神嶌 敏弘, 赤穂 昭太郎 "参加システムの嗜好パターンが異なる場合の集団協調フィルタリング",
人工知能学会研究会資料, SIG-FPAI-A702-03 (2007)

参加システムの嗜好パターンが異なる場合の集団協調フィルタリング

Collective Collaborative Filtering Whose Participants Are Mixed-Minded

神嶌 敏弘 [1]，赤穂 昭太郎
Toshihiro Kamishima and Shotaro Akaho

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

Abstract: Collaborative filtering (CF) is a technique to recommend items based on preference pattern among the like-minded users. Therefore, to perform collaborative filtering, the sufficient size of users' preference data is required, but collecting sufficient users is often difficult. To overcome this problem, preference data are collected from multiple sites while keeping users' privacy, and CF is performed based on this collected data. Such a framework is often referred as privacy-preserving CF. However, such type of CF suffer from another issue. Because preference pattern is induced from collected data, specific pattern of each participant site is weaken. In this report, we discuss this problem.

1 はじめに

本論文では，複数の参加サイトからのデータを集めて推薦をする集団協調フィルタリングを扱う．ここで，各ローカルシステムごとの特徴を考慮した推薦用モデルを獲得する方法について考察する．

推薦システムは利用者が好むであろうアイテムを予測し，それを利用者に提示することで，情報過多の問題に対処するためのシステムである．協調フィルタリング [10] は，この推薦システムを実現するための枠組みの一つである．これは,「口コミ」の過程を自動化したもので，過去の嗜好パターンが類似している利用者群を見つけ出し，彼らが好むものを推薦するものである．協調フィルタリングは，1994 年の GroupLens の方法 [12] 以来多様な手法が開発され，実際にいろいろな場面で利用されている．

しかし，運用上の問題はいくつも存在する．その中で，推薦システムの利用者数の少なさに起因する問題に注目する．大規模なネットショッピングモールでは，十分な数の利用者がシステムを利用している．だが，中小規模のサイトでは，利用する顧客数も少ないので，必然的にシステムの利用者数も少なくなる．このような状況では次の二つの問題を生じる．一つは，だれにも評価されてない，すなわち，好きかどうかの情報を与えられていないアイテムが生じることが頻繁に生じる．これは，各利用者はアイテム集合の一部しか評価せず，また，評価されるアイテム群は一部のものに集中する傾向があるためである．他の利用者の意見を参考にして推薦アイテムを決める協調フィルタリングでは，こうした誰にも評価されていないアイテムは推薦の対象には決してなることはなく，被覆率が低下する．もう一つの問題は，現在利用している利用者（活動利用者）と類似した嗜好をもつデータベース中の利用者（標本利用者）が見つからない問題である．協調フィルタリングは，いわゆるロングテールと呼ばれる少数派の嗜好パターンの利用者を結びつけ，それらのニッチな要求にも対応できると言われることがある．だが，実際には，それが実現されるには，ニッチな嗜好を持つある程度の利用者がシステムを利用していることが前提となる．利用者数が少ないシステムでは，こうした少数派のパターンは，それと類似した嗜好をもつ利用者がシステム内にいないため，ノイズのように扱われてしまい，その嗜好が適切に推薦に反映されることはない．

こうした問題を解決するには，小規模なシステムの利用者のデータを集積して，十分な数のデータを集めて予測を行えばよい．だが，ここで別の問題が生じる．利用者の嗜好データは，一般に，秘匿すべき個人情報である．よって，これらをそのまま集積して計算することは認められない．こうした状況に対応するため，プライバシー協調フィルタリングが提案されている．この枠組みでは，利用者集団の嗜好のパターンからは個人それぞれの嗜好パターンは復元できないため個人情報ではないとの前提に立つ．そして，個人情報である，各利用者の嗜好データを秘匿したまま，全体の嗜好パターンを表すモデルを獲得できれば，個人情報の漏洩はないと考えられる．Canny は，こうしたモデルを，個人の嗜好データを暗号化したままで各種の演算を行う，安全な計算 (secure computation) という技術を用いて獲

