超スパースなデータに対する潜在クラスモデルを用いた推薦システムに関する研究

情報数理応用研究　　5212C022-6　坂本俊輔
指導教員　後藤正幸

A Study on Recommender System Using Latent Class Model for High Dimensional and Sparse Data

SAKAMOTO Shunsuke

1　研究背景・目的

近年，情報技術の進展により，EC サイト等の Web サービスで扱われる情報やアイテムの数が膨大になっている．ユーザの嗜好の多様化も伴い，ユーザの嗜好を満たした情報やアイテムを自動で推薦するシステムの効果は大きい．このような推薦システムの代表的な手法として，ユーザ間の過去の購買履歴情報を用いて推薦を行う協調フィルタリングがあり，確率モデル [1], [2] やベクトル空間モデル [3] を用いた手法など，様々な手法が既に提案されている．特に確率モデルの中でも，Aspect Model [1], Latent Dirichlet Allocation(以下 LDA) [2], [4] に代表される潜在クラスモデルの有効性を示す研究が数多くなされてきている．その理由としては，潜在クラスによってユーザの嗜好の異質性やアイテムの特徴の多様性を表現できるためである．特に LDA はベイズ統計の枠組みを導入し，潜在クラスへの所属確率分布自体を確率的に生成している．これにより，分布が学習データのみに依存する Aspect Model の問題点を解決している [5], [6].

他方，EC サイトは一般的に多くのアイテムを扱っていることから，購買履歴データは非常に高次元，かつスパースである．一例として，あるアパレル関係の EC サイトにおける 1 年間の購買データでは，アイテム数 5,168 に対しユーザ 1 人の平均購入点数は 13 点程度であり, データ密度は 0.25 %である．現実の EC サイトではこのようにデータ密度が 1 %以下であることが多く，推薦システムの構築にはこれらの超スパースなデータに対しても安定して推薦を可能にする必要がある．しかし，超スパースなデータに対して LDA を適用しようとした場合，学習データがモデルのパラメータ数に対して相対的に少ない状況となり，しばしばその学習が収束しないという問題が生じる．これは LDA の確率モデルでは，各潜在クラスに対し，全ユーザと全アイテムの所属確率が割り当てられるため，潜在クラスの数を変化させると，そのパラメータ数は，ユーザ数とアイテム数の合計に比例して大きくなることに起因する．そのため潜在クラスへの所属確率の事後分布を推定するのに十分な学習データ数が確保できないと，学習が収束しなくなってしまう．

上記問題の解決策として，LDA のパラメータ数に対して学習データが多い状況を実現することができれば，パラメータの推定が可能になると考えられる．すなわち，クラスタリングによって得られる類似ユーザの購買履歴データを統合すれば，上記の状況が実現できる可能性がある．しかし，一般的に EC サイトでは大量にアイテムを買う少数のユーザが存在する一方，殆ど購買を行わないユーザが大多数であるという傾向を持つ．そのため，単純にクラスタリングを適用しただけでは前者がサイズの小さいクラスタを，後者がサイズの大きいクラスタを構成し，サイズの小さいクラスタに関するパラメータの推定が困難となってしまう．LDA の学習を可能とするためのクラスタリングの場合，推定に十分なパラメータ数を確保するだけでなく，各クラスタのサイズが均一になっていることが望ましい．

以上の議論から，本研究では学習データに対し，階層的にクラスタリングを行う手法を導入することで，超スパースなデータの下でも LDA の学習を可能とさせる．サイズの大きいユーザクラスタに対して階層的にクラスタリングを行うことで各クラスタサイズを均一に近付け，モデルの学習を成功させる方法を提案する．提案手法を超スパースであるアパレル関係の EC サイトにおける 1 年間の購買データに適用し，提案手法の有効性を示す.

