SCIS 2012 The 29th Symposium on
Cryptography and Information Security
Kanazawa, Japan, Jan. 30 - Feb. 2, 2012
The Institute of Electronics,
Information and Communication Engineers

アイテム依存の摂動化によるプライバシー保護情報推薦
Privacy-Preserving Recommendation via Item-Variant Perturbation

望月 安菜* 菊池 浩明*
Anna Mochizuki Hiroaki Kikuchi

あらまし 一定確率でアイテムの評価データをランダマイズし，推薦者のプライバシーを保護した情報推薦方式が提案されている．しかし，従来方式の多くは，全てのアイテムに対して一定の確率でランダマイズを行なっているので，推薦の精度を左右する重要なアイテムや，機微なアイテムに異なる摂動化レベルを割り当てることができない．そこで本研究では，アイテムごとに必要十分な維持確率を与えることによって，推薦精度を向上させる情報推薦方式を提案する．

キーワード プライバシー保護, 情報推薦, Slope One

1 はじめに

情報推薦の主流は，複数のユーザによって複数のアイテムが評価付けされているデータベースにおいて，他のユーザの値を基に評価されていないアイテムの評価値を予測する協調フィルタリング (Collaborative Filtering) である．しかし，これらの情報推薦には，不正なサービス事業者によるプライバシー漏洩という問題が挙げられる．そこでプライバシーを守るため，準同型性を満たした公開鍵暗号を使った個人情報を秘匿する研究がある．しかし，暗号は，プライバシー保護は出来るが，大きな計算コストがかかる．そこで，本研究では，個人のプライバシーを保護しながら，暗号化をせずにユーザに応じた情報推薦を行うことを目的とする．

H. Polat らは，加法摂動化による協調フィルタリング方式を提案している [3]．彼らの研究では，オリジナルデータ X に一様分布の乱数 R を加えた $Y = X + R$ について平均値 $\sum_i Y_i = \sum_i X_i + \sum_i R_i \approx \sum_i X_i$ であることを仮定したナイーブな推薦方式である．従って，Z. Huang らによって，Y を主成分分析 (PCA) することで，加えた乱数ノイズを取り除くことが出来ることが指摘されており [4]，その安全性は低い．

菊池ら [6] は，分析対象のデータには，分析に重要で変化させたくない属性値を保持することで，通常の撹乱・再構築に比べ計算量を削減する手法を提案している．

S. Zhang ら [5] は，全てのアイテムを一様にアイテム不変に摂動化することは，精度の面でも，プライバシーの面でも効率が悪い事を指摘し，特異値分解 (SVD) を用いた摂動化方式を提案している．

本研究では，アイテムに依存した維持確率を使用し，再構築の計算コストと推薦精度を向上した情報推薦方式について検討する．

2 要素技術

2.1 摂動化と再構築

再構築問題 (Reconstruction Problem) とは，摂動化された $Y_1 = X_1 + R_1$，確率変数 Y から，真の値 X の確率分布を見積もる問題である．R.Agrawal and R. Srikant [1] によって最初に発表された摂動化アルゴリズムである．秘匿したい情報に意図的にランダムノイズを乗せて格納されたデータのプライバシーを保護する．

例えば，年齢が $x = 20$ 代であるという情報をそのまま渡す代わりに，一様分布 (またはガウス分布) の乱数 r を加え，$y = x + r = 30$ のように歪んだ値 y を登録する．30 という属性値を持った顧客がいても，本当に 30 代なのか乱数で 20 代から歪まされたのか，第三者には区別がつかない．

暗号化による方法と異なり，時間のかかる暗号ではなく，計算も各パーティで独立に計算できる．通信効率も計算効率も高い．大規模なデータベースにおいても適用可能である．

* 東海大学大学院 工学研究科 情報理工学専攻 〒 259-1292 神奈川県平塚市北金目四丁目 1 番 1 号. Course of Information Science and Engineer, Graduate School of Engineering, Tokai university 4-1-1, Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan. {cream_18_puff,kikn}@cs.dm.u-tokai.ac.jp

