Vol.2012-MBL-63 No.3
2012/8/30

同行者依存のトピック発見モデル

深澤佑介[†1,2]　太田順[†2]

コンテキストはユーザの興味・嗜好に影響する非常に重要な要因である。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（会話相手、一緒に行動する人）に注目し、同行者依存のトピック発見モデルを提案する。ベイジアン階層プロセスによってモデル化を行い、Collapsed Gibbs Sampling に基づきモデルの推論を行う。Twitter から同行者依存の投稿データを抽出し、提案モデルと LDA の比較実験を実施した。従来手法とは同行者の予測精度の観点で比較評価し提案手法の優位性を示した。また、質的評価も行い、妥当な同行者のトピックのモデル化が行われていることを確認した。

Companion Dependent Topic Discovery Model

YUSUKE FUKAZAWA[†1,2]　JUN OTA[†2]

Context is understood as an important factor that affects user's preferences or topics occurred. Unlike other models that considers context of time and location, we focus on companion of users (friends, wife, husband etc.) as the most important factor to determine the topic of conversation occurred. To find the topics under the context of companion, we extend LDA(Latent Dirichlet Allocation) model by introducing latent companion class into document layer and latent switch variable into word layer. The latent companion class has a probability distribution over words, topics and the companion that is associated with each document. The switch variable is used as a document specific probabilistic distribution to judge which class (background, latent companion class and latent preference class) each word comes from for generating words in each token. We conduct experiments on two data sets, and they show that the proposed model can capture the topics dependent on context of companion, and we show it is useful as a generative model in the analysis of the topic change depending on context of companion.

1. はじめに

近年、ベイズ推定を用いた様々な文書生成モデル（document generative model）が提案されている。文書生成モデルの最も基礎的なモデルは 2006 年に Bleiet al.によって提案された LDA（Latent Dirichlet allocation）[1]があり、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。既存のクラスタリング手法（K-means や LSA(Latent Semantic Analysis)）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（Context）を考慮していない。たとえば、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、それぞれのトピックが異なるにも関わらず､それを区別することができない｡現在、LDA を拡張する形で様々な Context を考慮した文書生成モデルが提案されている。

まず、Context について定義する。ユーザの Context 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、Context 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。Adomavicious et al.は、情報推薦やトピック抽出において重要な Context とは、「時間」「場所」「同行者」であると定義している[13]。また、著者らは、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 et al.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[16]。これらのことから、Context には、「時間」「場所」「同行者」の 3 つが重要な要素であるといえる。

過去の Context を考慮した文書生成モデルとして､3 つの Context の中で「時間」「場所」に応じた文書生成モデルが非常に多く提案されている。しかしながら「同行者」を考慮した文書生成モデルについては提案されていない。そこで本研究では､「同行者」を考慮した文書生成モデルを提案する。

提案モデルは二つの特徴を有する。第一に、文書のトピックは同行者によって決まるトピック（例：家族と一緒に夕食）と、同行者に関係なく自分の趣味嗜好で決まるトピックがあるが、提案モデルでは切り分けて考慮することができる。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。第二に、文書中で使われる単語について、同行者によって決まる単語／ユーザの興味（トピック）によって決まる単語（(同行者に無関係)）／ユーザの興味や同行者に無関係に決まる単語（a,the など）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単語を明示的に獲得可能である。この結果を利用することで、ユーザが明示的に同行者について述べていなくても単語の集合からユーザの同行者を確率的に推定することが可能である。

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文書生成モデルを提案する。4 章にて同行者付きの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 Context に応じた文書生成モデル

Context 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 et al.は、DTM(Dynamic Topic Model) を提案している[9]。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 et al.は、TOT（Topics Over Time）を提案している[8]。このモデ

†1 株式会社 NTT ドコモ
NTT DOCOMO, Inc.
†2 東京大学
Tokyo University

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[10]。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。

位置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。Einstein et al.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[11]。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。Liangjie et al.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位置の違い（Tweet する位置の違い）を考慮した潜在トピックの推定モデルを提案している[12]。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文書生成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 et al.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[14]。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（ユーザも同行者もこのサービスを利用していることが前提）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 et al.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[16]。しかしながら、これらは全ユーザの位置情報の収集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[15]。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（この論文ではレビュー）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。既存のクラスタリングを適用、各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。Yize らの手法は、ユーザが書いた文章のみから同行者を推定するため、コストが安く現実的である。一方、クラスタリングをする際に同行者のみを考慮し、ユーザの興味・嗜好などを考慮していないことからノイズが載る可能性が高い。次節にて詳述する。

2.3 LDA の適用

Yize らの手法において既存のクラスタリングを LDA に適用した場合について説明する。同一同行者のもとで書かれた文書集合を一つの文書として扱うことで LDA を適用可能である。LDA を適用した場合、同行者を考慮した文書生成モデルは以下のようになる。

1. Draw C (number of companions) multinomials of topic classes θ_c from Dirichlet prior α, one for each companion c;
2. Draw Z (number of topics) multinomials φ_z from Dirichlet prior β, one for each topic z;
3. For each token i in companion c:
   a) Draw topic z_{ci} from multinominal θ_c
   b) Draw word w_{ci} from multinominal $\varphi_{z_{ci}}$

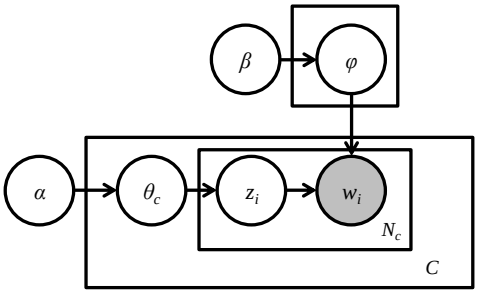

Fig. 1: LDA model tuned for companion dependent topic modeling

このモデルにより、類似の同行者を集めたトピッククラスタおよび各トピッククラスタごとに特徴となる単語集合を求めることが出来る。しかしながら、以下の問題点が発生する。

1） 上記のモデルでは同行者のみが文書のトピックを生成することを仮定している。しかしながら、文書のトピックは同行者によってのみ決まるわけではなく、ユーザの興味嗜好も大きく影響する。

2） 上記のモデルでは、文書中の単語は同行者クラスタのみによってのみ生成されると仮定している。文書中で使われる単語は、同行者によって決まる単語／ユーザの興味（トピック）によって決まる単語（(同行者に無関係)）／ユーザの興味や同行者に無関係に決まる単語（a,the など）があるが、考慮されていない。

