

# アイテム間類似度に基づくプライバシー保護協調フィルタリングの提案

多田 美奈子<sup>†</sup>

菊池 浩明<sup>‡</sup>

<sup>†</sup> 東海大学大学院総合理工学研究科  
259-1292 神奈川県平塚市北金目 1117

lukia@cs.dm.u-tokai.ac.jp

<sup>‡</sup> 東海大学情報通信学部通信ネットワーク工学科  
259-1292 神奈川県平塚市北金目 1117

kikn@tokai.ac.jp

あらまし 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一つとして、情報推薦システムがある。情報推薦システムにおいて利用者のプライバシー情報をサービス運営者に漏らさずに推薦を行う方法が提案されているが、利用者数に応じて大きな計算量が掛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ー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<sup>†</sup>

Hiroaki Kikuchi<sup>‡</sup>

<sup>†</sup> Graduate School of Science and Technology, Tokai University  
1117 Kitakaname Hiratsuka-Shi Kanagawa 259-1292 Japan

lukia@cs.dm.u-tokai.ac.jp

<sup>‡</sup> Department of Communication and Network Engineering, Tokai University  
1117 Kitakaname Hiratsuka-Shi Kanagawa 259-1292 Japan

kikn@tokai.ac.jp

**Abstract** A recommendation system enables us to take information from huge datasets about tastes effectively. Some cryptographical protocols for computing privacy-preserving recommendation without leaking the privacy of users are proposed. However, the large overhead depending on the number of users is current issue.

In this paper, we propose an efficient scheme by using item-item similarities for providing a prediction of arbitrary values of rating. We show the performance and the accuracy of our proposed scheme based on a numerical experiment.

## 1 はじめに

今日では多くの Web ショップが存在し、多くのユーザがこれを利用している。様々な商品が購入できるようになったが、選択肢が多く、目的の商品を見つけることが困難となっている。多

くの情報の中から自分の目的とする情報を取り出すためには、ある程度の情報リテラシーが必要となる。そのため初心者ユーザには使いづらく、利用に二の足を踏んでしまうことも起こっている。

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一つである。Amazon.com[4]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ー情報をサービス提供者がすべて持っているため、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ー情報が流出したりするという危険性がある。

そこで、この問題を解決する方法として、推薦システムにおける手法の一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ー情報を秘匿したまま行う試みがされている。Canny は、各ユーザが準同型暗号を用いて評価値を秘匿したまま共役勾配法を実行し、特異値に分解された評価行列を求める手法を提案している [6]。木澤らは、ユーザ同士の類似度を元に予測評価を行う方式を提案している [2, 3]。これらの方式では Canny の方式に比べて計算時間や通信コストが削減されているが、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あった。

一方、Sarwar らは密度がまばらな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元に予測評価を行う方式のほうが精度がよいことを示している [7]。プライバシー保護の観点から考えても、ユーザ間の類似度よりもアイテム間の類似度のほうが都合がよい。しかも、効率の悪い類似度計算を初回のみ計算して公開すれば、二回目以降のコストを削減できる効果も得られる。

それゆえ本論文では、アイテム間の類似度を利用したプライバシー保護協調フィルタリング方式を提案する。本方式は、精度がよく、計算コストが削減できる可能性がある事前計算に向いていることを示す。また、公開実験データ [5] を用いて、提案方式の精度と性能を明らかにする。

## 2 準備

### 2.1 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  $U = \{u_1, u_2, \dots, u_n\}$  とし、アイテムの集合を  $I = \{i_1, i_2, \dots, i_m\}$  とする。ユーザ  $u_j$  がアイテム  $i_k$  を評価した値を  $r_{j,k}$  とし、未評価であることを、 $r_{j,k} = \phi$  と表す。推薦システムは、ユーザ  $i$  におけるアイテム  $k$  の予測評価値  $P_{i,k}$  を求めることを目的とする。

### 2.2 協調フィルタリング方式 (CF 方式)

CF (Collaborative Filtering) 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 [1]。この方式は、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ので、ユーザは思いがけない嗜好を発見することが期待できる。

