情報処理学会論文誌 Vol. 51 No. 9 1554–1562 (Sep. 2010)

アイテム間類似度に基づく プライバシ保護協調フィルタリングの提案

多田 美奈子[†1] 菊池 浩明[†2]

嗜好に関する巨大なデータセットから自分に必要な情報を効率的に取り出す手段の1つとして，情報推薦システムがある．これまでにも情報推薦システムにおいて利用者のプライバシ情報をサービス運営者に漏らさずに推薦を行う方法がいくつか提案されているが，利用者数に応じて大きな計算量がかかることが課題になっていた．そこで，本論文では，対象となるアイテム間の類似度を利用して評価値を予測することで，コストを削減できることを示す．提案するアイテムベース方式のプライバシ協調フィルタリングを従来のユーザベース方式と精度や性能の観点で比較し，その有効性を示す．

Proposal on a Privacy-Preserving Collaborative Filtering Protocol Based on Similarity between Items

Minako Tada[†1] and Hiroaki Kikuchi[†2]

A recommendation system enables us to effectively extract knowledge from huge dataset about personal preference. Some cryptographical protocols for computing privacy-preserving recommendation without leaking the privacy of users have been proposed. However, the large overhead depending on the number of users is one of the current issues. In this paper, we propose an efficient scheme based on item-item similarities for providing a prediction of arbitrary values of rating. We show the performance and the accuracy of our proposed scheme in comparison with the existing schemes based on a numerical experiment.

†1 東海大学連合大学院理工学研究科総合理工学専攻
Course of Science and Technology, Graduate School of Science and Technology, Tokai University
†2 東海大学情報通信学部通信ネットワーク工学科
Department of Communication and Network Engineering, School of Information and Telecommunication Engineering, Tokai University

1. はじめに

情報推薦システムは，嗜好に関する巨大なデータセットから自分に必要な情報を効率的に取り出す手段の1つである．たとえば，Amazon.com[1)]では，購入履歴や商品評価，商品ページの閲覧履歴などからユーザの嗜好を分析し，商品の推薦を行っている．しかし，既存の推薦サービスは，利用者の購入履歴，評価などのプライバシ情報をサービス提供者が持っているため，管理が不十分であったり，悪意のある管理者によってプライバシ情報が流出したりする危険性がある．

そこで，推薦システムにおける手法の1つである協調フィルタリング（以下CF方式）を，利用者のプライバシ情報を秘匿したまま行う試みがされている[2)–7)]．CF方式ではユーザ間の嗜好の類似性に基づいて評価値を予測するので，プライバシ情報へのアクセスが必要であった．そこでCannyは，固有ベクトルで次元の削減されたコミュニティ間の類似度に変換して，推薦に用いる手法を提案している[2)]．この方式では，各ユーザが準同型性暗号を用いて評価値を暗号化し，秘匿したまま共役勾配法を実行して固有値を求め，特異値に分解された評価行列を計算する．ただし，固有値の計算のためには，全ユーザを同期させて行列値を逐次的に更新することが必要であり，大規模なユーザ間で実現するのは現実的ではない．木澤らは，ユーザ同士の類似度をもとに予測評価を行う方式を提案している[3)–5)]．この中で木澤らは，全ユーザ間の類似度の暗号文を使うとコストが高いため，ユーザ集合をサンプリングしたり，アイテム間をクラスタリングしたりして，計算量を減らしている．また，Ahmadらは，変形ElGamal暗号によるCannyの方法の改良について提案している[6)]．そのほかに，Katzenbeisserらは，健康管理システムの応用として，準同型性暗号による個人プロファイルマッチングシステムを提案している[7)]．しかしながら，これらのユーザ間の類似度を用いる方式では，利用者の数に応じて大きな計算量がかかるという問題が残っていた．

そこで我々は，大きなコストとプライバシの問題が生じるユーザ間類似度に代わり，アイテム間類似度を導入する．アイテム間の類似度はプライバシ情報を含まないので，その取扱いが容易で，暗号化処理のコストを大きく下げる．しかも，最初にアイテム間類似度によるCFを提案したSarwarらの文献8)によると，密度が低い評価値のデータセットを利用した場合，アイテム間の類似度をもとに予測評価を行うアイテム間類似度CF方式のほうが精度が良い．したがって，アイテム間類似度CF法には，次にあげる利点がある．

1) アイテム間類似度にはユーザ固有のプライバシがなく，類似度計算コストが低い．

2) アイテム間類似度は共通の情報であり，公開することができる．ひとたび公開された

類似度を用いれば，それ以降の推薦値計算には計算済みの類似度を活用すればよい．

3)　大規模なユーザ間での推薦では，多くの場合，ユーザ間類似度に基づく評価よりも精度が高い．

本論文では，アイテム間類似度に基づくプライバシ保護 CF 方式を提案し，公開実験データ[9),10)] を用いて，提案方式の精度と性能を明らかにする．

本論文の構成は次のとおりである．まず，2 章で CF の基本モデルといくつかの要素技術と，ユーザ間とアイテム間の類似度の定義を行う．次に，3 章で提案方式の具体的なアルゴリズムを示す．最後に 4 章では，提案方式の処理コストと精度を公開データセットに基づいて評価するための実験方法とその結果について述べる．5 章で結論と今後の課題を議論する．