このように、LDA を単に適用しただけでは、上記の問題が発生する。そこで、本稿では、1，2 の要素を考慮した同行者依存のトピック発見モデルを提案する。

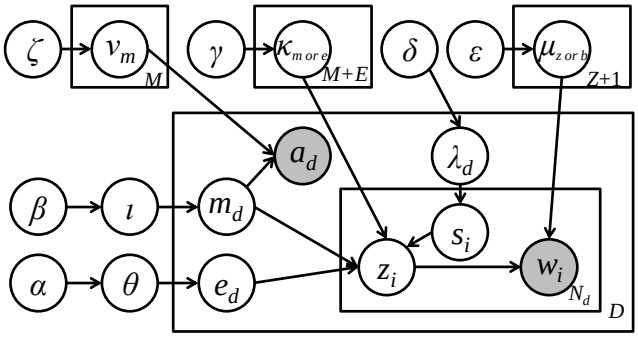

Fig. 2: Proposed graphical model

3. 提案モデル

3.1 同行者のモデリング

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、同行者およびユーザの興味嗜好の両方から決定されるとする。そのためここでは、各文書ごとに潜在同行者クラス（latent companion class）および、潜在嗜好クラス（latent preferecne class）を定義し文書生成モデルに組み込む。また、2)の問題を解決するため、各単語ごとに、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=2）
(イ) ユーザの興味によって決まる単語（s=1）
(ウ) ユーザの興味や同行者に無関係に決まる単語（s=0）

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、学習によって自動的に学習する。具体的には、その単語の属する文書からのトップダウンと、単語からのボトムアップの両方の学習を同時に行う。トップダウンとは、文書ごとにスイッチ変数の潜在クラスを定義し、文書の傾向によってその文書に含まれる各単語のスイッチ変数を学習する（文書の属する潜在スイッチクラスがアの傾向が 1、イの傾向が 2、ウの傾向が 4 の場合ウが 1/2 の確率で選ばれる）。ボトムアップとは、同じ単語が別の文書で利用されている場合は別の文章で学習されたスイッチ変数の値の分布に基づき学習を進する（単語 A が別の文書でアが 3 回、イが 2 回、ウが 1 回の場合アが 1/2 の確率で選ばれる）。Fig. 2 に提案するグラフィカルモデルを示す。パラメータの定義を Table 1 に示す。

Table 1: Definition of variables in the model

Variable	Meaning
E	number of latent preference class
M	number of latent companion class
Z	number of topics
D	number of documents
N_d	number of words of each document d
e_d	the preference class associated with document d
m_d	the latent companion class associated with document d
z_i	topic associated with ith token
a_d	the companion associated with document d
s_i	the switch associated with the ith token
w_i	the ith token
θ	the multinomial distribution of preference classes
ι	the multinomial distribution of latent companion classes specific to companion a_d ($\iota \mid \beta \sim$ Dirichlet(β))
υ_m	the multinomial distribution of companion specic to latent companion class m ($\upsilon_m \mid \zeta \sim$ Dirichlet(ζ))
$\kappa_{m \text{ or } e}$	the multinomial distribution of topics specic to latent companion class m or latent preference class e ($\kappa_{m \text{ or } e} \mid \gamma \sim$ Dirichlet(γ))
$\mu_{z \text{ or } b}$	the multinomial distribution of words specic to topic z or background topic b ($\mu_{z \text{ or } b} \mid \varepsilon \sim$ Dirichlet(ε))
λ_d	the multinomial distribution of switch variable specific to document d ($\lambda_d \mid \delta \sim$ Dirichlet(δ))
α	the fixed parameters of symmetric Dirichlet priors on the distributions of θ
β	the fixed parameters of symmetric Dirichlet priors on the distributions of ι
ζ	the fixed parameters of symmetric Dirichlet priors on the distributions of υ_m
γ	the fixed parameters of symmetric Dirichlet priors on the distributions of $\kappa_{m \text{ or } e}$
δ	the fixed parameters of symmetric Dirichlet priors on the distributions of λ_d
ε	the fixed parameters of symmetric Dirichlet priors on the distributions of $\mu_{z \text{ or } b}$

3.2 提案モデルの推論

提案モデルは、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[17]を利用することが可能である。まずは、提案モデルの文書生成プロセス（Generative Process）を述べる。

1. Draw multinomial θ from Dirichlet prior α;
2. Draw multinomial ι from Dirichlet prior β;
3. Draw M multinomials υ from Dirichlet prior ζ, one for each document d;
4. Draw $M+E$ multinomials $\kappa_{m \text{ or } e}$ from Dirichlet prior γ, one for each latent companion class m or preference class e;
5. Draw D multinomials κ from Dirichlet prior γ, one for each document d;
6. Draw $Z+1$ multinomials $\mu_{z \text{ or } b}$ from Dirichlet prior ε, one for each topic z or background topic b;
7. For each document d:
   a) Draw preference class e_d from multinominal θ
   b) Draw latent companion class m_d from multinominal ι
   c) Draw companion a_d from multinominal υ_{m_d}
   d) For each token i in document d:
      i) Draw switch variable r_{di} from multinominal λ_d;
      if r_{di} =0
         a) Draw word w_{di} from multinominal μ_b
      if r_{di} =1
         a) Draw topic z_{di} from multinominal κ_{e_d}
         b) Draw word w_{di} from multinominal $\mu_{z_{di}}$
      if r_{di} =2
         a) Draw topic z_{di} from multinominal κ_{m_d}
         b) Draw word w_{di} from multinominal $\mu_{z_{di}}$

提案モデルは、ベイジアンの階層プロセス（Baysian Hierarchical Process）とみなすことができる。推論を行うためには、各クラスの条件付確率を求める必要がある。まず、全文書の結合分布は以下のような混合分布となる。

$$p(\mathbf{e},\mathbf{m},\mathbf{s},\mathbf{z},\mathbf{a},\mathbf{w},\theta,\iota,\kappa,\mu,\lambda,\upsilon,;\alpha,\beta,\gamma,\varepsilon,\delta,\varsigma)$$

$$= p(\theta,\mid\alpha)\times\prod_d^D p(e_d\mid\theta)\times p(\iota\mid\beta)\times\prod_d^D p(m_d\mid\iota)$$

$$\times\prod_d^D p(\lambda_d\mid\delta)\times\prod_d^D\prod_i^{N_d} p(s_{di}\mid\lambda_d)\times\prod_j^{E+M} p(\kappa_j\mid\gamma)\times\prod_d^D\prod_i^{N_d} p(z_{di}\mid s_{di},\kappa_{e_d},\kappa_{m_d})$$

$$\times\prod_l^{Z+1} p(\mu_l\mid\varepsilon)\times\prod_d^D\prod_i^{N_d} p(w_{di}\mid s_{di},\mu_{z_{di}})\times\prod_m^M p(\upsilon_m\mid\varsigma)\times\prod_d^D p(a_d\mid\upsilon_{m_d})$$