#### 2.2.1 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方式である。

ユーザ  $i$  におけるアイテム  $k$  の予測評価値  $P_{i,k}$  は次式で与えられる。

$$P_{i,k} = \bar{r}_i + \frac{\sum_{j \in U_k} s(u_i, u_j)(r_{j,k} - \bar{r}_j)}{\sum_{j \in U_k} |s(u_i, u_j)|} \quad (1)$$

ここでアイテム  $k$  に評価を行ったユーザの集合を  $U_k = \{i \in U | r_{i,k} \neq \phi\}$  とする。 $\bar{r}_i$  はユーザ  $i$  が行った評価の平均値とする。また、 $s(u_i, u_j)$  はユーザ間類似度である。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。

$$s(u_i, u_j) = \frac{\sum_{k=1}^n r_{i,k} r_{j,k}}{\sqrt{r_{i,1}^2 + \dots + r_{i,n}^2} \sqrt{r_{j,1}^2 + \dots + r_{j,n}^2}} \quad (2)$$

### 2.2.2 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  $s(i_k, i_j)$  を用いて、予測評価値  $P_{i,k}$  は以下の式で求める。

$$P_{i,k} = \bar{r}_k + \frac{\sum_{j \in I_i} s(i_k, i_j)(r_{i,j} - \bar{r}_j)}{\sum_{j \in I_i} |s(i_k, i_j)|} \quad (3)$$

ここでユーザ  $i$  が評価を行ったアイテムの集合を  $I_i = \{j \in I | r_{i,j} \neq \phi\}$  とする。 $\bar{r}_k$  はアイテム  $k$  の評価の平均値とする。

### 2.3 準同型性暗号

準同型性暗号は、以下の性質を満たす。

$$E[m_1]E[m_2] = E[m_1 + m_2] \quad (4)$$

$$E[m_1]^{m_2} = E[m_1 m_2] \quad (5)$$

加法の準同型性を満たす暗号方式としては Paillier 暗号や変形 ElGamal 暗号が知られている。提案方式では加法の準同型性を利用して、暗号化したままで加算を行う。

## 3 提案方式

### 3.1 概要

提案方式では、被評価ユーザが他ユーザが評価したアイテムの暗号文を入手し、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。計算した後で暗号文を復号して公開し、それらのアイテムの平均評価値とノルムをからアイテム同士の類似度を求め、これより予測評価値を求める。暗号化した状態で行われるのは、平均評価値とノルムの計算、および類似度であり、最終的な予測評価値は暗号化なしに計算できる。

### 3.2 プロトコル

未評価の予測評価値  $P_{i,k}$  を次の手順で求める。

1. 各ユーザ  $j = 1, \dots, n$  はアイテム  $k = 1, \dots, m$  について、

$$A_{j,k} = E(r_{j,k})$$

$$B_{j,k} = E(e_{j,k})$$

$$C_{j,k} = E(r_{j,k}^2)$$

を計算する。ただし、ここで、 $e_{j,k}$  は

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

で定めるフラグとする。全ての  $k \neq \ell \in I$  について、

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

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

2. 被推薦者<sup>1</sup>  $u_i$  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  $k = 1, \dots, m$  について以下を計算する。

$$E(n\bar{r}_k) = \prod_{j=1}^n A_{j,k} = E\left(\sum_{j=1}^n r_{j,k}\right)$$

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

$$E\left(\sum_{j=1}^n r_{j,k}^2\right) = \prod_{j=1}^n C_{j,k}$$

ただし、 $n_k = |I_k|$ 、すなわち、 $k$  番目のアイテム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. 閾値以上のユーザの合意の上で、暗号文  $E(n\bar{r}_k), E(n_k), E(n\bar{r}_k^2), E(\sum_{j=1}^n r_{j,k}r_{j,\ell})$  を次のように分散復号し、

$$D(E(n\bar{r}_k))/n = \bar{r}_k$$

<sup>1</sup> 推薦値を要求するユーザだけではなく、この処理はどのユーザが行ってもよく、一度計算すればそれ以降の推薦に継続して利用できる

平均値を求める． $n_k, r_k^2, r_{j,k}r_{j,\ell}$  について  
も同様とする．

4. ユーザ  $u_i$  は復号された値を使って，平均  
値，ノルム，類似度を求め，その結果を  
公開する．

$$\begin{aligned}\bar{r}_k &= \frac{D(E(n\bar{r}_k))}{n_k} \\ |r_k| &= \sqrt{r_{1,k}^2 + \cdots + r_{n,k}^2} \\ s(i_k, i_\ell) &= \frac{D(E(\sum_{j=1}^n r_{j,k}r_{j,\ell}))}{|r_k||r_\ell|}\end{aligned}$$