2　推薦システム

推薦システムとは，購買履歴からユーザの嗜好を推定し，アイテムを推薦するシステムのことである [8]．いま，ユーザを $u \in \{1, \cdots, U\}$, アイテムを $i \in \{1, \cdots, I\}$ とする．このとき，ユーザがアイテムを購買した場合は 1, 未購買の場合は 0 をとる購買履歴データを要素とする行列を $\boldsymbol{R} = [R_{u,i}]_{U \times I}$ と定義し ($1 \leq u \leq U$, $1 \leq i \leq I$), 未購買アイテムの中からユーザが好むと予測されるアイテムを推薦する.

3　潜在クラスモデルに基づく推薦システム

3.1　Latent Dirichlet Allocation

推薦システムにおける LDA は，Aspect Model にベイズ統計の枠組みを導入し，潜在クラスへの所属確率を表わすパラメータをディリクレ分布から確率的に生成するように拡張された潜在クラスモデルである．潜在クラスによってユーザの嗜好やアイテムの特徴の多様性を表現することができる．いま，K を潜在クラス数として，潜在クラスを $k \in \{1, \cdots, K\}$ で表わす．LDA では，ユーザ u がアイテム i を購買する確率 $P(i|u)$ を以下の式で表す.

$$P(i|u) = \sum_{k=1}^{K} \int \int \theta_{u,k} P(\theta_{u,k}|\alpha_k) \phi_{k,i} P(\phi_{k,i}|\beta_i) d\theta_{u,k} d\phi_{k,i}. \quad (1)$$

ここで，$\theta_{u,k}$ は，ユーザ u が潜在クラス k に所属する確率を表し，これをまとめて $\boldsymbol{\theta} = [\theta_{u,k}]_{U \times K}$ と表す．また $\phi_{k,i}$ は，潜在クラス k の下でアイテム i が生起する確率を表し，$\boldsymbol{\phi} = [\phi_{k,i}]_{K \times I}$ とする．また，式 (1) における $\theta_{u,k}$ と $\phi_{k,i}$ はそれぞれ任意の正の値をとるパラメータ $\boldsymbol{\alpha} = (\alpha_k)$, $\boldsymbol{\beta} = (\beta_i)$ を持つディリクレ分布に従うと仮定する．これらのディリクレパラメータによって，ユーザの嗜好やアイテムの特徴に適した様々な分布を仮定することができる．以上のモデルの構造から，LDA では各ユー

ザ，アイテムは全ての潜在クラスへの所属確率を持つという特徴を持つ．

3.2 学習・予測アルゴリズム

LDA の学習・予測アルゴリズムを以下に示す．

Step1) 各ハイパーパラメータ $\boldsymbol{\alpha}, \boldsymbol{\beta}$ を用いて，$\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ の初期値を生成する．

Step2) $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ の現在値からギブスサンプリング [7] を行って，潜在クラスへの所属確率の事後分布を近似する．

Step3) Step2 の結果から各ディリクレ分布のハイパーパラメータを更新し，再び $\boldsymbol{\theta}$ と $\boldsymbol{\phi}$ を生成する．

Step4) Step2, 3 を繰り返し，値が収束したら (1) 式の値が大きい順にアイテムを推薦する．

□

3.3 超スパースなデータにおける LDA の問題点

超スパースなデータに対する LDA の挙動を明らかにするため，某アパレル関係の EC サイトにおける 1 年間の購買データ (データ密度 0.25 %) に適用したときのパラメータ推定の事例を図 1 に示す．図 1 より，対数尤度は収束せず，パラメータの学習ができていないことがわかる．超スパースなデータの下では，学習データがモデルのパラメータ数に対して相対的に少なくなってしまい，安定した推定結果が得られずに LDA の学習が収束しないことが原因であると考えられる．従って，各ユーザに適切なアイテムを推薦できないことになる．