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 θ、κ、υ、μ、λ、ι を積分消去する。上述の式を以下のように積分の形に変形する。

$$= \int_\theta p(\theta,\mid\alpha)\prod_d^D p(e_d\mid\theta)d\theta$$

$$\times\int_\iota p(\iota\mid\beta)\times\prod_d^D p(m_d\mid\iota)d\iota$$

$$\times\int_\kappa\prod_j^{E+M} p(\kappa_j\mid\gamma)\times\prod_d^D\prod_i^{N_d} p(z_{di}\mid s_{di},\kappa_{e_d},\kappa_{m_d})d\kappa$$

$$\times\int_\lambda\prod_d^D p(\lambda_d\mid\delta)\times\prod_d^D\prod_i^{N_d} p(s_{di}\mid\lambda_d)d\lambda$$

$$\times\int_\mu\prod_l^{Z+1} p(\mu_l\mid\varepsilon)\times\prod_d^D\prod_i^{N_d} p(w_{di}\mid\mu_{z_{di}})d\mu$$

$$\times\int_\upsilon\prod_m^M p(\upsilon_m\mid\varsigma)\times\prod_d^D p(a_d\mid\upsilon_{m_d})d\upsilon$$

式変形を繰り返すことにより、θ、κ、υ、μ、λ、ι を消去した分布を以下の式に得られる。積分消去の式変形に関しては Appendix I に記載する。

$$p(\mathbf{e},\mathbf{m},\mathbf{s},\mathbf{z},\mathbf{a},\mathbf{w};\alpha,\beta,\gamma,\varepsilon,\delta,\varsigma)$$
$$=\frac{\Gamma(\sum_e^E\alpha_e)}{\prod_e^E\Gamma(\alpha_e)}\frac{\prod_e^E\Gamma(n_e+\alpha_e)}{\Gamma(\sum_e^E n_e+\alpha_e)}\times\frac{\Gamma(\sum_m^M\beta_m)}{\prod_m^M\Gamma(\beta_m)}\frac{\prod_m^M\Gamma(n_m+\beta_m)}{\Gamma(\sum_m^M n_m+\beta_m)}$$
$$\times\prod_j^{E+M}\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{j,z}+\gamma_z)}{\Gamma(\sum_z^Z n_{j,z}+\gamma_z)}\times\prod_d^D\frac{\Gamma(\sum_l^L\delta_l)}{\prod_l^L\Gamma(\delta_l)}\frac{\prod_l^L\Gamma(n_{d,l}+\delta_l)}{\Gamma(\sum_l^L n_{d,l}+\delta_l)}$$
$$\times\prod_l^{Z+1}\frac{\Gamma(\sum_i^V\varepsilon_i)}{\prod_i^V\Gamma(\varepsilon_i)}\frac{\prod_i^V\Gamma(n_{l,i}+\varepsilon_i)}{\Gamma(\sum_i^V n_{l,i}+\varepsilon_i)}\times\prod_m^M\frac{\Gamma(\sum_a^A\varsigma_a)}{\prod_a^A\Gamma(\varsigma_a)}\frac{\prod_a^A\Gamma(n_{m,a}+\varsigma_a)}{\Gamma(\sum_a^A n_{m,a}+\varsigma_a)}$$

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 潜在嗜好クラス

文書 d について潜在嗜好クラスが f となるときの条件付確率を以下のように求める。ここで、潜在嗜好クラスの影響を受ける潜在トピッククラスも潜在パラメータであることから文書 d の i 番目の単語の潜在嗜好クラスを g とする。

$$p(e_d=f,z_{d,i}=g\mid\mathbf{e}_{\setminus d},\mathbf{z}_{\setminus d};\alpha,\gamma)$$
$$=\frac{p(\mathbf{e},\mathbf{z};\alpha,\gamma)}{p(\mathbf{e}_{\setminus d},\mathbf{z}_{\setminus di};\alpha,\gamma)}$$
$$=\frac{\frac{\Gamma(\sum_e^E\alpha_e)}{\prod_e^E\Gamma(\alpha_e)}\frac{\prod_e^E\Gamma(n_e+\alpha_e)}{\Gamma(\sum_e^E n_e+\alpha_e)}}{\frac{\Gamma(\sum_e^E\alpha_e)}{\prod_e^E\Gamma(\alpha_e)}\frac{\prod_e^E\Gamma(n_{e\setminus d}+\alpha_e)}{\Gamma(\sum_e^E n_{e\setminus d}+\alpha_e)}}\times\frac{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{f,z}+\gamma_z)}{\Gamma(\sum_z^Z n_{f,z}+\gamma_z)}}{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{f,z\setminus d}+\gamma_z)}{\Gamma(\sum_z^Z n_{f,z\setminus d}+\gamma_z)}}$$
$$=\frac{n_{f/d}+\alpha_f}{\sum_e^E n_{e/d}+\alpha_e}\frac{n_{f,g\setminus d}+\gamma_g}{\sum_z^Z n_{f,z\setminus d}+\gamma_z}$$

ここで、$n_{e\setminus d}$ は潜在嗜好クラス f に割り当てられた文書の数（文書 d に割り当てられた潜在嗜好クラスは除く）を表している。また、$n_{f,g\setminus d}$ は潜在嗜好クラス f に割り当てられた文書の中で、潜在トピッククラスに g が割り当てられた単語の数を表す（ただし、文書 d に割り当てられた潜在嗜好クラスは除く）。

3.2.2 潜在同行者クラス

文書 d について潜在同行者クラスが h、同行者が y となるとき条件付確率を以下のように求める。詳細な導出過程は Appendix II に記載する。ここで、潜在嗜好クラスの影響を受ける潜在トピッククラスも潜在パラメータであることから文書 d の i 番目の単語の潜在嗜好クラスを g とする。

