子育て支援 QA サイトにおける潜在嗜好変化の時系列推定

住谷祐太[1]　富川雄斗[2]　伊藤尚紀[2]　高橋里司[3]
[1] 電気通信大学大学院 情報理工学研究科 情報学専攻　[2] 電気通信大学 情報理工学域 I 類
[3] 電気通信大学大学院 情報理工学研究科 情報・ネットワーク工学専攻
{sumiya,tomikawa,ito,stakahashi}@uec.ac.jp

概要

ソーシャルメディアを活用した子育て情報の収集が盛んになっている．特に子育てに関する QA コミュニティでは，子育て中の母親が質問を気軽に投稿できるため，投稿されるテキストには一般的な情報だけでなく，ユーザが抱える悩みや嗜好が含まれていることが多い．それらの特徴量を把握することはソーシャルメディアマーケティングにおいて重要な技術課題となる．そこで本研究では，ユーザの潜在嗜好を捉えるための新しいトピックモデルを提案する．提案手法では，ユーザの年齢区分や子供の月齢などの補助情報の活用や，嗜好の時間依存性を考慮する．実験では令和 3 年度データ解析コンペティションでコネヒト株式会社から提供されたママリのデータを使用し，手法の有効性を検証する．

1　はじめに

子育て情報の収集・共有を行う QA コミュニティの一つに，コネヒト株式会社が提供するママリが存在する．ママリでは匿名性を利用して知人には相談できないような悩みを質問できるため，投稿される質問はユーザの潜在嗜好を把握する上で有益な情報だと言える．子育て特有の事情として，ユーザの嗜好は属性や子供の月齢，時間によって異なると考えられ，従来のユーザ全体を解析の対象とした場合では十分に嗜好を把握できない問題がある．

関連研究として，ユーザの質問文から潜在嗜好を推定する手法にトピックモデルが存在し，自然言語の潜在的意味解析に広く用いられている．特に Blei et al.［1] が提案した LDA (Latent Dirichlet Allocation) は，一つの文書に複数の潜在トピックが存在すると仮定し，そのトピックの分布を離散分布としてモデル化する代表的な潜在変数モデルである．

一方で新生児の母親が主なユーザである QA コミュニティ上での質問は，文書となる質問文以外に質問者の年齢や居住地域，そして事前に付与される質問カテゴリタグなどが付与され，それに応じて質問の意図が大きく異なると想定される．また子供を育てるユーザの場合，月齢に応じて質問内容が変化していくことも考慮する必要がある．本研究では，ユーザの潜在嗜好を捉えるためにこれらの補助情報を利用した新しいトピックモデルを提案する．

2　提案手法

QA コミュニティ上での課題を解決する新しいトピックモデルとして，SCTTM (Supervised Correspondence Topic Tracking Model) を提案する．SCTTM では，LDA に対して三つの拡張を行う．はじめに，最初の質問時における子供の月齢を考慮するようにトピックを学習させるため，連続変数を考慮したトピックを推定できる sLDA [2] を組み込む．sLDA は，学習過程のトピックを説明変数にとり，それらで月齢を線形回帰したときの当てはまりが良くなるようにトピックと偏回帰係数を逐次的に学習する．

次に，離散値ラベルの補助情報も考慮してトピックを学習させるため，Corr-LDA [3] を取り入れる．Corr-LDA を取り入れることにより，各トピックの単語分布の推定と同様に，補助情報の分布も推定する．ここで補助情報の生成に用いられるトピックは，必ず単語を生成したトピックとなるようにモデル化するため，単語と補助情報の対応関係を適切に学習することができる．

最後に，ユーザの嗜好変化を考慮するための手法として，TTM [4] を用いたパラメータ推定を行う．TTM は時間情報が付けられた文書集合から，時間変化するトピックを推定する手法であるため，質問文からユーザ個人の潜在嗜好と時間発展を適切に追跡することができる．また TTM はオンライン学習が可能な点から，日々蓄積されるユーザの質問データを効率的に学習できる利点がある．本研究では，

最初の時刻 $t=1$ で月齢と補助情報に沿ったトピックを学習した後，時刻 $t>1$ 以降はそれらのトピックが時間依存して変化するように学習させる．

2.1 トピックの生成

全ユーザ数を U，総時刻を T，全ユーザで観測される総単語の種類を V，トピック数を K として，ある時刻 t においてあるユーザ u が生成した文書の単語の組を $\boldsymbol{w}_{tu}=(w_{tu1},\cdots,w_{tuN_{tu}})$ と表記する．ここで N_{tu} は時刻 t においてユーザ u が生成した文書に含まれる単語数である．