2.2 摂動化

簡単な数値例を用いて再構築アルゴリズムの原理を示す．確率変数 A が表 1 の分布に従って与えられているとする．

表 1: 真の確率分布 $P(A)$

a	0	1	2	3
$P(A=a)$	0.1	0.3	0.1	0.5

ここで，A の分布を秘匿する為に，表 2 の条件付確率 $P(B|A)$ に従って，A の値を変化（摂動化）させた結果を B とおく．

表 2: 条件付確率 $P(B|A)$, 維持確率 $p=0.4$

$B\backslash A$	0	1	2	3
0	0.4	0.2	0.2	0.2
1	0.2	0.4	0.2	0.2
2	0.2	0.2	0.4	0.2
3	0.2	0.2	0.2	0.4

ここで，維持確率 $p=0.4$ は，A を変化させない確率の大きさであり，変化させるときは一様な確率で分布させることにする．こうして摂動化した結果を表 3 で示す．オリジナルの分布では $A=3$ が最頻度で生じていたのに対して，値の差が小さくなりどの値も同じくらい確からしい．

表 3: 摂動化した確率分布 $P(B)$

b	0	1	2	3
$P(B=b)$	0.22	0.26	0.22	0.3

2.3 再構築アルゴリズム

再構築アルゴリズムは，この $P(B|A)$ と摂動化後の確率分布 $P(B)$ だけを与えて，オリジナルの分布 $P(A)$ を近似することを目的とする．初期値を $P^0(A)=P(B)$ で与える．事後確率の i 番目の近似値は，

$$\begin{aligned} P^i(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P^{i-1}(A)}{\sum_{a\in A} P(B|A=a)P^{i-1}(A=a)} \end{aligned}$$

と近似され，この値を用いて A の事後確率の第一近似値は

$$P^1(A)=\sum_{b\in B} P^0(A|B=b)P(B=b)$$

で与えらられる．こうして，逐次的に近似を繰り返し，$P^{i+1}(A)=P^i(A)$ と収束した分布を再構築された P^* とする．この数値例の場合の第二近似値までの結果を表 4 で示す．徐々に真の分布へ近づいていることが分かる．

表 4: 再構築された確率分布の第一近似 $P^1(A)$ と第二近似 $P^2(A)$

a	0	1	2	3
$P^1(A=a)$	0.22	0.26	0.22	0.31
$P^2(A=a)$	0.21	0.26	0.21	0.33

2.4 Slope One

D. Leniel ら [2] によって提案された Slope One はアイテムベースの情報推薦アルゴリズムである．シンプルなアルゴリズムと高い性能で商用にも採用されている．Slope One とは，アイテム間の相関に傾き 1 の一次式 $f(x)=x+b$ を用いているところからその名が付いている．特異値分解などの既存の推薦方式と比較して，アイテム間平均差分に基づいて推薦を行うので実装も容易で処理性能も高い．

簡単な数値例を用いて Slope One アルゴリズムを解説する．ここで，評価値の定義域は 1 から 5 の離散値とし，“0” は欠損値を表す．

	I_1	I_2	I_3	I_4	I_5
U_1	-	1	4	2	1
U_2	-	-	-	-	2
U_3	-	-	1	1	-
U_4	5	1	-	-	3
U_5	2	-	2	3	5

この関係を，評価値行列

$$R_{5\times 6}=(r_{i,j})=\begin{pmatrix} 0 & 1 & 4 & 2 & 1 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 1 & 0 \\ 5 & 1 & 0 & 0 & 3 \\ 2 & 0 & 2 & 3 & 5 \end{pmatrix}$$

で表す．Slope One はシンプルに，差分の平均値で評価値を与える．すなわち，アイテム i_1 の評価値は，アイテム i_2 とその間の差分の平均 δ_{i_1,i_2} から，$r_{i_1}=\delta_{i_1,i_2}+r_{i_2}$ と定義される．

平均差分は，その両方のアイテムとも評価を与えているユーザについて求める．i_2 の i_3 による類似度は i_1 によるものよりも高いので，評価値はより大きく影響を受ける．

上の例では，どちらのアイテムも同じ数のユーザによって評価されているので，単純に 2 で割って平均を取って

いるが，欠損値がある場合はこの限りではない．そこで，重みを考える．アイテム a と b の平均差分 $\delta_{a,b}$ を，

$$\overline{\delta_{a,b}} = \frac{\Delta_{a,b}}{\phi_{a,b}} = \frac{\sum_i \delta_{i,a,b}}{\phi_{a,b}} = \frac{\sum_i (r_{i,a} - r_{i,b})}{\phi_{a,b}} \quad (1)$$