[1]連絡先：神嶌 敏弘 http://www.kamishima.net/

得する方法を示した [4]．だが，安全な計算の計算量は大きく大規模・高速化には困難が伴う．そこで，暗号化による厳密なプライバシー保護ではなく，もう少し緩い状況を想定する．個人から送信された嗜好データは，直接集めるのではなく，小規模サイトで運用されるサイトに一度蓄積されるものとする．各サイト内で，これらの個人情報は安全に管理されると仮定する．そして，各サイトごとに，ローカルな嗜好パターンのモデルを計算し，それらを集積して推薦を行うことを考える．ローカルなモデルからは，すでに個人情報は回復できないので，もはや個人情報ではないものとここでは考える．このような枠組みでは安全な計算は不要になるので，計算資源の制限はなくなる．こうして集めたローカルモデルを要約して大域的なモデルを生成すれば，多数の利用者の意見を反映した推薦ができる．ここでは各サイトが集まってフィルタリングを実行することに重点があるので，この枠組みを集団協調フィルタリングと呼ぶ．

本論文では，Hofmann の，probabilistic latent semantic analysis (pLSA) [8] モデル（aspect モデルとも呼ばれる）を利用した協調フィルタリング [9] を対象にこの集団協調フィルタリングを考える．この pLSA モデルの計算には，潜在変数がある場合の最適化問題をとく EM アルゴリズムが用いられる．この EM アルゴリズムは，各反復ごとに同期は必要ではあるが，一般に分散環境で並列に計算できることが知られていた [13, 7]．pLSA を用いた協調フィルタリングを，EM アルゴリズムを並列実行させて解く試みとして Das らの研究 [5] がある．

しかし，この枠組みでもさらなる問題がある．中小サイトはそれぞれ特徴あるアイテムを扱い，また，利用者集団にも固有の特徴があるだろう．そうした集団に，集積した大域的なモデルを適用しても，十分にその特殊性を反映したモデル構築はできない．本論文では，こうした特殊性をもつモデルの獲得を行うサイト適応型集団協調フィルタリングについて考察する．

2 節では，pLSA を用いた基本的な CF とその分散環境での実行について，3 節では，各参加サイトの利用者に適合させたモデルを分散環境で獲得する手法について論じる．4 節はまとめである．

2 pLSA による協調フィルタリングとその分散環境での実行

2.1 pLSA による協調フィルタリング

pLSA による協調フィルタリング [9] の，数学的定義と問題設定から始める．

利用者とアイテムをそれぞれ確率変数 x と y で表す．x は $\mathcal{X} = \{1, \ldots, i, \ldots, n\}$ 中の値を，y は $\mathcal{Y} = \{1, \ldots, j, \ldots, m\}$ 中の値をとる多値の確率変数である．ここで，このモデルで重要な役割をはたす潜在変数 z を導入する．これも多値変数で $\mathcal{Z} = \{1, \ldots, l\}$ 中の値をとり，潜在的な嗜好のパターンを表す．

ここでは，評価値を使わない，変数 x と y の共起関係だけを使うモデルを示す．嗜好データを，利用者 i がアイテム j を購入した場合を考える．これは，$x = i$ という事象と，$y = j$ という事象が共起していることに相当する．このモデルでは購入しないという行為が無関係なので，未評価と不支持が区別できない暗黙的評価での問題が生じない利点がある．利用者 i がアイテム j をどれくらい好きかを，利用者が i であったとき，アイテム j を好む確率 $\Pr[y = j | x = i]$ の大きさで測る．この条件付確率を潜在変数を導入して次のように表すのが pLSA モデルである．

$$\Pr[x, y] = \sum_{z \in Z} \Pr[x|z] \Pr[y|z] \Pr[z] \tag{1}$$

このモデルでは z が与えられたときに x と y が条件付独立であることを仮定して，モデルのパラメータの総数を減らしている．モデルのパラメータは $\theta = (\{\Pr[z|x]\}, \{\Pr[y|z]\}, \{\Pr[z]\})$ であるが，これらは最尤推定で求める．すなわち，N 個の共起データ $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^{N}$ に対する次の対数尤度を最大にするように求める．

$$\mathcal{L}(\mathcal{D}; \theta) = \textstyle\sum_{(i,j) \in \mathcal{D}} \ln \Pr[x = i, y = j; \theta]$$