$$p(m_d=h,z_{d,i}=g,a_d=y\mid\mathbf{m}_{\setminus d},\mathbf{z}_{\setminus d},\mathbf{a};\beta,\gamma,\varsigma)$$
$$=\frac{p(\mathbf{m},\mathbf{z},\mathbf{a};\beta,\gamma,\varsigma)}{p(\mathbf{m}_{\setminus d},\mathbf{z}_{\setminus d},\mathbf{a}_d;\beta,\gamma,\varsigma)}$$
$$=\frac{\frac{\Gamma(\sum_m^M\beta_m)}{\prod_m^M\Gamma(\beta_m)}\frac{\prod_m^M\Gamma(n_m+\beta_m)}{\Gamma(\sum_m^M n_m+\beta_m)}}{\frac{\Gamma(\sum_m^M\beta_m)}{\prod_m^M\Gamma(\beta_m)}\frac{\prod_m^M\Gamma(n_{m\setminus d}+\beta_m)}{\Gamma(\sum_m^M n_{m\setminus d}+\beta_m)}}\times\frac{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{h,z}+\gamma_z)}{\Gamma(\sum_z^Z n_{h,z}+\gamma_z)}}{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{h,z\setminus d}+\gamma_z)}{\Gamma(\sum_z^Z n_{h,z\setminus d}+\gamma_z)}}$$
$$\times\frac{\frac{\Gamma(\sum_a^A\varsigma_a)}{\prod_a^A\Gamma(\varsigma_a)}\frac{\prod_a^A\Gamma(n_{m,a}+\varsigma_a)}{\Gamma(\sum_a^A n_{m,a}+\varsigma_a)}}{\frac{\Gamma(\sum_a^A\varsigma_a)}{\prod_a^A\Gamma(\varsigma_a)}\frac{\prod_a^A\Gamma(n_{m,a\setminus d}+\varsigma_d)}{\Gamma(\sum_a^A n_{m,a\setminus d}+\varsigma_a)}}$$
$$=\frac{n_{h\setminus d}+\beta_h}{\sum_m^M n_{m\setminus d}+\beta_m}\frac{n_{h,g\setminus d}+\gamma_g}{\sum_z^Z n_{h,z\setminus d}+\gamma_z}\frac{n_{h,y\setminus d}+\varsigma_y}{\sum_a^A n_{h,a\setminus d}+\varsigma_a}$$

ここで、$n_{h\setminus d}$ は潜在同行者クラス h に割り当てられた文書の数（文書 d に割り当てられた潜在嗜好クラスは除く）を表している。また、$n_{h,g\setminus d}$ は潜在同行者クラス h に割り当てらた文書の中で、潜在トピッククラスに g が割り当てられた単語の数を表す（ただし、文書 d に割り当てられた潜在同行者クラスは除く）。また、$n_{h,y\setminus d}$ は潜在同行者クラス 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(文書 d に割り当てられた潜在同行者クラスは除く）を表している。

3.2.3 スイッチ変数と潜在トピッククラス

文書 d の i 番目の単語 v について潜在トピッククラスが b（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。

$$p(s_{di}=0,z_{di}=b,w_{di}=v\mid\mathbf{s}_{\setminus di},\mathbf{z}_{\setminus di},\mathbf{w};\delta,\varepsilon)$$
$$=\frac{p(\mathbf{s},\mathbf{z},\mathbf{w};\delta,\varepsilon)}{p(\mathbf{s}_{\setminus di},\mathbf{z}_{\setminus di},\mathbf{w}_{\setminus di};\delta,\varepsilon)}$$
$$=\frac{\frac{\Gamma(\sum_l^L\delta_l)}{\prod_l^L\Gamma(\delta_l)}\frac{\prod_l^L\Gamma(n_{d,l}+\delta_l)}{\Gamma(\sum_l^L n_{d,l}+\delta_l)}}{\frac{\Gamma(\sum_l^L\delta_l)}{\prod_l^L\Gamma(\delta_l)}\frac{\prod_l^L\Gamma(n_{d,l\setminus di}+\delta_l)}{\Gamma(\sum_l^L n_{d,l\setminus di}+\delta_l)}}\times\frac{\frac{\Gamma(\sum_i^V\varepsilon_i)}{\prod_i^V\Gamma(\varepsilon_i)}\frac{\prod_i^V\Gamma(n_{b,i}+\varepsilon_i)}{\Gamma(\sum_i^V n_{b,i}+\varepsilon_i)}}{\frac{\Gamma(\sum_i^V\varepsilon_i)}{\prod_i^V\Gamma(\varepsilon_i)}\frac{\prod_i^V\Gamma(n_{b,i\setminus di}+\varepsilon_i)}{\Gamma(\sum_i^V n_{b,i\setminus di}+\varepsilon_i)}}$$
$$=\frac{n_{d,0\setminus di}+\delta_0}{\sum_l^L n_{d,l\setminus di}+\delta_i}\times\frac{n_{b,v\setminus di}+\varepsilon_v}{\sum_i^V n_{b,i\setminus di}+\varepsilon_i}$$

ここで、$n_{d,0\setminus di}$ 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、$n_{b,v\setminus di}$ は潜在トピッククラスに b が割り当てられた単語 v の数を表す（ただし、文書 d の i 番目の単語 v について割り当てられた潜在トピッククラスは除く）。

文書 d の i 番目の単語 v について潜在トピッククラスが z（文書 d の i 番目の単語のスイッチ変数は 1）となるとき条件付確率を以下のように求める。

$$p(s_{di}=1,z_{di}=k,w_{di}=v\mid f,\mathbf{s}_{\setminus di},\mathbf{z}_{\setminus di},\mathbf{w};\delta,\varepsilon,\gamma)$$
$$=\frac{p(f,\mathbf{s},\mathbf{z},\mathbf{w};\delta,\varepsilon)}{p(f,\mathbf{s}_{\setminus di},\mathbf{z}_{\setminus di},\mathbf{w}_{\setminus di};\delta,\varepsilon)}$$
$$=\frac{\frac{\Gamma(\sum_l^L\delta_l)}{\prod_l^L\Gamma(\delta_l)}\frac{\prod_l^L\Gamma(n_{d,l}+\delta_l)}{\Gamma(\sum_l^L n_{d,l}+\delta_l)}}{\frac{\Gamma(\sum_l^L\delta_l)}{\prod_l^L\Gamma(\delta_l)}\frac{\prod_l^L\Gamma(n_{d,l\setminus di}+\delta_l)}{\Gamma(\sum_l^L n_{d,l\setminus di}+\delta_l)}}\times\frac{\frac{\Gamma(\sum_i^V\varepsilon_i)}{\prod_i^V\Gamma(\varepsilon_i)}\frac{\prod_z^Z\Gamma(n_{f,z}+\gamma_z)}{\Gamma(\sum_z^Z n_{f,z}+\gamma_z)}}{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{f,z\setminus di}+\gamma_z)}{\Gamma(\sum_z^Z n_{f,z\setminus di}+\gamma_z)}}$$
$$\times\frac{\frac{\Gamma(\sum_i^V\varepsilon_i)}{\prod_i^V\Gamma(\varepsilon_i)}\frac{\prod_i^V\Gamma(n_{k,i}+\varepsilon_i)}{\Gamma(\sum_i^V n_{k,i}+\varepsilon_i)}}{\frac{\Gamma(\sum_i^V\varepsilon_i)}{\prod_i^V\Gamma(\varepsilon_i)}\frac{\prod_i^V\Gamma(n_{k,i\setminus di}+\varepsilon_i)}{\Gamma(\sum_i^V n_{k,i\setminus di}+\varepsilon_i)}}$$
$$=\frac{n_{d,1\setminus di}+\delta_1}{\sum_l^L n_{d,l\setminus di}+\delta_l}\times\frac{n_{f,k\setminus di}+\gamma_k}{\sum_z^Z n_{f,z\setminus di}+\gamma_z}\times\frac{n_{k,v\setminus di}+\varepsilon_v}{\sum_i^V n_{k,i\setminus di}+\varepsilon_i}$$

