DBSJ Journal, Vol.9, No.3
February 2011

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

Collaborative Filtering based on Topic Models Considering an Account Shared by Multiple Users

甲谷 優♥　岩田 具治♦
藤村 考♥

Yutaka KABUTOYA　Tomoharu IWATA
Ko FUJIMURA

1つのアカウントの複数人のユーザによる共有を考慮することにより推薦システムの精度を改善するための手法を提案する．VODなどのサービスでは，複数人，例えば家族が1つのアカウントを共有しているケースが多い．そのような場合，各アカウントの購買履歴から個々人の嗜好を推定することができなくなり，結果として推薦システムの精度が低下してしまう．そこで我々は，Probabilistic Lantent Semantic Analysis (PLSA)を基にアカウントを共有する潜在ユーザ毎の購買行動をモデル化し，このモデルに基づいて推薦を行う．本稿では，映像評点の実ログデータと，そのデータの 2 つのアカウントを結合した仮想ログデータの 2 種類の実験データを用いて，提案法が高精度にアカウントを共有するユーザを予測できること，従来の推薦アルゴリズムよりも高精度に推薦できることを示す．

We propose a probabilistic topic 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 purchase history of the account, thus it is difficult to accurately recommend items to those who share an account. We tackle this problem by assuming latent users sharing an account and establish a model by extending Probabilistic Latent Semantic Analysis (PLSA). Experiments on real log datasets from online movie services and artificial datasets created from these real datasets by combining the purchase histories of two accounts demonstrate high prediction accuracy of users and higher recommendation accuracy than conventional methods.

1. はじめに

近年，Amazonをはじめとして多くのWebサービスが推薦システムを重要視している．これは，推薦システムがeコマースの売上げ向上や，Webサイト内のクリック率の改善などWeb サービスにおけるユーザの満足度向上に有効であるためである．

それゆえに，推薦システムを実現する協調フィルタリングアルゴリズムが数多く提案されてきた．しかしNetflix Prizeの優勝チームのメンバであるKoren[9]とTosher[15]は，それら従来法にはアカウントが複数人に共有されている場合精度が低下するという問題点が存在することを指摘した．これは，従来の協調フィルタリングアルゴリズムによって推定される嗜好がアカウントによって関連付けられており，個々のユーザに関連付けられているとは限らないことが原因である．

ここで，我々が取り組むべき課題のケーススタディを与える．ビデオオンデマンド(VOD)サービスにて，ある家族に映像を推薦する場合を考える．家族は父親，母親，息子の3人から成り立っており，1つのアカウントを共有している．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．従来の協調フィルタリングアルゴリズムによって推定されるこのアカウントの嗜好は，3つの独立した嗜好の混合ということになり，適切な映像を推薦することが困難である．

また2 人で共有されているアカウントについて，それら個々の2 人と，2 人が一緒に利用する場合で嗜好が異なる場合，そのアカウントに対しては3 人のユーザに対する推薦とみなすべきである．ゆえに，アカウントを共有する実際のユーザが何人かわかっても推薦の精度が低下する可能性がある．

そこで本論文では，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する．提案モデルは，Probabilistic Latent Semantic Analysis (PLSA) モデル[6]の拡張となっている．PLSA においてはアカウント毎に1つの潜在トピック比率（嗜好）を持つと仮定しているが，提案モデルでは潜在トピック比率はアカウントが同じ場合でも潜在ユーザ毎に異なると仮定している．

さらに，提案モデルでは個々の購買の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ため，潜在ユーザは購買した商品と購買時刻により推定される．すなわち，実際の利用シーンにおいては，提案法では時刻ごとに異なる潜在ユーザ向けの推薦リストが出力されることになる．

実験では，映像評点に関する実ログデータと，そのデータセット中の2つのアカウントの履歴データを組み合わせて複数人による1 アカウントの共有を仮想的に作り出した人工データの2 種類のデータセットを用いた．ただし，実ログデータ中のそれぞれのアカウントは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．まず，提案モデルが適切に潜在ユーザを推定できているかを評価するために，人工データに対し提案法を適用し，個々の履歴が実ログデータにおけるどちらのアカウントによるものか予測した．さらに，潜在ユーザの推定が推薦システムの改善に有効であるか検証するために，人工データ，実ログデータに対して提案法による推薦精度を従来の協調フィルタリングアルゴリズムのものと比較した．

♥ 正会員　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　{kabutoya.yutaka, fujimura.ko}@lab.ntt.co.jp
♦ 非会員　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　iwata@cslab.kecl.ntt.co.jp

2. 提案法

2.1 表記法

今，履歴データとして，U 個のアカウントと，各アカウントについて購買した商品と，購買時刻のペアの集合 $(\boldsymbol{i}_u, \boldsymbol{t}_u)$ が与えられたとする．$\boldsymbol{i}_u = \{i_{um}\}_{m=1}^{M_u}$ はアカウント u により購買された商品の集合，$\boldsymbol{t}_u = \{t_{um}\}_{m=1}^{M_u}$ はアカウントu の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．

表 1　表記法

Table 1　Notation.