2. 準　　備

2.1 プライバシの定義

推薦システムにおけるプライバシ情報を定義する．

アイテム評価値　アイテム評価値は以下の情報が含まれる．

(1)　ユーザが評価したアイテム

(2)　評価点

(3)　ユーザの平均点

予測評価をするアイテム　これからどのアイテムを評価しようとしているか．

匿名性　誰がアイテムを評価しているのか．

インターネット販売サイトの例で考える．アイテム評価値の (1) はユーザが買ったアイテムのリストである．(2) はユーザがあるアイテムに対してつけた評価値である．(3) はユーザのアイテムにつけた評価値の平均点であり，人によっては高めになったり，低くなったりする．また，予測評価をするアイテムは，これから何を買おうとしているかということであり，ユーザやアイテムによっては他のユーザには知られたくない情報である．一方，匿名性は誰がそのアイテムを買ったのかという情報である．本論文ではこれらの情報をプライバシ保護協調フィルタリングにおけるプライバシ情報と定義する．

2.2 モ デ ル

本論文で提案する推薦システムのモデルを定義する．n をユーザ数，m をアイテム数とする．全ユーザの集合を $U = \{u_1, u_2, \ldots, u_n\}$ とし，アイテムの集合を $I = \{i_1, i_2, \ldots, i_m\}$ とする．ユーザ u_j がアイテム i_k を評価した値を $r_{j,k} > 0$ の自然数とする．また未評価であるとき，$r_{j,k} = 0$ とする．評価はすべてのアイテムについて行われるわけではなく，多くの評価値が未評価値であることを仮定する．

推薦システムは，被推薦者ユーザ u_i が未評価のアイテム i_k の予測評価値 $P_{i,k}$ を他のユーザの評価値に基づいて求めることを目的とする．

2.3 協調フィルタリング方式（CF 方式）

CF 方式は，嗜好の類似する他ユーザの評価値を参考にし，未評価のアイテムの評価値（欠損値）を予測する方式であるこれによりアイテムの内容を見ることなく評価値のみを参考としてユーザの隠れた嗜好を求められる．

2.3.1 ユーザベース方式

ユーザベース方式は，ユーザ間の類似度から評価値を予測する．ユーザ u_i におけるアイテム i_k の予測評価値 $P^U_{i,k}$ は次式で与えられる．

$$P^U_{i,k} = \overline{r_{i,\cdot}} + \frac{\sum_{j \in U_k} s(u_i, u_j)(r_{j,k} - \overline{r_{j,\cdot}})}{\sum_{j \in U_k} |s(u_i, u_j)|} \tag{1}$$

ここで，アイテム i_k に評価を行ったユーザの集合を $U_k = \{i \in U | r_{i,k} \neq 0\}$，$\overline{r_{i,\cdot}}$ はユーザ u_i が行った評価の平均値とする．また，$s(u_i, u_j)$ はユーザ u_i と u_j 間の類似度である．類似度には様々な定義があるが，ここでは次式で定義されるコサイン尺度を考える．

$$s(u_i, u_j) = \frac{\sum_{k=1}^{m} r_{i,k} r_{j,k}}{\sqrt{r^2_{i,1} + \cdots + r^2_{i,m}}\sqrt{r^2_{j,1} + \cdots + r^2_{j,m}}} \tag{2}$$

他の類似度を求める方法には，ユークリッド距離を利用する方法があるが，プライバシを保護しながらの類似度計算は，計算量や通信量などのコストが高くなってしまう．

2.3.2 アイテムベース方式

アイテムベース方式は，アイテム間の類似度を計算し，評価値を予測する方式である．アイテム間の類似度 $s(i_i, i_j)$ を用いて，予測評価値 $P^I_{i,k}$ は以下の式で求める．

$$P^I_{i,k} = \overline{r_{\cdot,k}} + \frac{\sum_{j \in I_i} s(i_k, i_j)(r_{i,j} - \overline{r_{\cdot,j}})}{\sum_{j \in I_i} |s(i_k, i_j)|} \tag{3}$$

ここで，ユーザ u_i が評価を行ったアイテムの集合を $I_i = \{j \in I | r_{i,j} \neq 0\}$ とする．$\overline{r_{\cdot,k}}$ はアイテム i_k の評価の平均値とする．また，$s(i_i, i_j)$ はアイテム i_i と i_j 間の類似度である．アイテム間の類似度はユーザ間の類似度と同様にコサイン尺度で考える．

$$s(i_i, i_j) = \frac{\sum_{k=1}^{n} r_{k,i} r_{k,j}}{\sqrt{r^2_{1,i} + \cdots + r^2_{n,i}}\sqrt{r^2_{1,j} + \cdots + r^2_{n,j}}} \tag{4}$$

2.4　準同型性暗号

準同型性暗号 E は，以下の加法準同型性を満たす．

$$E[m_1]E[m_2] = E[m_1 + m_2], \tag{5}$$

$$E[m_1]^{m_2} = E[m_1 m_2] \tag{6}$$

加法準同型性を満たす暗号方式としては Paillier 暗号[11] や変形 ElGamal 暗号[12] が知られている．どちらも，複数の鍵管理ユーザで秘密鍵を分散管理し，定められたしきい値以上の管理ユーザ間の協力の下で復号を実現するしきい値復号が可能である．本方式では，評価者が鍵管理ユーザを兼ねるが，実際の復号処理は代表ユーザによって代行される．