ここで、$n_{d,1\setminus di}$ 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、$n_{f,k,\setminus di}$ は潜在嗜好クラスに f が割り当てられた文書の中で潜在トピック k を持つ単語の数を表す（ただし、文書 d の i 番目の単語 v について割り当てられた潜在トピッククラスは除く）。$n_{k,v\setminus di}$ は潜在トピッククラスに z が割り当てられた単語 v の

数を表す（ただし、文書 d の i 番目の単語 v について割り当てられた潜在トピッククラスは除く）。

文書 d の i 番目の単語 v について潜在トピッククラスが z（文書 d の i 番目の単語のスイッチ変数は 2 ）となるとき条件付確率を以下のように求める。

$$p(s_{di}=2, z_{di}=k, w_{di}=v \mid h, \mathbf{s}_{\setminus d}, \mathbf{z}_{\setminus di}, \mathbf{w}; \delta, \varepsilon, \gamma)$$

$$= \frac{p(h, \mathbf{s}, \mathbf{z}; \delta, \varepsilon)}{p(h, \mathbf{s}_{\setminus di}, \mathbf{z}_{\setminus di}; \delta, \varepsilon)}$$

$$= \frac{\frac{\Gamma(\sum_l^L \delta_l)}{\prod_l^L \Gamma(\delta_l)} \frac{\prod_l^L \Gamma(n_{d,l} + \delta_l)}{\Gamma(\sum_l^L n_{d,l} + \delta_l)}}{\frac{\Gamma(\sum_l^L \delta_l)}{\prod_l^L \Gamma(\delta_l)} \frac{\prod_l^L \Gamma(n_{d,l\setminus di} + \delta_l)}{\Gamma(\sum_l^L n_{d,l\setminus di} + \delta_l)}} \times \frac{\frac{\Gamma(\sum_i^V \varepsilon_i)}{\prod_i^V \Gamma(\varepsilon_i)} \frac{\prod_z^Z \Gamma(n_{h,z} + \gamma_z)}{\Gamma(\sum_z^Z n_{h,z} + \gamma_z)}}{\frac{\Gamma(\sum_z^Z \gamma_z)}{\prod_z^Z \Gamma(\gamma_z)} \frac{\prod_z^Z \Gamma(n_{h,z\setminus di} + \gamma_z)}{\Gamma(\sum_z^Z n_{h,z\setminus di} + \gamma_z)}}$$

$$\times \frac{\frac{\Gamma(\sum_i^V \varepsilon_i)}{\prod_i^V \Gamma(\varepsilon_i)} \frac{\prod_i^V \Gamma(n_{k,i} + \varepsilon_i)}{\Gamma(\sum_i^V n_{k,i} + \varepsilon_i)}}{\frac{\Gamma(\sum_i^V \varepsilon_i)}{\prod_i^V \Gamma(\varepsilon_i)} \frac{\prod_i^V \Gamma(n_{k,i\setminus di} + \varepsilon_i)}{\Gamma(\sum_i^V n_{k,i\setminus di} + \varepsilon_i)}}$$

$$= \frac{n_{d,2\setminus di} + \delta_2}{\sum_l^L n_{d,l\setminus di} + \delta_i} \frac{n_{h,k\setminus di} + \gamma_k}{\sum_z^Z n_{h,z\setminus di} + \gamma_z} \frac{n_{k,v\setminus di} + \varepsilon_i}{\sum_i^V n_{k,i\setminus di} + \varepsilon_i}$$

ここで、$n_{d,2\setminus di}$ は文書 d 内でスイッチ変数 2 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、$n_{h,k,\setminus di}$ は潜在嗜好クラスに h が割り当てられた文書の中で潜在トピック k を持つ単語の数を表す（ただし、文書 d の i 番目の単語 v について割り当てられた潜在トピッククラスは除く）。$n_{k,v\setminus di}$ は潜在トピッククラスに z が割り当てられた単語 v の数を表す（ただし、文書 d の i 番目の単語 v について割り当てられた潜在トピッククラスは除く）。

4. データセットの構築

データは同行者を含む Twitter の投稿文を対象とする。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。データ抽出には検索エンジン Bing を用いた。なお対象となる同行者は英語の学習サイト「 http://usefulenglish.ru/vocabulary/jobs-professions-occupations」から抽出した。同行者のデータセットは以下の 2 つを用意した。

・データセット１：ビジネス・学校における同行者を対象とし、「with 同行者」の形式を投稿文中に含む Twitter の投稿を 5 件づつ抽出した。対象となる同行者の一部を Table 2 に示す。同行者の総数は計 138 個である。総単語数は計 629 個である。

・データセット２：プライベートにおける同行者を対象とし、「with 同行者」の形式を投稿文中に含む Twitter の投稿を 10 件づつ抽出した。対象となる同行者の一部を Table 3 に示す。同行者の総数は計 79 個である。総単語数は計 617 個である。

Table 2: Companions by professions and school used to create dataset 1

Category	**Proffesion**
Management	president, vice-president, executive officer
Office	office clerk, receptionist, secretary, typist, stenographer;
Banks	banker, accountant, bookkeeper, economist, teller, cashier, auditor;
Medicine	doctor, physician, family doctor, general practitioner;
Restaurants	chef, head cook, cook
Sales and stores	salesperson, salesman, saleswoman, salesgirl, salesclerk, cashier;
Art and creative work	musician, composer, singer, dancer,artist, painter, film director, producer, actor, actress, cameraman;
School and college	principal, dean, professor, teacher, student, pupil;
Construction	engineer, technician, mechanic;
Science	scientist, scholar, researcher, explorer;
Law and order	judge, lawyer, attorney, legal adviser;
Other	expert, specialist, consultant, adviser;

Table 3: Companions by professions and school used to create dataset 2

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

Perplexity とはモデルのにおける全単語の対数尤度を反映したものであり、モデル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。提案モデルの Perplexity は次式で表される。

$$Perplexity = \exp\left(-\frac{1}{\sum_{d=1}^{M} N_d} \sum_{d \in D} \sum_{i \in d} \log\left(\lambda_0 \mu_{bi} + \sum_z^Z \lambda_1 \kappa_{ez} \mu_{zi} + \sum_z^Z \lambda_2 \kappa_{mz} \mu_{zi} \right) \right)$$

