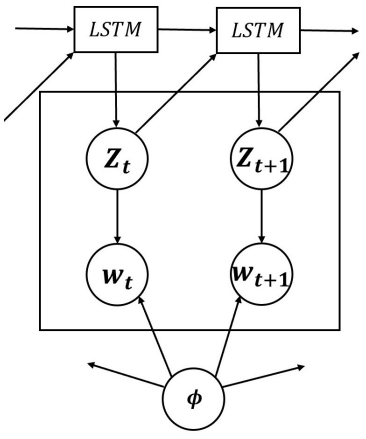

図 3: LLA のグラフィカルモデル

LLA の変数やパラメータの中で，実際に観測される変数は閲覧ページ 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 stochastic EM(SEM) を用して推定する．従来の EM は z の条件付き分布を推定するが，パラメータ数が多くなった場合の高速化のため z の条件付き分布の不偏推定量を求める. LLA における SEM アルゴリズムの各ステップの更新式は以下の式 (1)–(3) で計算される．

SE-Step:

$$\pi_k = \phi_{w_{d,t}} P_L\left(z_{d,t} = k \mid z_{d,1:t-1}, \xi_L\right) \tag{1}$$

Sample $z_{d,t} \sim \text{Categorical}(\pi)$

M-Step:

$$\pi_{wk} = \frac{n_{wk} + \beta}{n_k + V\beta} \tag{2}$$

$$\frac{\partial L}{\partial \xi_L} = \sum_{d \in D} \sum_{t=1}^{N_d} \frac{\partial \log P_L\left(z_{d,t} \mid z_{d,1:t-1}, \xi_L\right)}{\partial \xi_L} \tag{3}$$

ただし，n_{wk} はページ w をトピック z_k に割り当てられている回数であり，n_k はトピック z_k に割り当てられたページの総回数である．また，V はページ種類数を表し，ξ_L は LSTM のパラメータを表す．

5. 提案モデル

従来の LLA は，LSTM モデルと LDA モデルの組み合わせにより閲覧履歴データの背後に潜在クラスを仮定してクラスタリングできる．しかし，文章データのような時系列データとは異なり，今回扱う対象は閲覧履歴データであるため，ページ遷移の時間間隔もユーザの意思決定行動の特徴を反映していると考えられる．また，LLA モデルでは閲覧履歴のみを用いて潜在クラスを学習しているが，購買に至る割合の向上を検討するためには，閲覧履歴のみではなく，購買履歴との関連性を考慮することが重要と考えられる．ここでは，2 ステップに分けて提案モデルの拡張の方向性を示す．

5.1. 時間間隔を考慮したモデルへの拡張

EC サイトにおける顧客の購買行動では，各閲覧ページにおける滞在時間にもユーザの思考状態を表われると考えられる．例えば，複数の商品を閲覧し時間をかけて吟味するユーザや購買する商品がアクセスする前から決まっているユーザでは，データとして表される閲覧行動は大きく異なる．そのため，閲覧時間間隔をモデルに反映することが重要と考えられる．そこで，本研究では，時間間隔も潜在クラスから確率的に生起する確率変数であると仮定し，モデルを提案する．

セッション d の t 時刻における時間間隔を $T_{d,t}$ と仮定し，トピック z_k の下での時間分布を Ω_k と表記する．Ω_k に対し，γ をパラメータとするディリクレ事前分布を仮定する．提案する生成プロセスとグラフィカルモデルは以下のようになる．

