2E3-4in

ディリクレ過程を用いたアカウントを共有するユーザの購買のモデリング

Modeling Multiple Users' Transactions over a Single Account with Dirichlet Process

甲谷 優[*1] 岩田 具治[*2] 内山 俊郎[*1] 藤村 考[*1]
Yutaka Kabutoya Tomoharu Iwata Toshio Uchiyama Ko Fujimura

[*1]日本電信電話株式会社 NTT サイバーソリューション研究所
NTT Cyber Solutions Laboratories, NTT Corporation

[*2]日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

We propose a Dirichlet process mixture model for enhancing recommender systems to handle multiple users that share a single account. In several web services, since multiple individuals may share one account (e.g. a family), user preferences cannot be estimated from a simple perusal of the transactions of the account, thus it is difficult to accurately recommend items to those who share an account. Furthermore, the numbers of users varies by account. We tackle this problem by assuming latent users sharing an account and establish a model to infer the number of latent users by combining Dirichlet process mixture model and Latent Dirichlet Allocation (LDA). Experiments on artificial datasets created from real log datasets from online movie services by combining the transactions of from one to five accounts demonstrate higher recommendation accuracy than conventional methods.

1. はじめに

近年，Amazon[*1] 等の多くのサービスで協調フィルタリングに基づく推薦システムが採用されている．そういった中，オンラインの DVD レンタル企業である Netflix は 100 万ドルの賞金をかけてレコメンデーション精度のコンテストを行った[*2]．その優勝チームの手法 [Koren 09, Tösher 09] はきわめて高いレコメンデーション精度を実現したが，アカウントが複数人に共有されている場合には精度が低下するという問題を残した．

アカウントの共有により協調フィルタリングアルゴリズムによるレコメンデーションが困難になるケースは多い．たとえばビデオオンデマンド (VOD) サービスで，視聴履歴のアカウントが世帯に紐付けられているような場合を考える．とある世帯が父親，母親，息子の 3 人から成り立っており，母親はよく昼にドラマの映像を，息子はよく夕方にアニメの映像を，父親はよく夜にスポーツの映像を見るとしよう．この場合，従来の協調フィルタリングアルゴリズムでは個々人ではなく世帯の嗜好を推定するため，ドラマ，アニメ，スポーツのいずれを推薦すべきかが判定できない．

この問題に対し我々 [甲谷 11] は，Probabilistic Latent Semantic Analysis (PLSA) モデル [Hofmann 99] を拡張し，1 つのアカウントを共有する複数人のユーザの独立した嗜好を分析し抽出するためのトピックモデルを提案した．この手法では潜在トピック（メモリベースの協調フィルタリング [Resnick 94] における嗜好) はアカウントからではなく**潜在ユーザ**から生成されるものと仮定しており，さらに購買時刻がその購買の潜在ユーザに依存するものと仮定している．

しかしながら [甲谷 11] では各アカウントの潜在ユーザ数は既知で，かつ全てのアカウントで同数としていたが，これは全ての世帯が同人数であると仮定していることになり，現実の問題にそぐわない．そこで本稿では，[甲谷 11] をディリクレ過程を用いてノンパラメトリックベイズモデルに拡張した，アカウント毎に潜在ユーザ数を自動的に決定可能なモデルを提案する．また，[甲谷 11] では EM アルゴリズムを用いた MAP 推定を行っていたが，より頑健な MCMC に基づくベイズ推定法を導出する．

実験では，複数人により 1 アカウントが共有されているデータ上での，提案法によるレコメンデーションの精度を検証した．実験データには，映像評点に関する実ログデータを，その中のいくつかの（ランダムに 1〜5）のアカウントの履歴データを組み合わせたものを用いた．ここで，実ログデータ中の元のアカウントは 1 人のユーザによってのみ利用されているものと仮定している．ベースラインとしては，[Resnick 94] をはじめとした従来の協調フィルタリングアルゴリズムと比較した．

連絡先: 甲谷優，日本電信電話株式会社 NTT サイバーソリューション研究所，神奈川県横須賀市光の丘 1-1，kabutoya.yutaka@lab.ntt.co.jp

∗1 http://www.amazon.com
∗2 http://www.netflixprize.com

2. 提案法

2.1 モデル

今，履歴データとして，U 個のアカウントと，各アカウントの各購買について，商品と，時刻のペアの集合 $(\boldsymbol{I},\boldsymbol{T})$ が与えられたとする．ここで，$\boldsymbol{I}=\{\{i_{um}\}_{m=1}^{M_u}\}_{u=1}^{U}$ は購買された商品の集合，$\boldsymbol{T}=\{\{t_{um}\}_{m=1}^{M_u}\}_{u=1}^{U}$ は購買時刻の集合を指す．このとき，本稿で用いる表記法を表 1 に示す．

