情報処理学会論文誌 Vol.55 No.1 413–424 (Jan. 2014)

同行者に応じたトピックモデル

深澤 佑介[1,a)] 太田 順[2]

受付日 2013年4月9日, 採録日 2013年10月9日

概要：コンテキストはユーザの興味・嗜好に影響する重要な要因の 1 つである．本稿では，コンテキストの中でも特に家族や同僚など同行者に注目し，同行者依存のトピックモデルを提案する．第一に同行者クラスを考慮したモデル（CTM）を提案する．次にスイッチ変数を導入し同行者依存の単語を自動学習する仕組みを取り入れたモデルを提案する（sCTM）．さらに，同行者依存の単語を Web 全体から事前学習し，それをモデル内に反映するモデルを提案する（fCTM）．それぞれのモデルは Collapsed Gibbs Sampling（CGS）に基づき推論を行う．Web から同行者依存の投稿データを抽出し，提案モデル間の比較実験を実施した．文書の予測精度（Perplexity）の観点で CTM（ベースライン）と fCTM，sCTM を比較評価し提案手法の優位性を示した．また，質的評価として fCTM の同行者のトピックに含まれる単語を確認し，妥当なモデル化が行われていることを確認した．

キーワード：トピックモデル，コンテキストアウェア，同行者，Web マイニング，Collapsed Gibbs Sampling

Companion-aware Topic Model

Yusuke Fukazawa[1,a)] Jun Ota[2]

Received: April 9, 2013, Accepted: October 9, 2013

Abstract: Context is understood as an important factor that affects topics to be generated. We focus on companion of users (friends, wife, husband etc.) as one of the most important factors to determine the topic. Different from location and time context, context of companion does not appear explicitly with the documents but appears inside the document in the form of contextual words (e.g. friends). To discriminate contextual words, topical words and background words from documents, and obtain both precise and discriminative topics, we propose three kinds of context aware topic models. Firstly, we introduce context class (CTM) to extract context dominant topics from text. Secondly, we introduce switch variable (sCTM) to discriminate background words from contextual and topical words. Thirdly, we introduce fixed Dirichlet parameter learned from the web to sCTM (fCTM). We conduct experiments on data set extracted from the web, and they show that the proposed model (sCTM and fCTM) can capture interpretable and discriminative sets of topics than baseline CTM from the view point of perplexity and KL-Divergence.

Keywords: topic model, context-aware, companion, Web mining, Collapsed Gibbs Sampling

1. はじめに

近年，ベイズ推定を用いた様々なトピックモデル（topic model）が提案されている．トピックモデルとは，文書がトピック（潜在的な意味）に基づき生成される過程を確率的に表現したモデルであり，文書生成モデルとも呼ばれる．文章にトピックモデルをあてはめることにより，文章に内在するトピックを推定もしくは予測することが可能になる．トピックモデルの最も基礎的なモデルは 2003 年に Blei らによって提案された LDA（Latent Dirichlet allocation）[1] であり，文書のクラスタリング，トピック抽出 [2], [3], [4]，情報推薦 [5], [6], [7] に利用される．LDA は K-means [8]，PCA（Principle Component Analysis）[9]

[1] 株式会社 NTT ドコモ
NTT DOCOMO, Inc., Yokosuka, Kanagawa 239–8536, Japan
[2] 東京大学人工物工学研究センター
Research into Artifacts, Center for Engineering, The University of Tokyo, Kashiwa, Chiba 277–8568, Japan
[a)] fukazawayuu@nttdocomo.co.jp

や pLSI（Probabilistic Latent Semantic Indexing）[10] などの既存のクラスタリング手法に比べ事前確率分布を仮定している点から，学習セットへの過学習を防ぐ効果がある．

LDA 単体では，文書の文脈情報（コンテキスト）を考慮していない．たとえば，短文投稿サイト Twitter において，野球場で投稿される Tweet と，コンサート会場で投稿される Tweet では，場所がトピックの生成に大きく影響を与えているにもかかわらず，トピックの生成過程で場所の影響を考慮することができない．この場合，LDA では，野球関連トピックとコンサート関連のトピックを分離することはできるが，トピックの発生確率が場所に関連付いていない．そのため，野球場で Tweet があった場合，野球関連のトピックとコンサート関連のトピックが発生する確率は同じと見なされ，文書の予測精度が低下する．野球場では野球関連のトピックが多く，コンサート会場ではコンサート関連のトピックが多く発生することが分かれば，文書の予測精度が向上する．このような理解に従って，現在，LDA を拡張する形で様々なコンテキストを考慮したトピックモデルが提案されている．

ユーザのコンテキストはユーザの趣味嗜好に影響を与える重要な要素として考えられ，コンテキストからユーザの趣味嗜好やトピックを推定する研究が数多くなされている．コンテキストには「時間」「場所」「同行者」「天気」「音」「温度」など様々なものが考えられる．トピックモデルとしてこういったすべての情報を扱うことで，トピックの推定精度が向上する可能性がある．しかしながら，すべてのコンテキスト情報が平等に入手できるとは限らず，また，トピックへの影響度の大小も偏りがあると考えられる．たとえば，時間や場所は比較的入手しやすいが，ユーザの周囲の音や温度などは入手しにくい．また，時間や場所はトピックへの影響度も高いが，周囲の音がトピックに影響することは時間や場所に比べて低いと考えられる．時間や場所は上記のとおり，入手可能性および影響度の高さからコンテキストの一部として非常に重要視されてきた．一方，情報推薦の分野では，時間，場所と同様に重要視されているコンテキストとして「同行者」がある．たとえば，Adomavicious らは，情報推薦やトピック抽出において重要なコンテキストとは，「時間」「場所」「同行者」であると定義している [11]．また，Fukazawa らは，ユーザの状況に応じたタスクの推薦手法を提案しているが，その状況とは，Adomavicious らの定義と同様，ユーザの時間，場所，同行者の 3 つの要素から決定される [12]．「同行者」は「音」「温度」などと同様に入手可能性が低いが，情報推薦などのアプリケーションにおいて「時間」「場所」とともに利用可能性が高いことからここでは「同行者」をユーザのトピックに影響を与えるコンテキストとして扱う．

従来，コンテキストを考慮したトピックモデルとして，3 つのコンテキストの中で「時間」「場所」に応じたトピックモデルが数多く提案されている．しかしながら，時間と場所が同一でも同行者の有無，同行者が誰かによってトピックは変わる．従来提案されたコンテキストを考慮したトピックモデルでは「同行者」をトピックを変化させるパラメータとして考慮したトピックモデルは提案されていない．そこで本研究では，同行者依存のトピックモデルを提案する．**表 1** に同行者に依存する文書の例を示す．表では，C が同行者，T がトピック，左枠に文書（Twitter サービス畳での Tweet からの抜粋）を示している．「夫」といる場合には，トピック「生活」に関する Tweet が投稿される傾向にあることが分かる．また，「兄弟」といる場合には，トピック「視聴」,「友達」といる場合にはトピック「遊び」に関する Tweet が投稿される傾向がある．上記のとおり同行者に応じて特定のトピックから文章が構成されていることが分かる．

表 1　同行者に依存するトピックの例
Table 1　Example set of tweets that show companion(C)-topic(T)-words dependency.