生成アルゴリズム

Step1 k は 1 から K まで

(a) ϕ_k を決める: $\phi_k \sim Dir(\beta)$

(b) Ω_k を決める: $\Omega_k \sim Dir(\gamma)$

Step2 D におけるすべての d に対して

(a) 初期化：LSTM の $s_{d,0} = 0$ とする

(b) t は 1 から N_d まで

i. s_t を更新する：

$$s_{d,t} = \textbf{LSTM}\left(z_{d,t-1}, s_{d,t-1}\right)$$

ii. θ を生成する：

$$\theta = softmax_K\left(W_p s_{d,t} + b_p\right)$$

iii. トピックを生成する：

$$z_{d,t} \sim Categorical\left(\theta\right)$$

iv. ページを生成する：

$$w_{d,t} \sim Categorical\left(\phi_{z_{d,t}}\right)$$

iv. 時間間隔を生成する：

$$T_{d,t} \sim Categorical\left(\Omega_{z_{d,t}}\right)$$

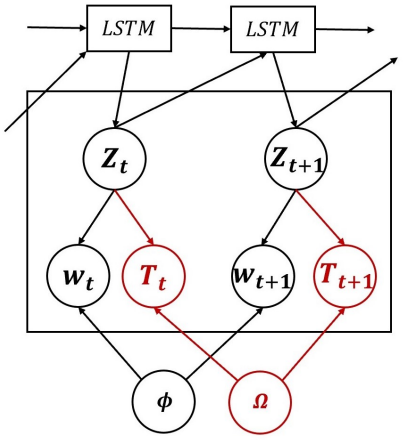

図 4: 提案手法のグラフィカルモデル

提案手法における SEM アルゴリズムは以下の式 (4)〜式 (7) のように更新される．

SE-Step:

$$\pi_k = \phi_{w_{d,t}} \Omega_{w_{d,t}} P_L\left(z_{d,t} = z_k \mid z_{d,1:t-1}, \xi_L\right) \tag{4}$$

Sample $z_{d,t} \sim \text{Categorical}(\pi)$

M-Step:

$$\pi_{wk} = \frac{n_{wk} + \beta}{n_k + V\beta} \tag{5}$$

$$\Omega_{wk} = \frac{n_{Tk} + \gamma}{n_{kT} + J\gamma} \tag{6}$$

$$\frac{\partial L}{\partial \xi_L} = \sum_{d \in D} \sum_{t=1}^{N_d} \frac{\partial \log P_L\left(z_{d,t} \mid z_{d,1:t-1}, \xi_L\right)}{\partial \xi_L} \tag{7}$$

ただし，n_{Tk} は時間間隔 T がトピック z_k に割り当てる回数になり，n_{kT} はトピック z_k に属する時間間隔数になる．また，J は時間間隔の種類数を表す．

5.2. 購買有無を考慮したモデルへの拡張

LLA では，閲覧履歴データのみを用いて類似セッション同士をソフトクラスタリングし，購買有無の結果を考慮していない．そのため，購買を促進させるための施策検討と直接的に結び付けることは難しいと考えられる．そこで提案手法では，閲覧履歴データの最後に購買有無の結果データを追加することで，閲覧履歴と購買履歴を紐づけてクラスタリングすることと考える．それにより，購買に至りやすいセッションと購買に至りにくいセッションそれぞれの特徴を抽出することができるため，企業側が購買のキーポイントとなるページを明らかにすることができ，販売促進のための施策立案に繋がると考えられる．

6. 実データ分析

6.1. 分析条件

本研究の有効性を検証するため，生花専用 EC サイト A により提供された閲覧履歴データと購買履歴データを用いる．対象データの概要を以下に示す．

データ期間： 2019 年 3 月-2020 年 2 月

説明変数： セッション番号，閲覧ページ，時間間隔データ，購買有無の結果

学習セッション数: 約 200,000 件

提案モデルの学習については，全セッションからランダムに 200,000 件をサンプリングした．潜在クラスの個数 K を経験的に 6 と設定し，バッチサイズを 200，エポック数を 10 と設定した．また，閲覧履歴のページ数を 25 で統一し，ユニット数は 50 と設定して，実験を行う．

6.2. 実験結果と考察

各潜在トピックの平均時間間隔（秒）と購買結果の分類を表 2 に示す．

表 2: 各潜在トピックの平均時間間隔と購買結果の分類

潜在 Topic	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
時間間隔（秒）	1134	5005	877	2667	798	698
購買結果	閲覧	閲覧	離脱	閲覧	購買	閲覧

表 2 により，Topic2 の平均時間間隔は著しく長くなっており，提案手法によって時間間隔を考慮したトピックが生成されていることがわかる．また，購買に至るセッションが所属する確率が高い Topic5 と離脱に至るセッションが所属する確率が高い Topic3 の詳細は表 3 で示す．