提案法の説明に入る前に，そのベースとなっている LDA について簡単に説明する．LDA では，各アカウントがトピック比率 $\boldsymbol{\theta}_u$ を持っており，そのアカウントの嗜好を表している．アカウント u の購買の度に，トピック z_{um} がトピック比率に従い選択され，商品 i_{um} がトピック z_{um} の商品出現確率 $\boldsymbol{\phi}_{z_{um}}$ から生成される．

本モデルでは 1 つのアカウントが複数人で共有されることを仮定している．アカウント毎に複数の潜在ユーザが存在し，各潜在ユーザがトピック比率 $\boldsymbol{\theta}_{uv}$ を持つ．潜在ユーザは，アカウントの購買の度にディリクレ過程の構成法の 1 つである Chinese restaurant 過程 [Aldous 83] により選択される．購買商品の生成はトピック比率がえ与えられた LDA と同様で，

表 1: 表記法

Symbol	Description
U	number of accounts
N	number of unique items
Z	number of topics
V	number of users sharing an account
M_u	number of items purchased by the uth account
i_{um}	mth item purchased by the uth account, $i_{um} \in \{1, \ldots, N\}$
t_{um}	time of mth transaction by the uth account
z_{um}	topic of the mth transaction by the uth account, $z_{um} \in \{1, \ldots, Z\}$
v_{um}	user of the mth transaction by the uth account, $v_{um} \in \{1, \ldots, \infty\}$

トピック z_{um} がトピック比率 $\boldsymbol{\theta}_{uv_{um}}$ に従い選択され，そして商品 i_{um} が各トピックの商品出現確率 $\boldsymbol{\phi}_{z_{um}}$ から生成される．さらに提案モデルでは購買時刻についても，ユーザ依存の平均 $\tau_{uv_{um}}$，分散 $\sigma^2_{uv_{um}}$ の正規分布により生成されるものと仮定している．

すなわち，提案モデルでは以下の過程により購買商品集合 $\boldsymbol{I}$，購買時刻集合 $\boldsymbol{T}$ が生成されるものとする．

1. For each topic $z = 1, \ldots, Z$:
   (a) Draw item probability $\boldsymbol{\phi}_z \sim \text{Dirichlet}(\beta)$
2. For each account $u = 1, \ldots, U$:
   (a) For each user $v = 1, \ldots, \infty$:
      i. Draw time variance $\sigma^2_{uv} \sim \text{InverseGamma}(\eta, \rho)$
      ii. Draw time mean $\mu_{uv} \sim \text{Normal}(\nu, \xi^{-1}\sigma^2_{uv})$
      iii. Draw topic proportions $\boldsymbol{\theta}_{uv} \sim \text{Dirichlet}(\alpha)$
   (b) For each transaction $m = 1, \ldots, M_u$:
      i. Draw user $v_{um} \sim \text{CRP}(\gamma)$
      ii. Draw time $t_{um} \sim \text{Normal}(\mu_{uv_{um}}, \sigma^2_{uv_{um}})$
      iii. Draw topic $z_{um} \sim \text{Multinomial}(\boldsymbol{\theta}_{uv_{um}})$
      iv. Draw item $i_{um} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{um}})$

ここで $\boldsymbol{\phi}_z$ はトピック z の商品分布，μ_{uv}，σ^2_{uv} はそれぞれアカウント u の潜在ユーザ v の購買時刻の平均と分散，$\boldsymbol{\theta}_{uv}$ はアカウント u，潜在ユーザ v のトピック比率を表す．

なお，提案法ではトピック分布，購買商品分布の事前分布として多項分布の共役事前分布であるディリクレ分布を，購買時刻分布の事前分布として正規分布の共役事前分布である正規-逆ガンマ分布を用いている．図 1 に提案モデルのグラフィカルモデルを示す．ここで，塗潰し円は観測変数，中抜き円は潜在変数，矢印は依存関係，矩形は繰り返しを表す．