Symbol	Description
U	number of accounts
N	number of unique items
Z	number of topics
V	number of latent users per an account
M_u	number of items purchased by the uth account
i_{um}	mth item purchased by the uth account
t_{um}	time-of-day of mth purchase by the uth account
z_{um}	topic of the mth purchase by the uth account
v_{um}	user of the mth purchase by the uth account

2.2 モデル

提案法の説明に入る前に，そのベースとなっているPLSAについて簡単に説明する．PLSAでは，各アカウントがトピック比率 $\boldsymbol{\xi}_u$ 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 $\boldsymbol{\varphi}_{z_{um}}$ 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，それら各ユーザがトピック比率 $\boldsymbol{\theta}_{uv}$ を持つ．アカウント u の購買の度に，ユーザ v_{um} がユーザ比率 $\boldsymbol{\psi}_u$ に従い選択される．購買した商品の生成はトピック比率が与えられたPLSAと同様で，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 $\boldsymbol{\varphi}_{z_{um}}$ から生成される．さらに購買した商品だけではなく，提案モデルでは購買毎にその時刻も生成される．購買時刻はユーザ依存の平均 $\tau_{uv_{um}}$，分散 $\sigma^2_{uv_{um}}$ の正規分布により生成される．

すなわち，提案モデルでは以下の過程によりアカウントの購買集合 $\{(\boldsymbol{i}_u, \boldsymbol{t}_u)\}_{u=1}^{U}$ が生成されるものとする．

1. For each topic $z = 1,\ldots,Z$:
   (a) Draw item probability $\boldsymbol{\varphi}_z \sim \text{Dirichlet}(\beta)$
2. For each account $u = 1,\ldots,U$:
   (a) Draw user proportions $\boldsymbol{\psi}_u \sim \text{Dirichlet}(\gamma)$
   (b) For each user $v = 1,\ldots,V$:
      i. Draw topic proportions $\boldsymbol{\theta}_{uv} \sim \text{Dirichlet}(\alpha)$
   (c) For each purchase $m = 1,\ldots,M_u$:
      i. Draw user $v_{um} \sim \text{Multinomial}(\boldsymbol{\psi}_u)$
      ii. Draw time-of-day $t_{um} \sim \text{Normal}(\tau_{uv_{um}}, \sigma^2_{uv_{um}})$
      iii. Draw topic $z_{um} \sim \text{Multinomial}(\boldsymbol{\theta}_{uv_{um}})$
      iv. Draw item $i_{um} \sim \text{Multinomial}(\boldsymbol{\varphi}_{z_{um}})$

ここで多項分布のパラメータである $\boldsymbol{\varphi}_z$，$\boldsymbol{\psi}_u$，$\boldsymbol{\theta}_{uv}$ は多項分布の共役事前分布であるディリクレ分布から生成されると仮定した．3章にて述べる実験では，ディリクレ分布のハイパーパラメータとして，$\alpha = 0.001$，$\beta = 0.001$，$\gamma = 0.001$ をそれぞれ用いている．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．

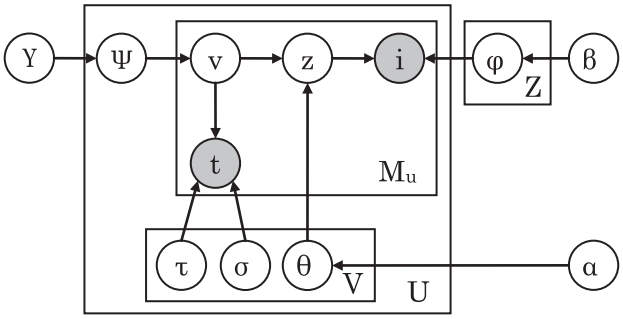


図 1　提案モデルのグラフィカルモデル

Fig.1 Graphical representation of the proposed model.

ユーザ比率集合，時刻の平均集合と標準偏差集合，トピック比率集合，商品出現確率集合が与えられたときのアカウントの購買履歴 $(\boldsymbol{i}_u, \boldsymbol{t}_u)$ の生成確率は

$$P(\mathbf{i}_u, \mathbf{t}_u \mid \boldsymbol{\psi}_u, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u, \boldsymbol{\Theta}_u, \boldsymbol{\Phi}) = \prod_{m=1}^{M_u} \sum_{v=1}^{V} P(v \mid \boldsymbol{\psi}_u) p(t_{um} \mid v, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) \sum_{z=1}^{Z} P(z \mid v, \boldsymbol{\Theta}_u) P(i \mid z, \boldsymbol{\Phi}), \tag{1}$$