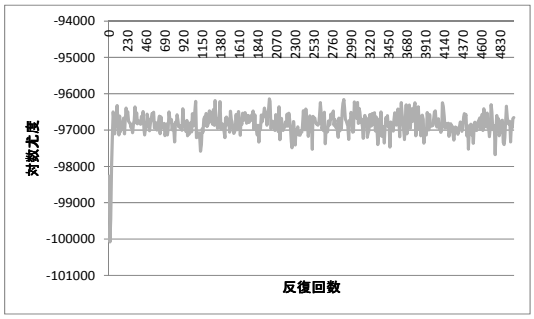

図 1:LDA のパラメータの収束

4 提案手法

上記の通り，推薦システムにおける LDA では，超スパースなデータに対して，学習データがモデルのパラメータ数と比べ相対的に少なくなり，学習が収束しない．このとき，LDA を用いた推薦システムでは各ユーザの購買確率をうまく推定できず，嗜好に適したアイテムを推薦できない．そこでユーザのクラスタリングによって類似ユーザの購買履歴データを統合し，LDA のパラメータ数を削減することを考える．しかし，単純にクラスタリングを適用しただけでは各クラスタサイズに偏りが生じてしまい，特にサイズの小さいクラスタに関して，購買確率を推定するために十分な学習データを確保できない恐れがある．これは一般的に EC サイトでは大量にアイテムを買う少数ユーザがいる一方，多く買わないユーザが大多数いる傾向を持ち，前者にはサイズの小さいクラスタが，後者にはサイズの大きいクラスタが形成されるためである．従って，各クラスタのサイズに偏りを生じさせないことが望まれる．そこで提案手法では階層的クラスタリングを行うことで LDA のパラメータ数の削減とともにクラスタサイズの均一化を図り，安定したパラメータ推定を行うことを目指す．

4.1 提案モデルの構造

提案手法では，ユーザクラスタの概念を導入することで，LDA のパラメータ数の削減を図る．ユーザクラスタはユーザの嗜好に従ってクラスタリングを行うことで生成され，各ユーザは 1 つのユーザクラスタに所属するものと仮定する．嗜好の似ているユーザを 1 つのユーザクラスタとしてまとめることにより，ユーザクラスタを 1 人の代表ユーザとしてみなすものとする．その上で提案モデルでは，ユーザクラスタの嗜好とアイテムの特徴を LDA により推定する．ユーザクラスタとアイテム間に潜在クラスを仮定することで，ユーザクラスタの嗜好の多様性を考慮できるようになる．いま，ユーザクラスタを $c \in \{1, \cdots, C\}$, ユーザクラスタ c が潜在クラス k に所属する確率を $\theta'_{c,k}$ とし，これをまとめて $\boldsymbol{\theta}' = [\theta'_{c,k}]_{C \times K}$ と表す．$\boldsymbol{\theta}'$ はディリクレパラメータ $\boldsymbol{\alpha}' = (\alpha'_k)$ から生成されるとする．ユーザクラスタ c がアイテム i を購買する確率 $P(i|c)$ を以下の式で表す．

$$P(i|c) = \sum_{k=1}^{K} \int\int \theta'_{c,k} P(\theta'_{c,k}|\alpha'_k) \phi_{k,i} P(\phi_{k,i}|\beta_i) d\theta'_{c,k} d\phi_{k,i}. \quad (2)$$

ユーザ u がアイテム i を購買する確率 $P(i|u)$ に関して，各ユーザが 1 つのユーザクラスタに所属するという仮定のため，$P(i|u)$ と $P(i|c)$ は同値である．

提案モデルの構造の例を図 2 に示す．四角はそれぞれユーザ集合，ユーザクラスタ集合，潜在クラス集合，アイテム集合を表し，ノードは各ユーザ，ユーザクラスタ，潜在クラス，アイテムを表す．ユーザクラスター潜在クラス間の全リンクに $\theta'_{c,k}$ の値が，潜在クラス―アイテム間の全リンクに $\phi_{k,i}$ の値が割り当てられる．各ユーザは 1 つのユーザクラスタに所属するため，各ユーザの購買確率は式 (2) で表される．