上記モデルにおける購買商品集合 $\boldsymbol{I}$ と購買時刻集合 $\boldsymbol{T}$，潜在ユーザ集合 $\boldsymbol{V} = \{\{v_{um}\}_{m=1}^{M_u}\}_{u=1}^{U}$，トピック集合 $\boldsymbol{Z} = \{\{z_{um}\}_{m=1}^{M_u}\}_{u=1}^{U}$ の完全尤度は下式で表される．

$$P(\boldsymbol{I}, \boldsymbol{T}, \boldsymbol{V}, \boldsymbol{Z} | \alpha, \beta, \gamma, \xi, \nu, \eta, \rho) = P(\boldsymbol{V}|\gamma) P(\boldsymbol{T}|\boldsymbol{V}, \xi, \nu, \eta, \rho) P(\boldsymbol{Z}|\boldsymbol{V}, \alpha) P(\boldsymbol{I}|\boldsymbol{Z}, \beta). \quad (1)$$

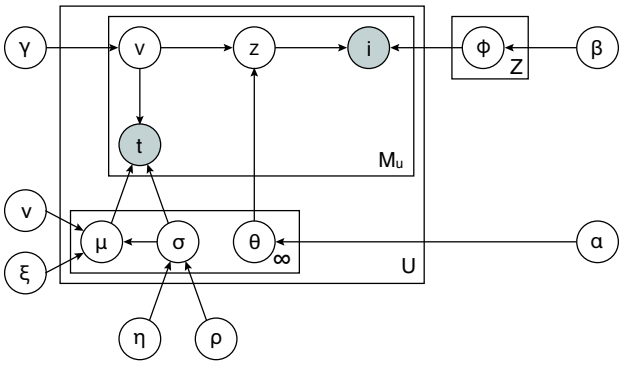


図 1: 提案モデルのグラフィカルモデル

2.2 学習

トピック集合 $\boldsymbol{Z}$ 及び潜在ユーザ集合 $\boldsymbol{V}$ は購買商品集合 $\boldsymbol{I}$ と購買時刻集合 $\boldsymbol{T}$ を入力として Collapsed ギブスサンプリングを用いて効率よく推定できる．アカウント u の m 番目の購買を生成するトピック z_j，$j = (u, m)$ のサンプリング確率は下式により計算できる．

$$P(z_j = k | \boldsymbol{I}, \boldsymbol{T}, \boldsymbol{Z}_{\backslash j}, \boldsymbol{V}) \propto (M_{uv_j k \backslash j} + \alpha) \cdot \frac{M_{ki_j \backslash j} + \beta}{M_{k \backslash j} + \beta N}, \quad (2)$$

ここで ここで M_{ki} はトピック k におけるアイテム i の出現回数，$M_k = \sum_{i=1}^{N} M_{ki}$ を表す．なお，$\backslash j$ はアカウント u の m 番目の購買を除いたときの回数もしくは変数である．また潜在ユーザ v_j のサンプリング確率は $w = \{1, \ldots, \hat{V}_{u\backslash j}\}$ において

$$\begin{aligned} &P(v_j = w | z_j = k, \boldsymbol{I}, \boldsymbol{T}, \boldsymbol{Z}_{\backslash j}, \boldsymbol{V}_{\backslash j}) \\ &\propto M_{uw\backslash j} \cdot \frac{M_{uwk\backslash j} + \alpha}{M_{uw\backslash j} + \alpha Z} \cdot \left(\frac{1 + \bar{\xi}_{uw\backslash j}}{\bar{\xi}_{uw\backslash j}} \right)^{-\frac{1}{2}} \cdot \frac{\Gamma(\frac{1+\bar{\eta}_{uw\backslash j}}{2})}{\Gamma(\frac{\bar{\eta}_{uw\backslash j}}{2})} \cdot \bar{\rho}_{uw\backslash j}^{\frac{\bar{\eta}_{uw\backslash j}}{2}} \\ &\cdot \left(\rho + \xi\nu^2 + C_{uw\backslash j} + t_j^2 - \frac{(T_{uw\backslash j} + t_j + \xi\nu)^2}{M_{uw\backslash j} + 1 + \xi} \right)^{-\frac{1+\bar{\eta}_{uw\backslash j}}{2}}, \end{aligned} \quad (3)$$

のように計算でき，また新規潜在ユーザのサンプリング確率は

$$\begin{aligned} &P(v_j = \hat{V}_{u\backslash j} + 1 | z_j = k, \boldsymbol{I}, \boldsymbol{T}, \boldsymbol{Z}_{\backslash j}, \boldsymbol{V}_{\backslash j}) \\ &\propto \gamma \cdot \frac{1}{Z} \cdot \left(\frac{1 + \xi}{\xi} \right)^{-\frac{1}{2}} \cdot \frac{\Gamma(\frac{1+\eta}{2})}{\Gamma(\frac{\eta}{2})} \cdot \rho^{\frac{\eta}{2}} \\ &\cdot \left(\rho + \xi\nu^2 + t_j^2 - \frac{(t_j + \xi\nu)^2}{1 + \xi} \right)^{-\frac{1+\eta}{2}}, \end{aligned} \quad (4)$$