C:husband T:live	loving life with husband / now living in Atlanta GA with husband / Lives in London with husband.
C:Brother T:watch	With brother in the cinema :)/ Movie night with brother/Watching boxing now with brother/ Watching doomsday prep show with brother
C:Friends T:enjoy	Driving with my two good friends to pick up lunch./ Took the last few days off to enjoy the holiday with friends/ hung out with friends early this morning/ will be running in Atlanta with friends

提案モデルは 3 つの特徴を有する．第 1 に，文書のトピックを決定する場合に同行者の要素を考慮することができる（例：家族と一緒に夕食）．これにより，同行者に依存して決まるトピックを高精度に推定可能である．第 2 に，文書中で使われる単語について，単語に対するスイッチ変数の導入により，同行者によって決まる単語/同行者に無関係に決まる単語を切り分けることができる．そのため，同行者によって決まるトピックに属する単語を高精度に推定可能である．第 3 に，同行者によって決まる単語を量の少ないコーパスの中から学習するのではなく，Web 全体から検索エンジンを利用して自動で抽出し，その結果をモデルで利用する．これにより，同行者に依存した単語をより高精度に推定可能である．

以下，2 章で関連研究について述べる．3 章にて同行者に応じたトピックモデルを提案する．4 章にて実験用データの構築方法について述べる．5 章で評価実験を行う．6 章で結論を述べる．

2. 関連研究

2.1 コンテキストに応じたトピックモデル

コンテキストには，時間，場所があるが，時間に応じた

トピックモデルとして，様々なトレンド解析モデルが提案されている．Blei らは，DTM（Dynamic Topic Model）を提案している [13]．このモデルでは，時間を一定の単位で量子化し，量子化された時間ごとのトピックを推定する．これにより，時間依存のトピックを抽出することができる．Wang らは，TOT（Topics Over Time）を提案している [14]．このモデルでは，各トピックに時間に関する確率分布を仮定し，それを推定する．これにより同時期に発生した複数のトピックを同時に推定することが可能になる．Kawamae は，TAM（Trend Analysis Model）を提案している [15]．このモデルでは，突発的に発生する単語を他の単語と区別する仕組みを導入することで，より高精度に時間に依存したトピック（トレンド）を推定することが可能である．

場所に応じたトピックモデルについても様々なモデルが提案されてきた．Eisenstein らは，位置が付与された文書集合から地理的に分布する潜在トピックの推定モデルを提案している [16]．このモデルでは，地理的にグローバルなトピックをまず生成し，そのトピックから地理的にローカルなトピックを生成している．Hong らは，上記に加え，ユーザ（文書の作成者）によって文書を作成する場所の違い（たとえば Twitter において Tweet する場所の違い）を考慮した潜在トピックの推定モデルを提案している [17]．ユーザの動線を考慮することにより文書からのユーザの位置推定精度が向上することを確認している．

上記のとおり，時間や場所を考慮したトピック解析モデルは提案されているが，コンテキストとして重要である同行者を考慮したトピックモデルは提案されていない．

2.2 同行者を考慮した情報提示

同行者を考慮した情報提示を行うためには，同行者を推定することが重要である．Sebastian らは，同行者およびユーザの場所の両方を考慮した情報提示を行うモバイルアプリケーション IYOUIT を提案している [18]．このモデルでは，ユーザの同行者をユーザの GPS 計測による位置情報および人間関係が記載された Social Ontology を用いて推定している．Fukazawa らはこの Social Ontology を利用し，位置履歴ではなくユーザの駅改札の通過履歴をもとに同行者を推定している [12]．しかしながら，これらは全ユーザの位置情報の収集が必要であることおよび人間関係をあらかじめ知っていることが前提でありコストが高い．

Yize らは，場所，時間，同行者を考慮した情報推薦システムを構築している [19]．位置情報や時間情報とは異なり，同行者に関する情報は明示的にユーザのログ（Yize らの論文では飲食店の口コミレビュー）に付与されていない．そのため，未知の文章から同行者を推定するための辞書を構築している．具体的には，まず，レビューサイトから，「with 同行者」という形式になっている文章を抽出し，その文章の正解同行者として付与する．既存の文書分類機能を持つソフトウェアを適用，同行者ごとに特有の単語を抽出し，同行者推定用の辞書を作成している．しかしながら，同行者をトピックを変化させるパラメータとして取り扱っておらず，同行者に応じたトピックを見つけることはできない．

以上のとおり，過去に提案されたグラフィカルモデルでは以下の問題点が発生する．

1) 同行者に依存したトピックが抽出できない．
2) 文書中で使われる単語は，同行者によって決まる単語/同行者に無関係に決まる単語があるが，考慮されていない．

そこで，本稿では，1)，2) の課題を解決する同行者依存のトピックモデルを提案する．

3. 提案モデル

3.1 同行者のモデリング

本節では上記であげた 1) および 2) の問題を解決するモデルを提案する．1) の問題を解決するため，提案モデルでは，文書のトピックは，同行者から決定されるとする．そのためここでは，文書ごとに同行者クラス（companion class）を定義しトピックモデルに組み込む．**図 1** に提案するグラフィカルモデル（CTM：Context-aware Topic Model）を示す．このモデルは LDA を拡張したものであり，LDA の文書の生成過程において同行者を考慮可能にする．既存の場所を考慮した LDA 拡張モデルにおいて，場所情報を同行者に置き換えたモデルといえる [16]（ただし，場所情報は緯度経度といった連続値をとるが，一方同行者は単語であるため離散値をとるといった違いが存在する）．パラメータの定義を**表 2** に示す．グラフィカルモデルとは，確率変数間の関係をグラフとして構造化したものである．グラフは確率変数，確率分布，確率分布のパラメータをあらわすノードおよびノード間の関係をあらわす有向リンクから構成される．網掛けの確率変数は既知の確率変数（a_d，w_{di}）であり，白抜きの確率変数は未知の確率変数（m_d，z_{di}），確率分布（ι，ν_m，κ_m，μ_z）および確率分布のパラメータ（β，ζ，α，ε）である．このモデルの目的は既知の変数（a_d，w_{di}）から未知の確率分布（ι，ν_m，κ_m，μ_z）を求めることである．

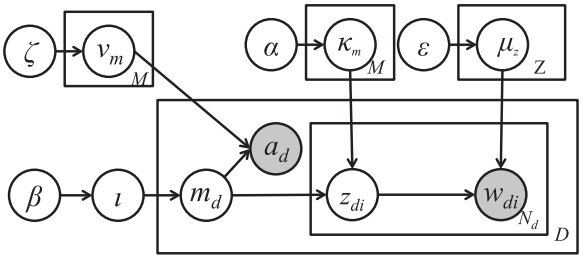

図 1 提案モデル（CTM）
Fig. 1 Proposed graphical model.

表 2 変数の定義
Table 2 Definition of variables in the model.

Variable	Meaning
M	同行者クラスの数
Z	トピック数
D	文書数
R	ユニークな単語数
N_d	文書 d の単語数
m_d	文書 d の同行者クラス
z_{di}	文書 d の i 番目の単語のトピック
a_d	文書 d の同行者
s_{di}	文書 d の i 番目の単語のスイッチ変数
w_{di}	文書 d の i 番目の単語
ι	ある文書の同行者クラスの分布 ($\iota \mid \beta \sim$ Dirichlet(β))
ν_m	同行者クラス m の同行者への分布 ($\nu_m \mid \zeta \sim$ Dirichlet(ζ))
κ_m	同行者クラス m のトピックへの分布 ($\kappa_m \mid \alpha \sim$ Dirichlet(α))
$\mu_{z\,or\,b}$	単語のトピック z もしくは 背景クラス b への分布 ($\mu_{z\,or\,b} \mid \varepsilon \sim$ Dirichlet(ε))
λ_d	文書 d のスイッチ変数への分布 ($\lambda_d \mid \delta \sim$ Dirichlet(δ))
β	ι のディリクレ事前分布のパラメータ
ζ	ν_m のディリクレ事前分布のパラメータ
α	κ_m のディリクレ事前分布のパラメータ
δ	λ_d のディリクレ事前分布のパラメータ
ε	$\mu_{z\,or\,b}$ のディリクレ事前分布のパラメータ

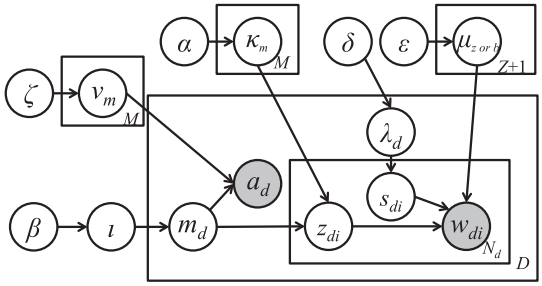

図 2 提案モデル（sCTM）
Fig. 2 Proposed graphical model.