2.5　Paillier 暗号

Paillier 暗号は，P. Paillier が提案した加法の準同型性を満たす暗号方式である[11]．

公開情報 $N = pq$ と $g \in Z^*_{N^2}$ を定義する．ここで，p と q は素数であり，g は位数が N の倍数であるような素数である．秘密情報として，$\lambda(N) = \mathrm{lcm}(p-1, q-1)$ を定義する．公開鍵は (N, g)，秘密鍵は N の素因数 p と q である．

平文 $M \in Z_N$ の暗号化は，乱数 $r \in Z_N$ を決め，

$$E(M) = g^M r^N \bmod N^2$$

と定める．暗号文 $c = E(M)$ の復号は，

$$\frac{L(c^\lambda \bmod N^2)}{L(g^\lambda \bmod N^2)} \bmod N = M$$

により行う．ただし，$L(u) = (u-1)/N$ とする．

Paillier 暗号は以下に示すように，加法の準同型性を満たす．

$$E(M_1)E(M_2) = (g^{M_1} r_1^N)(g^{M_2} r_2^N) = g^{M_1+M_2}(r_1 r_2)^N = E(M_1 + M_2)$$

変形 ElGamal 暗号では復号した値 g^M から平文 M を求める方法が総当たりでの計算以外にはない．そのため，平文空間が広い場合には Paillier 暗号の方が適している．

3. 提 案 方 式

3.1　概　　要

提案方式では，全ユーザが全アイテムについての評価値を暗号化して同報する．代表ユーザが，全アイテムの暗号文から，加法の準同型性を用いて，アイテム評価値の平均値，およびノルムを求める．これらの値は暗号化したまま加算した後で復号して公開し，アイテムの平均評価値とノルムからアイテムどうしの類似度を求め公開する．暗号化した状態で行われるのは，平均評価値とノルム，および類似度の計算であり，最終的な予測評価値は平文の状態で計算できる．

提案方式で利用する暗号系は，同じ平文を持つ別の暗号文を作成する．このため，確定的暗号を利用すると，復号せずに平文を予想できる場合があるため，確率暗号を用いる必要がある．また，本方式は平文空間が広いため，復号効率を考慮して Paillier 暗号を利用する．

3.2　プロトコル

未評価の予測評価値 P^I_{i,k_0} を次の手順で求める．

(1)　各ユーザ $j = 1, \ldots, n$ はアイテム $k = 1, \ldots, m$ について，

$$A_{j,k} = E(r_{j,k}),$$
$$B_{j,k} = E(e_{j,k}),$$
$$C_{j,k} = E(r^2_{j,k})$$

を計算する．ただし，ここで，$e_{j,k}$ は

$$e_{j,k} = \begin{cases} 1 & \text{if } r_{i,k} \neq 0, \\ 0 & \text{otherwise} \end{cases}$$

で定めるフラグとする．すべての $k \neq \ell \in I$ となる組 k，ℓ について，

$$D_{j,k,\ell} = E(r_{j,k} r_{j,\ell})$$

を計算し，$A_{j,k}$，$B_{j,k}$，$C_{j,k}$，$D_{j,k,\ell}$ を同報する．

(2)　被推薦ユーザ u_i はアイテムごとの評価値の平均，ノルム，アイテム間の類似度を求めるために暗号化したまま，アイテム $k = 1, \ldots, m$ について以下を計算する．この処理は推薦値を要求するユーザだけではなく，どのユーザが行ってもよく，1 度計算すればそれ以降の推薦に継続して利用できる．

$$E(n_k \overline{r_{\cdot,k}}) = \prod_{j=1}^{n} A_{j,k} = E\left(\sum_{j=1}^{n} r_{j,k}\right),$$

$$E(n_k) = \prod_{j=1}^{n} B_{j,k} = E\left(\sum_{j=1}^{n} e_{j,k}\right),$$

$$E(||\boldsymbol{r}_k||^2) = \prod_{j=1}^{n} C_{j,k} = E\left(\sum_{j=1}^{n} r^2_{j,k}\right),$$

ただし，$n_k = |I_k|$, すなわち，アイテム i_k を評価したユーザ数とする．また，$k \neq \ell \in I$ について，

$$E\left(\sum_{j=1}^{n} r_{j,k}r_{j,\ell}\right) = \prod_{j=1}^{n} D_{j,k,\ell}$$

を求める．

(3)　しきい値以上のユーザの合意のうえで，暗号文 $E(n\overline{r_{\cdot,k}})$，$E(n_k)$，$E(||\boldsymbol{r}_k||^2)$，$E(\sum_{j=1}^{n} r_{j,k}r_{j,\ell})$ を分散復号する．

(4)　ユーザ u_i は復号された値を使って，平均値，ノルム，類似度を次のように求め，その結果を公開する．

$$\overline{r_{\cdot,k}} = \frac{D(E(n_k\overline{r_{\cdot,k}}))}{n_k},$$
$$||\boldsymbol{r}_k|| = \sqrt{r_{1,k}^2 + \cdots + r_{n,k}^2},$$
$$s(i_k, i_\ell) = \frac{D(E(\sum_{j=1}^{n} r_{j,k}r_{j,\ell}))}{||\boldsymbol{r}_k|| \cdot ||\boldsymbol{r}_\ell||}$$

