データ拡張による Biterm Topic Model の解釈性向上法に関する研究

1X18C067-0　西田有輝
指導教員　後藤正幸

1. 研究背景と目的

EC サイト上に蓄積されている購買履歴データの特徴として，購買数の少ないライトユーザが多く含まれるという点が挙げられる．こうしたデータに対して，Latent Dirichlet Allocation[1]（以下，LDA）などの代表的なトピックモデルを適用した場合，ユーザごとの少ない購買データからトピックを推定するため，推定精度が低下してしまう．そのため，ライトユーザが多く含まれる購買データに対してもトピックを適切に推定可能なモデルとして，Biterm Topic Model[2]（以下，BTM）が提案されている．BTM は，ユーザの購買履歴に含まれる 2 つのアイテムのペア（バイターム）に同一のトピックを仮定したモデルであり，同時購買確率が高いバイタームを重視してトピックを学習する．しかし実際には，同時購買確率は低くとも，アイテム x が購買されたもとでアイテム y が購買される条件付購買確率が高いバイタームの方が，マーケティング施策立案などのビジネス目的では重要性が高い．こうしたバイタームを重視した学習が可能となれば，ユーザの嗜好に関連する 2 つのアイテムが同一のトピックに所属しやすくなり，ユーザの嗜好を捉えたトピックの推定と，応用面での有用な分析結果が期待できる．

そこで，本研究では，条件付購買確率が高いアイテムのペアを関連性のあるバイタームと定義し，これらのバイタームを重視した新たな学習方法を提案する．加えて，人工データと実データに対して提案手法を適用し，その有効性を検証する．

2. Biterm Topic Model

BTM は，ユーザの購買履歴に含まれる 2 つのアイテムのペアをバイタームと定義し，このバイタームごとに同一のトピックを仮定したモデルである．BTM のグラフィカルモデルを図 1，モデル式を式 (1) に示す．

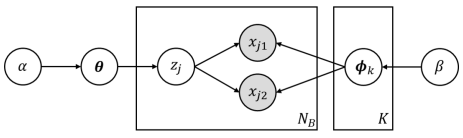

図 1: BTM のグラフィカルモデル

$$P(\boldsymbol{\mathcal{B}}|\boldsymbol{\theta},\boldsymbol{\Phi}) = \prod_{j=1}^{N_B}\sum_{k=1}^{K}\theta_k\phi_{kx_{j1}}\phi_{kx_{j2}} \tag{1}$$

ここで，K はトピック数，N_B は総バイターム数を表す．BTM は，ユーザ集合全体のトピック分布 $\boldsymbol{\theta}$ をもち，バイターム集合 $\boldsymbol{\mathcal{B}}$ に含まれる各バイターム（2 つのアイテム x_{j1}, x_{j2} のペア）は，$\boldsymbol{\theta}$ に従って選択されたトピック z_j に対応するアイテム分布 $\boldsymbol{\phi}_{z_j}$ に従って生成される．また，α と β はそれぞれトピック分布とアイテム分布が従うディリクレ分布のハイパーパラメータを表す．BTM では，崩壊型ギブスサンプリングにより，パラメータ $\boldsymbol{\theta}, \boldsymbol{\Phi}$ を推定する．

BTM では，ユーザがもつバイタームのトピックを集計することで各ユーザのトピック分布 $\boldsymbol{\theta}_u$ を推定でき，1 人のユーザが複数のトピックのもとで購買するという状況をモデル化している．さらに，ユーザ集合全体のトピック分布 $\boldsymbol{\theta}$ を推定するため，十分なユーザ数が存在する場合，ライトユーザが多く含まれるデータに対しても適切な推定が可能である．

3. 提案手法

3.1. 概要

BTM では，同時購買確率が高いバイタームを重視してトピックを学習する．そのため，多くのユーザに購買される人気アイテムを含むバイタームは，同一のトピックで購買されるアイテム（トピックアイテム）のペアで構成されたバイタームよりも同時購買確率が高く，重視して学習されてしまう．しかし，トピックアイテム同士で構成されたバイタームは，条件付購買確率が高く，ユーザの嗜好を表現していると考えられる．従って，このような関連性のあるバイタームを重視して学習することで，ユーザの嗜好を捉えた解釈性の高いトピックの推定が可能となる．