提案モデルのパラメータ値を Table 4 に示す。ここでは、潜在同行者クラスの数 M の最適値を探す。データセット 1 に対し(M, Z) =(20,40), (30,50), (40,60)、データセット 2 に対し(M, Z) =(15,30), (25,40), (35,50)のパラメータで比較した。提案モデル、LDA ともに Gibbs Sampling の Iteration 回数=20 で実行した。Fig. 3、Fig. 4 にそれぞれデータセット１、データセット 2 における Perplexity を示す。図に示すとおり、データセット 1 に対し(M, Z) = (40,60)、データセット 2 に対し(M, Z) = (35,50)が最も精度が高く、次節以降の検証ではこのパラメータを利用する。LDA についても同様にチューニングを行った。

Table 4: Parameters sets

Variable	Value
E	Dataset1:20 Dataset2:15
M	Dataset1:20,30,40 Dataset2:15,25,35
Z	Dataset1:40,50,60 Dataset2:30,40,50

α	1/E(Proposed model)
β	1/M(Proposed model)
ζ	1/ Num of Companion
γ	1/Z
δ	1/3
ε	1/Num Of Unique Words

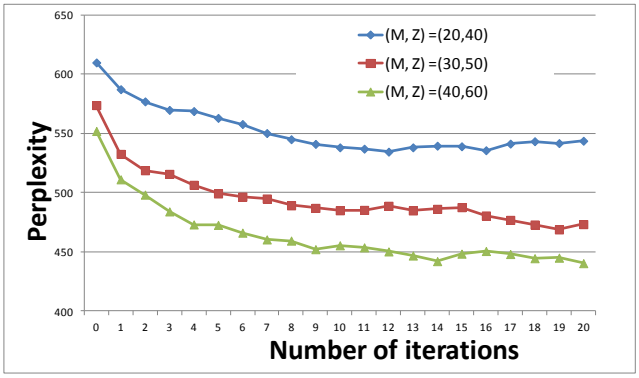

Fig. 3:Perplexity of data set of companions by professions

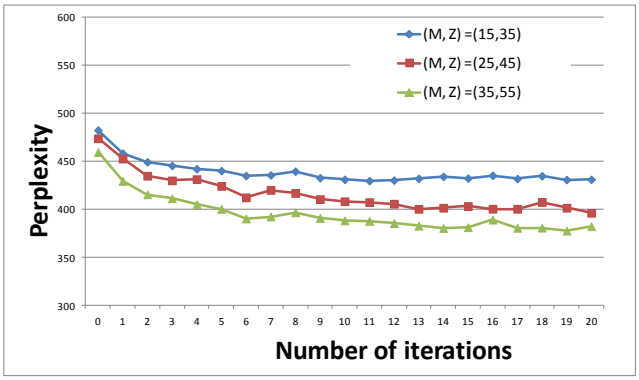

Fig. 4: Perplexity of data set of companions by family

5.2 予測精度による評価

本節では、提案モデルと LDA の同行者予測精度を比較する。具体的には、Twitter の投稿内容から提案モデルを用いて同行者を予測し、正解データと比較することで同行者の推定精度を評価する。まず、学習データとは別にテストデータを用意する。テストデータはデータセット１、データセット２で抽出したデータと重複しないよう、「with 同行者」の形式を投稿文中に含む Twitter の投稿を 100 件づつ抽出した。ここではテストデータからは同行者を表す単語を削除している。提案モデルの予測精度は以下の手順により計算を実施した。各文書 d に対し以下の計算を行う。

１） 文書 d 内で sw_{di}=2 となっている単語 i のみを抽出する。LDA は本ステップは不要。

２） １で抽出された単語に対し学習データから学習した μ_z（LDA の場合は φ）を用いてその単語の潜在トピッククラスの分布を計算する。

３） 各文書の潜在トピッククラスの分布を単語の潜在トピッククラスの分布の総和により求める。

４） 各文書の潜在同行者クラスの分布を、学習データから学習した κ_m（LDA の場合は θ）を用いて求め、最大となる潜在同行者クラスを予測（C_p）とする。

５） 各同行者が属する潜在同行者クラスを、学習データから学習した ν_m（LDA の場合は θ）を用いて求め、最大となる潜在同行者クラスを予測（C_t）とする。

６） C_p =C_t となる場合、正解個数 T に 1 を追加する。

すべての文書について上記を実施し、T/全文書数を計算する。結果を Table 5 に示す。

Table 5: Precision of prediciton of companion

	Dataset1	Dataset2
LDA	14.2%	13.5%
Proposed model	18.0%	18.2%

表に示すとおり、LDA に比べ提案手法の同行者クラスの推定精度が高いことが分かる。これは提案モデル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点、ユーザの同行者に伴って発生する嗜好と同行者に非依存の嗜好を分離できていることの効果が表れたといえる。

5.3 質的評価

データセット１およびデータセット２の学習結果をそれぞれ Table 6、Table 7 に示す。表では、各潜在同行者クラスと、それに対応する潜在トピッククラスの対応関係を示している。各潜在同行者クラスにはそのクラスに属する同行者の集合、および各潜在トピッククラスにはそのクラスに属する単語が記載されている。潜在同行者クラスの同行者（例：bride, fiance）と対応する潜在トピッククラスの単語（例：engaged, groom）関には密接な関係があり、提案モデルによって妥当な分類結果が得られていることが分かる。

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更なる向上を目指す。また、その他のコンテキスト（時間や位置）も考慮した文書生成モデルのモデル化を目指す。

参考文献

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

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

[3] L. AlSumait, D. Barbara, and C. Domeniconi, “On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking,” *Proc. of ICDM*, pp. 3-12, 2008.

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

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

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

[7] Y. Chen, D. Pavlov, and J. F. Canny, “Large-scale behavioral targeting,” *Proc. of KDD*, pp 209-218, 2009.

[8] X. Wang and A. McCallum, “Topics over time: a non-markov continuous-time model of topical trends,” *Proc. of KDD*, pages 424-433, 2006.

[9] D. Blei and J. Lafferty, “Dynamic topic models,” 23:113-120, 2006.

[10] N. Kawamae, “Trend analysis model: trend consists of temporal words, topics, and timestamps,” *Proc. of WSDM*, 317-326, 2011.

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

[12] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," *Proc. of WWW*, 769-778, 2012.

[13] G. Adomavicius and A. Tuzhilin, "Context-Aware Recommender Systems," *Recommender Systems Handbook*, pp.217-253, 2011.

[14] S. Böhm, J. Koolwaaij, M. Luther, B. Souville, M. Wagner and M. Wibbels, "Introducing IYOUIT," *Proc. of International Semantic Web Conference*, pp.804-817, 2008.