ただし，ここで，$\boldsymbol{r}_k$ は k 番目の評価値ベクトルとする．

(5)　最後に被推薦ユーザ u_i は式 (3) により予測評価値を計算する．

ステップ (1)–(4) までは一度実行すればよく，ステップ (5) だけは独立にどのユーザでも実行できる．アイテム間の類似度の平均は，ユーザの平均値になった時点でプライバシ情報ではなく，全ユーザで共有できる統計量と見なすことができるからである．

3.3 計　算　例

表 1 の例について計算例を示す．被推薦ユーザ u_2 は i_3 の評価値 $P_{2,3}$（表中の*の値）を予測する．

(1)　各ユーザ u_1 から u_5 はアイテム $k = 1, \ldots, 5$ について，それぞれ計算する．ここでは u_1 の計算例を説明する．

$$A_{1,3} = E(r_{1,3}) = E(5),$$
$$A_{1,4} = E(r_{1,4}) = E(3),$$
$$A_{1,1} = A_{1,2} = A_{1,5} = E(0),$$
$$B_{1,3} = E(1) = B_{1,4},$$
$$B_{1,1} = E(0) = B_{1,2} = B_{1,5},$$
$$C_{1,3} = E(r_{1,3}^2) = E(5^2),$$
$$D_{1,1,2} = E(r_{1,1}r_{1,2}) = E(0),$$
$$\vdots$$
$$D_{1,4,5} = E(r_{1,4}r_{1,5}) = E(0).$$

(2)　被推薦ユーザ u_2 はアイテムごとの評価値の平均，ノルム，アイテム間の類似度を次のように求める．ここでは特にアイテム i_1 についての計算例を示すが，他のアイテムも同様に計算する．

$$E(n_1\overline{r_{\cdot,1}}) = E(0)E(3)E(2)E(1)E(0) = E(6),$$
$$E(n_1) = E(0)E(1)E(1)E(1)E(0) = E(3),$$
$$E(||\boldsymbol{r}_1||^2) = E({r_{1,1}}^2)\cdots E({r_{5,1}}^2) = E(9+4+1) = E(14),$$
$$E(\boldsymbol{r}_1 \cdot \boldsymbol{r}_3) = E(r_{1,1}r_{1,3})\cdots E(r_{5,1}r_{5,3}) = E(2\cdot 2).$$

(3)　閾値以上のユーザの合意のうえで，$n_1\overline{r_{\cdot,1}}$，$n_1$，$||\boldsymbol{r}_1||^2$，$s(i_1, i_2), \ldots, s(i_4, i_5)$ を分散復号する．

(4)　ユーザ u_2 は復号した値を使って，平均値とノルム，アイテム間類似度を求める．

$$\overline{r_{\cdot,1}} = \frac{n_1\overline{r_{\cdot,1}}}{n_1} = \frac{6}{3} = 2,$$
$$||\boldsymbol{r}_1|| = \sqrt{||\boldsymbol{r}_1||^2} = \sqrt{14}$$
$$s(i_1, i_3) = \frac{\boldsymbol{r}_1 \cdot \boldsymbol{r}_3}{||\boldsymbol{r}_1|| \cdot ||\boldsymbol{r}_3||} = \frac{4}{\sqrt{420}} = 0.2,$$
$$s(i_3, i_5) = \frac{\boldsymbol{r}_3 \cdot \boldsymbol{r}_5}{||\boldsymbol{r}_3|| \cdot ||\boldsymbol{r}_5||} = \frac{13}{\sqrt{1350}} = 0.35.$$

(5)　最後にユーザ u_2 は以下を計算する．

$$P_{2,3}^I = 2.67 + \frac{0.2(3-2) + 0.35(2-3.67)}{0.20 + 0.35} = 1.98$$

また，ユーザベース方式であれば，予測評価値は以下のように求められる．

表 1　評価値行列 R
Table 1　Rating matrix R.

	i_1	i_2	i_3	i_4	i_5
u_1	0	0	5	3	0
u_2	3	0	*	0	2
u_3	2	0	2	0	4
u_4	1	0	0	0	0
u_5	0	3	1	0	5
$\overline{r_{\cdot,k}}$	2	3	2.67	3	3.67
$\lvert\boldsymbol{r}_k\rvert$	$\sqrt{14}$	3	$\sqrt{30}$	3	$\sqrt{45}$
$s(i_3, i_k)$	0.20	0.18	-	0.91	0.35

$$P^{U}_{2,3} = 2.5 + \frac{0.79(2-2.7) + 0.47(1-3)}{0.79+0.47} = 1.32$$

4. 評価と考察

4.1 実験方法

ユーザの評価データには GroupLens[9] で配布されている “MovieLens Data Sets” を用いた．ユーザ $n = 943$ 人の $m = 1{,}682$ アイテムに対する 100,000 個の評価値で構成されている．ここから無作為に 100 個を未評価値として選び，ユーザ間類似度 CF 法と提案方式で予測し，MAE（Mean Absolute Error）を導出した．**表 2** に実験環境を示す．