2) の問題を解決するため，CTM を拡張して単語ごとに，以下を分類するスイッチ変数を定義する．

(ア) 同行者によって決まる単語（$s=0$）

(イ) そのほかの要因で決まる単語（$s=1$）

スイッチ変数はあらかじめ決定されているわけではなく，自動的に学習する．**図 2** に提案するグラフィカルモデル（sCTM：switch introduced Context-aware Topic Model）を示す．sCTM では，図 1 に示す CTM にスイッチ変数 s_{di}，確率分布 λ_d および確率パラメータ δ が追加されている．ここで，$s_{di}=0$ の場合は同行者によって決まる単語と見なされ，単語 w_{di} のトピックは z の中から選ばれる．一方，$s_{di}=1$ の場合はその他の要因で決まる単語と見なされ，背景クラス b が選ばれる．つまり，トピックの多項分布の μ はトピック z と背景クラス b の 2 つから構成される．トピック z は同行者によって変化するため Z 個持ちうるが，背景クラスは同行者以外の単語を区別さえすればよいので，1 個のみでよい．δ は，ディリクレ事前分布のパラメータであり，δ から λ_d への有向リンクは，パラメータ δ に従うディリクレ分布から多項分布 λ_d を選ぶことを意味する．

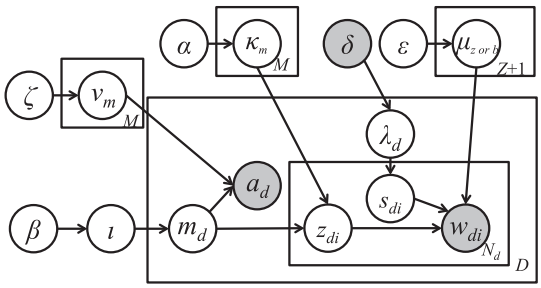

図 3 提案モデル（fCTM）
Fig. 3 Proposed graphical model.

図 2 では，スイッチ変数は，コーパスから自動的に学習されるが，コーパスが少なく，同行者に依存する単語かどうか精度高く分類できない可能性がある．そこで，Web 全体から検索エンジンを用いて，同行者に依存する単語を事前に分類し，その結果を図 2 の同行者依存トピックのモデルで利用する方法を提案する．この場合，スイッチ変数のディリクレ事前分布（パラメータ δ）は既知となる．**図 3** にそのグラフィカルモデル（fCTM：fixed switch variable introduced Context-aware Topic Model）を示す．δ の事前の決定の仕方の詳細については，3.3 節にて説明する．

3.2 提案モデルの推論

提案モデル（CTM，sCTM，fCTM）は，LDA を拡張したモデルであるため，LDA の推論で利用される Collapsed Gibbs Sampling（CGS）[20] を利用することが可能である．代表して sCTM の文書生成過程を述べる．

1. パラメータ β のディリクレ事前分布から多項分布 ι を選ぶ．
2. 各同行者クラス m についてパラメータ ζ のディリクレ事前分布から同行者クラスの数 M 個分の多項分布 ν_m を選ぶ．
3. 各同行者クラス m についてパラメータ α のディリクレ事前分布から同行者クラスの数 M 個分の多項分布 κ_m を選ぶ．
4. 各トピック z もしくは背景クラス b についてディリクレ事前分布 ε から多項分布 μ_z もしくは μ_b を選ぶ．

5. 各文書 d について以下の処理を行う．
   a) 多項分布 ι から同行者クラス m_d を選ぶ．
   b) 多項分布 ν_{m_d} から同行者 a_d を選ぶ．
   c) ディリクレ事前分布 δ から多項分布 λ_d を選ぶ．
   d) 文書 d の単語 i について以下の処理を行う．
      i) 多項分布 λ_d からスイッチ変数 s_{di} を選ぶ．
      もし $s_{di}=1$ である場合
         a) 多項分布 μ_b から単語 w_{di} を選ぶ．
      もし $s_{di}=0$ である場合
         b) 多項分布 κ_{m_d} からトピック z_{di} を選ぶ．
         c) 多項分布 $\mu_{z_{di}}$ から単語 w_{di} を選ぶ．

CGS にて推論を行うためには，まず，次式で表される全文書の結合分布を求める．

$$\begin{aligned}
&p(\mathbf{m},\mathbf{s},\mathbf{z},\mathbf{a},\mathbf{w},\iota,\kappa,\mu,\lambda,\upsilon;\beta,\alpha,\varepsilon,\delta,\zeta)\\
&= p(\iota|\beta)\times\prod_m^M p(\upsilon_m|\zeta)\times\prod_j^M p(\kappa_j|\alpha)\times\prod_l^{Z+1} p(\mu_l|\varepsilon)\\
&\quad\times\prod_d^D p(m_d|\iota)\times\prod_d^D p(a_d|\upsilon_{m_d})\times\prod_d^D p(\lambda_d|\delta)\\
&\quad\times\prod_d^D\prod_i^{N_d} p(s_{di}|\lambda_d)\times\prod_d^D\prod_i^{N_d} p(w_{di}|s_{di},\mu_b)\\
&\quad\times\prod_d^D\prod_i^{N_d} p(z_{di}|s_{di},\kappa_{m_d})\times\prod_d^D\prod_i^{N_d} p(w_{di}|s_{di},\mu_{z_{di}})
\end{aligned}\tag{1}$$

この結合分布は，上記で述べた文書の生成過程において，すべての文書，同行者クラス，同行者，トピックおよび単語についてその発生確率を結合することにより求めることができる．式 (1) を説明するにあたり，数式を「×」で分割し，出現順に第 1 項，第 2 項 $\cdots$ とする．式 (1) 第 1 項は，上記の文書の生成過程 1 において同行者クラスの分布が ι である条件付き確率（パラメータ β が条件）に該当する．式 (1) 第 2 項は，上記の文書の生成過程 2 において，同行者クラスの同行者への分布が ν である条件付き確率（パラメータ ζ が条件）をすべての同行者クラスについて計算および結合した確率に該当する．同様にして式 (1) 第 3 項は上記の文書の生成過程 3 における条件付き確率をすべての同行者クラスについて計算および結合した確率に該当する．また，式 (1) 第 4 項は，上記の文書の生成過程 4 において，トピックもしくは背景トピックから単語への分布が μ である条件付き確率（パラメータ ε が条件）をすべてのトピックおよび背景トピックについて計算および結合した確率に該当する．同様にして，式 (1) 第 5，6，7 項は上記の文書の生成過程 5a)，5b)，5c) における条件付き確率をすべての文書について計算および結合した確率に該当する．また，式 (1) 第 8，9，10，11 項は上記の文書の生成過程 5d-i)，5d-i-a)，5d-i-b)，5d-i-c) における条件付き確率をすべての文書および文書に登場する単語について計算および結合した確率に該当する．

Collapsed Gibbs Sampling では，まず，解析的に直接求めることができないパラメータ κ，ν，μ，λ，ι を積分消去する．上述の式を用いて式 (2) のように積分の形に変形する．

$$\begin{aligned}
&p(\mathbf{e},\mathbf{m},\mathbf{s},\mathbf{z},\mathbf{a},\mathbf{w},\iota,\kappa,\mu,\lambda,\upsilon;\beta,\alpha,\varepsilon,\delta,\zeta)\\
&=\int_\iota\int_\upsilon\int_\kappa\int_\lambda\int_\mu p(\mathbf{m},\mathbf{s},\mathbf{z},\mathbf{a},\mathbf{w},\iota,\kappa,\mu,\lambda,\upsilon;\beta,\alpha,\varepsilon,\delta,\zeta)\\
&\quad d\mu d\lambda d\kappa d\upsilon d\iota\\
&=\int_\iota p(\iota|\beta)\times\prod_d^D p(m_d|\iota)d\iota\\
&\quad\times\int_\upsilon\prod_m^M p(\upsilon_m|\zeta)\times\prod_d^D p(a_d|\upsilon_{m_d})d\upsilon\\
&\quad\times\int_\kappa\prod_j^M p(\kappa_j|\alpha)\times\prod_d^D\prod_i^{N_d} p(z_{di}|s_{di},\kappa_{m_d})d\kappa\\
&\quad\times\int_\lambda\prod_d^D p(\lambda_d|\delta)\times\prod_d^D\prod_i^{N_d} p(s_{di}|\lambda_d)d\lambda\\
&\quad\times\int_\mu\prod_l^{Z+1} p(\mu_l|\varepsilon)\times\prod_d^D\prod_i^{N_d} p(w_{di}|s_{di},\mu_{z_{di}})\\
&\quad\times\prod_d^D\prod_i^{N_d} p(w_{di}|s_{di},\mu_b)d\mu
\end{aligned}\tag{2}$$

次に κ，ν，μ，λ，ι を消去した分布を得る．ここでは，式 (2) の第 3 項および第 4 項にある ν を積分消去する式変形について説明する．式 (3) に示す変形を行う．

$$\begin{aligned}
&\int_\upsilon\prod_m^M p(\upsilon_m|\zeta)\times\prod_d^D p(a_d|\upsilon_{m_d})d\upsilon\\
&\quad=\prod_m^M\int_\upsilon p(\upsilon_m|\zeta)\times\prod_d^D p(a_d|\upsilon_{m,d})d\upsilon
\end{aligned}\tag{3}$$