潜在変数 z があるため，この最尤推定は EM アルゴリズム [6, 3] によって行う．EM アルゴリズムは十分統計量の計算である E ステップと，各パラメータの推定である M ステップを交互に実行する．具体的に E ステップは次の確率を計算する．

$$\Pr[z|x, y] = \frac{\Pr[z] \Pr[x|z] \Pr[y|z]}{\sum_{z'} \Pr[z'] \Pr[x|z'] \Pr[y|z']} \tag{2}$$

一方，M ステップでは次式を計算する．

$$\Pr[x|z] = \frac{\sum_y n(x,y)\Pr[z|x,y]}{\sum_{x',y} n(x',y)\Pr[z|x',y]} \quad (3)$$

$$\Pr[y|z] = \frac{\sum_x n(x,y)\Pr[z|x,y]}{\sum_{x,y'} n(x,y')\Pr[z|x,y']} \quad (4)$$

$$\Pr[z] = \frac{\sum_{x,y} n(x,y)\Pr[z|x,y]}{\sum_{x,y,z'} n(x,y)\Pr[z'|x,y]} \quad (5)$$

ただし，$n(x,y)$ は，$x=i$ かつ $y=j$ となる $\mathcal{D}$ 中の対の数である．以上の手続きを反復し，収束した $\Pr[x|z]$, $\Pr[y|z]$，および $\Pr[z]$ が計算できれば，$\Pr[y|x]$ は次式で計算できる．

$$\Pr[y|x] = \frac{\sum_z \Pr[z]\Pr[x|z]\Pr[y|z]}{\sum_{y',z} \Pr[x|z]\Pr[y'|z]\Pr[z]} \quad (6)$$

利用者 i に対しては，次のアイテム y^* を推薦すればよい．

$$y^* = \arg\max_{y\in\mathcal{Y}} \Pr[y|x=i] \quad (7)$$

2.2 pLSA 計算の並列化

前節の EM アルゴリズムの計算を分散環境で行うことを考えよう [5]．簡単のため，利用者のデータは二つのサイトで分散されて保持されているものとする．ここで，$n = n_1 + n_2$ とし，サイト 1 では利用者 $\mathcal{X}_1 = \{1,\ldots,n_1\}$ のに対するデータ $\mathcal{D}_1$ を，サイト 2 では残りの利用者 $\mathcal{X}_2 = \{n_1+1,\ldots,n\}$ に対するデータ $\mathcal{D}_2$ が保持されているとしても一般性を失わない．パラメータ $\Pr[z]$ と $\Pr[y|z]$ は個人 x とは関係ない，すなわち個人情報を含まないのでサイト 1 と 2 で共有する．一方，$\Pr[x|z]$ は個人情報なので，サイト 1 では $x \in \mathcal{X}_1$ のみ，サイト 2 では $x \in \mathcal{X}_2$ のもののみを保持しているとする．

この状況で，サイト 1 では，$x \in \mathcal{X}_1$ なる x については，式 (2) を計算するのに必要な量は全て知っていることになる．サイト 2 でも同様に E ステップを実行できる．次に M ステップについて考える．まず，式 (3) の分子は，$x \in \mathcal{X}_1$ なる x については，E ステップで計算した $\Pr[z|x,y]$ と，自身が保持している $\mathcal{D}_1$ から求めた $n(x,y)$ を使えば計算できる．しかし，分母は $x' \in \mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ についての総和であるから，$x' \in \mathcal{X}_2$ についての，$n(x',y)$ や $\Pr[z|x',y]$ の値が必要になる．ここで，簡単な関係

$$\sum_{x'\in\mathcal{X},y} n(x',y)\Pr[z|x',y] = \sum_{x'\in\mathcal{X}_1,y} n(x',y)\Pr[z|x',y] + \sum_{x'\in\mathcal{X}_2,y} n(x',y)\Pr[z|x',y] \quad (8)$$

表 1: テスト集合の利用者上の $\Pr[y|x]$ の総和

テスト集合	平均確率	人数
全体	**0.0695**	189
20 歳未満	0.0664 ×	77
20 歳代	0.0747 ○	332
30 歳代	0.0706 ○	240
40 歳代	0.0593 ×	168
50 歳以上	0.0610 ×	125
60 歳以上	0.0559 ×	31