トピックモデルでは，ユーザ毎にトピック分布 $\boldsymbol{\theta}_{tu}=(\theta_{tu1},\cdots,\theta_{tuK})^\top\in\mathbb{R}^K$ が与えられ，各要素 θ_{tuk} は時刻 t でユーザ u の単語にトピック k が割り当てられる確率を表す．すなわち $\theta_{tuk}\geq 0$，$\sum_k\theta_{tuk}=1$ を満たす．各トピック k には固有の単語分布 $\boldsymbol{\phi}_{tk}=(\phi_{tk1},\cdots,\phi_{tkV})^\top\in\mathbb{R}^V$ が存在し，各要素 ϕ_{tkv} は時刻 t のトピック k で単語 v が生成される確率を表す．ここで $\phi_{tkv}\geq 0$，$\sum_v\phi_{tkv}=1$ を満たす．

生成過程における各単語 w_{tun} は，トピック分布 $\boldsymbol{\theta}_{tu}$ にしたがってトピック z_{tun} $(n=1,\cdots,N_{tu})$ が割り当てられ，その単語分布 $\boldsymbol{\phi}_{tz_{tun}}$ にしたがって生成されると仮定する．

2.2 トピックの推定

はじめに，ユーザが投稿した質問単語ごとにトピックをサンプリングする．本研究では崩壊型ギブスサンプリング [6] を用いて，トピック分布，単語分布，補助情報分布のパラメータを周辺化した周辺同時分布により単語トピックのサンプリング確率を求める．時刻 $t=1$ において，ユーザ u の単語 n に割り当てられるトピック z_{tun} が k となる確率は，周辺同時分布にベイズの定理を適用することで式 (1) のように求めることができる．式中の $\backslash tun$ は，時刻 t でユーザ u の単語 n に割り当てられるトピックを除くことを意味する．また N_{tkn} は時刻 t においてトピック k に単語 n が割り当てられた個数，N_{tk} はその単語毎の総和を表し，N_{tuk}，M^i_{tuk} はそれぞれ時刻 t でユーザ u に割り当てられた単語トピック，補助情報 i のトピック k の個数を表す．さらに時刻 $t=1$ では，ユーザ u が持つ連続変数 y_u がガウス分布 $\mathcal{N}(\boldsymbol{\eta}^\top\widetilde{\boldsymbol{z}}_{tu},\sigma^2)$ に従い生成されると仮定する．$\boldsymbol{\eta}\in\mathbb{R}^K$ は線形回帰パラメータ，$\widetilde{\boldsymbol{z}}_{tu}=\left(\frac{N_{tu1}}{N_{tu}},\cdots,\frac{N_{tuK}}{N_{tu}}\right)^\top\in\mathbb{R}^K$ は時刻 $t=1$ におけるユーザ u のトピック割合，σ^2 は分散パラメータを表す．

$$
\begin{aligned}
&p(z_{tun}=k\mid \mathrm{Z}_{\backslash tun},\mathrm{W}_t,\boldsymbol{\alpha},\beta,\boldsymbol{\eta},\sigma^2)\\
&\propto(N_{tuk\backslash tun}+\alpha_k)\frac{N_{tkw_{tun\backslash tun}}+\beta}{N_{tk\backslash tun}+\beta V}\\
&\times\left(\frac{N_{tuk\backslash tun}+1}{N_{tuk\backslash tun}}\right)^{\sum_i M^i_{tuk}}\\
&\times\exp\left\{\frac{\eta_k}{N_{tu}\sigma^2}\left(y_u-\boldsymbol{\eta}^\top\widetilde{\boldsymbol{z}}_{\backslash tun}-\frac{\eta_k}{2N_{tu}}\right)\right\}.
\end{aligned}
\tag{1}
$$

$\boldsymbol{\alpha}\in\mathbb{R}^K$, β はディリクレ分布のパラメータである．便宜上，$\boldsymbol{\theta}_{tu}$，$\boldsymbol{\phi}_{tk}$ を行列表現したものを $\boldsymbol{\Theta}_t$，$\boldsymbol{\Phi}_t$ と表し，$\boldsymbol{w}_{tu}$，$\boldsymbol{z}_{tu}$ をそれぞれ要素毎にまとめた集合を W_t，Z_t と表す．なお式 (1) で条件付き独立となる項およびその変数については省略している．