となる．ここで，$P(v \mid \boldsymbol{\psi}_u) = \psi_{uv}$，$\boldsymbol{\psi}_u = \{\psi_{uv}\}_{v=1}^{V}$，$\boldsymbol{\tau}_u = \{\tau_{uv}\}_{v=1}^{V}$，$\boldsymbol{\sigma}_u = \{\sigma_{uv}\}_{v=1}^{V}$，$P(z \mid v, \boldsymbol{\Theta}_u) = \theta_{uvz}$，$\boldsymbol{\Theta}_u = \{\boldsymbol{\theta}_{uv}\}_{v=1}^{V}$，$P(i \mid z, \boldsymbol{\Phi}) = \phi_{zi}$，$\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_z\}_{z=1}^{Z}$ であり，式(1)の右辺の第2 項は以下に示す正規分布により与えられる．

$$p(t_{um} \mid v, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) = \frac{1}{\sqrt{2\pi}\sigma_{uv}} \exp\left(-\frac{|t_{um} - \tau_{uv}|^2}{2\sigma_{uv}^2} \right), \tag{2}$$

ここで，$|t_{um} - \tau_{uv}|$ は2 時刻 t_{um} と τ_{uv} の距離を表す．たとえば，$|2:00:00 - 23:00:00| = 3:00:00$．

2.3 パラメータ推定

提案モデルの未知パラメータは，最大事後確率(MAP)推定により求めることができる．未知パラメータはユーザ比率集合 $\boldsymbol{\Psi} = \{\boldsymbol{\psi}_u\}_{u=1}^{U}$，時刻の平均集合 $\boldsymbol{T} = \{\boldsymbol{\tau}_u\}_{u=1}^{U}$ と標準偏差集合 $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}_u\}_{u=1}^{U}$，トピック比率集合 $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_u\}_{u=1}^{U}$，商品出現確率集合 $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_z\}_{u=1}^{U}$ であり，$\boldsymbol{\Lambda} = \{\boldsymbol{\Psi}, \boldsymbol{T}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}\}$ で表現する．購買集合 $\{(\mathbf{i}_u, \mathbf{t}_u)\}_{u=1}^{U}$ が与えられたときの未知パラメータ集合 $\boldsymbol{\Lambda}$ の対数尤度は

$$L(\mathbf{\Lambda} \mid \mathbf{U}) = \sum_{u=1}^{U}\sum_{m=1}^{M_u} \log \sum_{v=1}^{V} P(v \mid \boldsymbol{\psi}_u) p(t_{um} \mid v, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) \sum_{z=1}^{Z} P(z \mid v, \boldsymbol{\theta}_u) P(i \mid z, \boldsymbol{\varphi}_z), \tag{3}$$

となる．なお，ユーザ数V，トピック数Zは既知とする．

未知パラメータ$\mathbf{\Lambda}$をMAP推定するには，事後確率を直接最大化するよりもEMアルゴリズム[4]を用いて以下に示すQ関数を最大化する方が容易である．

$$\begin{aligned} Q(\mathbf{\Lambda} \mid \hat{\mathbf{\Lambda}}) = & \\ & \sum_{u=1}^{U}\sum_{m=1}^{M_u}\sum_{v=1}^{V}\sum_{z=1}^{Z} P(v, z \mid u, m; \mathbf{\Lambda}) \log P(v \mid \boldsymbol{\psi}_u) p(t_{um} \mid v, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) P(z \mid v, \boldsymbol{\theta}_u) P(i \mid z, \boldsymbol{\varphi}_z) \\ & + \sum_{u=1}^{U} \log p(\boldsymbol{\psi}_u) + \sum_{u=1}^{U}\sum_{v=1}^{V} \log p(\boldsymbol{\theta}_{vu}) + \sum_{z=1}^{Z} \log p(\boldsymbol{\varphi}_z), \end{aligned} \tag{4}$$

ここで$\mathrm{P}(v, z \mid u, m; \mathbf{\Lambda})$は$u$番目のアカウントの$m$番目の購買に対するユーザとトピックの事後確率を表す．Eステップでは，ベイズ則に従いユーザとトピックの事後確率を計算する．

$$P(v, z \mid u, m; \mathbf{\Lambda}) = \frac{P(v \mid \boldsymbol{\psi}_u) p(t_{um} \mid v, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) P(z \mid v, \boldsymbol{\Theta}_u) P(i \mid z, \boldsymbol{\varphi}_z)}{\sum_{v'=1}^{V} P(v' \mid \boldsymbol{\psi}_u) p(t_{um} \mid v', \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) \sum_{z'=1}^{Z} P(z' \mid v', \boldsymbol{\Theta}_u) P(i \mid z', \boldsymbol{\varphi}_z)}.$$

