Slope One を用いた摂動化プライバシー保護情報推薦方式

望月 安菜 †　　菊池 浩明 †

† 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
cream_18_puff,kikn@cs.dm.u-tokai.ac.jp

あらまし プライバシーを保護したまま，一定確率で評価値をランダマイズする摂動化の処理により秘匿したデータから再構築を行う新しい情報推薦方式を提案する．提案方式では，主流の協調フィルタリングではなく，アイテム間類似度を評価値の差分で定めるアイテムベース推薦方式の Slope One を用いることにより，ランダマイズで生じる誤差を小さくする．

Perturbation based Privacy Preserving for "Slope One" Recommendation Method

Anna Mochizuki†　　Hiroaki Kikuchi†

†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

Abstract This paper proposes a new perturbation method for privacy preserving in recommendation. The proposed method improves accuracy of recommendation based on the well-known item-based collaborative filtering with simplified item-item similarity defined by difference of rating values.

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 [7]. 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．

提案方式は，暗号化に代え，摂動化と呼ばれるランダマイズレスポンスを使用する．データマイニング時に人工的に加えたノイズの影響をベイズ推定によって取り除く．情報提供者の持つ情報にノイズを加える．このデータの解析を行った後に，ノイズ除去の処理をほどこし，解析結果を同定する．この過程を再構築という．このように摂動化を使用した手法は，暗号化と比較し計算コストが小さい．また暗号化に比べて実装が容易である．しかし，暗号化がほぼ厳密に正しい結果を得ることに対して，摂動化の学習結果は近似解である．安全性の保証観点からは，暗号化の方が厳密である．そこで，情報推

薦を行う際に使用していた協調フィルタリングを使用するのではなく，アイテム間類似度を評価値の差分で定めるアイテムベース推薦方式の Slope One[2] によって推薦値を求める．本研究の特徴として，新たな方式である摂動化 Slope One を行う点にある．

2 Slope One

D. Leniel and A. Maclachlan [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．

2.1 概要

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．$* = \frac{(\frac{(4-2)+(5-2)}{2}+1)+(\frac{(4-4)+(5-4)}{2}+4)}{2} = 4.0$ により評価が与えられる．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，

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

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．相対生起行列 **relative occurrence matrix**:

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

で与える．

この時，ユーザ u のアイテム x に対する Slope One では，

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

により予測値を求める．

$r_{4,2}$ 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) を用い 0.67 となる．

3　準備

3.1　関連研究

H. Polat ら [6] 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．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，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 [5]，その安全性は低い．

そこで，本研究では，単純な PCA による解析が困難なランダマイズレスポンス方式を用いて，摂動化を行う．安全性は向上するが,[6] の様な単純な協調フィルタリングでは精度が期待できない．

3.2　摂動化と再構築

再構築問題 (Reconstruction Problem) とは，摂動化された $Y_1 = X_1 + R_1$ 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 $x = 20$ 代であるという個人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，$y = x + r = 30$ の様に歪んだ値 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 40 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，時間のかかる暗号化はなく，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．

オリジナルデータ X をランダマイズレスポンスによって摂動化を行い，偽データ Y を作成する．評価値の集合を $V = \{1, 2, \cdots, v\}$, 維持確率を p とする．評価値 x の摂動化 y は，

$$y = \begin{cases} x & \text{確率 } p \\ a \in V & \text{確率 } 1 - p \end{cases} \tag{3}$$