時刻 $t>1$ でも同様にトピックのサンプリング確率を式 (2) のように求めることができる．式中の $\hat{\boldsymbol{\Theta}}_{t-1},\hat{\boldsymbol{\Phi}}_{t-1}$ はそれぞれ $\boldsymbol{\Theta}_{t-1},\boldsymbol{\Phi}_{t-1}$ の推定値である．

$$
\begin{aligned}
&p(z_{tun}=k\mid \mathrm{Z}_{\backslash tun},\mathrm{W}_t,\boldsymbol{\alpha}_t,\boldsymbol{\beta}_t,\hat{\boldsymbol{\Theta}}_{t-1},\hat{\boldsymbol{\Phi}}_{t-1})\\
&\propto\left(N_{tuk}+\alpha_{tu}\hat{\theta}_{t-1,uk}\right)\\
&\times\frac{N_{tkw_{tun}\backslash tun}+\beta_{tk}\hat{\phi}_{t-1,kw_{tun}}}{N_{tk\backslash tun}+\beta_{tk}}\\
&\times\left(\frac{N_{tuk\backslash tun}+1}{N_{tuk\backslash tun}}\right)^{\sum_i M^i_{tuk}}.
\end{aligned}
\tag{2}
$$

$\boldsymbol{\alpha}_t\in\mathbb{R}^U$, $\boldsymbol{\beta}_t\in\mathbb{R}^K$ はディリクレ分布のパラメータを表し，それ以外に条件付き独立となる項およびその変数については省略している．

3 評価実験

提案手法について実データを用いた検証を行う．検証では子供の月齢を登録している質問数が 5 回，10 回，20 回のユーザをそれぞれ 250 人ずつ無作為に抽出し，計 750 人のユーザを用いる．各ユーザが投稿した質問単語を 9:1 の比率で訓練単語とテスト単語に分割し，トピック数 K を 5 から 20 まで 5 刻みに増やした時の各々について，月齢予測に必要なパラメータとトピック分布 $\boldsymbol{\Theta}_t$，単語分布 $\boldsymbol{\Phi}_t$ を学習する．また，[5] では，Perplexity [1] を用いた最適なトピック数 $K=25$ を求めて，実験を行なっている．

3.1 月齢の予測精度の有効性

最初の質問時 $(t=1)$ における，各ユーザの子供の月齢を SCTTM による逐次的な線形回帰で予測し，その精度を検証する．モデルはユーザを訓練用とテスト用に 8:2 で分けて，各トピック数毎に MSE を用いて 5 分割交差検証で評価し，表 1 の結果を得

表 1　子供の月齢の予測結果：訓練，テストはそれぞれのデータでの MSE 値を表す．

	LDA		RandomForest		LightGBM		SCTTM	
トピック数	訓練	テスト	訓練	テスト	訓練	テスト	訓練	テスト
5	658.522	681.336	104.643	781.514	159.933	825.258	629.174	678.055
10	646.163	714.964	101.231	808.721	89.994	835.817	477.400	580.834
15	641.037	692.885	101.330	735.052	70.126	755.990	276.611	426.931
20	632.810	704.458	99.673	776.600	**61.745**	818.892	153.497	**219.463**

表 2　平均 Top-N-accuracy(%) による質問単語の予測結果

		質問数 5			質問数 10			質問数 20		
	トピック数	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
LDA	5	13.750	23.000	31.000	12.555	21.555	27.333	11.842	20.000	26.421
	10	13.750	23.500	29.750	13.222	20.888	27.555	11.736	19.157	25.157
	15	13.750	23.500	30.000	11.333	19.555	25.666	11.263	19.052	25.473
	20	13.250	22.500	29.500	10.777	19.555	26.555	11.684	19.894	25.631
TTM	5	14.250	23.750	31.500	13.222	22.333	29.333	12.473	20.631	27.210
	10	15.000	24.250	33.250	13.222	22.444	29.444	13.421	22.473	28.894
	15	15.250	24.000	32.750	13.222	22.888	30.555	13.631	21.684	28.263
	20	15.250	25.250	32.000	13.444	21.555	28.777	13.842	21.947	27.842
CTTM	5	13.000	22.500	30.250	11.666	19.888	27.333	11.736	20.368	27.052
	10	14.000	23.000	31.250	12.888	21.888	28.666	11.631	20.052	26.894
	15	15.250	22.000	30.500	12.777	21.888	28.111	11.578	19.842	26.631
	20	13.500	22.750	30.250	13.111	21.000	27.111	12.210	20.421	27.000
SCTTM	5	15.250	24.750	34.250	**15.666**	**25.111**	**32.444**	14.263	**24.000**	**31.421**
	10	14.750	23.750	31.750	14.444	23.666	30.333	13.894	23.526	30.315
	15	14.500	24.750	32.750	14.111	23.000	29.777	13.578	22.473	28.894
	20	**16.250**	**27.250**	**34.500**	14.222	23.555	30.000	**14.526**	23.421	29.210