また，利用する暗号系は 2,048 ビットの Paillier 暗号とする．

4.2 推薦精度

図 1 はアイテム数 m を 1,682 個，評価データのユーザ数 n を 0 から 943 人まで変化させたときの提案方式の精度を示している．MAE の測定をそれぞれのポイントで 10 回行い，平均をとりプロットした．MAE は 0.82 から 0.66 へと次第に改善され，ユーザ数が 943 人のとき MAE が最良となった．評価データ全体に対する利用の割合が増えると精度が改善されるのは理にかなっている．また，全ユーザ数の約 1/5 の 200 人のときの精度は，最良だった MAE0.66 のおよそ 90%の精度となる．つまり，全体のユーザ数の 1/5 程度で最も良いときの 90%の精度が得られるということである．

一方，従来のユーザベース CF 方式との比較のため，ユーザ数 n を固定してアイテム数 m を変化させたときの精度を測定した．**図 2** は m を 200 から 1,682 まで変化させたときの精度を示している．n が 943 のとき，この図では $m < n$，$m \approx n$，$m > n$ の 3 つのケースがある．提案方式が，ユーザベース CF 方式よりもつねに精度が良いのが興味深い．ユーザベース CF 方式の精度は 0.68 からあまり変化しない．一方，アイテムベースはアイテム数が少ないとき精度が高いが，予測評価に利用するアイテム数が，少なすぎるからと考えられる．

表 3 に方式 3) と提案方式の MAE，標準偏差を示す．この表から，ユーザベース方式と

表 2　実験環境
Table 2　Specification of evaluation environment.

OS	OS Windows 7 Professional 32 bit
CPU	Inter(R) Core(TM)2 Duo CPU P 8700 @ 2.53 GHz（2 CPUs）
Software	Java(TM) SE Runtime Environment build 1.6.0_17-b04

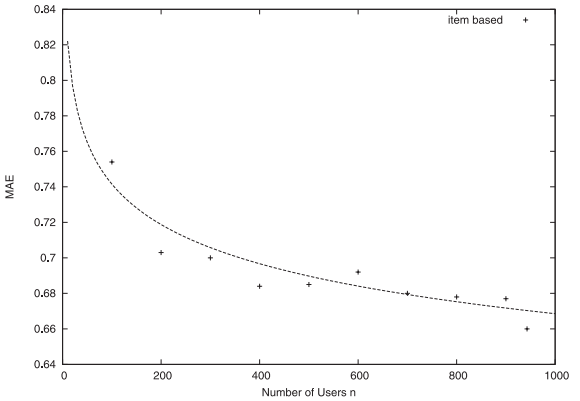

図 1　ユーザ数 n についての平均精度
Fig. 1　MEA with regards to number of users n.

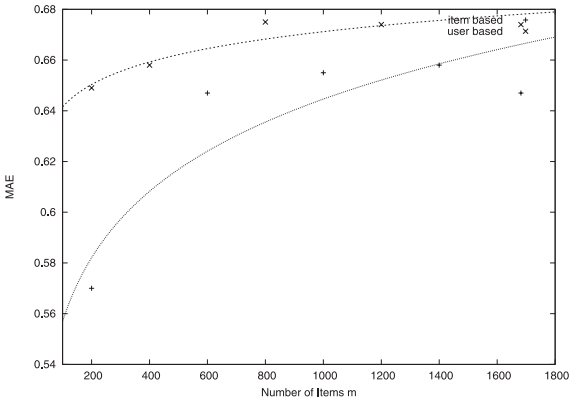

図 2　アイテム数 m についての平均精度
Fig. 2　MAE with regards to number of items m.

表 3　平均精度
Table 3　Average accuracy.

	ユーザ間類似度 CF 方式[3]	提案方式
平均誤差（MAE）	0.68	0.65
標準偏差	0.062	0.058

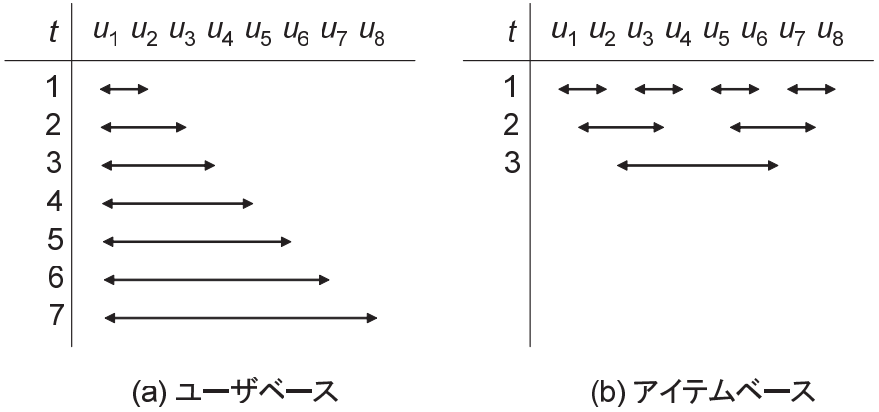

図 3　通信処理イベントのタイムチャート
Fig. 3　Time-line chart of communication events.