から，サイト 2 から右辺第 2 項の値が送信されていれば計算可能である．この値は，サイト 2 の利用者の嗜好パターンを表してはいるが，個々の利用者の情報はほとんど復元できないため，個人情報ではないとみなす．なお，これが個人情報にあたるとする場合でも，この和を求める計算のみを安全な計算によって行うことで，これらの情報を秘匿することも可能である．一方，サイト 2 では，サイト 1 から式 (8) の第 1 項を受信すれば，やはり式 (3) は計算できる．同様のことが式 (4) と (5) についてもいえる．よって，各反復で M ステップの実行前に式 (8) の右辺の項のような値を幾つかサイト間で同期すれば，残りの計算は全てサイト内で実行できる．以上のような手順で，各サイトで分散保持されているデータから pLSA を用いた推薦が実行できる．

3 サイト適応型集団協調フィルタリング

3.1 従来の並列化の問題点

1 節では，中小のサイトが集団的に運用する協調フィルタリングについて論じた．こうした環境では (1) 各サイトごとに利用者の層や，扱われているアイテムに散らばりがあり，また (2) データは広域のネットワークに分散して保持されているといった問題がある．

まず，(1) の利用者層の散らばりの問題を示すために，協調フィルタリングの代表的なベンチマークである MovieLens データ [11] を対象に予備実験を行った．このデータでは，943 人の利用者が，1682 種の映画について採点法で 5 段階の 10 万個の評価値を与えている．評価値が 5 か 4 なら，映画について好意的な評価があるとする．すなわち，利用者 i が，映画 j を 5 と評価したなら $(x=i, y=j)$ のデータとなるが，評価が 3 以下なら無視する．このデータを，同じ利用者ごとにまとめ，利用者を訓練用とテスト用に分割する．訓練用利

用者のデータは全て学習に利用するが，テスト用の利用者のデータは半分だけを学習用に利用する．この学習データに 2.1 節の方法を適用し，モデルのパラメータを獲得する．予測精度を評価するために，テスト用の各利用者 i の残りのデータに含まれる映画 j について $\Pr[y=j|x=i]$ を求め，それらの各利用者ごとの総和を計算し，全テスト利用者について平均をとる．

$$\frac{1}{|\mathcal{X}_t|}\sum_{i\in\mathcal{X}_t}\sum_{(i,j)\in\mathcal{D}_t}\Pr[y=j|x=i] \tag{9}$$

ただし，$\mathcal{X}_t$ はテスト用利用者で，$\mathcal{D}_t$ はテスト用データである．潜在的に肯定的に良い評価される映画のうち実際に評価されたものの割合 α が，全利用者について一定と仮定する．そして，50%をの評価を学習用に用いたことから，完全な予測ができれば，この評価スコアは最大値 0.5α になる．

実験結果を表 1 に示す．若干の平滑化などの改良があるが詳細は省略する．「テスト集合」の列にはテスト用利用者を選択した基準を示した．全体とは，利用者の 20%をランダムにサンプリングしてテスト用利用者とした場合の結果である．これは，データ全体の平均的な利用者を反映したベースラインである．それ以外では，年齢によってテスト用利用者を選択した．なお，60 歳以上のグループは 50 歳以上の部分集合である．テスト用ではない利用者は全て訓練用とした．それぞれの条件に該当する利用者の人数は「人数」の列に示した．「平均確率」は，各実験での，式 (9) のスコアであり，ベースラインである「全体」の結果を上回るものに○を，そうでないものに×を付けてある．この結果から，人数が多くこのデータの中核となっている 20〜30 代の利用者では予測精度は良いが，少数派の集団では悪くなる．さらに，人数の少ないほどよりこの傾向が強くなることが分かる．すなわち，たとえ複数のサイトから大量にデータを集積できたとしても，大域的に均一なモデルを用いたのでは，少数派のサイトの利用者には良い推薦ができないことを示唆している．以上のことから，複数サイトによる集団協調フィルタリングでは，各サイトに適応したモデルを，集積したデータから獲得する必要がある．