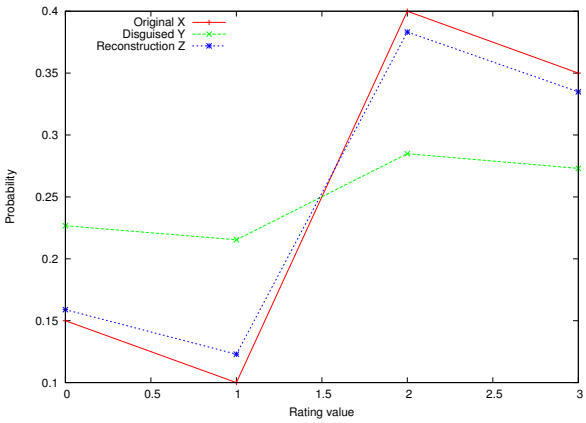

図 1: 再構築された評価値 Z の分布

4　提案方式

4.1　アイデア

オリジナルデータ X の評価値行列 R^X を，維持確率 p でランダマイズレスポンスによって摂動化を行い，偽データ Y の評価値行列 R^Y を作成する．この偽データ 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．

4.2　提案方式 - 摂動化 Slope One

Slope One を行う際に使用する共生起行列 $\Phi_{i,j}$ と平均差分行列 $\Delta_{i,j}$ を求める．偽データの評価値行列 R^Y のまま Slope One を行うと大きな誤差が起こるため，再構築を行い R^X に近似した R^Z によって，Slope One で情報推薦を行う．

4.2.1　共生起行列

維持確率 p より，欠損値数の予測を行う．摂動化を行って生成した行列 R^Y，維持確率 p よ

りオリジナルデータに期待欠損数を予測する．n 個の要素を持つオリジナルデータ X の欠損値数を表す確率変数を K_X，摂動化したデータで観測した欠損値数を表す確率変数を K_Y とする．摂動化を行うと欠損値を維持する確率は p，欠損値から評価値へと変化する確率は $1-p$ である．同様に，評価値を維持する確率は $1-\frac{1-p}{v}$ であり，評価値から欠損値へと変化する確率は $\frac{1-p}{v}$ である．これより，真の欠損値数 k_x の時に，摂動化データに k_y 個の欠損値が生じる条件付き確率は

$$P(K_Y|k_x) = \sum_{j=0}^{K_X} \binom{K_X}{j} (1-p)^j p^{K_X-j}$$
$$\binom{N-K_X}{K_Y-j} (1-\frac{1-p}{v})^{K_Y-j} (\frac{1-p}{v})^{N-K_X-K_Y+j}$$

で与えられる評価値を $V=\{1,2,3\}$，$p=0.4$，$n=4$ とした時の k_Y の確率分布を図 2 に示す．$k_x=0$ であっても，$k_y>0$ となる確率があることが分かる．この分布から欠損値数の期待値を $E[K_Y|K_X] = \sum_{k\in V} P(k_Y = R|K_X)$，最尤値を $L[K_Y|K_X] = \arg\max_{k\in V} kP(K_Y=k|K_X)$ で求めることができる．図 2 から算出した期待値と最尤値を図 5 に示す．興味深いことに両者は必ずしも一致しない．

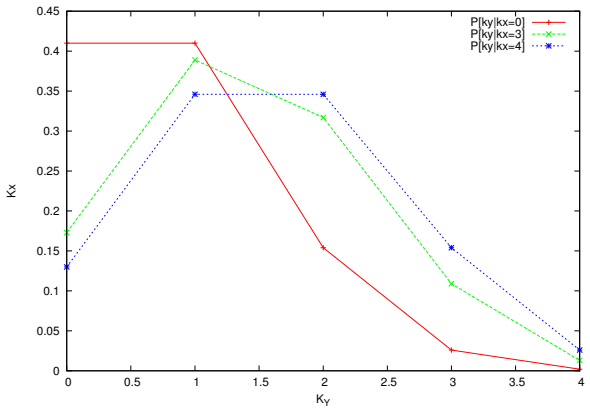

図 2: 摂動化の欠損値数 $P(K_Y|K_X)$ の確率分布

事前確率 $P(K_Y|K_X)$ より，再構築アルゴリズムよりベイズ推定を行うことで事後確率 $P(K_X|K_Y)$ を求める．図 2 の分布から再構築した $P(K_X|K_Y)$