Mステップでは，$\sum_{v=1}^{V} \psi_{uv} = 1$，$\sum_{z=1}^{Z} \theta_{uvz} = 1$，$\sum_{i=1}^{N} \phi_{zi} = 1$という制約のもと，$Q(\mathbf{\Lambda} \mid \hat{\mathbf{\Lambda}})$を，$\psi_{uv}$，$\tau_{uv}$，$\sigma_{uv}$，$\theta_{uvz}$，$\phi_{zi}$に関して最大化することにより，各未知パラメータの推定値$\hat{\psi}_{uv}$，$\hat{\tau}_{uv}$，$\hat{\sigma}_{uv}$，$\hat{\theta}_{uvz}$，$\hat{\phi}_{zi}$を求める．

$$\hat{\phi}_{zi} = \frac{\sum_{u=1}^{U}\sum_{m=1}^{M_u}\sum_{v=1}^{V} I(i_{um} = i) P(v, z \mid u, m; \mathbf{\Lambda}) + \beta}{\sum_{u=1}^{U}\sum_{m=1}^{M_u}\sum_{v=1}^{V} P(v, z \mid u, m; \mathbf{\Lambda}) + \beta N}, \tag{6}$$

$$\hat{\psi}_{uv} = \frac{\sum_{m=1}^{M_u}\sum_{z=1}^{Z} P(v, z \mid u, m; \mathbf{\Lambda}) + \gamma}{\sum_{v'=1}^{V}\sum_{m=1}^{M_u}\sum_{z=1}^{Z} P(v', z \mid u, m; \mathbf{\Lambda}) + \gamma V}, \tag{7}$$

$$\hat{\tau}_{uv} = \frac{\sum_{m=1}^{M_u}\sum_{z=1}^{Z} P(v, z \mid u, m; \mathbf{\Lambda}) t_{um}}{\sum_{m=1}^{M_u}\sum_{z=1}^{Z} P(v, z \mid u, m; \mathbf{\Lambda})}, \tag{8}$$

$$\hat{\sigma}_{uv}^2 = \frac{\sum_{m=1}^{M_u}\sum_{z=1}^{Z} P(v, z \mid u, m; \mathbf{\Lambda}) \left| t_{um} - \hat{\tau}_{uv} \right|^2}{\sum_{m=1}^{M_u}\sum_{z=1}^{Z} P(v, z \mid u, m; \mathbf{\Lambda})}, \tag{9}$$

$$\hat{\theta}_{uvz} = \frac{\sum_{m=1}^{M_u} P(v, z \mid u, m; \mathbf{\Lambda}) + \alpha}{\sum_{z'=1}^{Z}\sum_{m=1}^{M_u} P(v, z' \mid u, m; \mathbf{\Lambda}) + \alpha Z}, \tag{10}$$

ここで$\mathrm{I}(\cdot)$は指示関数，すなわちAが真ならば$\mathrm{I}(A) = 1$，偽ならば$\mathrm{I}(A) = 0$を表す．式(3)で表わされる対数尤度が収束するまでEステップとMステップを交互に繰り返すことにより，$\mathbf{\Lambda}$の局所最適解を推定することができる．

3. 実験

3.1 データセット

EachMovie，MovieLens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．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と考え，ランダムに選択した2アカウントの履歴データを組み合わせ仮想的に2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．組み合わせた実データ中の2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．

表2　データセット

Table 2　Datasets.

	# of accounts	# of items	# of purchases
EachMovie	7,077	1,249	569,171
MovieLens	471	1,152	88,826

3.2 ユーザの推定

3.2.1 評価尺度

本実験の目的は提案モデルがどれくらいの精度で1アカウントを共有するユーザを推定できるかを評価することである．人工データ中の各履歴データがそのアカウントを構成する実データ中における2アカウントのどちらによるものかを予測した．

ここで人工データ中のアカウントuの購買mが実データにおいてはユーザ$\bar{v}_{um} (\in \{1,2\})$番目によるものであるとする．アカウント$u$の購買$m$がユーザ$\bar{v}_{um}$によるものである確率を

$$P(v = \bar{v}_{um} \mid u, m) = \sum_{z=1}^{Z} P(v = \bar{v}_{um}, z \mid u, m; \mathbf{\Lambda}), \tag{11}$$

のように算出した．ただし，トピック数Zは$\{10,20,\ldots,100\}$の10種類を用いた．人工データの各アカウントは2人のユーザによって利用されているので，ユーザ数Vは2とし，$\mathrm{P}(v = \bar{v}_{um} \mid u, m) > 0.5$である場合正しくユーザを推定できたものとみなした．

3.2.2 結果

図2に提案モデルの(a) EachMovieデータと(b) MovieLensデータに対するユーザの予測精度を示す．ベースラインは各アカウントについて購買数が多い方のユーザを選択し続けた場合の予測精度，すなわち購買数が多いユーザの購買の割合を表している．両方のデータセット，すべてのトピック数について提案モデルの方がベースラインよりも高精度にユーザを推定できた(符号検定，$p < .01$)．