次に (2) のデータが広域分散している問題について論じる．データが広域分散していると，各サイト間の通信帯域は狭くなり，大量のデータを頻繁に送信することは難しくなる．しかし，2.2 節の方法では，$\sum_{x\in\mathcal{X}_k} n(x,y)\Pr[z|x,y]$ を共有するために各反復ごとに同期をとる必要が生じてしまう．データ量は比較的少ないが，データの同期が終了するまで多くのサイトでは待ち時間が生じ，効率的な計算はできない．そこで，各サイトで局所的に 2.1 節の pLSA を実行し，求めたパラメータを一つの中心サイトに集めて，集団協調フィルタリングを実行する枠組みを想定する．pLSA のパラメータには，$\Pr[x|z]$，$\Pr[y|z]$，および $\Pr[z]$ がある．このうち最初の $\Pr[x|z]$ は利用者個人に依存した個人情報なので，中心サイトには送信できない．よって，$\Pr[y|z]$ と $\Pr[z]$ のみを中心サイトに送信し，改良したパラメータ $\hat{\Pr}[y|z]$ と $\hat{\Pr}[z]$ を返してもらうことで，集団協調フィルタリングを実現する枠組みを考える．

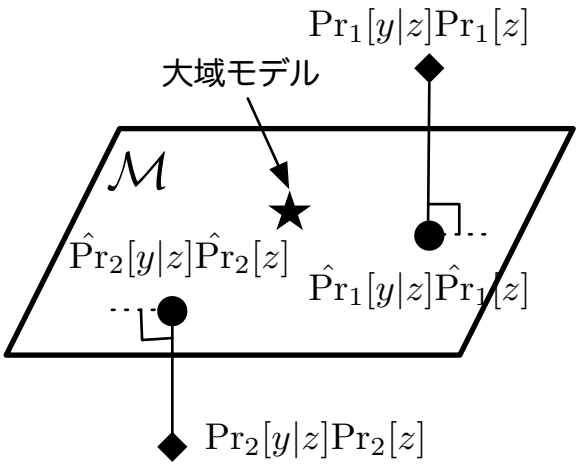

図 1: 分布パラメータの次元縮約

パラメータ $\Pr[y|z]$ と $\Pr[z]$ の高次元空間．この空間中の 1 点は同時分布 $\Pr[y,z]=\Pr[y|z]\Pr[z]$ に相当する．

まとめると，個人情報ではない局所モデルのパラメータのみを中心サイトに送信し，かつ，中心サイトでは各サイトに適応させたモデルを集めたパラメータ群から獲得し，各サイトに返すような枠組みが必要になる．これをサイト適応型集団協調フィルタリング (site adaptive collective CF; SACCF) と呼ぶ．本稿の残りで，この SACCF を実現するアイデアを 2 種類示す．なお，簡潔さのため，サイトが二つの場合で説明する．

3.2 分布パラメータの次元縮約による方法

一つ目の方法は，赤穂による分布のパラメータ空間での次元縮約 [2] を用いるものである．

この方法では各サイトからパラメータ $\Pr_k[y|z]$ と $\Pr_k[y|z]$ が中心サイトに集められる．ただし，$k=\{1,2\}$．アイテム数を m，潜在変数のとりうる値の数 l として，パラメータ $\Pr_k[y|z]$ は ml 個，$\Pr_k[y|z]$ は l 個存在する．そして図 1 のような ml^2 次元のパラメータの空間を考えると，同時分布 $\Pr_k[y,z]=\Pr_k[y|z]\Pr_k[z]$

