事例研究［論文］

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

住谷　祐太，富川　雄斗，伊藤　尚紀，高橋　里司

1. はじめに

近年日本における女性の社会進出が進んでいるが，子育てについては母親への負担が依然として高い傾向にある [1]．特に 3 世帯同居世帯の割合が減少し，日常的に祖父母が育児に対して協力する機会が失われつつある [2]．その反面，スマートフォンなどの普及により，インターネットを活用して子育て情報を入手する母親は 7 割を超えている [1]．

スマートフォンを利用する母親が増えて以降，子育て情報をより集めやすくするための QA コミュニティの需要が高まっている．コネヒト株式会社が提供するママリもその一つであり，育児などに関する質問をユーザが投稿し，それに対して別のユーザが回答をすることで悩みの迅速な解決を促している．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うえで有益な情報だといえる．しかしながら，ユーザの嗜好は属性や子供の月齢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．

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ことに加え，得られるトピックに対して自動でラベル付けを行う．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．

すみや　ゆうた
電気通信大学大学院情報理工学研究科情報学専攻
とみかわ　ゆうと，いとう　なおき
電気通信大学情報理工学域 I 類コンピュータサイエンスプログラム
たかはし　さとし
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
〒 182–8585 東京都調布市調布ケ丘 1–5–1
受付 22.7.15　採択 22.10.20

2. 関連研究

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. [3] 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．LDA においては，文書中の各単語のトピックはその文書のトピック分布から確率的に割り当てられ，その割り当てられたトピックの単語分布に従って単語が生成されると仮定する．

LDA は文書に含まれる単語のみから潜在的意味解析を行うモデルであるが，多くの文書データは文書のみの情報に留まらず，その文書のカテゴリタグや評価区分などの離散値ラベルが付与されることがある．そのような補助情報をモデルに取り入れる手法として，Blei and Jordan [4] が提案した Corr-LDA (Correspondence LDA) は，単語を生成したトピックモデルのみから補助情報の分布を推定するように LDA を拡張している．Corr-LDA では，補助情報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になるため，単語と補助情報の関係を適切にモデル化することができる．さらに Mcauliffe and Ble [5] が提案した sLDA (supervised-LDA) は，文書に含まれる単語トピックのベクトルから，補助情報を線形回帰で予測できるように LDA を拡張している．sLDA では補助情報として連続変数を適用できるため，文書と連続変数が対応して観測されるような潜在的意味解析に有用である．

一方で，文書の時間発展を考慮したトピックモデルも提案されている．Blei and Lafferty [6] が提案した DTM (Dynamic Topic Model) では，文書集合全体のトピックの時間発展を追跡できるように LDA を拡張させ，そのハイパーパラメータに時間依存性を追加して

いる．また Iwata et al. [7] が提案した TTM (Topic Tracking Model) では，文書集合全体の時間発展ではなく，個別の文書に着目したときのトピック分布とその単語分布が時間発展するように LDA を拡張している．

3. 提案手法

本節では，まず 3.1 節において提案手法である SCTTM について説明し，続く 3.2 節以降で SCTTM を用いた学習と解釈性のためのトピックの自動ラベル付けについて説明する．

3.1 SCTTM

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，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．これらの影響から，単に質問データ全体に対して LDA を適用するだけでは，地域差，時間軸などを考慮した適切な潜在嗜好変化を捉えることができない．そこで本研究では，ユーザの嗜好を適切に捉えられるように LDA と 2 節のモデルを取り入れた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 を組み込む．sLDA は，学習過程のトピックを説明変数に取り，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 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 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．

3.2 生成過程

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，ユーザごとにトピック分布 $\boldsymbol{\theta}_{tu}=(\theta_{tu1},\cdots,\theta_{tuK})^\top\in\mathbb{R}^K$ 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．ここで $\theta_{tuk}\geq 0$，$\sum_k\theta_{tuk}=1$ 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 $\boldsymbol{\theta}_{tu}$ に従ってトピック z_{tun} $(n=1,\cdots,N_{tu})$ が割り当てられ，その単語分布 $\boldsymbol{\phi}_{tz_{tun}}$ に従って生成されると仮定する．

3.2.1 時刻 $t=1$ における生成過程

SCTTM は LDA を拡張させたモデルである．LDA では，トピック分布 $\boldsymbol{\theta}_{tu}$ はカテゴリ分布のパラメータであることから，その共役事前分布である以下のディリクレ分布に従って生成されると仮定する．

$$P(\boldsymbol{\theta}_{tu}|\boldsymbol{\alpha})\propto\prod_k\theta_{tuk}^{\alpha_k-1}. \tag{1}$$

式 (1) において，$\boldsymbol{\alpha}\in\mathbb{R}^K$ はデータから推定されるディリクレ分布のハイパーパラメータで，$\boldsymbol{\alpha}>\boldsymbol{0}$ を満たす．また各トピックの単語分布も同様にカテゴリ分布のパラメータであることから，以下のディリクレ分布に従って生成されると仮定する．

$$P(\boldsymbol{\phi}_{tk}|\beta)\propto\prod_v\phi_{tkv}^{\beta-1}. \tag{2}$$

式 (2) において，$\beta\in\mathbb{R}$ はデータから推定されるディリクレ分布のハイパーパラメータで，$\beta>0$ を満たす．LDA では，ユーザの過去の文書から得られる潜在嗜好は考慮せず，時刻 t で観測された文書の潜在嗜好のみを推定する．SCTTM の時刻 $t=1$ におけるトピック分布 $\boldsymbol{\theta}_{tu}$ と単語分布 $\boldsymbol{\phi}_{tk}$ については，前の時刻の分布が存在しないため，式 (1) と式 (2) で示した LDA と同様のディリクレ事前分布に従うとする．

SCTTM の時刻 $t=1$ では，LDA に対して次の二つ

の拡張を行う．一つ目として，各トピックが特定の連続変数に従うようにsLDAを適用する．ユーザuがもつ連続変数y_uが ガウス分布$\mathcal{N}(\boldsymbol{\eta}^\top \tilde{\boldsymbol{z}}_{tu}, \sigma^2)$に従って生成されると仮定する．ここで$\boldsymbol{\eta} \in \mathbb{R}^K$は$K$次元の線形回帰パラメータ，$\tilde{\boldsymbol{z}}_{tu} = \left(\frac{N_{tu1}}{N_{tu}}, \cdots, \frac{N_{tuK}}{N_{tu}}\right)^\top \in \mathbb{R}^K$は時刻$t=1$におけるユーザ$u$のトピック割合を表す．ユーザのトピック分布と連続変数の関係を線形回帰で予測するため，ガウス分布の平均$\boldsymbol{\eta}^\top \tilde{\boldsymbol{z}}_{tu}$と分散$\sigma^2$をデータから学習する必要がある．