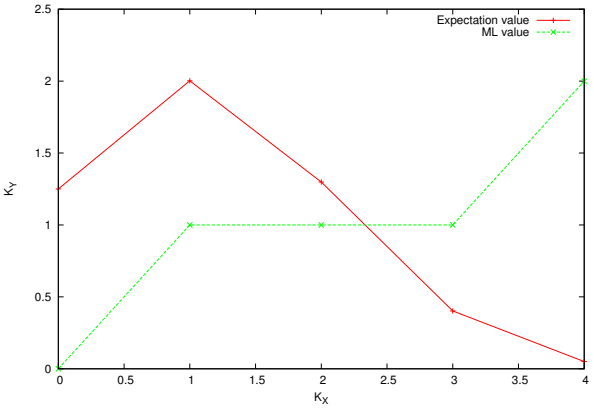

図 3: $P(K_Y|K_X)$ の期待値・最尤値の分布

の確率分布を図 4 に示す．

$$P(K_X|K_Y) = \frac{P(K_Y|K_X)P(K_X)}{\sum_X P(K_Y|K_X)P(K_X)}$$

$P(K_X|K_Y)$ より，期待値 $E[K_X|K_Y]$ と最尤値 $L[K_X|K_Y]$ をそれぞれ求めたものを図 5 に示す．例えば，$K_Y=1$ の時，期待値は $E[K_X]=1.9$ であり，　最尤値は $L[K_X]=1$ である．Φ は共通に評価している欠損でないアイテム数なので，K_Y が与えられた時の共生起数の期待値は $E[\Phi_X] = N - E[K_X|K_Y]$ で与えられる．

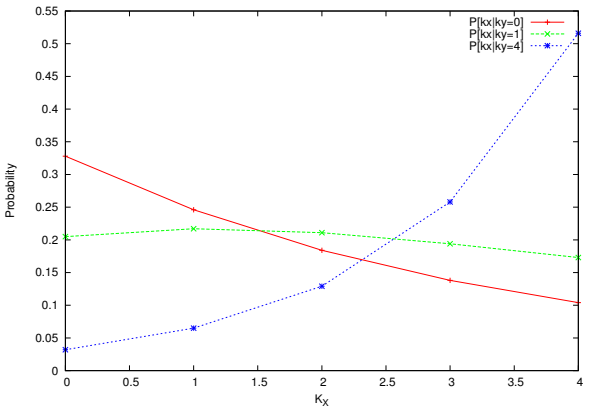

図 4: 真の欠損値数 $P(K_X|K_Y)$ の確率分布

4.2.2　平均差分行列

摂動化評価値行列 R^Y と維持確率 p から定まる条件付き確率 $P(Y|X)$ より，摂動化 Slope One のための Δ_R を求める．偽データの Δ_Y より，再

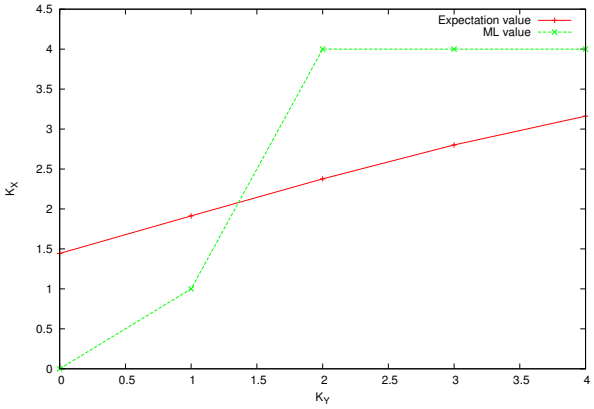

図 5: 欠損値数の期待値 $E[K_X|K_Y]$・最尤値 $L[K_X|K_Y]$ の分布

構築の際のベイズ推定で得られた条件付き確率 $P(X|Y)$ を用いて，ある列における差分の総和 Δ_Y から，真の差分総和 Δ_X を予測する．

$$P(\Delta_Y|\Delta_X) = \sum_{\Delta=\Delta_X} \sum_{\delta\in\Delta} P(\Delta_Y|\delta) \quad (4)$$