[15] Y. Li, J. Nie, Y. Zhang, B. Wang, B. Yan and F. Weng, "Contextual Recommendation based on Text Mining," *Proc. of COLING*, pp.692-700, 2010.

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

[17] T. L. Griffiths and M. Steyvers , "Finding scientific topics," *Proc. of the National Academy of Sciences of the United States of America*, 2004.

Table 6: Data set for companions by professions

Latent companion class	co1	co2	co3	co4	co8
associated companion	girlfriend mistress	bride fiance	child daughter	husband my family	parents grandmother
The highest topic	class19	class21	class12	class16	class24
associated words	night divorce tips club feelings lovely food sxsw star lives	favorite engaged cool groom cooking set meet peter drive bonds	support separated photos really rtfqp grandchild read funny proud mess	single god hate business john school guy wine collapse singing	home visit following date learn long waiting office marriage boss

Table 7: Data set for companions by professions

Latent companion class	co22	co26	co28	co30	co0
associated companion	lawyer journalist	film director singer	economist buyer	politician president	teacher adviser
The highest topic	class27	class4	class42	class9	class10
associated words	partners reporter recording line blogspot rific blogtalkradio service obama young	studio website principal future dexter org george song lil nurul	buyer inspector gold investment married google victoria week office review	right tonight president radio early couple website ebay nutkinnb conversation	best lock told obsessed training head course abc latest bryan

Appendix I

v の条件付き確率の式から v を積分消去する式変形について記載する。

$$\int_{\upsilon}\prod_{m}^{M}p(\upsilon_m\mid\varsigma)\times\prod_{d}^{D}p(a_d\mid\upsilon_{m_d})d\upsilon$$
$$=\prod_{m}^{M}\int_{\upsilon}p(\upsilon_m\mid\varsigma)\times\prod_{d}^{D}p(a_d\mid\upsilon_{m_d})d\upsilon$$

$p(v_m|\varsigma)$をディリクレ分布に置換することにより、以下の式を得る。

$$=\prod_{m}^{M}\int_{\upsilon}\frac{\Gamma(\sum_{d}^{D}\varsigma_d)}{\prod_{d}^{D}\Gamma(\varsigma_d)}\prod_{d}^{D}\upsilon_{m_d}^{\ \varsigma_d-1}\prod_{d}^{D}p(a_d\mid\upsilon_{m_d})d\upsilon$$

$p(a_d|v_{md})$を多項分布に置換することにより以下の式を得る。

$$=\prod_{m}^{M}\int_{\upsilon}\frac{\Gamma(\sum_{d}^{D}\varsigma_d)}{\prod_{d}^{D}\Gamma(\varsigma_d)}\prod_{d}^{D}\upsilon_{m_d}^{\ \varsigma_d-1}\prod_{d}^{D}\upsilon_{m_d}^{\ n_{d,m}}d\upsilon$$
$$=\prod_{m}^{M}\int_{\upsilon}\frac{\Gamma(\sum_{d}^{D}\varsigma_d)}{\prod_{d}^{D}\Gamma(\varsigma_d)}\prod_{d}^{D}\upsilon_{m_d}^{\ n_{d,m}+\varsigma_d-1}d\upsilon$$

上記の式はディリクレ分布×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だす。

$$=\prod_{m}^{M}\frac{\Gamma(\sum_{d}^{D}\varsigma_d)}{\prod_{d}^{D}\Gamma(\varsigma_d)}\frac{\prod_{d}^{D}\Gamma(n_{d,m}+\varsigma_d)}{\Gamma(\sum_{d}^{D}n_{d,m}+\varsigma_d)}\int_{\upsilon}\frac{\Gamma(\sum_{d}^{D}n_{d,m}+\varsigma_d)}{\prod_{d}^{D}\Gamma(n_{d,m}+\varsigma_d)}\prod_{d}^{D}\upsilon_{m_d}^{\ n_{d,m}+\varsigma_d-1}d\upsilon$$
$$=\prod_{m}^{M}\frac{\Gamma(\sum_{d}^{D}\varsigma_d)}{\prod_{d}^{D}\Gamma(\varsigma_d)}\frac{\prod_{d}^{D}\Gamma(n_{d,m}+\varsigma_d)}{\Gamma(\sum_{d}^{D}n_{d,m}+\varsigma_d)}$$

Appendix II

潜在同行者クラスに関する GibbsSampling の更新式を導出する。まず、各確率分布をディリクレ分布に置換する。

$$p(m_d=h,z_{d,i}=g,a_d=y\mid\mathbf{m}_{\setminus d},\mathbf{z}_{\setminus d},\mathbf{a};\beta,\gamma,\varsigma)$$
$$=\frac{p(\mathbf{m},\mathbf{z},\mathbf{a};\beta,\gamma,\varsigma)}{p(\mathbf{m}_{\setminus d},\mathbf{z}_{\setminus di},\mathbf{a}_d;\beta,\gamma,\varsigma)}$$
$$=\frac{\frac{\Gamma(\sum_m^M\beta_m)}{\prod_m^M\Gamma(\beta_m)}\frac{\prod_m^M\Gamma(n_m+\beta_m)}{\Gamma(\sum_m^M n_m+\beta_m)}}{\frac{\Gamma(\sum_m^M\beta_m)}{\prod_m^M\Gamma(\beta_m)}\frac{\prod_m^M\Gamma(n_{m\setminus d}+\beta_m)}{\Gamma(\sum_m^M n_{m\setminus d}+\beta_m)}}\times\frac{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{h,z}+\gamma_z)}{\Gamma(\sum_z^Z n_{h,z}+\gamma_z)}}{\frac{\Gamma(\sum_z^Z\gamma_z)}{\prod_z^Z\Gamma(\gamma_z)}\frac{\prod_z^Z\Gamma(n_{h,z\setminus d}+\gamma_z)}{\Gamma(\sum_z^Z n_{h,z\setminus d}+\gamma_z)}}$$
$$\times\frac{\frac{\Gamma(\sum_a^A\varsigma_a)}{\prod_a^A\Gamma(\varsigma_a)}\frac{\prod_a^A\Gamma(n_{m,a}+\varsigma_a)}{\Gamma(\sum_a^A n_{m,a}+\varsigma_a)}}{\frac{\Gamma(\sum_a^A\varsigma_a)}{\prod_a^A\Gamma(\varsigma_a)}\frac{\prod_a^A\Gamma(n_{m,a\setminus d}+\varsigma_d)}{\Gamma(\sum_a^A n_{m,a\setminus d}+\varsigma_a)}}$$
$$=\frac{\frac{\prod_m^M\Gamma(n_m+\beta_m)}{\Gamma(\sum_m^M n_m+\beta_m)}}{\frac{\prod_m^M\Gamma(n_{m\setminus d}+\beta_m)}{\Gamma(\sum_m^M n_{m\setminus d}+\beta_m)}}\times\frac{\frac{\prod_z^Z\Gamma(n_{h,z}+\gamma_z)}{\Gamma(\sum_z^Z n_{h,z}+\gamma_z)}}{\frac{\prod_z^Z\Gamma(n_{h,z\setminus d}+\gamma_z)}{\Gamma(\sum_z^Z n_{h,z\setminus d}+\gamma_z)}}\times\frac{\frac{\prod_a^A\Gamma(n_{m,a}+\varsigma_a)}{\Gamma(\sum_a^A n_{m,a}+\varsigma_a)}}{\frac{\prod_a^A\Gamma(n_{m,a\setminus d}+\varsigma_d)}{\Gamma(\sum_a^A n_{m,a\setminus d}+\varsigma_a)}}$$