3.3 推薦

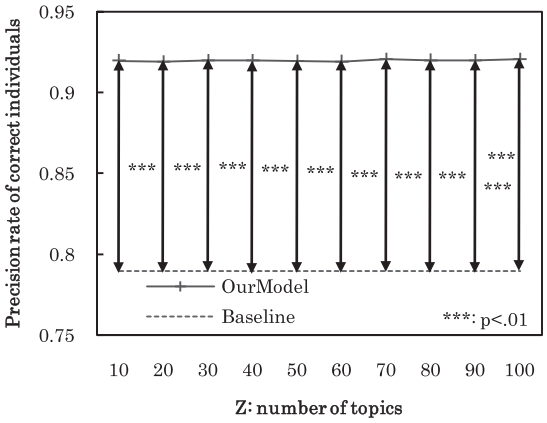
3.3.1 設定

本評価の目的はアカウントを共有する潜在ユーザを予測することで推薦精度が向上するかどうかを検証することであり，人工データセットだけでなく実データセットも用いた．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，$\mathrm{P}(i \mid u)$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．

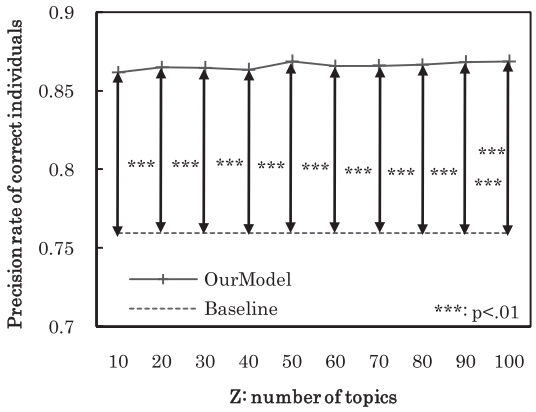
3.3.2 評価尺度

各手法による推薦精度として，トップn正答率を算出する．トップn正答率は，たとえば岩田[16]が用いている評価尺度であり，全てのアカウントを訓練データ中における購買数に関わらず平等に評価することができる．

トップn正答率は，下式により算出する．



(a) EachMovie.



(b) MovieLens.

図 2　提案モデルのユーザの予測精度
Fig.2 Prediction rates of the proposed model.

$$A = \frac{\left|\left\{u \mid u \in \mathbf{U} \wedge \bar{i}_u \in \hat{\mathbf{I}}_u\right\}\right|}{U}, \qquad (12)$$

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 $\mathrm{P}(i \mid u)$ が最も高い n 件の商品集合をそれぞれ表す．

3.3.3 比較手法

以下に示す 5 つの推薦アルゴリズムを用いて，提案法を評価する．

1. **OurModel**：提案モデルに基づく推薦．アカウントが商品を購買する確率は

$$P(i \mid u) \propto \sum_{v=1}^{V} P(v \mid \boldsymbol{\psi}_u) p(\bar{t}_u \mid v, \boldsymbol{\tau}_u, \boldsymbol{\sigma}_u) \cdot \sum_{z=1}^{Z} P(z \mid v, \boldsymbol{\Theta}_u) P(i \mid z, \boldsymbol{\Phi}), \qquad (13)$$

のように算出できる．ここで $\bar{t}_u$ はアカウントの最後の購買時刻，すなわちテストデータの時刻を表す．

2. **Unigram**：ユニグラムモデル．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独立に選択されることを仮定したモデルである．

$$P(i \mid u) \propto P(i) = \frac{\sum_{u=1}^{U} \sum_{m=1}^{M_u} I(i_{um} = i)}{\sum_{u=1}^{U} M_u}, \qquad (14)$$

要するに，購買された数が多い商品から順に，すべてのアカウントに同じように推薦するモデルである．

3. **UserCF**：ユーザベースの協調フィルタリング．Resnick ら[10]により提案された推薦システム “GroupLens” に採用されたアルゴリズムである．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．

$$UserSim(u,u') = \frac{\sum_{i=1}^{N}(r_{ui} - \bar{r}_u)(r_{u'i} - \bar{r}_{u'})}{\sqrt{\sum_{i=1}^{N}(r_{ui} - \bar{r}_u)^2}\sqrt{\sum_{i=1}^{N}(r_{u'i} - \bar{r}_{u'})^2}}, \qquad (15)$$

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，$r_u = M_u / N$ はアカウントの購買率を表す．このときアカウント u が商品を購買する確率を

$$P(i \mid u) \propto \bar{r}_u + \frac{\sum_{u' \in \mathbf{U}_{\setminus u}} UserSim(u,u')(r_{u'i} - \bar{r}_{u'})}{\sum_{u' \in \mathbf{U}_{\setminus u}} \left|UserSim(u,u')\right|}, \qquad (16)$$