で与える．平均差分行列 (average difference matrix) は，

$$\overline{\Delta_{5\times5}} = (\overline{\delta_{i,j}}) = \begin{pmatrix} 0 & 4 & 0 & -1 & -0.5 \\ -4 & 0 & -3 & -1 & -1 \\ 0 & 3 & 0 & 0.33 & 0 \\ 1 & 1 & -0.33 & 0 & -0.5 \\ 0.5 & 1 & 0 & 0.5 & 0 \end{pmatrix}$$

と定義する．ここで共生起数 $\phi_{a,b}$ は両方のアイテムを評価しているユーザの数である．共生起行列 (relative occurrence matrix) は

$$\Phi_{5\times5} = (\phi_{i,j}) = \begin{pmatrix} 2 & 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 & 2 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 1 & 3 & 3 & 2 \\ 2 & 2 & 2 & 2 & 4 \end{pmatrix}$$

で与える．

この時，ユーザ u のアイテム x に対する評価値の予測を Slope One では，

$$\begin{aligned} r_{u,x} &= \frac{\sum_{a|a\neq x}(\overline{\delta_{x,a}} + r_{u,a})\phi_{x,a}}{\sum_{a|a\neq x}\phi_{x,a}} \\ &= \frac{\sum_{a|a\neq x}(\Delta_{x,a} + r_{u,a}\phi_{x,a})}{\sum_{a|a\neq x}\phi_{x,a}} \end{aligned} \quad (2)$$

により求める．$r_{4,2}$ について Slope One を行うと

$$r_{4,2} = \frac{\sum_{j=1,5}(\overline{\delta_{2,j}} + r_{4,j})\phi_{2,j}}{\sum_{j=1,5}\phi_{2,j}} = 1.67$$

となり，その誤差は，平均絶対誤差 (Mean Absolute Error:MAE) で 0.67 である．

3 提案方式

3.1 アイデア

オリジナルデータ X の評価値行列 R^X について共生起行列 Φ を求め，アイテムごとの欠損値数を考慮して維持確率 p_i を決める．アイテム毎に randomized response によって摂動化を行い，摂動化データ R^Y と p についてベイズ推定を行い，真のデータ X の再構築を行う．再構築の過程で，得られた条件付き確率 $P(X|Y)$ を用いて，推薦精度を向上させることを試みる．

- アイテム不変の摂動化

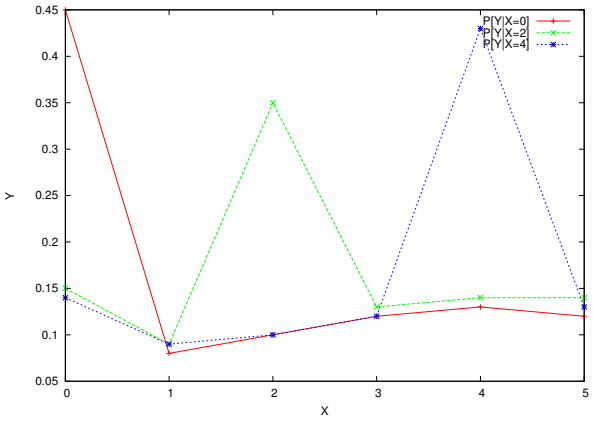

図 1: 摂動化 Y の $P(Y|X)$ の確率分布

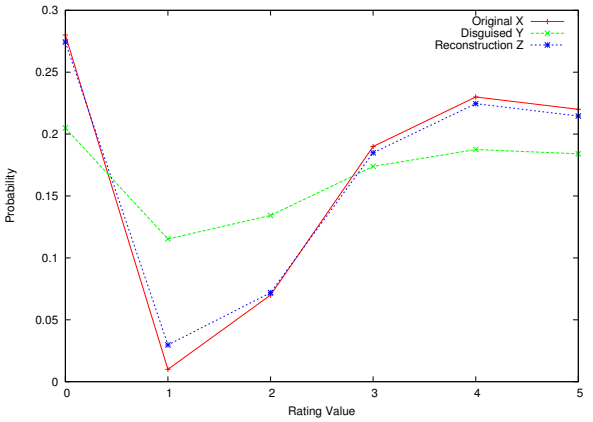

図 2: 摂動化と再構築による評価値分布の変化 (維持確率 $p = 0.4$)