そこで提案手法では，バイターム集合の中から関連性のあるバイタームを抽出し，それらを複製することで新たな学習用バイタームを構築する．これにより，関連性のあるバイタームを重視した学習が可能となる．また，関連性の指標として，式 (2) で定義されるリフト値を用いる．

$$\mathrm{lift}(x,y) = \frac{\hat{P}(y|x)}{\hat{P}(y)} = \frac{U(x,y)/U(x)}{U(y)/U} \tag{2}$$

ここで，U はユーザ数，$U(x)$ はアイテム x を購買したユーザの数を表す．リフト値は，あるアイテム x が他のアイテム y の購買確率の推定値をどのくらい向上させるのかを表し，1 以上の場合には，アイテム x がアイテム y の購買確率を向上させたことを示すため，2 つのアイテム間には関連性があると判断できる．

3.2. 学習用バイタームの構築

提案手法では，下記の手順で学習用バイタームを構築し，BTM による学習を行う．

1) 各ユーザの購買履歴の中で，2 つの購買アイテムの全ての組み合わせをバイタームと定義し，ユーザ全体で N_B 組の元バイタームを得る．

2) リフト値が閾値 l 以上となる N_L 種類のバイタームを抽出後，各バイタームに対して一律に o 倍し，合計 $N_L \times o$ 組の追加バイタームを得る．

3) N_B 組の元バイタームに $N_L \times o$ 組の追加バイタームを加えた，N_B' 組のバイタームを学習用バイタームとする．

4. 人工データを用いた評価実験

4.1. 人工データの生成

本実験では，少ない購買の中でトピックアイテムと人気アイテムが混在するような人工データを下記の手順で生成する．

1) 特定のユーザとアイテムを高確率で生成するための K 個のトピックグループと，全ユーザが高確率で購買するア

イテムを生成するための 1 個の人気グループを仮定する．各トピックグループにはユーザが U_T 人，アイテムが I_T 個，人気グループにはアイテムが I_P 個所属する．

2) トピック k のアイテム分布 $\boldsymbol{\phi}_k$ において，トピックグループ k のアイテムの購買確率を t_k，人気グループのアイテムの購買確率を p_k，それ以外のトピックグループのアイテムの購買確率を $1-t_k-p_k$ とする．また，各グループ内でアイテムは等確率で購買されるとする．

3) トピックグループ k のユーザのトピック分布 $\boldsymbol{\theta}_u$ において，トピック k への所属確率を w_k，トピック k 以外の各トピックへの所属確率を $(1-w_k)/(K-1)$ とする．

4) 購買数の最小値を 2，最大値を R とし，式 (3) で定義されるパレート分布から購買数 N_u をサンプリングする．ここで，b はパレート分布の形状パラメータを表す．

$$f(N_u, b) = \frac{b}{(N_u - 1)^{b+1}} \tag{3}$$

5) LDA の生成過程に従って，購買数が式 (3) で与えられる購買履歴データを生成する．

4.2. 実験条件

人工データの生成では，$K = 5, U_T = 200, I_T = 10, I_P = 5, w_k = 0.8, R = 20, b = 1.3$ とした．本実験では，トピックグループのアイテムの購買確率 t_k と，人気グループのアイテムの購買確率 p_k を変化させ，提案手法の挙動を確認する．また，t_k, p_k, w_k は全ての k に対して同じ値を設定した．また，ハイパーパラメータは $\alpha = 0.1, \beta = 0.1, l = 2, o = 15$，比較手法として LDA，BTM を用いた．

4.3. 実験結果と考察

t_k,p_k を変化させたときの Coherence，Perplexity を表 1 に示す．ここで，Coherence は値が大きいほどトピックに一貫性があること，Perplexity は値が小さいほどモデルの予測性能が高いことを意味する．また，$t_k = 0.50$,$p_k = 0.45$ のときの各バイタームに対するリフト値をヒートマップで可視化した結果を図 2 に示す．