二つ目として，Corr-LDAによる補助情報を考慮したトピックの学習を行う．時刻tにおいてM_t種類の補助情報が観測されているとき，補助情報i $(i = 1, \cdots, M_t)$におけるユーザuの観測値の組を$\boldsymbol{x}^i_{tu} = (x^i_{tu1}, \cdots, x^i_{tuM^i_{tu}})$と定義する．ここで$M^i_{tu}$は補助情報$i$の観測数である．補助情報$i$の取り得る観測値の種類を$S^i$とすると，各トピック$k$には固有の補助情報分布$\boldsymbol{\psi}^i_{tk} = (\psi^i_{tk1}, \cdots, \psi^i_{tkS^i})^\top \in \mathbb{R}^{S^i}$が存在し，その各要素$\psi^i_{tks}$は時刻$t$のトピック$k$で補助情報$i$の観測値$s$が生成される確率を表す．ここで$\psi^i_{tks} \geq 0, \sum_s \psi^i_{tks} = 1$を満たす．各トピックの補助情報$i$の分布はカテゴリ変数のパラメータであることから，以下のディリクレ分布に従って生成されると仮定する．

$$P(\boldsymbol{\psi}^i_{tk} | \gamma^i_t) \propto \prod_s (\psi^i_{tks})^{\gamma^i_t - 1}. \tag{3}$$

式(3)において，$\gamma^i_t \in \mathbb{R}$はディリクレ分布のハイパーパラメータで，$\gamma^i_t > 0$を満たす．

生成過程における補助情報iの各観測値x^i_{tus}は，$\tilde{\boldsymbol{z}}_{tu}$のカテゴリ分布に従って補助情報トピック$y^i_{tum}$ $(m = 1, \cdots, M^i_{tu})$が割り当てられ，その補助情報分布$\boldsymbol{\psi}^i_{ty^i_{tum}}$に従って生成されると仮定する．補助情報の観測値の生成に用いられるカテゴリ分布のパラメータは，トピック割合$\tilde{\boldsymbol{z}}_{tu}$であるため，必ず単語を生成したトピックが割り当てられるように学習される．SCTTMの最初の時刻$t=1$における生成過程を次に示す．また提供データを当てはめたときのグラフィカルモデルを図1aに示す．

1. For トピック $k = 1, \cdots, K$
   (a) 単語分布を生成 $\boldsymbol{\phi}_{tk} \sim \mathrm{Dirichlet}(\beta)$
   (b) For 補助情報 $i = 1, \cdots, M_t$
      i. 補助情報分布を生成
      $\boldsymbol{\psi}^i_{tk} \sim \mathrm{Dirichlet}(\gamma^i_t)$
2. For ユーザ $u = 1, \cdots, U$
   (a) トピック分布を生成 $\boldsymbol{\theta}_{tu} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$
   (b) 連続変数を生成 $y_u \sim \mathcal{N}(\boldsymbol{\eta}^\top \tilde{\boldsymbol{z}}_{tu}, \sigma^2)$
   (c) For 単語 $n = 1, \cdots, N_{tu}$
      i. 単語トピックを生成
      $z_{tun} \sim \mathrm{Categorical}(\boldsymbol{\theta}_{tu})$
      ii. 単語を生成
      $w_{tun} \sim \mathrm{Categorical}(\boldsymbol{\phi}_{tz_{tun}})$
   (d) For 補助情報 $i = 1, \cdots, M_t$
      i. For 観測値 $m = 1, \cdots, M^i_{tu}$
         A. 補助情報トピックを生成
         $y^i_{tum} \sim \mathrm{Categorical}(\tilde{\boldsymbol{z}}_{tu})$
         B. 補助情報を生成
         $x^i_{tum} \sim \mathrm{Categorical}(\boldsymbol{\psi}^i_{ty^i_{tum}})$

3.2.2 時刻 $t > 1$ における生成過程

ユーザの潜在嗜好は時刻ごとに大きく変化しないと仮定し，時刻tのトピック分布$\boldsymbol{\theta}_{tu}$と単語分布$\boldsymbol{\phi}_{tk}$は，1時刻前の推定値$\hat{\boldsymbol{\theta}}_{t-1,u}$,$\hat{\boldsymbol{\phi}}_{t-1,k}$に依存して変化すると考える．これを定式化するために，TTMではトピック分布の事前分布として，以下のディリクレ分布を仮定する．

$$P(\boldsymbol{\theta}_{tu} | \hat{\boldsymbol{\theta}}_{t-1,u}, \alpha_{tu}) \propto \prod_k \theta^{\alpha_{tu}\hat{\theta}_{t-1,uk}-1}_{tuk}. \tag{4}$$

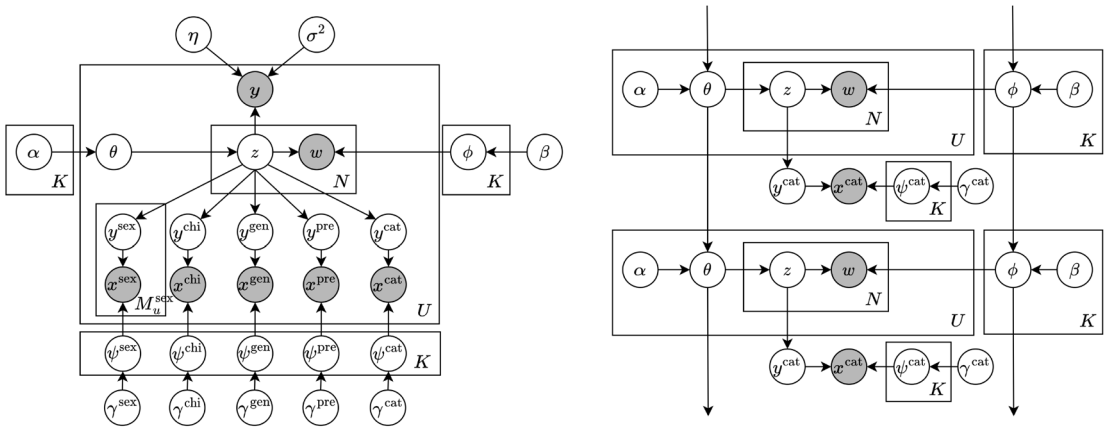
式(4)において，α_{tu}はユーザの興味がどれくらい変化しにくいかを表すハイパーパラメータである．α_{tu}が大きいほど，興味の分散が小さく，前時刻に対する嗜好変化が小さいことを意味する．この精度α_{tu}をデータから学習することで，個々のユーザの嗜好変化の度合いを推定することができる．単語分布についても同様に，事前分布として以下のディリクレ分布を仮定する．

$$P(\boldsymbol{\phi}_{tk} | \hat{\boldsymbol{\phi}}_{t-1,k}, \beta_{tk}) = \prod_v \phi^{\beta_{tk}\hat{\phi}_{t-1,kv}-1}_{tkv}. \tag{5}$$