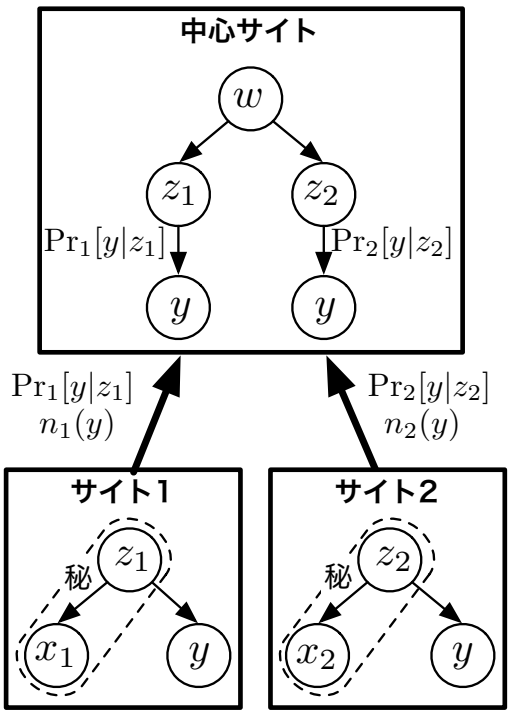

図 2: 大域潜在変数を導入したモデル

のモデルはこの空間中の 1 点で表せるようになる．情報幾何 [1] では，この空間での適切な距離や射影の概念が導入されている．

各サイトの特徴を無視して，各サイトから集めたモデル（図中の◆）のいわば平均にあたる点を求めると大域モデル（図中の★）が得られる．だが，これでは 3.1 節で述べたように，各サイトの特徴は無視され，少数派の利用者で構成されるサイトでは良い推薦を受けられない．そこで，元の ml^2 次元より低次元のある部分空間 $\mathcal{M}$ を導入する．回帰分析における回帰直線のように，各サイトのモデルと，この $\mathcal{M}$ の間の損失の総和が最小になるように，この $\mathcal{M}$ を選ぶ．そして，この部分空間へ，各サイトのモデル $\Pr_k[y|z]\Pr_k[z]$ を射影した $\hat{\Pr}_k[y|z]\hat{\Pr}_k[z]$（図中の●）を考える．すると，このモデル $\hat{\Pr}_k[y|z]\hat{\Pr}_k[z]$ は，全体に共通する性質を表したモデルの部分空間である $\mathcal{M}$ 上で，最もサイトのモデル $\Pr_k[y|z]\Pr_k[z]$ に適応したものといえる．あとは，このモデルの $\hat{\Pr}_k[y|z]$ と $\hat{\Pr}_k[z]$ を元のサイトに返し，そのサイトで $\Pr_k[y|z]$ と $\Pr_k[z]$ の代わりに利用して，推薦を行う．

3.3 大域潜在変数を導入する方法

次に，図 2 のように，各サイトでの潜在変数 z_1 と z_2 が大域的な潜在変数 w に依存するモデルを提案する．このモデルでも潜在変数があるので，EM アルゴリズムによってパラメータを推定する．

各サイト k からはデータ $\mathcal{D}_k$ に含まれるアイテムの個数 $n_1(y)$ と，分布パラメータ $\Pr_1[y|z_k]$ を中心サイトに集める．これらを使って中心サイトのモデルのパラメータを求よう．中心サイトでサイト 1 からのデータが生じる確率は次式で表せる．

$$\Pr[y] = \sum_{z_1,z_2,w} \Pr_1[y|z_1]\Pr[z_1|w]\Pr[z_2|w]\Pr[w] \quad (10)$$

ここで，添え字のある Pr で表した，$\Pr_1[y|z_1]$ はパラメータではなく，サイト 1 から送られてきたものであり，定数である．するとサイト 1 でのデータの中心サイトのモデルでの対数尤度は次式になる．

$$\log \mathcal{L}_1 = \sum_y n_1(y) \times \log\left[\sum_{z_1,z_2,w} \Pr_1[y|z_1]\Pr[z_1|w]\Pr[z_2|w]\Pr[w]\right] \quad (11)$$

ここで，仮に z_1，z_2，および w の値が既知であるとすると対数尤度は次式になる．

$$\log \mathcal{L}_1 = \sum_y \sum_{z_1,z_2,w} n_1(y)\Pr'[z_1,z_2,w|y] \times \log\Big[\Pr_1[y|z_1]\Pr[z_1|w]\Pr[z_2|w]\Pr[w]\Big] \quad (12)$$

ただし，$\Pr'$ は直前の E ステップで求めた値である．この対数尤度はサイト 2 についても同様であり，EM アルゴリズムの Q 関数は $Q = \log \mathcal{L}_1 + \log \mathcal{L}_2$ となる．M ステップではこの Q 関数を最大化するようなパラメータ $\Pr[z_1|w]$，$\Pr[z_2|w]$，および $\Pr[w]$ を計算すればよい．次にサイト 1 についての E ステップについて述べる．ここでは，前の M ステップで求めたパラメータ $\Pr[z_1|w]$，$\Pr[z_2|w]$，$\Pr[w]$ と，サイトからの $\Pr_1[y|z_1]$ を用いて $\Pr'[z_1,z_2,w|y]$ を次式で求める．

$$\Pr'[z_1,z_2,w|y] = \frac{\Pr[y,z_1,z_2,w]}{\sum_{z_1,z_2,w}\Pr[y,z_1,z_2,w]}$$