比較して，アイテムベース方式での精度が遜色ないことが分かる．

4.3 並列処理

図 3 は，提案プロトコルの並列処理による通信処理のイベントの関係を表している．$n = 8$ のユーザにおいて，時刻 t で生じる通信を矢印で表している．ユーザ u_1 を被推薦ユーザとする．

(a) のユーザベースプロトコルでは，u_1 は全ユーザとの間でユーザ間の類似度を秘匿計算するために，u_1 がボトルネックとなり，n に比例する通信時間がかかる．一方，(b) の提案プロトコルでは，ユーザ u_1 と u_2 の通信を行う間に u_3，u_4 間と u_5，u_6 間，u_7，u_8 間の通信を並列に行うことが可能である．したがって，プロトコルのステップ (2) で全ユーザの積は，$\log(n)$ 時間で求まる．最適な通信ノードを求めるのは自明ではないが，P2P のようなアーキテクチャを仮定しておけば $O(\log n)$ の計算量を実現可能である．

4.4 通信コスト

予測評価値を得る際に必要な暗号文によって，通信コストを比較した．通信コストはユーザ数 n とアイテム数 m に依存する．**表 4** に，各ユーザの通信量と全プロトコルを実行するのにかかる通信時間の計算量を整理する．ここで，公開鍵暗号の暗号文の長さを 2,048 ビット（512 バイト）として通信コストの見積もっており，4.3 節で述べた並列処理による効率化を仮定している．

表 4　通信コスト
Table 4　Communication cost.

	ユーザ間類似度 CF 方式[3]	提案方式（アイテム間）
各ユーザ	$3m^2 + m + 2m$	$\frac{1}{2}m(m-1) + 3m$
全体	$O(m^2 n)$	$O(m^2 \log n)$

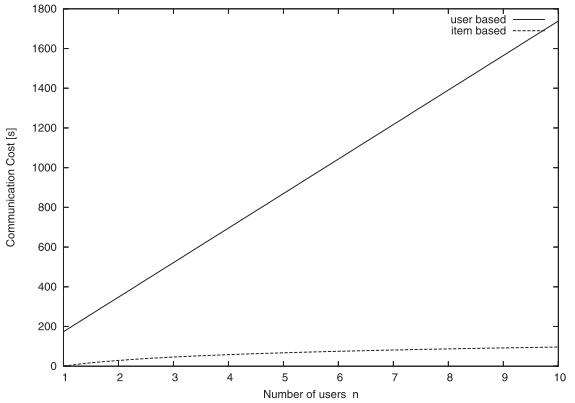

図 4　ユーザ数 n における通信処理コスト
Fig. 4　Communication cost with regards to number of users n.

ユーザ数 n に対する全体での処理時間を**図 4** に示す．ネットワークの伝送速度を $B = 100$ [Mbps] を仮定して，全体の推薦処理が終わるまでの処理時間を表している．つねに，提案方式の方が処理が小さいことが分かる．

4.5 計算コスト

表 5 に計算コストの見積りを示す．計算コストは公開鍵暗号の暗号化，復号，べき乗剰余演算数に比例して生じる．処理時間は，表 2 の環境での実測値である．

並列処理は，ユーザベースとアイテムベースの両方に効果があり，評価値の暗号化は全ユーザが並列に行うものとする．復号処理は，被推薦者などの代表ユーザにかかる処理であり，全体で 1 回でよい．

図 5 は，アイテム数 $m = 1{,}682$ のときのユーザ数 n に対する総処理時間を示す．ここで，$|N^2| = 2{,}048$ ビット Paillier 暗号の暗号化，復号，べき乗剰余演算を仮定している．提案方式はユーザ数 n に比例しないので従来のユーザベースの方式よりはるかに処理時間が短い．

表 5　計算コスト
Table 5　Computation cost.

	ユーザ間類似度 CF 方式[3]	提案方式	処理時間 [s]
（各ユーザ）暗号化	$2m^2+m$	$\frac{1}{2}m(m-1)+3m$	1.14
（代表ユーザ）復号	$2m$	$\frac{1}{2}m(m-1)+3m$	1.54
（代表ユーザ）べき乗剰余	$2n(m-1)$	0	0.57
	$O(m^2+n)$	$O(m^2)$	

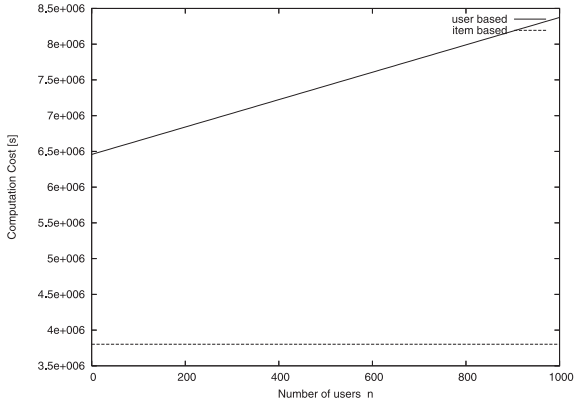

図 5　ユーザ間類似度 CF と提案方式の計算コスト
Fig. 5　Computation cost in user-based CF and the proposed scheme.