ただし，ここで， $r_k$  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ー情報ではなく，  
全ユーザで共有できる統計量とみなすことが出  
来るからである．

### 3.3 計算例

表 1: 利用者評価値

|                  | $i_1$       | $i_2$ | $i_3$       | $i_4$ | $i_5$       |
|------------------|-------------|-------|-------------|-------|-------------|
| $u_1$            |             |       | 5           | 3     |             |
| $u_2$            | 3           |       | *           |       | 2           |
| $u_3$            | 2           |       | 2           |       | 4           |
| $u_4$            | 1           |       |             |       |             |
| $u_5$            |             | 3     | 1           |       | 5           |
| $\bar{r}_i$      | 2           | 3     | 2.67        | 3     | 3.67        |
| $ r_j $          | $\sqrt{14}$ | 3     | $\sqrt{30}$ | 3     | $\sqrt{45}$ |
| $s(i_3, i_{k'})$ | 0.20        | 0.18  | -           | 0.91  | 0.35        |

表 1 の例について計算例を示す．被推薦者  $u_2$   
は  $i_3$  の評価値  $P_{2,3}$  を予測する．

1. 各ユーザ  $u_1$  から  $u_5$  は  $k = 1, \dots, 5$  に  
ついて，

$$A_{j,k} = E(r_{j,k})$$

$$\begin{aligned}B_{j,k} &= E(e_{j,k}) \\ C_{j,k} &= E(r_{j,k}^2)\end{aligned}$$

を計算する．また，すべての  $k \neq l \in I$  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．

例えば  $u_1$  は， $i_3$  と  $i_4$  を評価しているの  
で以下の計算を行う．尚，以下はアイテ  
ム  $i_3$  についての計算である．これをアイ  
テム  $i_1$  から  $i_5$  まで繰り返す．

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

2. 被推薦者  $u_2$  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求め  
るために暗号化したまま以下を計算する．  
ここではアイテム  $i_1$  の平均，ノルム，ア  
イテム間類似度について計算例を示す．

$$\begin{aligned}E(n\bar{r}_1) &= E(0)E(3)E(2) \\ &\quad E(1)E(0) \\ E(\bar{n}_1) &= E(0)E(1)E(1) \\ &\quad E(1)E(0) \\ E(\sum_{j=1}^n r_{j,1}^2) &= E(r_{1,1}^2) \cdots E(r_{5,1}^2) \\ E(s(i_1, i_3)') &= E(r_{1,1}r_{1,3}) \cdots \\ &\quad E(r_{5,1}r_{5,3}) \\ E(s(i_3, i_5)') &= E(r_{1,3}r_{1,5}) \cdots \\ &\quad E(r_{5,3}r_{5,5})\end{aligned}$$

3. 閾値以上のユーザの合意の上で， $\bar{r}_1, n_1, |r_1|$   
， $s(i_2, i_3), s(i_1, i_3), s(i_3, i_4), s(i_3, i_5)$  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．

$$\begin{aligned}\bar{r}_1 &= \frac{\bar{r}_1}{n_1} = \frac{6}{3} = 2 \\ |r_1| &= \sqrt{|i_1|'} = \sqrt{14} \\ s(i_1, i_3) &= \frac{s(i_1, i_3)'}{|r_1||r_3|} = \frac{4}{\sqrt{420}} = 0.2 \\ s(i_3, i_5) &= \frac{s(i_3, i_5)'}{|r_3||r_5|} = \frac{13}{\sqrt{1350}} = 0.35\end{aligned}$$

5. 最後にユーザ  $u_2$  は以下を計算する．

$$\begin{aligned}P_{2,3} &= 2.67 \\ &+ \frac{0.2(3-2) + 0.35(2-3.67)}{0.20 + 0.35} \\ &= 1.98\end{aligned}$$