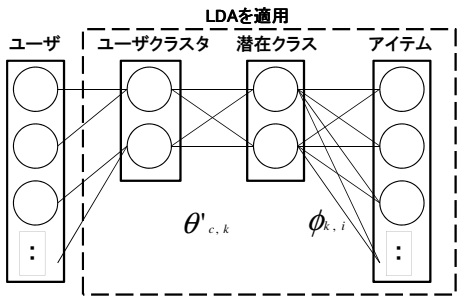

図 2:提案モデルの構造

4.2 クラスタリングの方法

提案モデルを超スパースなデータに適用するために，望ましいクラスタリング方法について以下で検討を行う．いま，ユーザクラスタ c に所属するユーザ数を $|U_c|$ とし，$|U_c|$ はクラスタサイズとする．また，ユーザクラスタに関する購買履歴データの行列を $\boldsymbol{R}' = [R'_{c,i}]_{C \times I}$ と定義する $(1 \leq c \leq C, 1 \leq i \leq I)$．$R'_{c,i}$ は，ユーザクラスタ c に所属する全ユーザがアイテム i を購買しなかった場合に 0 を，1 人でも購入があった場合に 1 をとる．ここで各ユーザクラスタ c の購買ベクトルを $\boldsymbol{R}'_c = (R'_{c,1}, R'_{c,2}, \cdots, R'_{c,I})$ と定義する．また，$\boldsymbol{R}' = (\boldsymbol{R}'_c)$ とする．各 $\boldsymbol{R}'_c$ の密度にばらつきが生じると $|U_c|$ が偏り，$\boldsymbol{R}'_c$ の密度の小さいユーザクラスタ c に関する購買履歴データが十分確保されず，パラメータの推定が収束しない恐れがある．そのため，超スパースなデータの下で LDA を用いて学習を行うためには，各ユーザクラスタ c における $\boldsymbol{R}'_c$ の密度が均一に

なっていることが望ましい．しかし，一般の EC サイトでは，非常に多くのアイテムを購買する少数のユーザが存在する一方，殆どアイテムを購買しないユーザが大多数である傾向を持つ．このようなデータに通常のクラスタリング手法を適用した場合，ごく少数の大量購買ユーザが小さいサイズのユーザクラスタを形成し，その他のユーザが大きいユーザクラスタサイズにまとめられるため，$|U_c|$ が偏ってしまう．その結果, サイズの小さいユーザクラスタにおけるデータ密度のスパースさはあまり改善されず，単にクラスタリングしただけでは LDA の学習が不可能となることがほとんどであると考えられる．そこでサイズが大きいユーザクラスタに対して順次階層的にクラスタリングを行い，$\boldsymbol{R}'_c$ の密度を均一に近づける．パラメータ数が学習データ数を上回るようにすることでパラメータの推定を可能にする．

4.3 クラスタリング・学習・予測アルゴリズム

提案手法のクラスタリング・学習・予測アルゴリズムを以下に示す．アルゴリズムは 8 つのステップから成り，Step1 から Step5 までがクラスタリングステップ，Step6 が学習ステップ，Step7, 8 が予測ステップである．

クラスタリングステップでは，4.2 節で述べた EC サイトにおけるユーザの傾向から，$\boldsymbol{R}'_c$ の密度を均一に近付けるために階層的にクラスタリングを行う．この際に k-means 法を用いる．いま，C_1 を単純にクラスタリングを行った際の分割個数，C_2 を階層的にクラスタリングを行った際の分割個数とする．ただし，C_1, C_2 はあらかじめ決めておく定数とする．Step1 では 1 回目のユーザクラスタ数を $k = C_1$ として k-means 法を適用する．次に条件を満たした場合のみ，Step3 で階層的クラスタリングを行い，C_2 個のユーザクラスタを追加生成する．以降学習ステップでは，Step6 で従来同様ギブスサンプリングにより $\boldsymbol{\theta}'$ と $\boldsymbol{\phi}$ の推定を行い，収束した値を用いて予測ステップ，Step7, 8 で $P(i|u)$ の値に従って推薦を行う．なお，Step5 における $\mathcal{U}_c$ はユーザクラスタ c に所属するユーザの集合を示す．