全てのアイテムを同じ維持確率によって摂動化を行う．

- アイテム依存の摂動化

各アイテムのスパース率によって維持確率を変える．

3.2 提案方式 - アイテム依存維持確率

アイテム i のスパース率 s_i によって，アイテムごとの維持確率を定める．アイテム i スパース率 s_i とは，アイテム i を評価済のユーザの密度で定める．データベースが疎であるとスパース率は高く，密であると低い値を示す．

アイテム i のスパース率と共生起数 ϕ_i には

$$\phi_i = 1 - s_i \quad (3)$$

の関係がある．共生起数は，アイテム毎に安全な電子投票プロトコルを用いて求め，全員で共有して持つ．スパース率 s_i のアイテム i を摂動化する維持確率 p_i は，要求する精度と守るべきプライバシーの両方の観点から定める．全アイテムでの目標精度 MAE^* とプライバシー ϕ_Y^* を固定し，再構築したデータ Z の $MAE(Z) < MAE^*$,$\phi_Z^* \geq \phi_Y^*$ を満たす最小の p_i を決める．

表 5: 2 種類のスパース率を持つ評価値データセット (Original Data)

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	5	4	5	5		3			3	2
u_2	5	3		5	3	4			5	
u_3	5	4	5	3	1	4	4		3	
u_4	5	5	5	4		3	4			
u_5	5	3	3	5	3	4		4		2
u_6	5		5	4	3	4			5	
u_7	5	4	5	4	3	3				3
u_8	5	3	4	3	2	3	3	2	4	
u_9	4	5	5	4	3	4				4
u_{10}	4	2	4	2		4	2	5	4	
p	0.8						0.1			

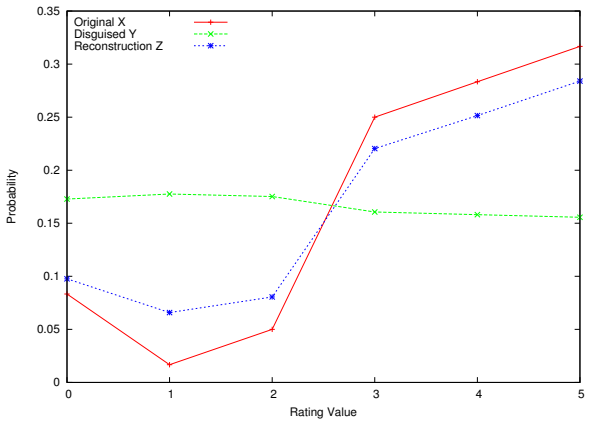

図 3: データセット A における評価値分布 ($p_A = 0.8$)

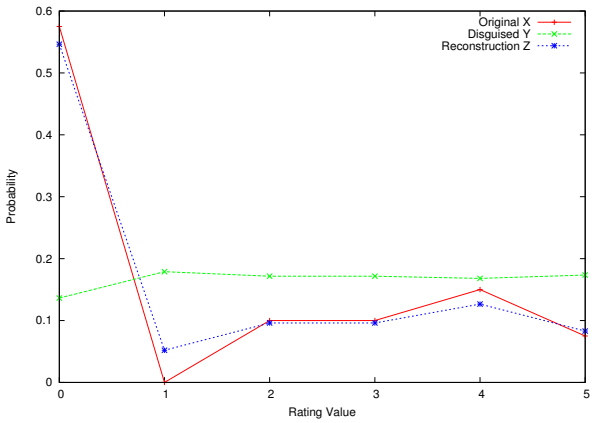

図 4: データセット B における評価値分布 ($p_B = 0.1$)