また，アイテムベース方式ではなく，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．

$$\begin{aligned}P_{2,3} &= 2.5 + \frac{0.79(2-2.7) + 0.47(1-3)}{0.79 + 0.47} \\ &= 2.5 + \frac{-1.49}{1.26} \\ &= 1.32\end{aligned}$$

## 4 評価と考察

提案方式の精度と，計算コスト，通信コストを評価する．また，方式の優位性について考察する．

### 4.1 実験方法

ユーザの評価データは GroupLens[5] で配布されている “MovieLens Data Sets” を用いた．ユーザ  $n = 943$  人の  $m = 1682$  アイテムに対する評価値で構成されている．評価値の合計は 100,100 個ほどである．ここから無作為に 100 個を選び，未評価値として，[3] 方式および提案方式にて予測した．また，それぞれの方式の MAE(Mean Absolute Error) を導出した．

### 4.2 推薦精度

表 2 に方式 [3] と提案方式の MAE，標準偏差を示す．比較すると，提案方式が精度がよく，また誤差も少ない．

表 2: 各手法の精度

|      | 方式 [3] | 提案方式 |
|------|--------|------|
| 平均誤差 | 0.79   | 0.77 |
| 標準偏差 | 0.67   | 0.64 |

### 4.3 コスト

表 3: 各手法のコスト

|               | [3] 方式              | 提案方式                               |
|---------------|---------------------|------------------------------------|
| 類似度の通信コスト     | $m(2n+1)$           | $3m(n+1) + \frac{1}{2}m(m-1)(n+1)$ |
| 類似度の通信オーダ     | $\mathcal{O}(mn)$   | $\mathcal{O}(mn)$                  |
| 類似度の計算コスト     | $mn^2$              | $3m(n+1) + \frac{1}{2}m(m-1)(n+1)$ |
| 類似度の計算オーダ     | $\mathcal{O}(mn^2)$ | $\mathcal{O}(m^2n)$                |
| CF の通信コスト     | $mn^2$              | $m$                                |
| CF の通信オーダ     | $\mathcal{O}(mn^2)$ | $\mathcal{O}(m)$                   |
| 暗号化 CF の計算コスト | $2(m-1)$            | 0                                  |
| 暗号化 CF の計算オーダ | $\mathcal{O}(mn)$   | 0                                  |
| 平文 CF の計算コスト  | 0                   | $2(m-1)$                           |
| 平文 CFF の計算オーダ | 0                   | $\mathcal{O}(m)$                   |
| 事前計算          | ×                   | ○                                  |

表 3 は，類似度，CF の通信コストと計算コストをあらわしたものである．計算コストはべき乗剰余演算の回数である．

提案方式の類似度の通信コストと計算コストは， $A_{j,k}$ ,  $B_{j,k}$ ,  $C_{j,k}$ ,  $D_{j,k,\ell}$  についてかかる．

また、CFの通信コストは、アイテム  $m$  個分の類似度を入手すると予測評価値が求められるため、通信量は  $m$ 、暗号化に必要なべき乗剰余演算はない。通常の積の計算をカウントすると、 $2(m-1)$  となる。

#### 4.4 考察

効率について考察する。

[3]方式と大きく違う点は、類似度を求めた後、予測評価値を計算する際、提案方式では暗号化されていない状態で計算できるところである。準同型性を用いて暗号化したまま行う計算は、加算は積算になったり、計算結果を復号する必要があったり、計算量の増加につながる。そのため、提案方式はコスト削減になっているといえる。

## 5 結論

アイテム間類似度を利用したプライバシー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。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。

## 参考文献

- [1] J. S. Breese, D. Heckrman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," In UAI, pp.43-52, 2004.
- [2] 木澤, 菊池, "プライバシー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.
- [3] 木澤, 磯崎, 菊池, "秘匿積集合プロトコルを利用したプライバシー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), 2009.

- [4] Amazon.com, <http://www.amazon.co.jp/>
- [5] Grouplens Data Sets, <http://grouplens.org/> (2009年8月参照)
- [6] J. Canny, "Collaborative Filtering with Privacy," IEEE Conf. on Security and Privacy, Oakland CA, May 2002.
- [7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl "Item-Based Collaborative Filtering Recommendation Algorithms," WWW10, Hong Kong, May 2001.