のように計算できる．ここで $\hat{V}_u$ は現在のアカウント u の潜在ユーザ数を，M_{uvz} はアカウント u の購買のうち潜在ユーザ v，トピック z が割り当てられたものの数，$M_{uv} = \sum_{z=1}^{Z} M_{uvz}$，

$$T_{uv} = \sum_{m=1}^{M_u} I(v_{um} = v) t_{um}, \quad (5)$$

$$C_{uv} = \sum_{m=1}^{M_u} I(v_{um} = v) t_{um}^2, \quad (6)$$

表 2: データセット

	MovieLens	EachMovie
number of accounts	311	4,742
number of items	1,152	1,249
number of transactions	79,503	507,105

$$\bar{\xi}_{uv} = M_{uv} + \xi, \tag{7}$$
$$\bar{\eta}_{uv} = M_{uv} + \eta, \tag{8}$$
$$\bar{\nu}_{uv} = \frac{T_{uv} + \xi\nu}{M_{uv} + \xi}, \tag{9}$$
$$\bar{\rho}_{uv} = \rho + C_{uv} + \xi\nu^2 - \bar{\xi}_{uv}\bar{\nu}_{uv}^2, \tag{10}$$

をそれぞれ表す．なお，$\Gamma(\cdot)$ はガンマ関数，$I(\cdot)$ は指示関数である．

全購買について，(2)，(3)，(4) によるサンプリングを十分な回数繰り返すことでトピック集合 $\boldsymbol{Z}$ と潜在ユーザ集合 $\boldsymbol{V}$ を推定することができる．

2.3 推薦

トピック集合 $\boldsymbol{Z}$ 及び潜在ユーザ集合 $\boldsymbol{V}$ を推定した後，提案法によって時刻 t におけるアカウント u の各アイテムの購買確率を以下のように算出することにより，推薦が可能になる．

$$\begin{aligned}
&P(i,t|u,\boldsymbol{I},\boldsymbol{T},\boldsymbol{V},\boldsymbol{Z},\alpha,\beta,\gamma,\xi,\nu,\eta,\rho) \\
&= \sum_{v=1}^{V_u}\sum_{z=1}^{Z} P(i,t,v,z|u,\boldsymbol{I},\boldsymbol{T},\boldsymbol{V},\boldsymbol{Z},\alpha,\beta,\gamma,\xi,\nu,\eta,\rho) \\
&= \sum_{v=1}^{V_u}\sum_{z=1}^{Z} \frac{M_{uv}}{M_u+\gamma}\cdot\frac{M_{uvz}+\alpha}{M_{uv}+\alpha Z}\cdot\frac{M_{zi}+\beta}{M_z+\beta N} \\
&\quad\cdot\left(\frac{1+\bar{\xi}_{uv}}{\bar{\xi}_{uv}}\right)^{-\frac{1}{2}}\cdot\frac{\Gamma(\frac{1+\bar{\eta}_{uv}}{2})}{\Gamma(\frac{\bar{\eta}_{uv}}{2})}\cdot\bar{\rho}_{uv}^{\frac{\bar{\eta}_{uv}}{2}} \\
&\quad\cdot\left(\rho+\xi\nu^2+C_{uv}+t^2-\frac{(T_{uv}+t+\xi\nu)^2}{M_{uv}+1+\xi}\right)^{-\frac{1+\bar{\eta}_{uv}}{2}}.
\end{aligned} \tag{11}$$

3. 実験

3.1 データセット

MovieLens，EachMovie の 2 データを用いて提案法の評価を行った．EachMovie データは Compaq Systems Research Center により提供されていたものであり，MovieLens データは MovieLens Research Project により現在も提供されている[*3]．両データとも本来は映画評点データであるが，購買情報とみなして実験を行った．また，両データに含まれる評価時刻を購買時刻とみなした．両データから購買数が 10 未満の商品と，5 未満のユーザは省いた．