次に $p(\nu_m|\zeta)$ をディリクレ分布に置換することにより，式 (4) を得る．

$$=\prod_m^M\int_\upsilon\frac{\Gamma\left(\sum_d^D\zeta_d\right)}{\prod_d^D\Gamma(\zeta_d)}\prod_d^D\upsilon_{m,d}^{\zeta_d-1}\prod_d^D p(a_d|\upsilon_{m,d})d\upsilon\tag{4}$$

さらに $p(a_d|\nu_{md})$ を多項分布に置換することにより式 (5) を得る．

$$\begin{aligned}
&=\prod_m^M\int_\upsilon\frac{\Gamma\left(\sum_d^D\zeta_d\right)}{\prod_d^D\Gamma(\zeta_d)}\prod_d^D\upsilon_{m,d}^{\zeta_d-1}\prod_d^D\upsilon_{m,d}^{n_{d,m}}d\upsilon\\
&=\prod_m^M\int_\upsilon\frac{\Gamma\left(\sum_d^D\zeta_d\right)}{\prod_d^D\Gamma(\zeta_d)}\prod_d^D\upsilon_{m,d}^{n_{d,m}+\zeta_d-1}d\upsilon
\end{aligned}\tag{5}$$

式 (5) はディリクレ分布 × 多項分布となっている．ディリクレ分布と多項分布は自然共役の関係にあることを利用し，ディリクレ分布の部分のみ切り出す．その過程を式 (6) に示す．

$$= \prod_m^M \frac{\Gamma\left(\sum_d^D \zeta_d\right)}{\prod_d^D \Gamma(\zeta_d)} \frac{\prod_d^D \Gamma(n_{d,m}+\zeta_d)}{\Gamma\left(\sum_d^D n_{d,m}+\zeta_d\right)} \times \int_v \frac{\Gamma\left(\sum_d^D n_{d,m}+\zeta_d\right)}{\prod_d^D \Gamma(n_{d,m}+\zeta_d)} \prod_d^D v_{m,d}^{n_{d,m}+\zeta_d-1} dv = \prod_m^M \frac{\Gamma\left(\sum_d^D \zeta_d\right)}{\prod_d^D \Gamma(\zeta_d)} \frac{\prod_d^D \Gamma(n_{d,m}+\zeta_d)}{\Gamma\left(\sum_d^D n_{d,m}+\zeta_d\right)} \quad (6)$$

そのほかの分布 κ，μ，λ，ι についても同様に積分消去を行うことにより，式 (7) を得ることができる．

$$\begin{aligned} & p(\mathbf{m},\mathbf{s},\mathbf{z},\mathbf{a},\mathbf{w};\beta,\alpha,\varepsilon,\delta,\zeta) \\ &= \frac{\Gamma\left(\sum_m^M \beta_m\right)}{\prod_m^M \Gamma(\beta_m)} \frac{\prod_m^M \Gamma(n_m+\beta_m)}{\Gamma\left(\sum_m^M n_m+\beta_m\right)} \\ &\quad \times \prod_j^M \frac{\Gamma(\sum_z^Z \alpha_z)}{\prod_z^Z \Gamma(\alpha_z)} \frac{\prod_z^Z \Gamma(n_{j,z}+\alpha_z)}{\Gamma\left(\sum_z^Z n_{j,z}+\alpha_z\right)} \\ &\quad \times \prod_d^D \frac{\Gamma\left(\sum_l^L \delta_l\right)}{\prod_l^L \Gamma(\delta_l)} \frac{\prod_l^L \Gamma(n_{d,l}+\delta_i)}{\Gamma\left(\sum_l^L n_{d,l}+\delta_i\right)} \\ &\quad \times \prod_l^{Z+1} \frac{\Gamma\left(\sum_i^V \varepsilon_i\right)}{\prod_i^V \Gamma(\varepsilon_i)} \frac{\prod_i^V \Gamma(n_{l,i}+\varepsilon_i)}{\Gamma\left(\sum_i^V n_{l,i}+\varepsilon_i\right)} \\ &\quad \times \prod_m^M \frac{\Gamma\left(\sum_a^A \zeta_a\right)}{\prod_a^A \Gamma(\zeta_a)} \frac{\prod_a^A \Gamma(n_{m,a}+\zeta_a)}{\Gamma\left(\sum_a^A n_{m,a}+\zeta_a\right)} \end{aligned} \quad (7)$$

次に，各潜在パラメータに関して Gibbs sampling の更新式を求める．

3.2.1 同行者クラス

Collapsed Gibbs Sampling では，各文書，各単語に対するサンプル（文書に対する同行者クラスや各単語に対するトピック）を全体の確率分布に従ってサンプリングして求める．文書，同行者，単語に対するサンプルは当初ランダムで与えられるが複数回サンプリングを繰り返すことにより，サンプルの変動が少なくなり収束する．サンプリングは 1 つの文書，単語に対して行う．具体的には，サンプリング対象となっている文書や単語のサンプルを除き，それ以外の文書や単語のサンプルから得られる確率分布を用いて，サンプリング対象の文書や単語のサンプリングを実行する．

本項では，文書 d の同行者クラスのサンプリングを行う方法を説明する．まず，文書 d について同行者クラスが h，同行者が y，文書 d のトピックが g となるとき条件付確率を計算する．ここで h だけでなく y，g も条件としたのは文書 d の同行者クラスは，図 1 から分かるように文書 d の同行者 a_d および文書 d のトピック z_d の生成に影響を与えるためである．なお，正確にはトピックは文書ではなく文書内の単語単位に割り振られており，明示的に文書に対してトピックは割り振られていない．ここでは文書内の単語に割り振られているトピックの中で最も多くの単語に割り振られているトピックを各文書に割り振られているトピックとする．文書 d について同行者クラスが h，同行者が y，文書 d のトピックが g となるとき条件付き確率は次式のとおり定義される．

$$p(m_d=h,z_d=g,a_d=y|\mathbf{m}_{\backslash d},\mathbf{z}_{\backslash d},\mathbf{a};\beta,\alpha,\zeta) = \frac{p(\mathbf{m},\mathbf{z},\mathbf{a};\beta,\alpha,\zeta)}{p(\mathbf{m}_{\backslash d},\mathbf{z}_{\backslash d},\mathbf{a};\beta,\alpha,\zeta)} \quad (8)$$

ここで，$\mathbf{m}_{\backslash d}$ はすべての文書に割り当てられたサンプルの同行者クラスの集合（文書 d に割り当てられた同行者クラスは除く）を表している．また，$\mathbf{z}_{\backslash d}$ はすべての単語に割り当てられたトピック（文書 d 内の単語に割り当てられたトピックは除く）の集合を表している．$\mathbf{a}$ はすべての文書に割り当てられた既知の同行者の集合を表している．式 (8) は m，z，a の結合分布であることから，各確率分布をディリクレ分布に置換し次式を得る．

$$\begin{aligned} &= \frac{\frac{\Gamma\left(\sum_m^M \beta_m\right)}{\prod_m^M \Gamma(\beta_m)} \frac{\prod_m^M \Gamma(n_m+\beta_m)}{\Gamma\left(\sum_m^M n_m+\beta_m\right)}}{\frac{\Gamma\left(\sum_m^M \beta_m\right)}{\prod_m^M \Gamma(\beta_m)} \frac{\prod_m^M \Gamma(n_{m\backslash d}+\beta_m)}{\Gamma\left(\sum_m^M n_{m\backslash d}+\beta_m\right)}} \times \frac{\frac{\Gamma\left(\sum_z^Z \alpha_z\right)}{\prod_z^Z \Gamma(\alpha_z)} \frac{\prod_z^Z \Gamma(n_{h,z}+\alpha_z)}{\Gamma\left(\sum_z^Z n_{h,z}+\alpha_z\right)}}{\frac{\Gamma\left(\sum_z^Z \alpha_z\right)}{\prod_z^Z \Gamma(\alpha_z)} \frac{\prod_z^Z \Gamma(n_{h,z\backslash d}+\alpha_z)}{\Gamma\left(\sum_z^Z n_{h,z\backslash d}+\alpha_z\right)}} \\ &\quad \times \frac{\frac{\Gamma\left(\sum_a^A \zeta_a\right)}{\prod_a^A \Gamma(\zeta_a)} \frac{\prod_a^A \Gamma(n_{h,a}+\zeta_a)}{\Gamma\left(\sum_a^A n_{h,a}+\zeta_a\right)}}{\frac{\Gamma\left(\sum_a^A \zeta_a\right)}{\prod_a^A \Gamma(\zeta_a)} \frac{\prod_a^A \Gamma(n_{h,a\backslash d}+\zeta_d)}{\Gamma\left(\sum_a^A n_{h,a\backslash d}+\zeta_a\right)}} \\ &= \frac{\prod_m^M \Gamma(n_m+\beta_m)}{\Gamma\left(\sum_m^M n_m+\beta_m\right)} \times \frac{\prod_z^Z \Gamma(n_{h,z}+\alpha_z)}{\Gamma\left(\sum_z^Z n_{h,z}+\alpha_z\right)} \times \frac{\prod_a^A \Gamma(n_{h,a}+\zeta_a)}{\Gamma\left(\sum_a^A n_{h,a}+\zeta_a\right)} \\ &\quad \frac{\prod_m^M \Gamma(n_{m\backslash d}+\beta_m)}{\Gamma\left(\sum_m^M n_{m\backslash d}+\beta_m\right)} \times \frac{\prod_z^Z \Gamma(n_{h,z\backslash d}+\alpha_z)}{\Gamma\left(\sum_z^Z n_{h,z\backslash d}+\alpha_z\right)} \times \frac{\prod_a^A \Gamma(n_{h,a\backslash d}+\zeta_a)}{\Gamma\left(\sum_a^A n_{h,a\backslash d}+\zeta_a\right)} \end{aligned} \quad (9)$$