$P(\Delta_Y|\Delta_X)$ の最尤値を Δ_R と定める．

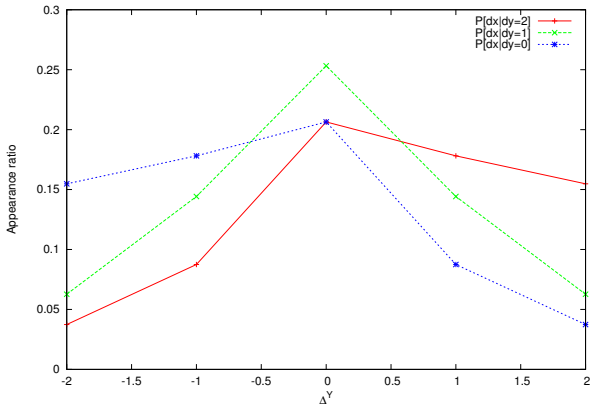

図 6: $P(\Delta_X|\Delta_Y)$ の分布

4.2.3 摂動化 Slope One

再構築を行った，共生起行列 Φ_Z と平均差分行列 Δ_Z が与えられた時，ユーザ u, アイテム i の予測値は

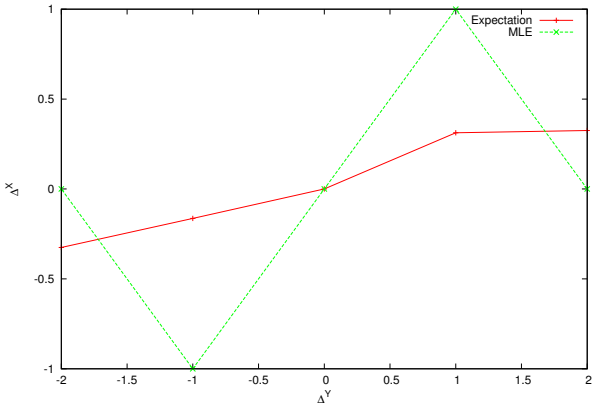

図 7: 真の差分 Δ_X・期待値 $E[\Delta_X|\Delta_Y]$・最尤値 $L[\Delta_X|\Delta_Y]$ の分布

$$\begin{aligned} r^Z_{u,i} &= \frac{\sum_i(\overline{\delta_{u,i}} + r_{u,i})\phi_{u,i}}{\sum_i \phi_{u,i}} \\ &= \frac{\sum_i(L[\Delta^X_i|\Delta^Y_i] + r_{u,i}(N - L[KX|K_Y]))}{\sum_i \phi_{u,i}} \end{aligned} \quad (5)$$

で与えられる．

4.3 数値例

本実験では，表 1 のオリジナルデータ R^X を評価値行列に，維持確率 $p = 0.4$ で摂動化した表 2 の偽データ R^Y を用いて，ユーザ数 $n = 4$, アイテム数 $m = 5$, 評価値 $V \in \{1, 2, 3\}$ とする．

偽データの評価値行列 R^Y と p,Slope One を行う際に使用する共生起行列 Φ_Y と平均差分行列 Δ_Y によって R^X を近似する R^Z を作成する．