式(5)において，β_{tk}は単語分布の一貫性を表すハイパーパラメータである．

補助情報については，時刻$t=1$のときと同様の方法で観測値を生成すると仮定する．このとき，補助情報の時間依存性は仮定せず，各時刻独立に推定する．時刻$t>1$以降のSCTTMの生成過程を次に示す．また提供データを当てはめたときのグラフィカルモデルを図1bに示す．

- For 時刻 $t = 2, \cdots, T$
  1. For トピック $k = 1, \cdots, K$



(a) 時刻 $t=1$ のとき　　(b) 時刻 $t>1$ のとき

図 1　SCTTM のグラフィカルモデル
色付きのノードは観測値を表す．補助情報の添字 sex, chi, gen, pre, cat はそれぞれ子供の性別，子供の数，世代，居住地域，質問カテゴリを表す．また観測値 y は子供の月齢を表す．複数の子供をもつユーザの場合，子供の性別も複数観測されるため，その観測数を M_u^{sex} と表す．それ以外の補助情報についてはユーザごとに一つ観測される．質問カテゴリのみは各時刻異なる値が観測されるため，時刻 $t>1$ でも補助情報分布を推定する．

(a) 単語分布を生成
$\boldsymbol{\phi}_{tk} \sim \text{Dirichlet}(\beta_{tk}\hat{\boldsymbol{\phi}}_{t-1,k})$
(b) For 補助情報 $i = 1, \cdots, M_t$
– 補助情報分布を生成
$\boldsymbol{\psi}^i_{tk} \sim \text{Dirichlet}(\gamma^i_t)$
2. For ユーザ $u = 1, \cdots, U$
(a) トピック分布を生成
$\boldsymbol{\theta}_{tu} \sim \text{Dirichlet}(\alpha_{tu}\hat{\boldsymbol{\theta}}_{t-1,u})$
(b) For 単語 $n = 1, \cdots, N_{tu}$
i. トピックを生成
$z_{tun} \sim \text{Categorical}(\boldsymbol{\theta}_{tu})$
ii. 単語を生成
$w_{tun} \sim \text{Categorical}(\boldsymbol{\phi}_{tz_{tun}})$
(c) For 補助情報 $i = 1, \cdots, M_t$
i. 補助情報トピックを生成
$y^i_{tu} \sim \text{Categorical}(\widetilde{\boldsymbol{z}}_{tu})$
ii. 補助情報を生成
$x^i_{tu} \sim \text{Categorical}(\boldsymbol{\psi}^i_{ty^i_{tu}})$

3.3　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 [8] 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．ここで各変数をまとめた記号を表 1 に示す．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 (6) 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 で単語 n が割り当てられた個数，N_{tk} はその単語ごとの総和を表し，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．

$$
\begin{aligned}
&p(z_{tun}=k \mid \mathbf{Z}_{\backslash tun}, \mathbf{W}_t, \mathbf{X}^\dagger_t, \mathbf{Y}^\dagger_t, \boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}^\dagger, \boldsymbol{\eta}, \sigma^2) \\
&\propto p(w_{tun} \mid \mathbf{W}_{\backslash tun}, \mathbf{Z}_t, \beta) \\
&\times p(z_{tun}=k \mid \mathbf{Z}_{\backslash tun}, \boldsymbol{\alpha}) \\
&\times p(\boldsymbol{y}^\dagger_{tu} \mid z_{tun}=k, \boldsymbol{z}_{tu\backslash tun}) \\
&\times p(y_u \mid z_{tun}=k, \boldsymbol{z}_{tu\backslash tun}, \boldsymbol{\eta}, \sigma^2) \\
&= (N_{tuk\backslash tun}+\alpha_k)\frac{N_{tkw_{tun\backslash tun}}+\beta}{N_{tk\backslash tun}+\beta V} \\
&\times \left(\frac{N_{tuk\backslash tun}+1}{N_{tuk\backslash tun}}\right)^{\sum_i M^i_{tuk}} \\
&\times \exp\left\{\frac{\eta_k}{N_{tu}\sigma^2}\left(y_u - \boldsymbol{\eta}^\top \widetilde{\boldsymbol{z}}_{\backslash tun} - \frac{\eta_k}{2N_{tu}}\right)\right\}. \quad (6)
\end{aligned}
$$

時刻 $t>1$ でも同様にトピックのサンプリング確率を式 (7) のように求めることができる．ここで $\hat{\boldsymbol{\Omega}}_{t-1} = \left\{\hat{\boldsymbol{\Theta}}_{t-1}, \hat{\boldsymbol{\Phi}}_{t-1}\right\}$ としてまとめている．

表 1　記号の定義

記号	意味
$\boldsymbol{\Theta}_t \in \mathbb{R}^{U\times K}$	$\boldsymbol{\theta}_{tu}$ を全ユーザでまとめた行列
$\boldsymbol{\Phi}_t \in \mathbb{R}^{K\times V}$	$\boldsymbol{\phi}_{tk}$ を全トピックでまとめた行列
$\mathbf{W}_t = \{\boldsymbol{w}_{tu}\}_{u=1}^{U}$	$\boldsymbol{w}_{tu}$ を全ユーザでまとめた集合
$\mathbf{Z}_t = \{\boldsymbol{z}_{tu}\}_{u=1}^{U}$	$\boldsymbol{z}_{tu}$ を全ユーザでまとめた集合
$\boldsymbol{\Psi}_t^i \in \mathbb{R}^{K\times M_{tu}^i}$	$\boldsymbol{\psi}_{tk}^i$ を全トピックでまとめた行列
$\boldsymbol{\Psi}_t^\dagger = \{\boldsymbol{\Psi}_t^i\}_{i=1}^{M_t}$	$\boldsymbol{\Psi}_t^i$ を全種類まとめた集合
$\mathbf{X}_t^i = \{\boldsymbol{x}_{tu}^i\}_{u=1}^{U}$	$\boldsymbol{x}_{tu}^i$ を全ユーザでまとめた集合
$\mathbf{X}_t^\dagger = \{\mathbf{X}_t^i\}_{i=1}^{M_t}$	$\mathbf{X}_t^i$ を全種類まとめた集合
$\mathbf{Y}_t^i = \{\boldsymbol{y}_{tu}^i\}_{u=1}^{U}$	$\boldsymbol{y}_{tu}^i$ を全ユーザでまとめた集合
$\mathbf{Y}_t^\dagger = \{\mathbf{Y}_t^i\}_{i=1}^{M_t}$	$\mathbf{Y}_t^i$ を全種類まとめた集合
$\boldsymbol{\gamma}_t^\dagger = \{\gamma_t^i\}_{i=1}^{M_t}$	γ_t^i を全種類まとめた集合