文書 d について同行者クラスが h，同行者が y，文書 d の i 番目の単語のトピックが g であることを利用し，式 (9) を式 (10) のように変形する．

$$\begin{aligned} &= \frac{\prod_{m\neq h}^M \Gamma(n_m+\beta_m)\Gamma(n_h+\beta_h)}{\Gamma\left(\sum_m^M n_m+\beta_m\right)} \Big/ \frac{\prod_{m\neq h}^M \Gamma(n_{m\backslash d}+\beta_m)\Gamma(n_{h\backslash d}+\beta_h)}{\Gamma\left(\sum_m^M n_{m\backslash d}+\beta_m\right)} \\ &\quad \times \frac{\prod_{z\neq g}^Z \Gamma(n_{h,z}+\alpha_z)\Gamma(n_{h,g}+\alpha_g)}{\Gamma\left(\sum_z^Z n_{h,z}+\alpha_z\right)} \Big/ \frac{\prod_{z\neq g}^Z \Gamma(n_{h,z\backslash d}+\alpha_z)\Gamma(n_{h,g\backslash d}+\alpha_g)}{\Gamma\left(\sum_z^Z n_{h,z\backslash d}+\alpha_z\right)} \\ &\quad \times \frac{\prod_{a\neq y}^A \Gamma(n_{h,a}+\zeta_a)\Gamma(n_{h,y}+\zeta_y)}{\Gamma\left(\sum_a^A n_{h,a}+\zeta_a\right)} \Big/ \frac{\prod_{a\neq y}^A \Gamma(n_{h,a\backslash d}+\zeta_y)\Gamma(n_{h,y\backslash d}+\zeta_y)}{\Gamma\left(\sum_a^A n_{h,a\backslash d}+\zeta_a\right)} \end{aligned} \quad (10)$$

ここで，$n_{h\backslash d}$ は同行者クラス h に割り当てられた文書の数（文書 d に割り当てられた同行者クラスは除く）を表している．また，$n_{h,g\backslash d}$ は同行者クラス h に割り当てられた文書の中で，トピックに g が割り当てられた単語の数を表す（ただし，文書 d 内のすべての単語へのトピックの割り当ては除く）．また，$n_{h,y\backslash d}$ は同行者クラス h に割り当てられた文書の中で同行者 y に割り当てられた文書の数（ただ

し，文書 d に割り当てられた同行者クラスは除く）を表している．

いま，更新対象の文書 d に割り当てられている同行者クラスは h であることから，文書 d に割り当てられている同行者クラス m（同行者クラスが h である場合を除く）の数（$n_{m(m\neq h)\cap d}$）は 0 となる．したがって，以下の等号が成り立つ．

$$\begin{aligned} n_{m(m\neq h)} &= n_{m(m\neq h)\backslash d} + n_{m(m\neq h)\cap d} = n_{m(m\neq h)\backslash d} + 0 \\ n_h &= n_{h\backslash d} + n_{h\cap d} = n_{h\backslash d} + 1 \end{aligned} \tag{11}$$

また，更新対象の文書 d に割り当てられている同行者クラスは h，トピックは g であることから，文書 d の同行者クラスが h およびトピックが z（トピックが g である場合を除く）となる場合（$n_{h,z(z\neq g)\cap d}$）は 0 となる．したがって，以下の等号が成り立つ．

$$\begin{aligned} n_{h,z(z\neq g)} &= n_{h,z(z\neq g)\backslash d} + n_{h,z(z\neq g)\cap d} = n_{h,z(z\neq g)\backslash d} + 0 \\ n_{h,g} &= n_{h,g\backslash d} + n_{h,g\cap d} = n_{h,g\backslash d} + 1 \end{aligned} \tag{12}$$

また，更新対象の文書 d に割り当てられている同行者クラス h，同行者は y であることから，文書 d の同行者クラスが h および同行者が a（同行者が y である場合を除く）となる場合（$n_{h,a(a\neq y)\cap d}$）は 0 となる．したがって，以下の等号が成り立つ．

$$\begin{aligned} n_{h,a(a\neq y)} &= n_{h,a(a\neq y)\backslash d} + n_{h,a(a\neq y)\cap d} = n_{h,a(a\neq y)\backslash d} + 0 \\ n_{h,y} &= n_{h,y\backslash d} + n_{h,y\cap d} = n_{h,y\backslash d} + 1 \end{aligned} \tag{13}$$

式 (11)，(12)，(13) を式 (9) に代入することにより次式を得る．

$$\begin{aligned} &= \frac{\prod_{m\neq h}^{M}\Gamma(n_{m\backslash d}+0+\beta_m)\Gamma(n_{h\backslash d}+1+\beta_h)}{\Gamma\left(\sum_m^M(n_{m\backslash d}+\beta_m)+1\right)} \Big/ \frac{\prod_{m\neq h}^{M}\Gamma(n_{m\backslash d}+\beta_m)\Gamma(n_{h\backslash d}+\beta_h)}{\Gamma\left(\sum_m^M n_{m\backslash d}+\beta_m\right)} \\ &\quad\times \frac{\prod_{z\neq g}^{Z}\Gamma(n_{h,z\backslash di}+0+\alpha_z)\Gamma(n_{h,g\backslash di}+1+\alpha_g)}{\Gamma\left(\sum_z^Z(n_{h,z\backslash di}+\alpha_z)+1\right)} \Big/ \frac{\prod_{z\neq g}^{Z}\Gamma(n_{h,z\backslash di}+\alpha_z)\Gamma(n_{h,g\backslash di}+\alpha_g)}{\Gamma\left(\sum_z^Z n_{h,z\backslash di}+\alpha_z\right)} \\ &\quad\times \frac{\prod_{a\neq y}^{A}\Gamma(n_{h,a\backslash d}+0+\zeta_a)\Gamma(n_{h,y\backslash d}+1+\zeta_y)}{\Gamma\left(\sum_a^A(n_{h,a\backslash d}+\zeta_a)+1\right)} \Big/ \frac{\prod_{a\neq y}^{A}\Gamma(n_{h,a\backslash d}+\zeta_y)\Gamma(n_{h,a\backslash d}+\zeta_y)}{\Gamma\left(\sum_a^A n_{h,a\backslash d}+\zeta_a\right)} \\ &= \frac{n_{h\backslash d}+\beta_h}{\sum_m^M n_{m\backslash d}+\beta_m} \times \frac{n_{h,g\backslash di}+\alpha_g}{\sum_z^Z n_{h,z\backslash di}+\alpha_z} \\ &\quad\times \frac{n_{h,y\backslash d}+\zeta_y}{\sum_a^A n_{h,a\backslash d}+\zeta_a} \end{aligned} \tag{14}$$