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

4. **ItemCF**：アイテムベースの協調フィルタリング．Sarwar ら[11]によって提案された手法であり，一般的にアカウント数よりも商品数が小であるためユーザベースのものよりも計算コストが少ないとされている．まず，2 商品がどのアカウントに購買されたかの類似度を adjusted cosine similarity を用いて

$$ItemSim(i,i') = \frac{\sum_{u=1}^{U}(r_{ui} - \bar{r}_u)(r_{ui'} - \bar{r}_{ii})}{\sqrt{\sum_{u=1}^{U}(r_{ui} - \bar{r}_u)^2}\sqrt{\sum_{u=1}^{U}(r_{ui'} - \bar{r}_u)^2}}, \qquad (17)$$

のように算出する．このときアカウントが商品を購買する確率はその商品とそのアカウントが購買した商品との類似度の和で算出される．

$$P(i \mid u) \propto \frac{\sum_{i' \in \mathbf{I}_{\setminus i}} r_{ui'} ItemSim(i,i')}{\sum_{i' \in \mathbf{I}_{\setminus i}} \left|ItemSim(i,i')\right|}, \qquad (18)$$

ここで $\boldsymbol{I}_{\setminus i} = \{1,\ldots,N\} - \{i\}$．

5. **PLSA**：Probabilistic latent semantic analysis[6]．PLSA モデルでは，トピック比率集合は $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_u\}_{u=1}^{U}$ から直接評価される．ここで $\xi_{uz} = P(z \mid \boldsymbol{\xi}_u)$ はアカウント u の z 番目のトピック比率を表す．$\boldsymbol{\xi}_u$ と $\boldsymbol{\Phi}$ が与えられたもとでの $\boldsymbol{i}_u$ の確率はのようになる．

$$P(\mathbf{i}_u \mid \boldsymbol{\xi}_u, \boldsymbol{\Phi}) \propto \prod_{m=1}^{M_u} \sum_{z=1}^{Z} P(z \mid \boldsymbol{\xi}_u) P(i_{um} \mid z, \boldsymbol{\Phi}). \qquad (19)$$

PLSA の未知パラメータ $\boldsymbol{Y}$ はトピック比率集合 $\boldsymbol{\Xi}$ と商品出現確率集合 $\boldsymbol{\Phi}$ であり，EM アルゴリズムを用いて以下に示す尤度を最大化することで求めることができる．

$$L(\mathbf{Y} \mid \mathbf{U}) = \sum_{u=1}^{U} \sum_{m=1}^{M_u} \log \sum_{z=1}^{Z} P(z \mid \boldsymbol{\xi}_u) P(i \mid z, \boldsymbol{\Phi}). \qquad (20)$$

E ステップでは，現在の推定値が与えられたもとでのトピック事後確率を計算する．

$$P(z \mid u, m; \mathbf{Y}) = \frac{P(z \mid \boldsymbol{\xi}_u) P(i \mid z, \boldsymbol{\Phi})}{\sum_{z'=1}^{Z} P(z' \mid \boldsymbol{\xi}_u) P(i \mid z', \boldsymbol{\Phi})}. \qquad (21)$$

M ステップでは，トピック比率と商品出現確率を下式

により推定する．

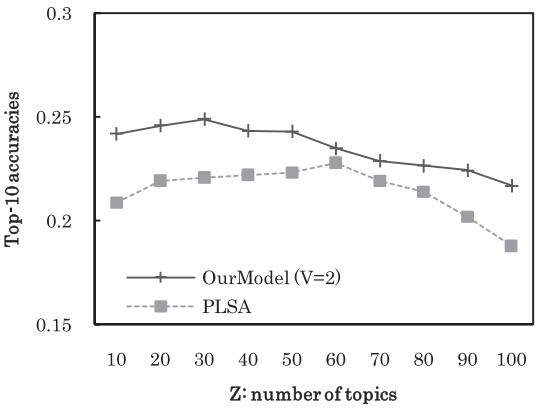
$$\hat{\xi}_{uz} = \frac{\sum_{m=1}^{M_u} P(z \mid u, m; \mathbf{Y})}{\sum_{z'=1}^{Z} \sum_{m=1}^{M_u} P(z' \mid u, m; \mathbf{Y})}, \quad (22)$$

$$\hat{\phi}_{zi} = \frac{\sum_{u=1}^{U} \sum_{m=1}^{M_u} I(i_{um} = i) P(z \mid u, m; \mathbf{Y})}{\sum_{u=1}^{U} \sum_{m=1}^{M_u} P(z \mid u, m; \mathbf{Y})}. \quad (23)$$