表 1: Original Data (R^X)

	i_1	i_2	i_3	i_4	i_5
u_1	2	2	3	1	
u_2	1	3	2		3
u_3	2		3	3	2
u_4	3	2	3	2	2

4.4 結果

オリジナルデータ，摂動化を行った偽データ，提案手法である再構築データのそれぞれで Slope

表 2: Disguised Data (R^Y)

	i_1	i_2	i_3	i_4	i_5
u_1	2	1		2	1
u_2	3	3	2		1
u_3	3	3	3	1	2
u_4	1	2	1	2	2

One を行い予測した値の MAE(Mean Absolute Error) と，協調フィルタリングによって予測した値の MAE との比較を表 3 に示した．$MAE^Z = \sum_{u,i} |r^X_{u,i} - r^Z_{u,i}|$ と定める．摂動化したデータにナイーブに協調フィルタリングを適用したものは，$MAE^Y = \sum_{u,i} |r^X_{u,i} - r^Y_{u,i}|$ である．偽データより提案手法である再構築データの方が誤差が 0.03 少ない．

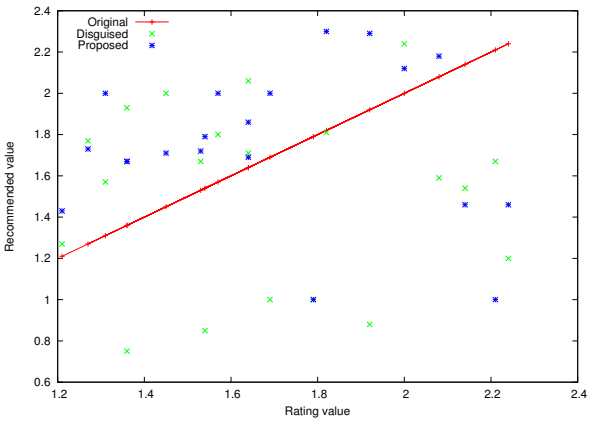

図 8: Slope One による推薦値の分布

表 3: Mean Absolute Error

	CF	Slope One
Original	0.97	0.73
Disguised	1.03	0.89
Proposed	1.01	0.86

4.5 考察

図 8 の散布図は，Slope One の推薦値を真のデータとした時の摂動化データのみによる推薦値と提案手法による推薦値を図示している．Slope One を行った際，再構築データはオリジナルデータに近づいていることが分かる．これは，表 3 からも分かるように，偽データより再構築データの方が誤差が 0.03 小さいことが表 3 から分かる．また，協調フィルタリングについても誤差を測り Slope One との比較を行った．全てのデータにおいて Slope One の方が誤差が少ない．最大で，0.24 の差がある．これら 2 点より，Slope One は，協調フィルタリングより精度が高く，摂動化との相性の良さが分かる．しかし，本結果で用いた評価行列は，オリジナルの推薦値の誤差が大きいことから人工的で歪んでいた可能性がある．

5 おわりに

プライバシーを保護したまま，摂動化したデータから再構築を行う新しい情報推薦方式を提案した．提案方式は，摂動化により一定確率で評価値をランダマイズされたデータが出来るためプライバシーが保護される．再構築によって，オリジナルデータへ近似させることで，情報推薦の精度を向上することを示した．情報推薦の主流は，協調フィルタリングであったが，Slope One を使って評価値を予測することで，より誤差の少ない推薦を行うことが期待出来る．

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 Online Rating-Based Collaborative Filtering”, Society for Industrial Mathematics, pp. 1-5, 2005.

[3] A. Basu, H. Kikuchi and J. Vaidya, “Privacy-preserving weighted Slope One for Item-based Collaborative Filtering”, IFIPTM 2011 Federated Workshop TP-DIS’11, pp. 1-12, 2011.

[4] 望月安菜, 菊池浩明, “摂動化によってプライバシーを保護した情報推薦方式”, DICOMO 2011, pp. 1-6, 2011.

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

[6] H. Polat and W. Du, “Privacy-Preserving Collaborative Filtering using Randomized Perturbation Techniques”, ICDM 2003, pp. 1-15, 2003.

[7] 青木良樹, 菊池浩明, “擬準同型性を満たす類似度による分散協調フィルタリングプロトコル”, SCIS 2011, pp. 1-6, 2011.