この計算をすべての同行者クラスについて行う．そして，同行者クラスごとの条件付き確率に従って文書 d の同行者クラスのサンプルを選ぶ．

3.2.2 スイッチ変数とトピック

文書 d の i 番目の単語 t についてトピックが k（文書 d の i 番目の単語のスイッチ変数は 0）となるとき条件付確率を以下のように求める．

$$\begin{aligned} &p(s_{di}=0, z_{di}=k, w_{di}=t|h, \mathbf{s}_{\backslash di}, \mathbf{z}_{\backslash di}, \mathbf{w}; \delta, \varepsilon, \alpha) \\ &= \frac{n_{d,0\backslash di}+\delta_1}{\sum_l^L n_{d,l\backslash di}+\delta_i} \times \frac{n_{h,k\backslash di}+\alpha_k}{\sum_z^Z n_{h,z\backslash di}+\alpha_z} \\ &\quad\times \frac{n_{k,t\backslash di}+\varepsilon_i}{\sum_r^R n_{k,r\backslash di}+\varepsilon_i} \end{aligned} \tag{15}$$

ここで，$n_{d,0\backslash di}$ は文書 d 内でスイッチ変数 0 に割り当てられた単語の数（文書 d の i 番目の単語に割り当てられたスイッチ変数は除く）を表している．また，$n_{h,k,\backslash di}$ は同行者クラスに h が割り当てられた文書の中でトピック k を持つ単語の数を表す（ただし，文書 d の i 番目の単語に割り当てられたトピックは除く）．$n_{k,t\backslash di}$ はトピックに k が割り当てられた単語 t の数を表す（ただし，文書 d の i 番目の単語に割り当てられたトピックは除く）．

文書 d の i 番目の単語 t についてトピックが b（文書 d の i 番目の単語のスイッチ変数は 1）となるとき条件付き確率を以下のように求める．

$$\begin{aligned} &p(s_{di}=1, z_{di}=b, w_{di}=t|\mathbf{s}_{\backslash di}, \mathbf{z}_{\backslash di}, \mathbf{w}; \delta, \varepsilon) \\ &= \frac{n_{d,1\backslash di}+\delta_0}{\sum_l^L n_{d,l\backslash di}+\delta_i} \times \frac{n_{b,t\backslash di}+\varepsilon_v}{\sum_r^R n_{b,r\backslash di}+\varepsilon_i} \end{aligned} \tag{16}$$

ここで，$n_{d,1\backslash di}$ は文書 d 内でスイッチ変数 1 に割り当てられた単語の数（文書 d の i 番目の単語に割り当てられたスイッチ変数は除く）を表している．また，$n_{b,t\backslash di}$ はトピックに b が割り当てられた単語 t の数を表す（ただし，文書 d の i 番目の単語について割り当てられたトピックは除く）．

3.3 同行者依存ワードの抽出

本節では，同行者に強く依存する単語（同行者依存ワード）を抽出する方法について述べる．一般に使われる単語をすべてカバーするのは非常に負荷が高いため，まずは品詞で絞ることとする．同行者と一緒にいるときは何らかの活動に参加していることが多く，その活動を表す品詞として動詞に着目する．Yize らも，同行者が強く関連する文書では名詞に比べ，動詞のほうが強く関連していると述べている．また，動詞は名詞に比べ種類が少なく，辞書の整備のコストが安い．たとえば「夫とハワイに旅行する」といった場合名詞（目的地）は無限にあり得るが，動詞は「旅行する」「旅する」など数種類をカバーするのみでよい．そこで，本研究では，同行者に依存する単語として動詞のみを対象とする．なお，以下の述べる手法は名詞にも対応可能であるが，本稿では動詞のみを同行者依存ワードとして扱った場合の有効性を示す．次に，動詞の中で同行者への依存度合いを計算する．ここでは，あらゆる同行者依存ワード y に対して，動詞 i と同行者 q 間の近さを計測し，掛け合わせることにより，動詞 i の同行者全体への依存度を計算する．まず，Yize らの研究で用いられてきた語彙統語パターン $b(i,q)$ =「動詞 i + with + 同行者依存ワード q」

（たとえば，play with father）といった語彙統語パターンを作成する．なお，「deal with」などの動詞句の場合，「deal with father」と「deal with the problem with father」のような異なる with の意味を持つものがあるが，動詞と同行者の依存度合いの観点では両方とも重要と見なし本稿では特別な対応は行わない．次に，同行者依存ワード q が含まれる文章の中で語彙統語パターン $b(i,q)$ の発生する条件付確率を計算する．

$$P(i|q) = \frac{pagecount(b(i,q))}{pagecount(q)} \tag{17}$$

ここで，$pagecount(k)$ は，検索エンジンでクエリ k を検索したページカウント数をとする．次に，すべての同行者 q に対する条件付確率を掛け合わせ，すべての同行者と動詞 i 間の近さの度合い $R(i)$ を計算する．

$$R(i) = \log\left(\prod_q P(i|q)\right) \tag{18}$$

さらに，正規化を次式のように行い，スイッチ変数を発生させるディリクレ事前分布のパラメータ δ_i とする．これは，動詞 i がスイッチ変数 $= 0$ となる確率をあらわしている．

$$\delta_i = \frac{R(i)}{\max_i(R(i))} \tag{19}$$

なお，検索エンジンとしてどのエンジンを利用してもよい．ただし，本稿ではプログラムからのアクセス制限が他エンジンに比べ軽減されている Bing を用いている．

4. データセットの構築

トピックモデルに内在する潜在クラスの確率分布を学習するためには学習データが必要となるが，時間と場所に応じたトピックモデルに関しては文書に正確な情報が付与されているデータを用いることが可能である．たとえば位置であれば GPS によって取得された緯度経度情報が利用可能である．一方，一般に明示的に同行者に関する情報が付与された文書情報はなく，同行者に応じたトピックモデルを学習するためには学習用データから構築する必要がある．ここでは，Web から同行者に関する情報を抽出することで学習データを構築する．Yize らの手法と同じく，「with 同行者」となっている投稿を抽出する．対象となる同行者は英語の学習サイト [22] から抽出した．対象となる同行者の一部を**表 3** に示す．同行者の総数は計 93 個である．Bing 検索エンジンを用い，表 3 に示す同行者ごとに「with 同行者」の形式でクエリを作成し，その検索結果からスニペット部分のみを抽出した．そして各スニペットを文書と扱う．結果として，総文書数は計 7,061 個，総ユニーク単語数は計 18,929 個からなる学習データを収集した．過去のグラフィカルモデルを用いた研究例に習い，ストップワードなどの前処理は行っていない [17]．また，fCTM では，このデータに対して 3.3 節で述べた手法を適用し，ディリクレ事前分布のパラメータを設定している．動詞は計 509 個抽出された．同行者に依存度のつよい動詞として上位から live，share，mend，deflect，eat，handle，encourage が抽出された．なお，本実験では英語の文書集合を対象にしているが，提案手法は他言語においても適用可能である．ただし，3.3 節で述べた語彙統語パターンを言語に応じて設計する必要がある．

表 3 データセット作成のための同行者
Table 3 Companions by family used to create dataset.

Category	Family and relatives
Family	husband, wife, spouse, father, mother, parents, son, daughter, child, children, brother, sister, siblings, twins;
Relatives	uncle, aunt; nephew, niece, cousin, first cousin, second cousin;
Relatives by marriage	in-laws, father-in-law, mother-in-law,brother-in-law, brothers-in-law, sister-in-law, sisters-in-law;
Age groups	child, baby, infant; boy, girl, teenager, adolescent; adult, grownup;
Marital status	fiance, bride, ex-husband, ex-wife, girlfriend, boyfriend, widower, widow;

5. 評価実験

5.1 Perplexity によるパラメータチューニング

ここでは，トピックモデルの文書の予測精度の評価指標の 1 つである Perplexity を利用し，同行者クラスの数（M）およびトピック数（Z）の最適値を探す．Perplexity とはグラフィカルモデルにおける全単語の対数尤度を反映したものであり，文書の予測精度を評価する手法として一般的に利用されている．Perplexity は尤度の逆数で表され低いほうが予測精度が高い．ここでは交差検証により Perplexity の計算を行う．まず，4 章で得られたデータセットの 1/3 をテストデータとし，残りを学習データとする．学習データを用いて分布 $\lambda_{d,}$，κ_m，および μ_z を推定する．そして，CGS の繰り返し回数 100 ごとに，学習データによって得られた分布を用いてテストデータの Perplexity を計算する．具体的には，テストデータの各文書の同行者クラスおよび各単語のトピックのサンプリングを 10 回行う．そして次式によって Perplexity を計算する．

$$PPX = \exp\left(-\frac{1}{\sum_{d=1}^{M} N_d}\right.$$

表 4 パラメータの定義
Table 4 Parameters sets.