表 1: t_k, p_k を変化させたときの評価指標の比較
（左：Coherence，右：Perplexity）

手法	(t_k, p_k)					
	(0.70,0.25)		(0.50,0.45)		(0.30,0.65)	
LDA	−83.7	26.2	−100.9	34.8	−110.2	49.0
BTM	−77.1	31.4	−96.5	41.0	−104.3	57.4
提案手法	**−73.5**	**24.2**	**−72.4**	**28.8**	**−84.8**	**44.7**

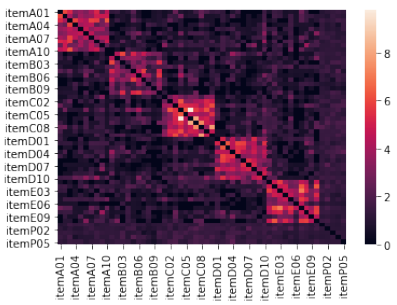

図 2: リフト値のヒートマップ

表 1 より，提案手法は従来手法よりも常に一貫性のあるトピックを推定可能で，かつ高い精度で予測可能であることが分かる．図 2 より，同一のトピックに所属するアイテムで構成された関連性のあるバイタームをリフト値によって抽出可能であることが視覚的にも確認できる．

5. 実データ分析

5.1. 分析条件

Amazon レビューデータ [3] を購買履歴データとみなして分析に用いる．期間は 2017 年 1 月〜12 月，ユーザ数は 16,623 人，アイテム数は 679 個，ユーザの平均購買数は 3.2 個である．また，ハイパーパラメータは $\alpha = 0.1, \beta = 0.1, l = 5, o = 2$，比較手法として LDA，BTM を用いた．

5.2. 分析結果と考察

トピック数 K を変化させた場合の Coherence を表 2，$K = 10$ のときに BTM と提案手法で共通するトピックのアイテム分布を比較した結果を表 3 に示す．

表 2 より，提案手法は従来手法よりも常にトピックに一貫性があることが分かる．表 3 を見ると，提案手法ではトピックに関連性のあるアイテムがより上位に位置していることが確認できる．以上より，提案手法はリフト値に基づきデータを拡張した学習用バイタームを用いて BTM を学習することで，解釈性の高いトピックの推定を可能にしているといえる．

提案手法の実応用として，推定した各ユーザのトピック分布 $\boldsymbol{\theta}_u$ を用いて，所属確率の高いトピックのアイテムを推薦したり，クーポンを発行するなどの施策が考えられる．

表 2: Coherence の比較

手法	$K = 5$	$K = 10$	$K = 15$
LDA	−179.0	−193.5	−198.7
BTM	−148.2	−150.0	−153.3
提案手法	**−138.2**	**−144.7**	**−141.2**

表 3: 共通トピックの比較

	トピック「冷凍食品」		トピック「調味料」	
	BTM	提案手法	BTM	提案手法
1	**アイスクリーム**	**アイスクリーム**	**コショウ**	**コショウ**
2	コーヒー	**ミートレス料理（冷凍）**	**シナモン**	**シナモン**
3	サラダ	**鶏肉料理（冷凍）**	コーヒー	**パプリカパウダー**
4	鶏肉	**ジェラート**	**ガーリックパウダー**	**クミン**
5	ソーセージ	バナナ	**パプリカパウダー**	**オニオンパウダー**
Coherence	−148.0	**−108.5**	−156.3	**−146.6**

6. まとめと今後の課題

本研究では，ライトユーザが多く含まれる購買履歴データに対してもトピックを適切に推定可能なモデルである BTM をベースとし，トピックの解釈性向上を目的とした新たな学習方法を提案した．また，人工データと実データの双方に対して提案手法の有効性を示した．今後の課題として，バイタームの追加個数の検討や，推定した各ユーザのトピック分布 $\boldsymbol{\theta}_u$ を用いたクラスタリングの実施などが挙げられる．

参考文献

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[2] X. Cheng, X. Yan, Y. Lan, J. Guo, “BTM: Topic Modeling over Short Texts,” *IEEE Transactions on Knowledge And Data Engineering*, Vol.26, No.12, pp.2928-2941, 2014.

[3] J. Ni, J. Li, J. McAuley, “Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects,” *Proceedings of the 2019 Conference on EMNLP and the 9th IJCNLP*, pp.188-197, 2019.