$$
\begin{aligned}
&p(z_{tun}=k \mid \mathbf{Z}_{\backslash tun}, \mathbf{W}_t, \mathbf{X}_t^\dagger, \mathbf{Y}_t^\dagger, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t^\dagger, \hat{\boldsymbol{\Omega}}_{t-1})\\
&\propto p(w_{tun} \mid z_{tun}=k, \mathbf{Z}_{\backslash tun}, \mathbf{W}_{\backslash tun}, \beta_{tk}, \hat{\boldsymbol{\phi}}_{t-1,k})\\
&\times p(z_{tun}=k \mid \mathbf{Z}_{\backslash tun}, \alpha_{tu}, \hat{\boldsymbol{\theta}}_{t-1,u})\\
&\times p(\boldsymbol{y}_{tu}^\dagger \mid z_{tun}=k, \boldsymbol{z}_{tu\backslash tun})\\
&= \left(N_{tuk} + \alpha_{tu}\hat{\theta}_{t-1,uk}\right)\\
&\times \frac{N_{tkw_{tun}\backslash tun} + \beta_{tk}\hat{\phi}_{t-1,kw_{tun}}}{N_{tk\backslash tun} + \beta_{tk}}\\
&\times \left(\frac{N_{tuk\backslash tun}+1}{N_{tuk\backslash tun}}\right)^{\sum_i M_{tuk}^i}.
\end{aligned}
\tag{7}
$$

次に補助情報トピックのサンプリング式を導出する．時刻 t のあるカテゴリ変数 i の補助情報トピック y_{tum}^i $(i=1,\cdots,M_t)$ が k となる確率について，周辺同時分布にベイズの定理を用いると式 (8) のように求めることができる．ここで M_{tks}^i は時刻 t の補助情報 i における補助情報トピック k の割当数，M_{tk}^i はその個数を表す．

$$
\begin{aligned}
&p(y_{tum}^i=k \mid \mathbf{W}_t^i, \mathbf{X}_t^i, \mathbf{Y}_{\backslash tum}^i, {\boldsymbol{\gamma}_t}^\dagger)\\
&\propto p(x_{tum}^i \mid \mathbf{X}_{\backslash tum}^i, \boldsymbol{Y}_t^i, \gamma_t^i)\\
&\times p(y_{tum}^i=k \mid \mathbf{Z}_t, \mathbf{Y}_{\backslash tum}^i)\\
&= N_{tuk}\frac{M_{tkx_{tum}^i\backslash tun}^i + \gamma_t^i}{M_{tk\backslash tum}^i + \gamma_t^i S^i}.
\end{aligned}
\tag{8}
$$

3.4　ハイパーパラメータの推定

各時刻におけるディリクレ分布のハイパーパラメータを，周辺同時分布の確率が最大となるように不動点反復法 [9] を用いて推定する．この手法は最適解が得られるような漸化式から初期点のパラメータを更新していき，最終的な解を得る．時刻 $t=1$ において，ハイパーパラメータ $\boldsymbol{\alpha}$，β の推定値は以下の更新式 (9)，(10) から得られる．

$$
\alpha_k^{\mathrm{new}} = \frac{\alpha_k \sum_u \Psi(N_{tuk}+\alpha_k) - U\Psi(\alpha_k)}{\sum_u \Psi\left(N_{tu} + \sum_{k'}\alpha_{k'}'\right) - U\Psi\left(\sum_{k'}\alpha_{k'}\right)},
\tag{9}
$$

$$
\beta^{\mathrm{new}} = \frac{\beta\sum_k\sum_v \Psi(N_{tkv}+\beta) - KV\Psi(\beta)}{V\sum_k \Psi(N_{tk}+\beta V) - KV\Psi(\beta V)}.
\tag{10}
$$

ここで $\Psi(\cdot)$ はディガンマ関数を表し，ガンマ関数 $\Gamma(x) = \int_0^\infty e^{-x}x^{s-1}dx$ に対して，$\Psi(x) = \frac{\partial \log\Gamma(x)}{\partial x}$ で定義される．また月齢予測のパラメータ $\boldsymbol{\eta}$，σ^2 の更新式は式 (11)，(12) で与えられる．

$$
\boldsymbol{\eta}^{\mathrm{new}} = \left(\sum_{u=1}^{U}\widetilde{\boldsymbol{z}_{tu}}\widetilde{\boldsymbol{z}_{tu}}^\top\right)^{-1}\left(\sum_{u=1}^{U} y_u\widetilde{\boldsymbol{z}_{tu}}\right),
\tag{11}
$$

$$
\sigma^{2,\mathrm{new}} = \frac{1}{U}\left(\sum_{u=1}^{U}{y_u}^2 - \boldsymbol{\eta}^\top\left(\sum_{u=1}^{U} y_u\widetilde{\boldsymbol{z}_{tu}}\right)\right).
\tag{12}
$$

時刻 $t>1$ についても同様に，不動点反復法からハイパーパラメータを推定すると，次式の更新式 (13)，(14) が得られる．

$$
\alpha_{tu}^{\mathrm{new}} = \alpha_{tu}\frac{\sum_k \hat{\theta}_{t-1,uk}A_{tuk}}{\Psi(N_{tu}+\alpha_{tu}) - \Psi(\alpha_{tu})},
\tag{13}
$$

$$
\beta_{tk}^{\mathrm{new}} = \beta_{tk}\frac{\sum_v \hat{\phi}_{t-1,kv}B_{tkv}}{\Psi(N_{tk}+\beta_{tk}) - \Psi(\beta_{tk})}.
\tag{14}
$$

ここで，$A_{tuk} = [\Psi(N_{tuk}+\alpha_{tu}\hat{\theta}_{t-1,uk}) - \Psi(\alpha_{tu}\hat{\theta}_{t-1,uk})]$，$B_{tkv} = [\Psi(N_{tkv}+\beta_{tk}\hat{\phi}_{t-1,kv}) - \Psi(\beta_{tk}\hat{\phi}_{t-1,kv})]$ である．

補助情報 i の補助情報分布のハイパーパラメータ γ_t^i については，時刻によらず同じ更新式 (15) を得る．

$$
\gamma_t^{i,\mathrm{new}} = \gamma_t^i\frac{\sum_k\sum_s \Psi(M_{tks}^i+\gamma_t^i) - KS^i\Psi(\gamma_t^i)}{S^i\sum_k \Psi(M_{tk}^i+\gamma_t^i S^i) - KS^i\Psi(\gamma_t^i S^i)}
\tag{15}
$$

3.5　トピック数の決定

本節では，最適なトピック数の決定において Perplexity [3] を用いた評価を行う．Perplexity は，式 (16) で定義されるように，ユーザの訓練単語の学習から得られたトピックモデルが，テスト単語をどれだけ精度良く予測できるかを評価する指標である．

$$
\begin{aligned}
&\mathrm{Perplexity}(\mathbf{W}_t^{\mathrm{test}} \mid \mathcal{M})\\
&= \exp\left(-\frac{\sum_u \log p(\boldsymbol{w}_{tu}^{\mathrm{test}} \mid \mathcal{M})}{\sum_u N_{tu}^{\mathrm{test}}}\right).
\end{aligned}
\tag{16}
$$