表 3: 特定なトピックに所属する確率 Top10 のページ

ランク	Topic 3 に属するページ内容	Topic 5 に属するページ内容
1	離脱	グーグル検索
2	ショッピングカートページ	オーダー確認完了ページ
3	お供え・お悔やみの献花	購買
4	購入をあきらめるページ	支払ページ
5	母の日関連ページ	サインイン確認ページ
6	サインイン完了ページ	配送関連ページ
7	開店祝い（customer）	購買完了ページ
8	マイページ	母の日
9	ユーザ ID 忘れ	クリスマス特集
10	お祝い：人気の高い順	法人注文関連ページ

表 3 は Topic3 と Topic5 の所属確率トップ 10 ページの内容を示している．これらのうち，Topic3 の 1 位である「離脱」と Topic5 の 3 位である「購買」は，購入するか否か，または購入する商品の選定という閲覧プロセスで見られるページではなく，購買の結果として示されるページである．

Topic 3 に属するページ内容の中には，購入をあきらめるページやユーザ ID 忘れなどの離脱の可能性があるページが含まれ，1 位の「離脱」と整合している．ショッピングカートページの順位が高いことは，「かご落ち」現象があるためと考えられる．また，サインイン完了ページが出現することから，ユーザは複数回閲覧し，吟味している可能性が高いと考えられる．

Topic 5 に属するページ内容の中には，購買と関連するページが複数出現しており，提案手法によって購買確率が高くなるページが集約できていることがわかる．また，イベント特集に関する購買確率が高く見られ，ユーザは商品をアクセスする前から用途を決めている方が購買確率が高くなる．

6.3. 応用

本節では，提案手法による分析結果が，具体的な販売促進施策にどう結び付くかについて考察する．提案手法により，閲覧ページにおける各ユーザの購買傾向が把握できることから，非購買トピックから購買トピックへの変換要素であるページを発見し，購買確率向上の施策が実施可能となるかについて検討してみる．表 4 に各ページにおける非購買トピックから購買トピックへ変換する回数のランクを示す．

表 4: Topic3 から Topic5 へ変換回数 Top10

ランク	ページ内容
1	購買
2	グーグル検索
3	検索キーワード：お祝い：人気の高い順
4	クリスマスの特集
5	会員登録情報変更
6	四十九日法要以降に贈る献花
7	母の日
8	検索キーワード：お祝い バラ：一致度順
9	当日配達特急
10	9 月の誕生花カレンダー

表 4 の 5 位である会員登録情報変更ページと 9 位である当日配達特急ページはユーザの意思の変換が推察できる内容である．また，検索キーワードが 2 回出現しており，検索という行動から購買確率が高まることが伺える．検索内容に注目すると，ユーザがどのような商品に興味を持つかがわかる．他にクリスマス特集や 9 月の誕生花カレンダーのような内容が出現したため，これらに関連したメルマガや UI 設計に通じて，購買確率向上に繋がると考えられる．

7. まとめと今後の課題

本研究では，EC サイトの閲覧履歴データに対し，購買履歴と閲覧時間間隔を考慮した改良型 Latent LSTM Allocation モデルを提案した．実データ分析の結果を通じ，提案モデルにより購買傾向が異なるグループを抽出できる．また，実際の購買に応用する際，トピック変換要素を探し，販売促進施策を検討するための材料となる．今後の課題としては，ユーザ今後の思考状態の変化をある程度推定できるようなモデルの構築などが考えられる．

謝辞

貴重な実データを提供頂いた花キューピット株式会社に厚く御礼を申し上げます．

参考文献

[1] Blei, D., Ng, A., and Jordan, M., “Latent Dirichlet Allocation”,*Journal of Machine Learning Research*, vol.3, pp.993–1022, 2003.

[2] Sepp Hochreiter; Jürgen Schmidhuber, “Long Short-Term Memory”, *Neural Computation*, vol.9(8), pp.1735–1780,1997.

[3] M. Zaheer, A. Ahmed, and A. J. Smola., “Latent LSTM Allocation: Joint Clustering and NonLinear Dynamic Modeling of Sequence Data”, *In ICML*, 2017.