各確率分布から決定済みのパラメータ部分のみを切り出す。

$$=\frac{\frac{\prod_{m\neq h}^M\Gamma(n_m+\beta_m)\Gamma(n_h+\beta_h)}{\Gamma(\sum_m^M n_m+\beta_m)}}{\frac{\prod_{m\neq h}^M\Gamma(n_{m\setminus d}+\beta_m)\Gamma(n_{h\setminus d}+\beta_h)}{\Gamma(\sum_m^M n_{m\setminus d}+\beta_m)}}\times\frac{\frac{\prod_{z\neq g}^Z\Gamma(n_{h,z}+\gamma_z)\Gamma(n_{g,z}+\gamma_g)}{\Gamma(\sum_z^Z n_{h,z}+\gamma_z)}}{\frac{\prod_{z\neq g}^Z\Gamma(n_{h,z\setminus d}+\gamma_z)\Gamma(n_{h,g\setminus d}+\gamma_g)}{\Gamma(\sum_z^Z n_{h,z\setminus d}+\gamma_z)}}$$
$$\times\frac{\frac{\prod_{a\neq y}^A\Gamma(n_{m,a}+\varsigma_a)\Gamma(n_{m,y}+\varsigma_y)}{\Gamma(\sum_a^A n_{m,a}+\varsigma_a)}}{\frac{\prod_{a\neq y}^A\Gamma(n_{m,a\setminus d}+\varsigma_y)\Gamma(n_{m,a\setminus d}+\varsigma_y)}{\Gamma(\sum_a^A n_{m,a\setminus d}+\varsigma_a)}}$$

更新対象の文書 d を切りだす。ここで、$n_{m\cap d}$は、文書 d に割り当てられている潜在同行者クラス m の数を表す。

$$=\frac{\frac{\prod_{m\neq h}^M\Gamma(n_{m\setminus d}+n_{m\cap d}+\beta_m)\Gamma(n_{h\setminus d}+n_{h\cap d}+\beta_h)}{\Gamma(\sum_m^M n_{m\setminus d}+n_{m\cap d}+\beta_m)}}{\frac{\prod_{m\neq h}^M\Gamma(n_{m\setminus d}+\beta_m)\Gamma(n_{h\setminus d}+\beta_h)}{\Gamma(\sum_m^M n_{m\setminus d}+\beta_m)}}$$
$$\times\frac{\frac{\prod_{z\neq g}^Z\Gamma(n_{h,z\setminus di}+n_{h,z\cap d}+\gamma_z)\Gamma(n_{g,z\setminus di}+n_{g,z\cap d}+\gamma_g)}{\Gamma(\sum_z^Z n_{h,z\setminus d}+n_{h,z\cap d}+\gamma_z)}}{\frac{\prod_{z\neq g}^Z\Gamma(n_{h,z\setminus d}+\gamma_z)\Gamma(n_{h,g\setminus d}+\gamma_g)}{\Gamma(\sum_z^Z n_{h,z\setminus d}+\gamma_z)}}$$
$$\times\frac{\frac{\prod_{a\neq y}^A\Gamma(n_{m,a\setminus d}+n_{m,a\cap d}+\varsigma_a)\Gamma(n_{m,y\setminus d}+n_{m,y\cap d}+\varsigma_y)}{\Gamma(\sum_a^A n_{m,a\setminus d}+n_{m,a\cap d}+\varsigma_a)}}{\frac{\prod_{a\neq y}^A\Gamma(n_{m,a\setminus d}+\varsigma_y)\Gamma(n_{m,a\setminus d}+\varsigma_y)}{\Gamma(\sum_a^A n_{m,a\setminus d}+\varsigma_a)}}$$

文書 d に関して数値化可能な部分を数値に変換し、更新式を得る。

$$= \frac{\frac{\prod_{m\neq h}^{M}\Gamma(n_{m\backslash d}+0+\beta_m)\Gamma(n_{h\backslash d}+1+\beta_h)}{\Gamma(\sum_{m}^{M}(n_{m\backslash d}+\beta_m)+1)}}{\frac{\prod_{m\neq h}^{M}\Gamma(n_{m\backslash d}+\beta_m)\Gamma(n_{h\backslash d}+\beta_h)}{\Gamma(\sum_{m}^{M}n_{m\backslash d}+\beta_m)}}$$

$$\times \frac{\frac{\prod_{z\neq g}^{Z}\Gamma(n_{h,z\backslash di}+0+\gamma_z)\Gamma(n_{g,z\backslash d}+1+\gamma_g)}{\Gamma(\sum_{z}^{Z}(n_{h,z\backslash d}+\gamma_z)+1)}}{\frac{\prod_{z\neq g}^{Z}\Gamma(n_{h,z\backslash d}+\gamma_z)\Gamma(n_{h,g\backslash d}+\gamma_g)}{\Gamma(\sum_{z}^{Z}n_{h,z\backslash d}+\gamma_z)}}$$

$$\times \frac{\frac{\prod_{a\neq y}^{A}\Gamma(n_{m,a\backslash d}+0+\varsigma_a)\Gamma(n_{m,y\backslash d}+1+\varsigma_y)}{\Gamma(\sum_{a}^{A}(n_{m,a\backslash d}+\varsigma_a)+1)}}{\frac{\prod_{a\neq y}^{A}\Gamma(n_{m,a\backslash d}+\varsigma_y)\Gamma(n_{m,a\backslash d}+\varsigma_y)}{\Gamma(\sum_{a}^{A}n_{m,a\backslash d}+\varsigma_a)}}$$

$$= \frac{n_{h\backslash d}+\beta_h}{\sum_{m}^{M}n_{m\backslash d}+\beta_m}\frac{n_{h,g\backslash d}+\gamma_g}{\sum_{z}^{Z}n_{h,z\backslash d}+\gamma_z}\frac{n_{h,y\backslash d}+\varsigma_y}{\sum_{a}^{A}n_{h,a\backslash d}+\varsigma_a}$$