ここで $\mathbf{W}_t^{\text{test}}$ は時刻 t での各ユーザのテスト単語の組 $\boldsymbol{w}_{tu}^{\text{test}}$ の集合，N_{tu}^{test} は時刻 t のユーザ u のテスト単語数，$\mathcal{M}$ は確率モデルを表す．式 (16) は単語の生成確率に関する負の対数尤度から計算され，値が小さいほどテスト単語を高い精度で予測できていることを意味する．提案手法の SCTTM では，単語の生成確率は式 (17) で計算される．

$$p(\boldsymbol{w}_{tu}^{\text{test}} \mid \mathcal{M}) = \prod_{n=1}^{N_{tu}^{\text{test}}} \sum_{k=1}^{K} \hat{\theta}_{tuk} \hat{\phi}_{tkw_{tun}}. \quad (17)$$

3.6 トピックの自動ラベル付け

本節では，トピックの内容を容易に把握しやすくするためのラベル付けの手法を示す．大町ら [10] は，時刻が変化するごとに LDA によって単語分布を推定し，そこからラベル付けを行っている．しかし，この手法では時系列変化に対応しているラベルを一意に定めることができず，ユーザ個人の嗜好全体の推定が困難である．そこで本研究では，SCTTM によって推定された単語分布を用いて，各時刻でラベリングを行い，そのラベルの平均をトピックのラベルとして付与する手法を提案する．これにより，時系列を考慮したラベルを一意に定め，ユーザ個人の嗜好全体を推定することができる．

3.6.1 単語のスコア付け

単語を抽出する際に，本研究では term-score [11] を時刻の次元 T まで拡張する．この指標は以下の式で定義される．

$$\text{term-score}_{tkv} = \hat{\phi}_{tkv} \log \frac{\hat{\phi}_{tkv}}{\prod_{k'=1}^{K} \hat{\phi}_{tk'v}}. \quad (18)$$

式 (18) は，トピック全体に単語 v が多く現れるほど，分母の $\prod_{k'=1}^{K} \hat{\phi}_{tk'v}$ が大きくなる．また特定のトピック k のみに単語 v が多く現れるほど，その生起確率である分子の $\hat{\phi}_{tkv}$ が大きくなる．

3.6.2 実装手順

トピックの自動ラベルを抽出するために，まず単語の分散表現ベクトルを Word2Vec [12] を用いて計算する．Word2Vec は 2 層のニューラルネットワークで構成されたモデルであり，入力した単語の分散表現ベクトルを求める際に利用される [13]．本研究では Wikipedia で事前学習したモデル [14] を利用し，以下の手順で自動ラベル付けを行う．

1. 時刻 t，トピック k について，単語分布 $\hat{\boldsymbol{\phi}}_{tk}$ の確率が高い上位 2 番目から 6 番目までの 5 単語を取り出す．
2. 取り出した 5 単語の分散表現ベクトルの平均を求め，最も類似した単語を事前学習済みの Word2Vec から抽出する．ここで，類似度の計算にはコサイン類似度を使用する．
3. すべての時刻 $t = 1, \cdots, T$，トピック $k = 1, \cdots, K$ についても同様にして，最も類似度の高い単語をそれぞれ抽出する．
4. 各時刻で得られた単語を分散表現ベクトルに直し，トピックごとに平均化する．
5. トピックごとに得られた平均分散表現ベクトルと最も類似する単語を，事前学習済みの Word2Vec から抽出し，自動ラベルとして割り当てる．

トピックへの寄与率が 1 番高い最上位の単語を除く理由は，抽象的な単語をラベルとして付与するためである [10]．この手順により，時系列の影響を考慮したうえでラベルを自動的に付与することができる．

4. 実データによる検証

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から提供された新生児・乳幼児の母親をメインユーザとする QA コミュニティ「ママリ」のデータを用いて，提案手法の有効性を確かめる．

4.1 提供データについて

提供されたデータには，ママリに登録されているユーザの情報，ユーザが投稿した質問文とそれに対する回答，検索履歴を含んだテキストデータが記録されている．ユーザ数は 290 万人，質問数は 540 万件，回答数は 3,500 万件，検索数は 1.2 億件であり，データサイズは合計約 20GB である．ユーザの情報には子供の性別，子供数，親の世代，居住する都道府県，質問カテゴリ，子供の月齢が存在し，投稿されている質問文は育児に関することや，妊娠に関することなど多岐にわたる．

4.2 前処理

実験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．なお質問数が均一ではないため学習時にはパディングを行う．

データの前処理は次の手順で行う．SCTTM ではデータの時間間隔が一定であるという制約があるため，質問回数を時間軸に取り，質問日時の古い順に時刻 $t = 1$, $t = 2$, ... とする．補助情報として用いる月齢に関しては，ユーザには複数の子供が存在する場合

があるため，第一子の月齢を用いることとする．質問文においては，形態素解析ツールである MeCab [15] を用いて質問文を単語に分割した後，名詞であり 2 文字以上かつひらがなだけではない単語を抽出した．条件に一致する単語数は 3,633 語であった．抽出した単語を用いて各質問文から（期間数 × ユーザ数 × 単語数）の三次元出現頻度行列を作成した．

4.3 SCTTM による推定結果

本節では，SCTTM によって推定された結果を以下に示す．まず，4.3.1 節でトピック数の決定手法について示す．次に，4.3.2 節で月齢の予測精度の有効性について示す．そして，4.3.3 節で SCTTM によって推定された単語分布と自動ラベル付けの結果を述べる．最後に，4.3.4 節で SCTTM によって推定された補助情報分布の推定結果を示す．

4.3.1 トピック数の決定

SCTTM では，トピック数を事前に決定したうえで推定する必要がある．本研究で用いたトピック数の決定は，まず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，各トピック数から最初の時刻 $t=1$ のトピック分布 $\boldsymbol{\Theta}_t$ と単語分布 $\boldsymbol{\Phi}_t$ を学習する．次に学習したパラメータとテスト単語を用いて 3.5 節の Perplexity を計算し，値が最小となるトピック数を採用するという手法である．実験ではトピック数 K を 5 から 30 まで 5 刻みに比較し，$K=25$ 個のトピック数を得た．以降ではトピック数を 25 と設定したうえでモデルの検証を行う．なお検証では，パラメータの推定に崩壊型ギブスサンプリングを用いる．このとき反復数を 1,000，最初に棄却する反復区間を 500 に設定する．