Step1) U 人のユーザを $k = C_1$ として k-means 法を適用する．$j = C_1$ とおく．

Step2) Step1 の結果から，各クラスタサイズに偏りが生じたら Step3 へ，そうでなければ Step5 に移る．

Step3) $|U_c|$ が最大のユーザクラスタ c に対してさらにクラスタリングを行う．$k = C_2$ として k-means 法を適用し，C_2 個のユーザクラスタに分割する．$j = j + C_2 - 1$ とする．

Step4) Step3 の結果から，各クラスタサイズに偏りがあれば Step3 に戻り，そうでなければ Step5 に進む．

Step5) ユーザクラスタの購買履歴 $\boldsymbol{R}'$ をクラスタリングの結果に従い構築する．各ユーザクラスタの購買履歴 $\boldsymbol{R}'_c$ の要素は以下の式で得られる．

$$\boldsymbol{R}'_c = \begin{cases} 1, & \sum_{u \in \mathcal{U}_c} R_{u,i} \geq 1, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Step6) $\boldsymbol{\theta}'$ と $\boldsymbol{\phi}$ をギブスサンプリングにより推定する．

Step7) ギブスサンプリングの結果から，$P(i|u)$ を式 (2) により算出する．

Step8) 購買履歴 $\boldsymbol{R}$ においてユーザ u が購買していないアイテムを式 (2) の値が大きい順にアイテムを推薦する．

□

5 実験

提案手法の有効性を検討するため，アパレル関係の EC サイトにおける 1 年間の購買データを用いて実験を行い，推薦精度の評価を行う．

5.1 実験条件

実験では，某アパレル関係の EC サイトにおける 1 年間の購買データ 14,059 件を用いた．ユーザ数は 835, アイテム数は 5,168 であり，80 %を学習データ，残り 20 %をテストデータとなるようランダムに分割した．データ密度は 0.25 %であり，超スパースなデータである．本実験では，パラメータの収束，推薦精度の評価，データ密度と推薦精度の関係について検討を行う．評価は各ユーザに購買確率が大きいアイテムを上位 N 件推薦し，そのうち実際に購買したアイテムの割合である Top-N 精度を用いる．比較手法として確率モデルの Naive Bayes, さらに潜在クラスを用いた Aspect Model, LDA 及び階層的にクラスタリングを行わない LDA を用いる．

5.2 実験 1：パラメータの収束

提案モデルにおけるパラメータ推定の結果を図 3, 比較として階層的にクラスタリングを行わなかった LDA のパラメータ推定の結果を図 4 に示した．$C_1 = 10$ とし，提案手法では $C_2 = 5$ を用いた．比較手法は提案手法で $C_1 = 10, C_2 = 0$ の場合に相当する．図 3 より，提案手法は対数尤度の変化率が確実に 0.1 %以下を満たしている状態 [9] を満たすため，パラメータ推定が行えていると考えられる．一方，図 4 より，階層的にクラスタリングを行わなかった場合は対数尤度は明らかに収束していないことから，階層的にクラスタリングを行う提案手法の有効性が示せた．