さらに，上記実データ中が本来は評点データであることから，その各アカウントは 1 ユーザにより利用されている可能性が高い．そこで，いくつかのアカウントの履歴データを組み合わせ仮想的に複数のユーザにより 1 アカウントが共有されている人工データを作成した．人工データ中の各アカウントは元のデータの最小で 1 つ，最大で 5 つのアカウントの履歴データをランダムに組み合わせたものである．このとき組み合わせた元のデータ中の 2 つ以上のアカウントが同じ商品を購買している場合，その履歴データのどちらかをランダムに選択し人工データ中から省いた．

表 2 に，データセットの概要を示す．

*3 http://www.grouplens.org/node/73

3.2 評価尺度

本稿では各手法の評価尺度として，トップ n 正答率を採用する．各アカウントについて，実験データ中で購買した商品の中で最も新しいものをテストデータとし，それ以外の商品を訓練データとする．訓練データからアカウント u の購買した商品を省いたものとテストデータの各商品について，$P(i|u)$ を算出し，その値が最も高い n 件を u への推薦リストとする．このときトップ n 正答率は下式で表される．

$$A = \frac{\left|\{u|u \in \boldsymbol{U} \wedge \bar{i}_u \in \hat{\boldsymbol{I}}_u\}\right|}{U}, \tag{12}$$

ここで $\bar{i}_u$ はテストデータである商品を，$\hat{\boldsymbol{I}}_u$ は u に推薦する商品集合，すなわち $P(i|u)$ が最も高い n 件の商品集合をそれぞれ表す．

3.3 比較手法

以下に示す 4 つの推薦アルゴリズムを用いて，提案法を評価する．

1. **OurModel**：提案法．ただし，時刻には各アカウントのテストデータの時刻を用いた．またモデルのハイパーパラメータはそれぞれ $\alpha = 0.001$，$\beta = 0.001$，$\gamma = 0.003$，$\xi = 1.0$，$\eta = 2.0$，ν は全購買時刻の平均，ρ は全購買時刻の分散をそれぞれ用いた．

2. **Unigram**：本モデルでは，すべてのアカウントが購買する商品は単一の多項分布から選択される．

$$P(i|u) \propto \frac{\sum_{u=1}^{U}\sum_{m=1}^{M_u} I(i_{um}=i)}{\sum_{i'=1}^{N}\sum_{u=1}^{U}\sum_{m=1}^{M_u} I(i_{um}=i')}. \tag{13}$$

3. **UserCF**：ユーザベースの協調フィルタリング [Resnick 94]．まず，2 アカウントの購買履歴の類似度をピアソンの積率相関係数によって算出する．

$$\mathrm{UserSim}(u,u') = \frac{\sum_{i=1}^{N}(r_{ui}-\bar{r}_u)(r_{u'i}-\bar{r}_{u'})}{\sqrt{\sum_{i=1}^{N}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i=1}^{N}(r_{u'i}-\bar{r}_{u'})^2}}. \tag{14}$$

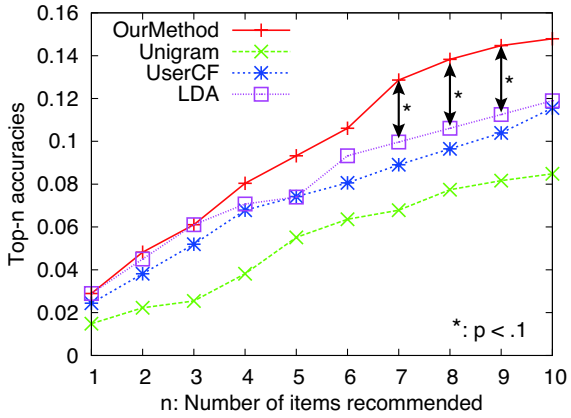
ここで，アカウント u が商品 i を購買したことがあれば $r_{ui} = 1$，そうでなければ $r_{ui} = 0$ とし，$\bar{r}_u = M_u/N$．このときアカウントが商品を購買する確率を

$$P(i) = \bar{r}_u + \frac{\sum_{u' \in \boldsymbol{U}_{\setminus u}} \mathrm{UserSim}(u,u')(r_{u'i}-\bar{r}_{u'})}{\sum_{u' \in \boldsymbol{U}_{\setminus u}} \left|\mathrm{UserSim}(u,u')\right|}, \tag{15}$$