Variable	Value
β	1/同行者クラスの総数
ζ	1/同行者総数
α	1/トピックの総数
δ	sCTM の場合 1/2. fCTM の場合，動詞は式(8)の学習結果，動詞以外は 1/2.
ε	1/ユニークな単語数

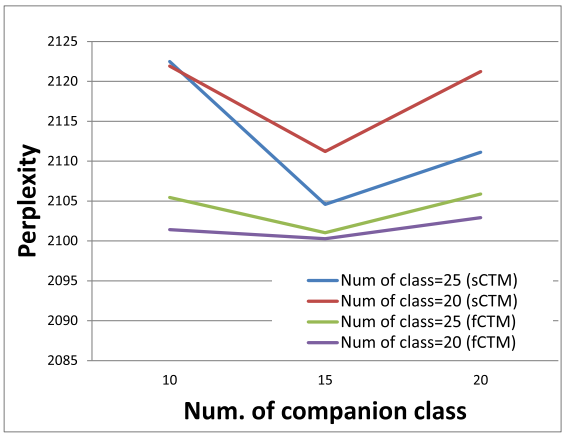

図 4 sCTM，fCTM における同行者クラス数のチューニング結果
Fig. 4 Tuning result of the number of companion class at sCTM and fCTM.

$$\times \sum_{d \in D} \sum_{i \in D} \log \frac{1}{S} \left(\lambda_0 \mu_{bi} + \sum_{z}^{Z} \lambda_1 \kappa_{mz} \mu_{zi} \right) \Bigg) \tag{20}$$

S はサンプルの数を表す（ここでは 10 である）．また，本実験においては初期のサンプルの割り当てをランダムに行っているが，偏りを軽減するため，初期のサンプルの割り当てを 10 回変えて異なる初期のサンプルデータから開始している．ここではその 10 回の実験の結果得られた Perplexity の平均値を示す．通常，Perplexity に与える初期値の違いによる影響は 1%以下であり，本実験においても 0.5%以下であった．実験に用いたマシンのスペックは CPU: Intel Xeon 2.66 GHz QuadCore，Memory: 2 GB ∗ 2 である．なお，ここでは，各ディリクレ事前分布のパラメータを**表 4** に示すとおりに設定した．

まず，トピック数を 20 および 25 で固定し，同行者クラスを 10，15，20 に変化させ Perplexity の変化の様子を観察した．その結果を**図 4** に示す．sCTM，fCTM ともに Perplexity が最小となる同行者クラス数（= 15）に設定する．

次に，同行者クラス数を 15 で固定し，トピック数を 10，15，20，25，30 に変化させ Perplexity の変化の様子を観察した．その結果を**図 5** に示す．sCTM，fCTM ともに

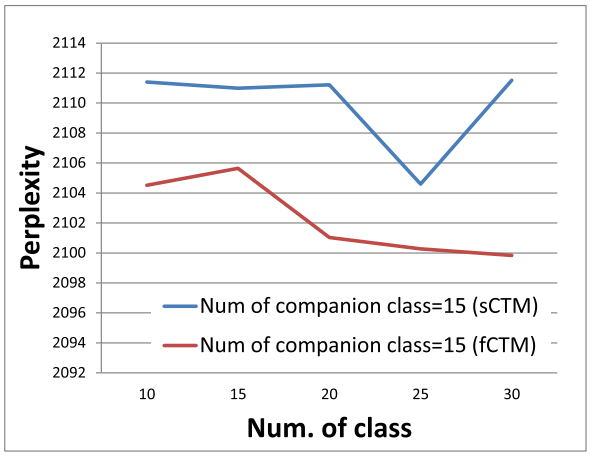

図 5 sCTM，fCTM におけるトピック数のチューニング結果
Fig. 5 Tuning result of the number of topic class at sCTM and fCTM.

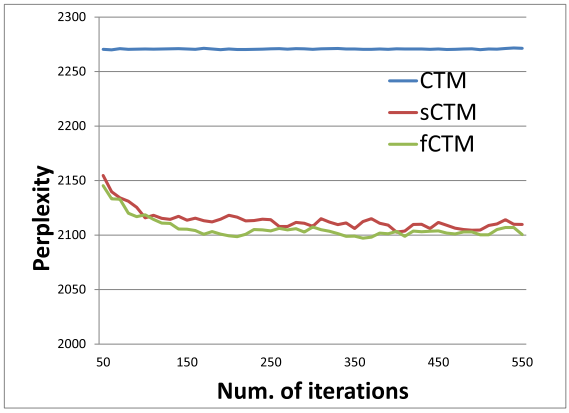

図 6 CTM，sCTM，fCTM 間の Perplexity の比較
Fig. 6 Perplexity comparison between CTM, sCTM and fCTM.

Perplexity が最小に近いトピック数（= 25）に設定する．ここで，トピック数 10 および 20 の場合の Perplexity とトピック数 25 の場合の Perplexity に有意差について t 検定を行った．その結果，前者は $p = 0.0042 < 0.05$，後者は $p = 0.0413 < 0.05$ となり，有意差があることが分かった．なお，本稿では Perplexity によりクラス数のチューニングを行っているが，文書生成過程で動的に最適なクラス数を求める手法も提案されている [21]．提案手法に動的なクラス数を求める手法を組み込むことにより，チューニングに要するコストを削減することが可能である．

5.2 Perplexity による文書予測精度評価

CTM，sCTM，fCTM の学習結果の Perplexity を比較する．なお，1 章で述べたように CTM は既存のトピックモデルである LDA に対し既存手法にしたがって同行者を考慮できるよう拡張したものである．そのため，ここでは CTM をベースラインとして設定し，そのほかの手法と比較する．**図 6** に結果を示す．図 6 に示すとおり，CTM に比べ sCTM および fCTM の Perplexity は低い．つまり，

表 5 CTM，sCTM，fCTM の KL-Divergence の平均値
Table 5 Average KL-divergence comparison of CTM, sCTM and fCTM.

CTM	sCTM	fCTM
0.960	1.368	1.415

fCTM，sCTM で学習した結果から生成される文章の予測精度は CTM に比べ高いことを意味する．これにより，sCTM および fCTM においてスイッチ変数により同行者に特有の単語を絞っていることの効果が表れたといえる．また，sCTM に比べ，fCTM の Perplexity は低い．これは，fCTM が事前に Web 全体の傾向から同行者に特有の単語のみに絞ってクラスタリングを行っており，結果として同行者に依存した単語から構成される文章を生成しているためと考えられる．

5.3 KL-Divergence による評価

一般にトピックモデルにおいて生成されたトピックどうしは意味的により分離していることが望ましい．トピック間の分離性能を量的に評価するため，KL-Divergence により評価を行った．クラス z_1 とクラス z_2 間の KL-Divergence は次式で表される．

$$KL(z_1, z_2) = \sum_{\mathbf{w}} p(w|z_1) \log \frac{p(w|z_1)}{p(w|z_2)} \tag{21}$$

ここで，KL-Divergence が大きいほどその 2 つのトピックは互いに分離性が高いといえる．一方，0 の場合は，2 つのトピックはまったく同一である．ここでは，CTM，sCTM，f CTM をそれぞれクラス数 25 のときの KL-Divergence を計算する．また，10 回初期のランダムなサンプルを変え，そのたびに KL-Divergence を計算し，10 回の平均値を計算する．結果を**表 5** に掲載する．表 5 において，t 検定を行い，CTM–sCTM 間では $p = 1.8 \times 10^{-73}$，CTM–fCTM 間では $p = 2.9 \times 10^{-46}$，CTM–sCTM 間では $p = 2.9 \times 10^{-7}$ の検定結果が得られた．表 5 から分かるように，fCTM，sCTM のほうが CTM に比べトピック間の分離性が高くなっている．これは，fCTM，sCTM においてはスイッチ変数により同行者に特有の単語を絞っているが CTM ではそのような機構がなくトピック間で共通の単語が混在しているためと考えられる．また，fCTM のほうが sCTM に比べトピック間の分離性能が高くなっている．これは，fCTM においては，あらかじめ，Web 全体から各動詞の同行者依存度合いを学習しモデルに組み込むことにより，同行者に特有の単語のみでのクラスタリングを可能にしているからである．そのため，トピック間の分離性能が高くなっていると考えられる．

5.4 質的評価

表 6 に fCTM の学習結果の一部を示す．表では，各同行者クラスと，それに対応するトピックの対応関係を示している．各同行者クラスにはそのクラスに属する同行者の集合，および単語が記載されている．各同行者クラスの単語は，各同行者クラスで最も属する確率の高いトピックを抽出，そのトピックに属する単語を表示している．表 6 から分かるように，同行者クラスの同行者（例：child，husband，parents）と対応するトピックの単語を確認すると，love，cope など同行者に関係のある単語が上位の単語として抽出されている．同行者クラスを観察すると，逆にどの立場の人がこれらの同行者依存のトピックを生成しているのか分かる．co0 の場合，child，wife があることから父親が家族といるときに発生するトピックである．同様に co6，co11 の場合はそれぞれ母親が家族といるときのトピック，子供が家族といるときのトピックであることが分かる．

表 6 提案モデル fCTM による同行者クラスの単語分布
Table 6 Results of distributions of words associated with each latent companion class by fCTM.