さらに，ユーザ間類似度 CF 方式と大きく違う点は，類似度を求めた後，その類似度は公開情報としてよいため，同じアイテムに対する予測評価ならば計算済みの類似度が再利用できるところである．そのため，表 5 に示されている計算コストは 2 回目以降の予測評価の際には 0 となるので，提案方式は非常に実用的である．

4.6　実用的な運用規模について

全体の処理時間は，通信にかかる処理と計算にかかる処理の和で表される．したがって，ネットワークの伝送速度 B，暗号文長 2,048bit を仮定すると総処理時間 $T(n,m)$ は，

$$\begin{aligned} T(n,m) &= (3m+m(m-1)/2)(2{,}048/B)\log n+(1.14+1.54)(3m+m(m-1)/2) \\ &= (3m+m(m-1)/2)((2{,}048/B)\log n+2.68) \\ &> (3m+m^2/2)2.68 \end{aligned}$$

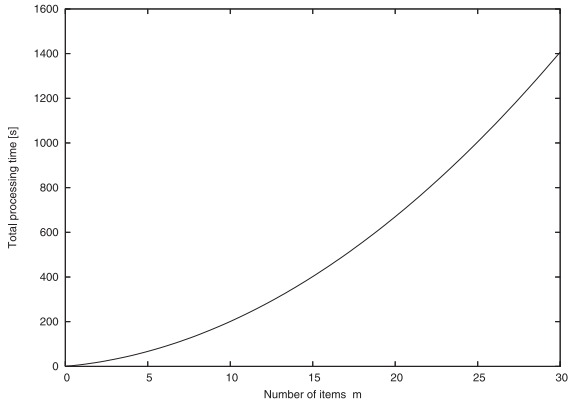

図 6　アイテム数 m についての総処理時間
Fig. 6　Total processing time with regards to number of items m.

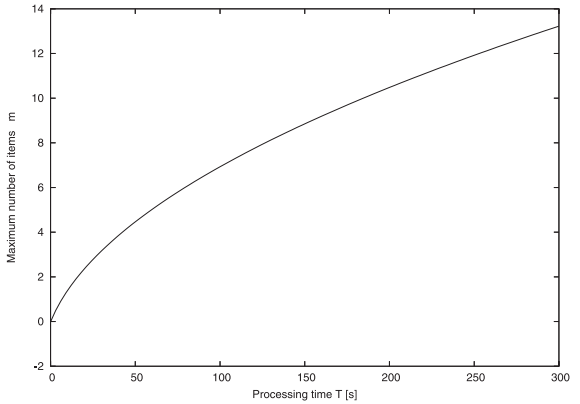

図 7　総処理時間についての最大アイテム数 m^*
Fig. 7　Maximum number of items m^* given total processing time T.

と与えられる．運用規模に関しては，アイテム数 m が支配的な要素であることが分かり，逆算すると許容される処理時間 T に対する最大アイテム数は $m^*=\sqrt{2T/2.68}$ で得られる．これを，m に対する関係は**図 6**，**図 7** で可視化する．$T=300\,[\mathrm{s}]=5\,[\mathrm{m}]$ のときで，$m^*=13$ である．

表 6　アイテム評価値 r についての評価値の平均 μ の出現頻度
Table 6　Frequency of mean rating μ given a rating r.

μ	$r=0$	$r=1$	$r=2$	$r=3$	$P(\mu)$
0	1	0	0	0	0.02
1	3	4	0	0	0.11
2	7	11	10	5	0.52
3	5	1	6	11	0.34
total	16	16	16	16	1

表 7　平均値 μ が与えられた時の条件付き確率 $P(r|\mu)$
Table 7　Conditional probability $P(r|\mu)$ given mean rating μ.

r	$P(\mu)$	$\mu=0$	$\mu=1$	$\mu=2$	$\mu=3$
0	0.25	1	0.429	0.212	0.227
1	0.25	0	0.571	0.333	0.045
2	0.25	0	0	0.303	0.273
3	0.25	0	0	0.152	0.455
$H(S)$	2	0	0.99	1.94	1.510

4.7 安　全　性

安全性について議論する．まず，これから予測評価を行うアイテムに関して述べる．予測評価を行う際に，すべての公開情報を入手することにより，予測評価の対象となるアイテムが推測できる確率は，ランダムに選んだ確率と等しい．次に，匿名性の保証については，本提案とは別の技術（たとえば仮名 ID など）を用いることで実現できる．

一方，アイテム評価値についての安全性は，要素技術の準同型性暗号の安全性にも依存するが，計算途中で公開される情報から一部の情報が漏れる可能性がある．提案方式は，各アイテムの平均値，ノルム，アイテム間の類似度を公開している．これらは，全ユーザで平均されているのでプライベートな情報ではないが，これらが分かることで各ユーザの持つ推薦値の部分的な情報が漏れる懸念がある．そこで，そのリスクの大きさを検討する．

ユーザ数 $n=3$，アイテム数 $m=3$ で，未評価 $r=0$，評価値 1，2，3 がすべて一様に生起するとき，すなわち，$P(r=0)=P(r=1)=P(r=2)=P(r=3)=1/4$ であるとき，平均値 μ の分布を考える[⋆1]．このときの，平均値 μ は**表 6** の頻度で生起する．平均値 μ は一様ではなく，2 を最頻とする分布をしている．この表は，μ と r に対する同時確率を与えている．

ベイズの定理により，平均値 μ が与えられたときの（特定の）評価値 r の条件付き確率は，

$$P(r|\mu)=\frac{P(\mu|r)P(r)}{P(\mu)}=\frac{P(\mu|r)P(r)}{\sum_{r'}P(\mu|r')P(r')}$$