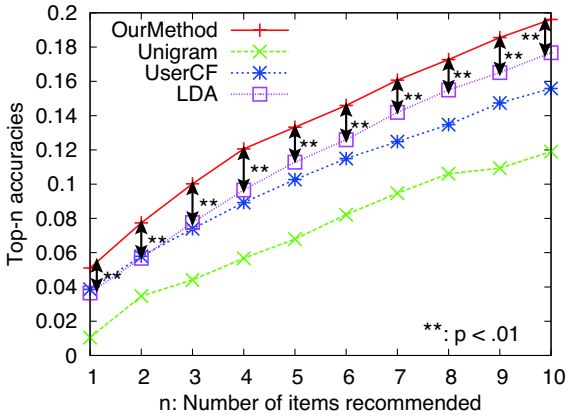
のように求める．ここで $\boldsymbol{U}_{\setminus u} = \{1,\ldots,U\} - \{u\}$．

4. **LDA**：Latent Dirichlet Allocation [Blei 03]．トピック集合 $\boldsymbol{Z}$ は，購買商品集合 $\boldsymbol{I}$ を入力として提案法と同様，Collapsed ギブスサンプリングを用いて推定できる．ユーザ u の m 番目の購買を決定するトピック z_j のサンプリング確率は下式により計算できる．

$$P(z_j = k|\boldsymbol{I},\boldsymbol{Z}_{\setminus j},\alpha,\beta) \propto \left(M_{uk\setminus j}+\alpha\right)\cdot\frac{M_{ki_j}+\beta}{M_k+\beta N}. \tag{16}$$



(a) MovieLens.



(b) EachMovie.

図 2: OurModel, Unigram, UserCF, LDA のトップ n 正答率

アカウント u が商品 i を購買する確率は，全購買について (16) によるサンプリングを十分な回数繰り返しトピックを推定した後，下式のように計算することができる．

$$P(i|u,\boldsymbol{I},\boldsymbol{Z},\alpha,\beta) = \sum_{z=1}^{Z} P(i,z|u,\boldsymbol{I},\boldsymbol{Z},\alpha,\beta) = \sum_{z=1}^{Z} \frac{M_{uz}+\alpha}{M_u+\alpha Z}\cdot\frac{M_{zi}+\beta}{M_z+\beta N}, \tag{17}$$

モデルのハイパーパラメータには $\alpha = 0.001$，$\beta = 0.001$ をそれぞれ用いた．

3.4 結果

提案法と LDA のトピック数は，各アカウントの 2 番目に新しい購買商品をバリデーションセットとみなし，テストデータとバリデーションセットを除くデータを用いて学習を行い，バリデーションセットを正解としたときのトップ 10 正答率を最大とするものを選択した．結果，MovieLens データに対しては $Z = 90$ の OurModel，$Z = 80$ の LDA が，EachMovie データに対しては $Z = 70$ の OurModel，$Z = 40$ の LDA が選択された．

次に，図 2 にテストデータに対する各手法のトップ n 正答率を示す．MovieLens データに対しては $7 \leq n \leq 9$ の場合において，EachMovie データに対してはすべての n について統計的に有意に OurModel の方が他の手法より高精度に推薦できた（$p < .1$, 符号検定）．その他の場合においても，統計的な有意差はなかったが，OurModel の方が高精度であった．

4. まとめ

本稿では，ユーザ数未知でアカウントを共有している状況での推薦法を提案した．提案法の有効性を検証するために，仮想的に 1 つのアカウントを 1 人から 5 人の複数人が利用している人工データを作成し，提案法が従来法よりも高精度に推薦できるかを検証した．結果，提案法の有効性を示すことができた．

今後は，本稿で導入したモデルと [甲谷 11] で導入したモデルについて，人工データの各履歴について，本来のアカウント情報を正解としたユーザの予測精度の観点から比較し，ディリクレ過程の有効性を検証する予定である．なお，被験者実験等による推薦結果の主観評価も今後の課題である．

参考文献

[Aldous 83] Aldous, D.: Exchangeability and related topics, *École d'Été de Probabilités de Saint-Flour*, Vol. XIII, pp. 1–198 (1983)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Hofmann 99] Hofmann, T.: Probabilistic latent semantic indexing, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, ACM (1999)

[Koren 09] Koren, Y.: Collaborative filtering with temporal dynamics, in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 447–456, ACM (2009)

[Resnick 94] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews, in *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pp. 175–186, ACM (1994)

[Tösher 09] Tösher, A. and Jahrer, M.: The BigChaos solution to the Netflix grand prize (2009), `http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf`

[甲谷 11] 甲谷 優, 岩田 具治, 藤村 考：複数人によるアカウントの共有を考慮したトピックモデルに基づく協調フィルタリング, 日本データベース学会論文誌, Vol. 9, No. 3, pp. 7–12 (2011)