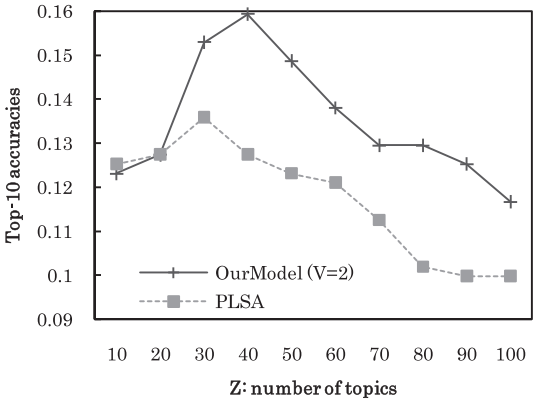
提案モデルと同じく，トピック比率および商品出現確率の事前分布としてディリクレ分布を用いた．
このときアカウントが商品を購買する確率は

$$P(i \mid u) \propto \sum_{z=1}^{Z} P(z \mid \xi_u) P(i \mid z, \Phi). \quad (24)$$

3.3.4 結果



(a) EachMovie.



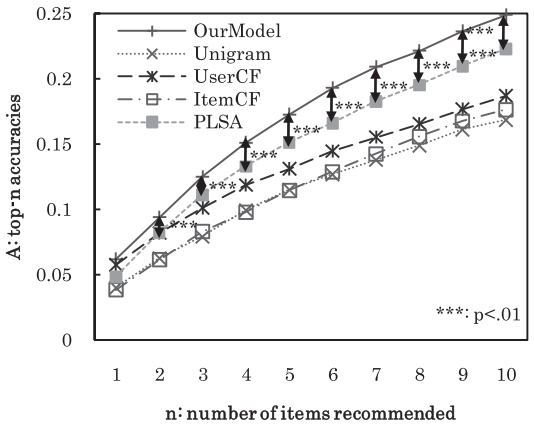
(b) MovieLens.

図 3 トピック数 Zを変化させたときのバリデーションセットに対するトップ 10 正答率

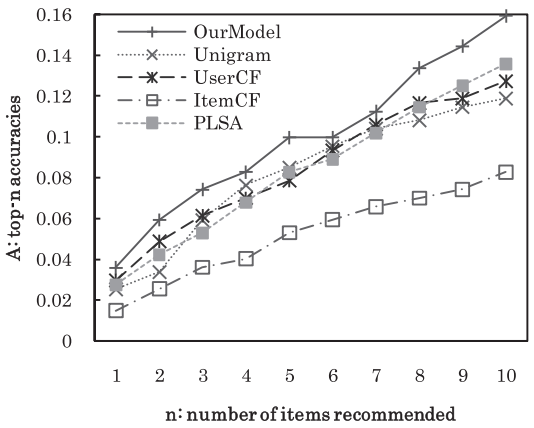
Fig.3 Top-10 accuracy scores for the validation set with different numbers of topics Z.

提案モデルのユーザ数，トピック数と PLSA のトピック数は以下のようにして求めた．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．各ユーザ数，トピック数におけるバリデーションセットに対するトップ 10 正答率を図 3 に示す．結果，EachMovie データに対しては$Z = 30$の提案モデル，$Z = 60$の PLSA が，MovieLens データに対しては$Z = 40$の提案モデル，$Z = 30$の PLSA が選択された．ここで，提案モデルの潜在ユーザ数Vは 2 としている．

次に，図 4 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．EachMovie データに対しては，統計的に有意に提案モデルの方が他の手法より高精度に推薦できた（$n \geq 2, p < .01$，符号検定）．また，MovieLens データに対しても，サンプル数が少なかったため統計的に有意ではなかったが，提案モデルの方が他の手法よりも高精度であった．すなわち，アカウントを共有するユーザの存在を考慮することが推薦精度の向上に有効であると言える．



(a) EachMovie.



(b) MovieLens.

図 4 テストデータに対するトップ n 正答率

Fig.4 Top-n accuracy scores for the test data.

4. 関連研究

時間情報を用いた推薦システムの精度改善に関する研究は多い．中でも，新しい購買履歴ほどユーザの現在の嗜好の推定に有効であるという仮説を置いた研究は多い．これは時間経過に伴う学習対象の変化を意味するコンセプトドリフトの一種で，“gradual” なものに関する研究であると言える．Adomavicius ら[1]は従来のメモリベースの協調フィルタリングをベースとして時間情報をはじめとした文脈情報を考慮した推薦手法を提案している．Sugiyama ら[13]は検索エンジンのパーソナライズに時間変化するユーザプロファイルを利用している．Ding ら[5]は類似度ベースの協調フィルタリング手法に時間の重みを考慮する手法を提案している．岩田ら[16]は購買履歴の新しさが推薦の精度に与える影響の大きさを最大エントロピーモデルとマルコフモデルの概念を元にモデル化している．Netflix Prize の優勝チームの一人である Koren[9]は時間に伴うユーザ嗜好の変化に着目している．このように，“gradual” なコンセプトドリフト

は推薦システムの精度改善によく利用されているが，別のタイプのコンセプトドリフト，例えば我々の手法のように“periodic”なコンセプトドリフトを考慮した研究は少ない．