$$\Pr[y,z_1,z_2,w] = \Pr_1[y|z_1]\Pr[z_1|w]\Pr[z_2|w]\Pr[w]$$

こうして，中心サイトでのモデルが求まれば，各サイトの潜在変数の事前分布を次式で計算する．

$$\Pr^{new}[z_1] = \sum_{z_2,w} \Pr[z_1]\Pr[z_2|w]\Pr[w] \quad (13)$$

こうして求めた $\Pr^{new}[z_1]$ を送り返せば，サイト 1 では，自身がもつ $\Pr_1[y|z_1]$ と $\Pr_1[x|z_1]$ によって推薦ができる．

この方法には，次回のパラメータ更新時に，$\Pr[z_1]$ を中心サイトのモデル $\Pr^{new}[z_1]$ に固定して $\Pr_1[y|z_1]$ と $\Pr_1[x|z_1]$ を更新し，更新した $\Pr_1[y|z_1]$ を中心サイトに送るという改良も考えられる．その他，サイト 1 と 2 で評価されるアイテムをそれぞれ y_1 と y_2 として，中心サイトの補助のもと $\Pr[y_1 = j, y_2 = j|x = i]$ を最大にするアイテム j を利用者 i に推薦する方法も考えられる．これは，サイトとは無関係に好きなアイテムは決まるという考えに基づくものである．

4 まとめ

本稿では，各サイトが，サイトの利用者の個人情報以外の情報を集積し，そのデータに基づいて，各サイトで利用する推薦用のモデルを学習する集団協調フィルタリングの問題を扱った．そして，各サイトの特徴を無視した大域的なモデルでは，推薦の予測精度が悪化するサイトがある場合があることを予備実験により示し，各サイトに適応させたモデルが必要であることを示した．また，広域的に分散している環境下で集団協調フィルタリングを実行するため，各サイトで局所的に計算したパラメータのみを中心サイトに送る枠組みを提案した．これら二つの特徴をもつサイト適応型集団協調フィルタリングのための手法として，パラメータ空間での次元縮約を導入する方法と大域的な潜在変数を導入する方法について考察した．今後はこれら二つの手法を実装して，データに適用し，有効性を検証する予定である．

参考文献

[1] 赤穂昭太郎．情報幾何と機械学習．計測と制御, Vol. 44, No. 5, pp. 299–306, 2005.

[2] 赤穂昭太郎. 情報幾何に基づく混合分布パラメータの次元縮小法. 電子情報通信学会技術研究報告, NC 2005–115, pp. 57–62, 2006.

[3] C. M. Bishop. パターン認識と機械学習 — ベイズ理論による統計的予測 (上下). シュプリンガー・ジャパン, 2007–2008. （元田 浩 他 訳）.

[4] J. Canny. Collaborative filtering with privacy. In *Proc. of the 2002 IEEE Symposium on Security and Privacy*, pp. 45–57, 2002.

[5] A. Das, M. Datar, A. Garg, and S. Rajaram. Google News personalization: Scalable online collaborative filtering. In *Proc. of The 16th Int'l Conf. on World Wide Web*, pp. 271–280, 2007.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society (B)*, Vol. 39, No. 1, pp. 1–38, 1977.

[7] G. Forman and B. Zhang. Distributed data clustering can be efficient and exact. *SIGKDD Explorations*, Vol. 2, No. 2, 2000.

[8] T. Hofmann. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence 15*, pp. 289–296, 1999.

[9] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proc. of the 16th Int'l Joint Conf. on Artificial Intelligence*, pp. 688–693, 1999.

[10] 神嶌敏弘. 推薦システムのアルゴリズム (1)〜(3). 人工知能学会誌, Vol. 22, No. 6 〜 Vol. 23, No. 2, 2007–2008.

[11] MovieLens data. http://www.grouplens.org/node/12#attachments.

[12] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of Netnews. In *Proc. of The Conf. on Computer Supported Cooperative Work*, pp. 175–186, 1994.

[13] B. Zhang, M. Hsu, and G. Forman. Accurate recasting of parameter estimation algorithms using sufficient statistics for efficient parallel speed-up. In *Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge Discovery*, pp. 243–254, 2000.