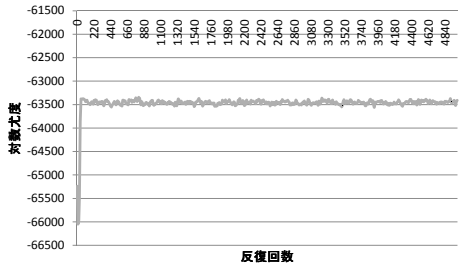

図 3:提案手法におけるパラメータ推定の結果

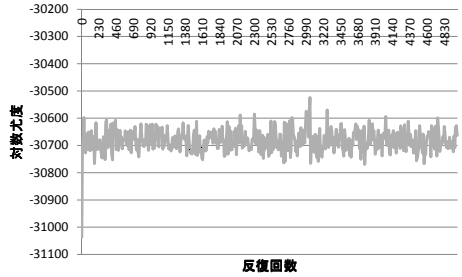

図 4:比較手法におけるパラメータ推定の結果

実験で用いたデータでは，835 人中わずか 4 人のユーザが 100 個以上のアイテムを購買している．このような少数の大量購買ユーザの存在により，1 回のクラスタリングでは少人数のユーザクラスタサイズが構成され $|U_c|$ に偏りが生じ，結果としてパラメータ推定が収束しないことにつながると考えられる．これに対し，サイズ $|U_c|$ が

最大のユーザクラスタに対して階層的にクラスタリングを行うことで，各 $\boldsymbol{R}'_c$ の密度を均一に近付け，超スパースなデータに対して提案モデルが望ましいクラスタリング結果を得ることができたと考えられる．

5.3 実験 2：Top-N 精度

Naive Bayes, Aspect Model, LDA 及び提案手法の N=1, 2, 3, 5, 10 における Top-N 精度を以下の図 5 に示す．図 5 における NB は Naive Bayes, AM は Aspect Model を指す．LDA の括弧内の値は (C_1, C_2) を示し，LDA(10,5) は提案手法を意味する．図 5 より全ての N に対し，提案手法の精度が勝っていることから，提案手法の有効性を示すことができた．

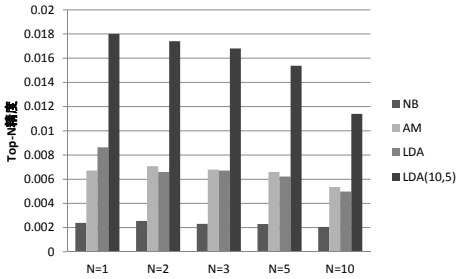

図 5:各モデルにおける Top-N 精度

Aspect Model が Naive Bayes より高い推薦精度を保っているのは，潜在クラスを用いることによってユーザの嗜好やアイテムの特徴の多様性をより考慮できたためと考えられる．また，提案手法が Naive Bayes, Aspect Model, LDA に比べて推薦精度が向上したのは，超スパースなデータに対してパラメータの推定が可能になったことで各ユーザへの適切なアイテムの推薦ができためと考えられる．ユーザクラスタはユーザの嗜好を基準に形成されるため，ユーザクラスタの嗜好を推定できたことで各ユーザの嗜好に合ったアイテムを推薦できるようになった．この手法により，LDA が超スパースなデータに対して学習可能となり，適切な推薦を実現し，精度が向上したと考えられる．

次にクラスタリングの違いによる性能の比較のため，階層的クラスタリングの結果を LDA に学習させる提案手法と，一括でクラスタリングした結果を LDA に用いる比較手法の Top-N 精度についての比較を図 6 に示す．図 5 と同様に，LDA の括弧内の値は (C_1, C_2) を示す．全ての N に対し，階層的にクラスタリングを行った場合の精度が勝っていることから，提案手法の有効性が示された．

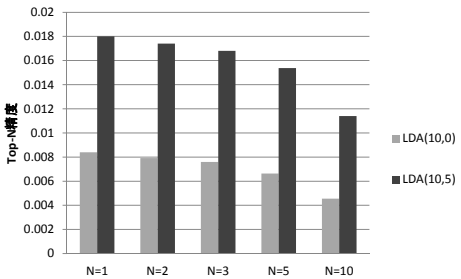

図 6:クラスタリングの違いによる Top-N 精度

実験 1 の結果から，非階層のクラスタリング $C_2 = 0$ ではパラメータが収束しないことから，各ユーザに適切なアイテムを推薦できていないため，精度が悪化したと考えられる．