で与えられる．**表 7** にこの条件付き確率の分布とそのときのエントロピー $H(S|\mu)$ を表す．平均値から漏れる情報量は，無条件のときのエントロピー $H(S)$ との差，

$$I=H(S)-H(S|\mu)$$

で定式化することができる．たとえば，$\mu=2$ が与えられときに漏れる情報量は $2-1.94=0.06$ [bit] であり，平均すると 1.65bit の情報である．

この情報量は，ユーザ数 n が多くなるに従って単調に減少する．十分なユーザ規模で運用している限り，プライバシのリスクはそれほど大きくないといえる．

⋆1 簡単のため，ユーザ u_i のアイテム j の評価値 $r_{i,j}=r$，アイテム j の平均値 $\overline{r_j}=\mu$ とおく．

5. 結　　論

アイテム間類似度を利用したプライバシ保護協調フィルタリング方式を提案し，効率と精度の観点より評価を行った．代表的な公開評価データによる評価実験の結果，精度については従来から用いられているユーザ間類似度を用いた CF 方式と比較しても問題ないことが分かった．また，通信コストや計算コストについても類似度の計算の際に並列処理が可能となるため，ユーザベース CF 方式より良いことが分かった．複数回予測評価値を計算する場合も，アイテム間類似度を再計算する必要はなく，実用性が高い．

今回は類似度を求める方法にコサイン尺度を利用したが，今後はより効率的な類似度の計算法の採用や，ユーザやアイテムの次数を削減する方法により，効率の改善を試みていきたい．

謝辞　多くの有益なコメントをいただいたモンクット王ラカバン工科大学の Sutheera Puntheeranurak 氏と東海大学大学院の木澤寛厚氏と匿名の査読者に感謝する．

参　考　文　献

1) Amazon.com. http://www.amazon.co.jp/ (2009 年 11 月参照)
2) Canny, J.: Collaborative Filtering with Privacy, *Proc. IEEE Conf. on Security and Privacy*, Oakland, CA, pp.45–57 (2002).
3) 木澤寛厚，菊池浩明：プライバシ協調フィルタリングにおける利用者評価行列の次元削減，コンピュータセキュリティシンポジウム 2008（CSS2008），pp.509–514 (2008).
4) 木澤寛厚，磯崎邦隆，菊池浩明：秘匿積集合プロトコルを利用したプライバシ協調フィルタリングの提案，2009 年暗号と情報セキュリティシンポジウム（SCIS2009），1B2-5,

pp.1–4 (2009).
5) Kikuchi, H., Kizawa, H. and Tada, M.: Privacy-Preserving Collaborative Filtering Schemes, *Proc. WAIS 2009, ARES 2009 Federated Workshop*, pp.911–916, IEEE Press (2009).
6) Ahmad, W. and Khokhar: An Architecture for Privacy Preserving Collaborative Filtering on Web Portals, *Proc. 3rd International Symposium on Information Assurance and Security*, IEEE Computer Society, pp.273–278 (2007).
7) Katzenbeisser, S. and Petkovic: Privacy-Preserving Recommendation Systems for Consumer Healthcare Services, *Proc. 2008 3rd International Conference on Availability, Reliability and Security* (*ARES 2008*), IEEE Computer Society, pp.889–895 (2008).
8) Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.: Item-Based Collaborative Filtering Recommendation Algorithms, *Proc. 10th International Conference on World Wide Web* (*WWW10*), Hong Kong, pp.285–295 (2001).
9) Grouplens Data Sets. http://grouplens.org/ (2009 年 8 月参照)
10) Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P. and Riedl, J.: GroupLens: An open architecture for collaborative filtering of netnews, *Proc. 1994 Computer Supported Collaborative Work Conference*, pp.175–186 (1994).
11) Paillier, P.: Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, *Advances in cryptology – EUROCRYPT'99*, LNCS, Vol.1592, pp.223–238, Springer-Verlag (1999).
12) El Gamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms, *Advances in Cryptology – CRYPTO'84*, LNCS, Vol.196, pp.10–18, Springer-Verlag (1985).

(平成 21 年 11 月 30 日受付)
(平成 22 年 6 月 3 日採録)

多田美奈子（正会員）

2002 年東海大学工学部電気工学科卒業．2004 年同大学院工学研究科博士前期課程修了．2004 年東芝ソリューション（株）入社，現在に至る．情報セキュリティの研究に従事．電子情報通信学会会員．

菊池　浩明（フェロー）

1988 年明治大学工学部電子通信工学科卒業．1990 年同大学院博士前期課程修了．1994 年同博士（工学）．1990 年（株）富士通研究所入社．1994 年東海大学工学部電気工学科助手．1995 年同専任講師．1999 年同助教授，2000 年同電子情報学部情報メディア学科助教授，2006 年同情報理工学部情報メディア学科教授．2008 年同情報通信学部通信ネットワーク工学科教授．1997 年カーネギーメロン大学計算機科学学部客員研究員．WIDE プロジェクト暗号メールシステム FJPEM の開発，認証実用化実験協議会（ICAT），IPA 独創情報技術育成事業等に従事．暗号プロトコル，ネットワークセキュリティ，ファジィ論理，ソフトコンピューティング等に興味を持つ．1990 年日本ファジィ学会奨励賞，1993 年情報処理学会奨励賞，1996 年 SCIS 論文賞，2010 年情報処理学会 JIP Outstanding Paper Award．電子情報通信学会，日本知能情報ファジィ学会，IEEE，ACM 各会員．情報処理学会フェロー．