3.3 実験

実験には，Movie Lens [8] のデータセット (ユーザ数 $u = 943$, アイテム数 $i = 1682$, 評価値数 100,000 件) からランダムに抽出し利用した表 5 のデータを用いる．ここで，i_1 から i_6 までを A ，i_7 から i_{10} までを B とおき，各々$p_A = 0.8$, $p_B = 0.1$ の維持確率で摂動化する．このデータをアイテム不変の維持確率 $p = 0.4$ で摂動化し，[9] で提案した手法で評価値の予測をした場合と，本提案のアイテム依存の維持確率で摂動化して評価した場合とを比較する．

3.4 実験結果

オリジナルデータ，全てのアイテムを同じ維持確率によって摂動化を行ったアイテム不変摂動化データ (item invariant)，アイテムのスパース率によって維持確率を変動させたアイテム依存摂動化データ (item dependence) のそれぞれで [9] の方式による Slope One を行い予測した値を比較する．MAE(Mean Absolute Error) による推薦値の誤差を表 6 に示す．図 1 は，摂動化したデータの確率分布，図 2 は，再構築によって補正された確率分布，図 3 と図 4 は表 5 の A と B に対応するデータに適用した摂動化と再構築の確率分布である．図 5 は，アイテム不変とアイテム依存の情報推薦の誤差の分布を示している．

表 6: Mean Absolute Error

	Original	item Inveriance	item Dependence
MAE	0.57	0.81	0.61

3.5 考察

図 5 により, アイテム依存の維持確率を使用することで，摂動化データはオリジナルデータに近い．表 6 により，MAE の値もそれを裏付けている．従って，アイテム依存の維持確率を使用した時より誤差の少ない情報推薦を行うことが示された．

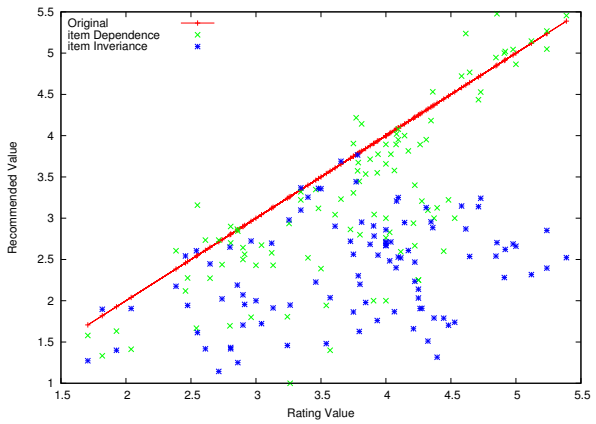

図 5: 再構築された評価値 Z の分布

4 おわりに

スパース率を利用したアイテム依存の摂動化，それに伴う情報推薦方式を提案した．提案方式は，アイテムごとに維持確率を定めるため再構築の計算コストが削減できた．今後の課題として，評価値の差分も考慮した維持確率を使用し，適した維持確率を使用する必要がある．また，誤差の少ない情報推薦を行う．

参考文献

[1] R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining”, ACM SIGMOD 2000, pp. 439-450, 2000.

[2] D. Leniel and A. Maclachlan, “Slope One Predictors for OnlineRating-Based CollaborativeFiltering”, Society for Industrial Mathematics, pp. 1-5, 2005.

[3] H. Polat and W. Du, “Privacy-Preserving Collaborative Filtering using Randomized Perturbation Techniques”, ICDM 2003, pp. 1-5, 2003.

[4] Z. Huang, W. Do and B. Chen, “Deriving Private Information from Randomized Data”, ACM SIGMOD 2005, pp. 37-48, 2005.

[5] S. Zhang, J. Ford and F. Makedon, “A Privacy-preserving Collaborative Filtering Scheme with Two-way Communication”, ACM EC’06, pp. 316-323, 2006.

[6] 菊池，五十嵐，千田，濱田, “属性値を保持する際に効果的な撹乱・再構築法”, Computer Security Symposium 2011, pp. 438-443, 2011.

[7] 豊田，宮川，側高，伊東, “匿名性グループ間の要素数の変化を比較可能な匿名化手法の実現”, Computer Security Symposium 2011, pp. 432-437, 2011.

[8] Grouplens Data Set, (http://grouplens.org/).

[9] 望月，菊池, “Slope One を用いた摂動化プライバシー保護情報推薦方式”, Computer Security Symposium 2011, pp. 379-384, 2011.