5.4 実験 3：データ密度と Top-N 精度の関係

階層的にクラスタリングを行う際，C_2 の値によってユーザクラスタの購買履歴 $\boldsymbol{R}'$ の密度は変化していく．$\boldsymbol{R}'$ の密度の変化が Top-N 精度にどのように影響を及ぼしているのか調べた．図 7 に $N = 10$ の場合の結果を示す．

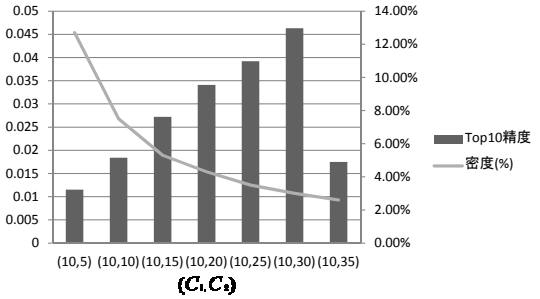

図 7:データ密度と Top-10 精度の関係

まず提案モデルの収束の可否について，$C_2 = 5$〜30 では収束したが，$C_2 = 35$ では再度収束しなくなった．C_2 の値が増加するとモデルのパラメータ数が増加し，学習データ数は減少するため，$\boldsymbol{R}'$ の密度が低下する．このため，$C_2 = 35$ の時点でパラメータ推定に十分な学習データが不足してしまったと考えられる．

推薦精度について，$C_2 = 5$〜30 の間は C_2 の値を増加させると精度は良くなるが，$C_2 = 35$ では悪化する．C_2 の増加に伴い，適切にユーザの嗜好を捉えたユーザクラスタが形成されたが，パラメータ推定がうまく行えないと適切な推薦が行われなくなったと考えられる．以上の議論から，ユーザクラスタの購買履歴 $\boldsymbol{R}'$ の密度に留意し，学習データ数がモデルのパラメータ数を下回らないよう C_2 を決めていく必要があるといえる．

6 まとめと今後の課題

本研究では，現実の推薦システムを想定した超スパースなデータに対して学習できない LDA に対して改善を行い，提案手法の有効性を示した．

今後の課題として，複数のユーザクラスタへの所属を許容し，拡張したモデルに適したパラメータ推定のアルゴリズムを構築することである．

参考文献

[1] Hofmann T., “Probabilistic Latent Semantic Analysis,” *UAI,* pp.289-296, 1999.

[2] Blei M.D., Ng Y.A., Jordan I.M., “Latent Dirichlet Allocation,” *Journal of Machine Learning Research, vol3,*33, pp.993-1022 2003.

[3] Resnick P., Iacovou N., Suchak M., Bergstrom P., Riedl J., “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” *in Proc. of The Conf. on Computer Supported Cooperative Work,* pp.175-186, 1994.

[4] 岩田具治, 渡辺晋司, 山田武士, 上田修功, “購買行動解析のためのトピック追跡モデル,” 電子情報通信学会 D, vol.J.93-D, no.6, pp.978-987, 2010.

[5] 貞光九月, 三品拓也, 山本幹雄, “混合ディリクレ分布を用いたトピックに基づく言語モデル,” 電子情報通信学会 D, vol.J,88-D, pp.1771-1779, 2005.

[6] 上田修功, 斉藤和巳, “多重トピックテキストの確率モデル : テキストモデル研究の最前線 (2),” 情報処理学会, 45(3), pp.282-289, Mar.2004.

[7] Griffiths L.T. Steyvers M., “Finding scientific topics,” *Proc. National Academy of Sciences*, vol.101, Suppl.1 pp.5228-5235, 2004.

[8] 神嶌敏弘, “推薦システムのアルゴリズム (2),” 人工知能学会誌, vol.23, no.1, pp.89-103, 2008.

[9] 横峯 樹, 江口 浩二, “混合メンバーシップ・ブロックモデルを用いた協調フィルタリング,” 情報処理学会研究報告, No.2010-FI-98, 2010.