トピックモデルには幅広い応用があり，多くの研究者の注目を集めている．Hofmann[6]によって提案されたProbabilistic Latent Semantic Analysis (PLSA)モデルは，そのもっとも代表的なものである．PLSAは情報検索や協調フィルタリングへの応用が提案されている[7]．Siら[12]はPLSAをベースとした協調フィルタリング用のモデルを提案している．Iwataら[8]はトピックモデルを用いた文書群の可視化手法を提案している．Bleiら[2]はトピックモデルを用いてWebページや画像データに対して自動的にアノテーションを付与している．本研究ではトピックモデルを用いてアカウントを共有する複数ユーザの購買をモデル化している．

5. まとめと今後の課題

本論文では，トピックモデルを用いて1つのアカウントを共有する複数人のユーザの購買をモデル化する手法を提案した．提案法の有効性を検証するために，2つの実験を行った．1つ目の実験では，実データの2つのアカウントを組み合わせて仮想的に1つのアカウントを2人が利用している人工データを作成し，提案モデルによって各アカウントを利用している各ユーザをどれくらい正確に予測できるか評価した．2つ目の実験では，提案モデルが従来法よりも高精度に推薦できるかを検証した．

本論文ではユーザ数を既知で，かつ全てのアカウントで同数としたが，ディリクレ過程[14]などのノンパラメトリックベイズモデルに拡張することで，アカウント毎にユーザ数を自動決定することが可能になる．また，未知パラメータもMAP推定ではなくLDA[3]のようにベイズ推定することでより頑健な推定が期待できる．被験者実験等による推薦結果の主観評価は今後の課題である．

[文献]

[1] Adomavicius, G., Sankaranarayanan, R., Sen, S. and Tuzhilin, A.: “Incorporating ontextual information in recommender systems using a multidimensional approach,” ACM Transactions on Information Systems, Vol.23, No.1, pp.103–145 (2005).

[2] Blei, D.M. and Jordan, M.I.: “Modeling annotated data,” Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, pp.127–134 (2003).

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

[4] Dempster, A., Laird, N. and Rubin, D.: “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistical Society, Series B, Vol.39, No.1, pp.1–38 (1977).

[5] Ding, Y. and Li, X.: “Time weight collaborative filtering,” Proceedings of the 14th ACM International Conference on Information and Knowledge Management, ACM, pp.485–492 (2005).

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

[7] Hofmann, T.: “Collaborative filtering via Gaussian probabilistic latent semantic analysis,” Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp.259–266 (2003).

[8] Iwata, T., Yamada, T. and Ueda, N.: “Probabilistic latent semantic visualization: topic model for visualizing documents,” Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp.363–371 (2008).

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

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

[11] Sarwar, B., Karypis, G., Konstan, J. and Reidl, J.: “Item-based collaborative filtering recommendation algorithms,” Proceedings of the 10th International Conference on World Wide Web, ACM, pp.285–295 (2001).

[12] Si, L. and Jin, R.: “Flexible mixture model for collaborative filtering,” Proceedings of the 20th International Conference on Machine Learning, AAAI Press, pp.704–711 (2003).

[13] Sugiyama, K., Hatano, K. and Yoshikawa, M.: “Adaptive web search based on user profile constructed without any effort from users,” Proceedings of the 13th International Conference on World Wide Web, ACM, pp.675–684 (2004).

[14] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M.: “Hierarchical Dirichlet processes,” Journal of the American Statistical Association, Vol.101, No.476, pp.1566–1581 (2006).

[15] T¨osher, A. and Jahrer, M.: “The BigChaos solution to the Netflix grand prize” (2009).

[16] 岩田具治，山田武士，上田修功：購買順序を効率的に用いた協調フィルタリング，情報処理学会論文誌: 数理モデル化と応用， Vol.49, No.SIG 4 (TOM20), pp.125‐133 (2008).

甲谷 優 Yutaka KABUTOYA

NTTサイバーソリューション研究所所属．2008年京都大学大学院情報学研究科博士前期課程修了．Webマイニングの研究開発に従事．日本データベース学会，人工知能学会，言語処理学会，各会員．

岩田 具治 Tomoharu IWATA

NTTコミュニケーション科学基礎研究所所属．2008年京都大学大学院情報学研究科システム科学専攻博士課程修了．機械学習，データマイニング，情報可視化の研究に従事．博士（情報学）．情報処理学会，電子情報通信学会，各会員．

藤村 考 Ko FUJIMURA

NTTサイバーソリューション研究所主幹研究員．電気通信大学大学院情報システム学研究科客員教授．1989年北海道大学大学院工学研究科博士課程修了．ソーシャルメディアからの知識抽出と可視化の研究開発に従事．工学博士．情報処理学会，電子情報通信学会，日本データベース学会，各会員．