4.3.2 月齢の予測精度の有効性

最初の質問時 ($t=1$) における，各ユーザがもつ子供の月齢を 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ユーザとテストユーザに 8:2 で分けて 5 分割交差検証で評価し，評価指標として MSE を採用した．なお SCTTM では線形回帰における偏回帰係数パラメータ $\boldsymbol{\eta}$ が反復数置きに更新されるため，テストユーザの評価における偏回帰係数パラメータは，テストユーザのデータから更新するのではなく，訓練時の最後の反復で得られたパラメータで固定する．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 [16]，LightGBM [17] の三つのモデルと比較している．一つ目の LDA は，2 節で示したように，子供の月齢情報を考慮しない教師なし学習であるため，まずトピック分布を学習し，その後月齢を重回帰で予測する．二つ目の RandomForest は，決定木による複数の弱学習器を統合させて目的変数を予測するアンサンブル学習である．三つ目の LightGBM は，誤差が最小となるように決定木を逐次的に作成していく GBDT (Gradient Boosting Decision Tree) の一種である．いずれの比較手法も，得られたトピック分布から事後的に月齢を回帰する手法であることから，トピックの学習とは別に予測される．

表 2 子供の月齢の予測結果

モデル	訓練 MSE	テスト MSE
LDA	634.331	707.990
RandomForest	101.506	762.656
LightGBM	54.484	804.326
SCTTM	**20.101**	**61.609**

表 2 は月齢の予測結果を表している．この結果から，訓練ユーザおよびテストユーザに関して，SCTTM がほかの手法に比べて MSE の精度が向上していることが確認できる．

4.3.3 単語分布の推定と自動ラベル付けの結果

自動ラベル付けの結果と時刻 $t=1$ における単語分布の推定値を表 3 に示す．ここで，各トピックの単語は生起確率の高い上位 4 件のみを示している．各単語の括弧内の数字は，その単語が生成される確率を表している．トピック番号の横に付与された単語とその括弧内の単語は，それぞれ提案手法による自動ラベルと $t=1$ における自動ラベルを表している．トピックモデルにおける自動ラベル付けは教師なし学習であるため，提案手法による結果と，$t=1$ における結果を比較する．表 3 より，まずトピック 1 に着目すると，妊娠や出産に関連する単語が上位に見られるため，妊娠や出産に関するトピックだと考えられる．次にトピック 2 に着目すると，赤ちゃんの子育てに関する単語が上位に現れているため，子育て関連のトピックだと判断できる．またトピック 6 では，妊娠だけでなく，検査，排卵，2 人という単語も見られるため，妊娠までに至る過程を表すトピックだといえる．

提案手法による自動ラベル付けの結果と時刻 $t=1$ における単語分布を比較する．トピック 5 や 6，17 に着目すると，ラベルが単語分布の上位に含まれている．また，トピック 2 や 8，11，12，16，18，19，20 に着目すると，ラベルの意味が上位の単語と近いと考えられる．これより，付与されたラベルが時刻 $t=1$ における一部のトピックの内容を表しているといえる．一方で，トピック 1 や 21，22 のように単語分布の内容を

表 3　自動ラベル付けと $t = 1$ における単語分布の推定結果

トピック	単語（上位 4 件）			
1：家政婦（妊娠）	確認 (0.032)	陣痛 (0.022)	出産 (0.019)	病院 (0.017)
2：寝床（仮眠）	赤ちゃん (0.037)	生後 (0.036)	毎回 (0.030)	授乳 (0.028)
3：くしゃみ（流産）	生理 (0.090)	妊娠 (0.053)	子宮 (0.022)	結婚 (0.019)
4：専業主婦（不眠症）	子供 (0.045)	下痢 (0.023)	最近 (0.022)	皆さん (0.020)
5：ミルク（乳児）	質問 (0.040)	ミルク (0.026)	病院 (0.026)	子ども (0.024)
6：妊娠（排卵）	妊娠 (0.058)	検査 (0.053)	排卵 (0.035)	2 人 (0.034)
7：小夜（往診）	妊娠 (0.026)	旦那 (0.025)	受精 (0.021)	一緒 (0.019)
8：母乳（授乳）	ミルク (0.091)	離乳食 (0.037)	生後 (0.023)	体温 (0.015)
9：女中（園児）	主人 (0.030)	2 人 (0.025)	子供 (0.023)	保育園 (0.016)
10：義姉（新妻）	息子 (0.052)	義母 (0.050)	子供 (0.033)	写真 (0.033)
11：乳児（分娩）	母乳 (0.071)	1 ヶ月 (0.039)	授乳 (0.029)	検診 (0.023)
12：流産（痛み）	妊娠 (0.079)	不安 (0.058)	赤ちゃん (0.050)	体重 (0.025)
13：介護（不登校）	病院 (0.055)	1 歳 (0.030)	分娩 (0.020)	保育園 (0.017)
14：帰宅（昼寝）	ママ (0.037)	仕事 (0.032)	お願い (0.020)	夜中 (0.014)
15：義母（バイク）	息子 (0.040)	シート (0.034)	チャイルド (0.030)	自転車 (0.026)
16：往診（診察）	出産 (0.144)	予定 (0.039)	ダメ (0.027)	検査 (0.019)
17：主人（女中）	抱っこ (0.057)	風呂 (0.034)	主人 (0.032)	子ども (0.023)
18：継母（父親）	旦那 (0.143)	実家 (0.049)	自分 (0.027)	両親 (0.020)
19：出産（幸せ）	今日 (0.047)	妊娠 (0.031)	子供 (0.020)	イヤ (0.019)
20：喀血（診療所）	出血 (0.047)	入院 (0.028)	オムツ (0.021)	トイレ (0.019)
21：食事（分娩）	予定 (0.038)	入院 (0.025)	お願い (0.023)	陣痛 (0.022)
22：ご馳走（病人）	風邪 (0.023)	手当 (0.019)	1 ヶ月 (0.017)	傷病 (0.015)
23：空腹（子猫）	痛み (0.048)	お腹 (0.028)	男の子 (0.027)	アドバイス (0.025)
24：出勤（新妻）	保育園 (0.054)	息子 (0.025)	2 人 (0.023)	幼稚園 (0.023)
25：育児（思い）	病院 (0.034)	仕事 (0.033)	気持ち (0.026)	1 人 (0.022)