た．本研究では月齢予測が有効であることを示すために，実験において月齢を考慮しない通常の LDA，RandomForest [7]，LightGBM [8] の三つのモデルと比較している．この結果から，訓練ユーザに関しては，LightGBM や RandamForest の MSE の精度が向上しているが，これは過学習を起こしているためであると考えられる．一方で，テストユーザに関しては SCTTM が他の手法に比べて MSE の精度が向上していることが確認できる．

3.2　質問単語の時系列推定の有効性

各時刻で質問される単語の推定精度を複数のモデルと比較しながら評価する．比較するモデルは，LDA，TTM，TTM に月齢以外の補助情報を加えた CTTM，提案手法の SCTTM の四つである．評価指標には次式で定義される Top-N-accuracy を採用する．

$$P(w_{tu} = v \mid \hat{\boldsymbol{\Theta}}_{t-1}, \hat{\boldsymbol{\Phi}}_{t-1}) = \sum_{k=1}^{K} \hat{\theta}_{t-1,uk} \hat{\phi}_{t-1,kv}. \qquad (3)$$

式 (3) は，現時刻 t で観測された単語 v をテスト単語とし，その生成確率を一つ前の時刻 $t-1$ の単語集合から推定されたトピック分布 $\hat{\boldsymbol{\Theta}}_{t-1}$ と単語分布 $\hat{\boldsymbol{\Phi}}_{t-1}$ で予測する．出力として，単語 v の生成確率が上位 N 件に含まれる割合を返す．この値が高いほど，精度良く単語を予測できていることを意味する．Top-N-accuracy を使い，質問回数が 5 回，10 回，20 回のユーザそれぞれについて $N = 1, 2, 3$ 件での平均を，トピック数を変えながら評価した結果を表 2 に示す．この結果から，時間依存性と補助情報，連続変数の月齢を考慮した SCTTM が他のモデルに比べて高い精度を示していることが確認できる．

4 おわりに

ユーザ毎に時間発展して観測される文書と，それらに対応する補助情報が観測される QA サイトにおけるソーシャルメディアマーケティングにおいて，ユーザの潜在嗜好を推定することは，広告推薦や新たなユーザコネクションの推薦において重要な技術課題である．本研究では，この技術課題を解決する新しいトピックモデルである SCTTM を提案した．実験では令和 3 年度データ解析コンペティションにおいてコネヒト株式会社のママリのデータを使い，月齢や補助情報を考慮しつつ，ユーザ毎に時間発展する嗜好変化を正しく推定できることを示した．また，抽出するトピック数を変化させた場合，少ないトピック数に対して時系列推定の精度は良いが，月齢予測の精度が悪いことを確認し，トレードオフの関係にあることを確認した．両者のバランスがよいトピック数として，20 が妥当であることを確認し，[5] での結果と比較すると Perplexy によってトピック数を決定することが最適とは限らないことが分かった．

本研究で提案した SCTTM は時間発展を考慮できる他のソーシャルメディアの文書データに対しても適用可能であり，汎用性の高いモデルであると言える．

謝辞

新生児・乳幼児の母親をメインユーザとするポータルアプリのデータを提供いただいた，コネヒト株式会社様およびデータ解析コンペティション運営の方々に感謝申し上げます．また，DB サーバの提供やコンペティション参加の支援をしていただいた電気通信大学情報工学工房および，学術技師の島崎様に感謝申し上げます．

参考文献

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, **3**, pp. 993–1022, 2003.

[2] J. Mcauliffe and D. Blei, “Supervised topic models,” In *Proceedings of Advances in Neural Information Processing Systems*, **20**, pp. 121–128, 2007.

[3] D. M. Blei and M. I. Jordan, “Modeling annotated data,” In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion R*etrieval, pp. 127–134, 2003.

[4] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, “Topic tracking model for analyzing consumer purchase behavior,” In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, **9**, pp. 1427–1432, 2009.

[5] 住谷 祐太, 富川 雄斗, 伊藤 尚紀, 高橋 里司, “SCTTM によるユーザ属性を考慮した 潜在嗜好変化の時系列推定”，オペレーションズ・リサーチ，**68**(2), 2023.

[6] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” In *Proceedings of the National Academy of Sciences*, **101**, pp. 5228–5235.

[7] L. Breiman, “Random forests,” *Machine Learning*, **45**, pp. 5–32, 2001.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” In *Proceedings of Advances in Neural Information Processing Systems*, **30**, pp. 3146–3154, 2017.