Latent companion class	co0	co6	co11
associated companion	child	child	father
	son	husband	sister
	wife	parents	brother
The highest topic	class4	class12	class0
associated words	chat	love	help
	study	cope	spend
	know	shower	contest
	need	beat	ask
	deal	hurt	make

表 7 sCTM による同行者クラスの単語分布
Table 7 Results of distributions of words associated with each latent companion class by sCTM.

Latent companion class	co5	co8	co14
associated companion	parents	daughter	daughter
	father	wife	son
	mother	son	brother
The highest topic	class19	class11	class16
associated words	vacation	enjoying	pregnant
	chat	aniston	red
	think	celebrate	enjoyed
	perfect	ready	golden
	share	jewelry	birthday

表 7 に sCTM の学習結果の一部を示す．fCTM と同様に，各同行者クラスともに，どの立場のトピックかが明確である．co5，co8，co14 それぞれ子ども，父親，両親が家

族といるときに発生するトピックであると考えられる．同行者クラスの同行者（例：co5 の parents，father）と対応するトピックの単語を確認すると，vacation，chat などの関係のある単語が登場しており妥当な結果といえる．

しかしながら sCTM では，Aniston といった人名が分類されている．人名はどのような同行者にもなりえるため同行者トピックを予測する際のノイズとなる可能性がある．たとえば，Aniston は mother にも daughter にもなりえる．このような単語が上位の単語として分類されてしまった原因として，モデルを学習する際に学習データのデータ量が少なく特定の文書の影響を受けているためと考えられる．fCTM では，事前に各動詞について Web 全体から検索エンジンを用いて同行者の依存度合いを学習しておき，モデルに組み込んでいる．そのため，同行者とは無関係の単語が登場していない．

Yize らは，飲食店口コミサイトから同行者推定用の辞書を構築している．Yize らは辞書を公開しておらず，「Friend」と一緒にいるときに発生する単語として「drink」をあげているのみであるため単純な比較はできないが，同行者ごとに特徴となる単語を抽出していると想定される．しかしながら，同行者ごとに抽出された単語集合内の単語間の関係については考慮されていない．つまり同一の同行者の特徴となる単語集合に複数の異なるトピックが含まれる可能性がある．提案手法では，同行者クラスとトピックを同時に考慮しているため，同一の同行者でも複数のトピックが発生する場合には，表 6 に示す Child のように複数のトピックを分離することが可能である．

上記の実験結果からユーザの同行者が分かれば，ユーザ–同行者間で発生するトピックを予測可能であることが分かった．また，同行者に依存したトピック推定の結果，そのユーザが投稿する文書（たとえば Tweet 投稿，Facebook 投稿）の内容についても予測可能である．逆に，ユーザの投稿した文書から同行者クラスを推定することも可能である．提案方式のアプリケーションとして，スケジューラの一機能として情報のレコメンデーションが考えられる．具体的には，スケジューラの入力内容から誰とどの時間帯に一緒にいるかを推定する．そして，そこで発生するトピックを予測し，当該トピックと関連する情報（店舗や情報コンテンツ）を提示する．また，Twitter と連動する情報提示アプリケーションも考えられる．具体的には，Tweet 内容からトピックおよび同行者クラスを推定し，推定された同行者クラスに属する別のトピックを提示することにより，発見性の高いレコメンデーションが可能になる．

6. 結論

本稿では，同行者依存のトピックの発見モデルを提案した．従来手法とは，同行者の予測精度の観点で評価し提案手法の優位性を示した．また，質的評価も行い，同行者のトピックの妥当なモデル化が行われていることを確認した．今後は，大規模なデータを用いて学習することにより予測精度のさらなる向上を目指す．また，その他のコンテキスト（時間や位置）も考慮したトピックモデルのモデル化を目指す．

参考文献

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, pp.993–1022 (2003).

[2] Andrzejewski, D., Zhu, X. and Craven, M.: Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors, *Proc. ICML* (2009).

[3] AlSumait, L., Barbara, D. and Domeniconi, C.: Online LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking, *Proc. ICDM*, pp.3–12 (2008).

[4] Ahmed, A. and Xing, E.P.: Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream, *Proc. Uncertainty in Artificial Intelligence* (*UAI*), pp.20–29 (2010).

[5] Wang, X., McCallum, A. and Wei, X.: Topical N-grams: Phase and Topic Discovery, with an Application to Information Retrieval, *Proc. ICDM*, pp.697–702 (2007).

[6] Ahmed, A., Low, Y., Aly, M., Josifovski, V. and Smola, A.J.: Scalable Distributed Inference of Dynamic User Interests for Behavioral Targeting, *Proc. KDD* (2011).

[7] Chen, Y., Pavlov, D. and Canny, J.F.: Large-scale behavioral targeting, *Proc. KDD*, pp.209–218 (2009).

[8] Duda, R.O., Hart P.E. and Stork, D.G.: Pattern classification, 2nd ed., Wiley (2000).

[9] Abdi, H. and Williams, L.J.: Principal component analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol.2, pp.433–459 (2010).

[10] Hofmann, T.: Probabilistic Latent Semantic Indexing, *Proc. SIGIR* (1999).

[11] Adomavicius, G. and Tuzhilin, A.: Context-Aware Recommender Systems, *Recommender Systems Handbook*, pp.217–253 (2011).

[12] Fukazawa, Y., Luther, M., Wagner, M., Tomioka, A., Naganuma, T., Fujii, K. and Kurakake, S.: Situation-aware Task-based Service Recommendation, *Proc. MobiSys* (2006).

[13] Blei, D. and Lafferty, J.: Dynamic topic models, Vol.23, pp.113–120 (2006).

[14] Wang, X. and McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends, *Proc. KDD*, pp.424–433 (2006).

[15] Kawamae, N.: Trend analysis model: trend consists of temporal words, topics, and timestamps, *Proc. WSDM*, pp.317–326 (2011).

[16] Eisenstein, J., O'Connor, B., Smith, N.A. and Xing, E.P.: A latent variable model for geographic lexical variation, *Proc. EMNLP*, pp.1277–1287 (2010).

[17] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J. and Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream, *Proc. WWW*, pp.769–778 (2012).

[18] Böhm, S., Koolwaaij, J., Luther, M., Souville, B., Wagner, M. and Wibbels, M.: Introducing IYOUIT, *Proc. International Semantic Web Conference*, pp.804–817 (2008).

[19] Li, Y., Nie, J., Zhang, Y., Wang, B., Yan, B. and Weng, F.: Contextual Recommendation based on Text Mining, *Proc. COLING*, pp.692–700 (2010).
[20] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. National Academy of Sciences of the United States of America* (2004).
[21] Ahmed, A. and Xing, E.P.: Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: With Applications to Evolutionary Clustering, *Proc. SIAM International Conference on Data Mining*, pp.219–230 (2008).
[22] Useful English, available from ⟨http://usefulenglish.ru/vocabulary/jobs-professions-occupations⟩.

深澤 佑介 （正会員）

2002 年東京大学工学部卒業．2004 年東京大学大学院工学系研究科修士課程修了．同年株式会社 NTT ドコモ入社．2011 年東京大学大学院工学系研究科博士後期課程修了．同年 10 月より東京大学人工物工学研究センターにて協力研究員兼任．IEEE，人工知能学会各会員．博士（工学）．

太田 順

1989 年東京大学大学院工学系研究科精密機械工学専攻修士課程修了．同年新日本製鐵（株）入社．1991 年東京大学工学部助手．1994 年同講師．1996 年東京大学大学院工学系研究科精密機械工学専攻助教授．2007 年同准教授．2009 年 4 月同教授．同年 6 月より東京大学人工物工学研究センター教授．この間 1996～1997 年 Stanford 大学 Center for Design Research 客員研究員，マルチエージェントロボット，大規模生産/搬送システム設計と支援，移動知，人の解析と人へのサービスの研究に従事．博士（工学）．