表 4　$t = 1$ における補助情報分布の推定結果

トピック	子供の性別	子供の人数	登録者の年齢	居住地域	質問カテゴリ
1	男の子 (0.773)	2 人 (0.559)	36 歳〜40 歳 (0.266)	兵庫県 (0.154)	妊娠・出産 (0.512)
2	男の子 (0.641)	1 人 (0.936)	31 歳〜35 歳 (0.316)	埼玉県 (0.152)	子育て・グッズ (0.823)
3	女の子 (0.621)	1 人 (0.878)	21 歳〜25 歳 (0.188)	東京都 (0.223)	妊娠・出産 (0.441)
4	女の子 (0.953)	1 人 (0.928)	26 歳〜30 歳 (0.487)	埼玉県 (0.181)	子育て・グッズ (0.330)
5	女の子 (0.637)	1 人 (0.881)	31 歳〜35 歳 (0.483)	北海道 (0.157)	子育て・グッズ (0.782)
6	男の子 (0.532)	1 人 (0.617)	26 歳〜30 歳 (0.366)	愛知県 (0.233)	妊活 (0.497)
7	女の子 (0.720)	2 人 (0.784)	26 歳〜30 歳 (0.343)	大阪府 (0.354)	妊娠・出産 (0.343)
8	男の子 (0.694)	1 人 (0.648)	31 歳〜35 歳 (0.374)	東京都 (0.157)	子育て・グッズ (0.669)
9	女の子 (0.601)	3 人 (0.532)	41 歳〜50 歳 (0.341)	広島県 (0.174)	子育て・グッズ (0.487)
10	男の子 (0.668)	1 人 (0.422)	26 歳〜30 歳 (0.231)	神奈川県 (0.192)	子育て・グッズ (0.481)
11	男の子 (0.803)	1 人 (0.632)	21 歳〜25 歳 (0.231)	大阪府 (0.183)	妊娠・出産 (0.446)
12	女の子 (0.763)	1 人 (0.572)	26 歳〜30 歳 (0.532)	埼玉県 (0.175)	妊娠・出産 (0.425)
13	女の子 (0.787)	1 人 (0.516)	21 歳〜25 歳 (0.135)	大阪府 (0.209)	子育て・グッズ (0.417)
14	女の子 (0.594)	2 人 (0.908)	31 歳〜35 歳 (0.412)	福岡県 (0.324)	子育て・グッズ (0.674)
15	男の子 (0.916)	1 人 (0.657)	31 歳〜35 歳 (0.213)	東京都 (0.196)	子育て・グッズ (0.479)
16	女の子 (0.731)	2 人 (0.536)	31 歳〜35 歳 (0.256)	千葉県 (0.184)	妊娠・出産 (0.421)
17	男の子 (0.716)	2 人 (0.526)	26 歳〜30 歳 (0.205)	静岡県 (0.131)	子育て・グッズ (0.238)
18	女の子 (0.417)	1 人 (0.820)	36 歳〜40 歳 (0.082)	宮崎県 (0.084)	家族・旦那 (0.621)
19	男の子 (0.583)	2 人 (0.630)	41 歳〜50 歳 (0.158)	長野県 (0.282)	子育て・グッズ (0.354)
20	女の子 (0.645)	2 人 (0.777)	36 歳〜40 歳 (0.160)	東京都 (0.253)	妊娠・出産 (0.709)
21	男の子 (0.636)	1 人 (0.559)	31 歳〜35 歳 (0.323)	兵庫県 (0.188)	妊娠・出産 (0.544)
22	男の子 (0.611)	1 人 (0.721)	36 歳〜40 歳 (0.206)	埼玉県 (0.124)	子育て・グッズ (0.775)
23	不明 (0.508)	1 人 (0.560)	21 歳〜25 歳 (0.283)	神奈川県 (0.153)	妊娠・出産 (0.725)
24	女の子 (0.674)	2 人 (0.603)	26 歳〜30 歳 (0.394)	愛知県 (0.261)	子育て・グッズ (0.327)
25	女の子 (0.660)	1 人 (0.851)	31 歳〜35 歳 (0.230)	千葉県 (0.186)	妊娠・出産 (0.686)

表すことができていないラベルも確認した．この原因は，表に示した単語分布は時刻 $t = 1$ のものだが，付与されたラベルはすべての時刻の単語分布を考慮しているためだと考えられる．実際に，括弧内の $t = 1$ における自動ラベルの結果を見ると，トピック 1 や 21，22 では，$t = 1$ における自動ラベルが適切にトピック

表 5　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		
	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$	$N=1$	$N=2$	$N=3$
LDA	14.500	22.750	30.000	11.333	20.667	26.889	11.842	19.105	24.684
TTM	13.500	23.000	29.750	12.667	21.778	27.778	11.474	19.053	24.895
CTTM	**16.010**	25.500	32.250	**15.000**	22.111	**29.556**	13.421	21.632	27.211
SCTTM	16.000	**26.500**	**34.000**	14.000	**22.667**	29.444	**14.526**	**23.158**	**29.368**

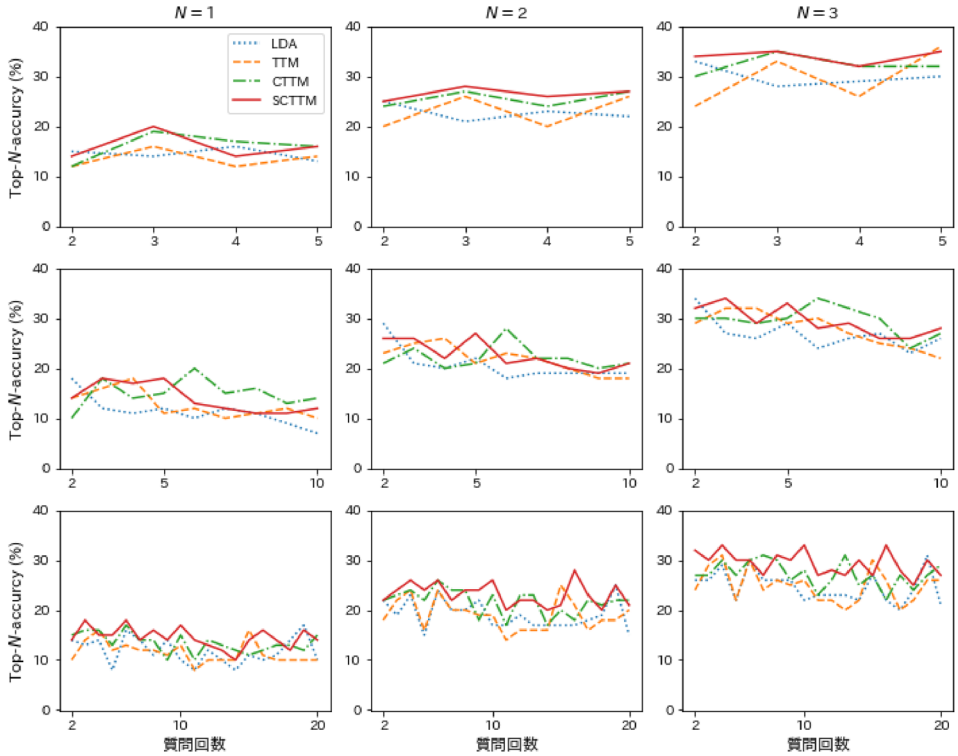


図 2　Top-N-accuracy による各時刻の質問単語予測の結果
各行は上から順に質問回数が 5 回，10 回，20 回のユーザに対応し，各列は左から順に $N=1,2,3$ の結果に対応する．各図において，横軸は質問回数，縦軸は Top-N-accuracy (%) の値を表し，各折れ線は凡例のモデルに対応する．Top-N-accuracy は前の時刻の分布から計算されるため，開始時刻はいずれも質問回数が 2 のときを始点としている．

の内容を表していると考えられる．

4.3.4　補助情報分布の推定結果

時刻 $t=1$ における補助情報分布の推定結果を表 4 に示す．ただし推定結果は，最も生起確率の高いもののみを示している．なお欠損値を取る確率が最も高い場合，次に確率が高い値を示している．また括弧内の値は，その補助情報が生成される確率を表す．表 4 より，性別や子供の人数に着目すると，ほとんどの値が 0.5 以上の確率で推定されている．またほかのカテゴリにおいても，それぞれのカテゴリ数に対して高い確率で推定されている．このことから，補助情報がばらつきなく推定されているといえる．

質問カテゴリに着目して，表 3 の単語分布の結果と比較する．多くの質問カテゴリは妊娠・出産，子育て・病院が多く，各トピックの上位単語もそれに近い単語が推定されていると考えられる．

4.4　質問単語の時系列推定の有効性

SCTTM では，トピック分布と単語分布が前の時刻の分布に依存して変化するように学習する．そこで各時刻で質問される単語をどれほど精度良く推定できるか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

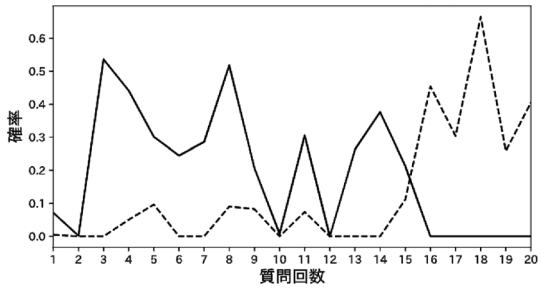


図 3　トピック分布における，あるユーザのトピック 6（実線）とトピック 24（破線）の変化

を採用する．

$$P(w_{tu} = v \mid \hat{\boldsymbol{\Theta}}_{t-1}, \hat{\boldsymbol{\Phi}}_{t-1}) = \sum_{k=1}^{K} \hat{\theta}_{t-1,uk} \hat{\phi}_{t-1,kv}. \tag{19}$$

式 (19) 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，$N = 1, 2, 3$ 件での平均を評価した結果を表 5 に示す．各モデルのトピック数は，SCTTM と同様に 25 に設定している．この表の結果から，時間依存性と補助情報を考慮した CTTM，SCTTM がほか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．また補助情報に連続変数の月齢も考慮した SCTTM の方が，CTTM に比べて全体的な予測精度が上回った．図 2 に，各時刻における比較手法の予測結果を示す．各行は上から順に質問回数が 5 回，10 回，20 回のユーザに対応し，質問回数ごとに横軸のスケールが異なる．各列は左から順に $N = 1, 2, 3$ の結果に対応する．各図において，横軸は質問回数，縦軸は Top-N-accuracy(%) の値を表し，折れ線はグラフの凡例モデルに対応する．

あるユーザのトピック 6 とトピック 24 の時系列変化を図 3 に示す．この図より，時刻が小さい場合はトピック 6（妊娠）と推定される確率が高いが，時間が進むにつれてトピック 24（育児）と推定される確率が高くなる．したがって，潜在思考を捉えることができていると考えられる．

5. おわりに

本研究では，ユーザごと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ことを考慮した，新しいトピックモデルである SCTTM を提案した．また得られたトピックについて，時系列変化を考慮した自動ラベル付けの手法を提案した．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，ユーザごと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．さらにトピックの自動ラベル付けにより，トピックの解釈も容易となった．提案手法を実データに適用する場合，ユーザ個人の補助情報が観測されているテキストデータであれば嗜好変化や補助情報の傾向を推定することができる．たとえば，QA コミュニティであれば Yahoo！知恵袋，他種類のコミュニティであれば Twitter などが挙げられる．提案手法をビジネスに適用することで，ユーザの嗜好変化の可視化から新しい施策を打ち出したり，異業種との提携を促すためのツールとして活用できると考える．

今後の課題として，以下の三つが挙げられる．一つ目は，トピックと単語分布を学習する際に，依存する過去の時刻を 2 時刻以上前に設定してみることである．本研究では直前の時刻のみを考慮して推定を行ったが，2 時刻以上前を考慮することで，より時系発展に頑丈な推定ができると考えられる．二つ目は，ほかの文書データを組み合わせることである．実験で用いたママリのデータには，質問以外にも検索や回答のデータも含まれていた．これらを組み合わせることで，質問，検索，回答間の関係も捉えられるような学習ができると考える．三つ目は，単語分布の推定において，各トピックの上位単語の重複を減らすことである．表 3 では，妊娠や子供，赤ちゃんなどの単語が複数上位に推定された．単語同士が重複しないように学習することで，より解釈性の高いトピックが得られると考えられる．

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，有用なコメントをいただいた 2 名の査読者の方々に感謝申し上げます．最後に，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] 東京都中央区,「中央区の子ども・子育てを取り巻く現状と課題」, https://www.city.chuo.lg.jp/kosodate/keikaku/kodomokosodatekeikakusakutei.files/Chapter02.pdf（2022 年 7 月 10 日閲覧）

[2] 厚生労働省政策統括官,「グラフでみる世帯の状況」, https://www.mhlw.go.jp/toukei/list/dl/20-21-h28.pdf （2022 年 7 月 10 日閲覧）

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

[4] 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 Retrieval*, pp. 127–134, 2003.

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

[6] D. M. Blei and J. D. Lafferty, "Dynamic topic models," In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pp. 113–120, 2006.

[7] 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.

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

[9] H. M. Wallach, "Topic modeling: Beyond bag-of-words," In *Proceedings of the 23rd International Conference on Machine Learning (ICML' 06)*, pp. 977–984, 2006.

[10] 大町凌弥，風間一洋，榊剛史，"単語の分散表現を用いた LDA のトピックラベリングと時系列可視化，" 第 12 回データ工学と情報マネジメントに関するフォーラム，2020.

[11] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining : Classification, Clustering, and Applications*, A. Srivastava and M. Sahami (eds.), CRC Press, pp. 101–124, 2009.

[12] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

[13] 森祥恭，高間康史，"Word2Vec で表現された単語の意味の可視化に関する検討，" 人工知能学会全国大会論文集 第 32 回，3F1OS12a02, 2018.

[14] M. Suzuki, Wikipedia Entity Vectors, https://github.com/singletongue/WikiEntVec （2022 年 7 月 10 日閲覧）

[15] 工藤拓，MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/ （2022 年 7 月 10 日閲覧）

[16] L. Breiman, "Random forests," *Machine Learning*, **45**, pp. 5–32, 2001.

[17] 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.

