垂直分割における通信効率の良い一致度の秘匿分散計算

青木 良樹† 菊池 浩明† 寺田 雅之‡ 石井 一彦* 関野 公彦*

†東海大学大学院工学研究科
259-1292 神奈川県平塚市北金目四丁目 1 番 1 号
ringo@cs.dm.u-tokai.ac.jp, kikn@tokai.ac.jp

‡株式会社 NTT ドコモ 先進技術研究所
239-8536 神奈川県横須賀市光の丘 3-6
teradama@nttdocomo.co.jp

* 株式会社 NTT ドコモ サービス & ソリューション開発部
239-8536 神奈川県横須賀市光の丘 3-6
{ishiikaz,sekino}@nttdocomo.co.jp

あらまし 本論文は三者間で垂直分割されたデータベース環境において，互の値を秘匿したまま通信効率の良い一致度の秘匿分散計算方式を提案する．n 人のユーザ間の類以度を計算する時，従来手法では n の二乗個の暗号文を通信する必要があり非常に効率が悪い．そこで，一致度という類以度の推移性を利用し，この n^2 のコストを，n に削減する近似方法を提案する．

Communication Efficient Distributed Concordance Evaluation in Vertical Partitioning

Yoshiki Aoki† Hiroaki Kikuchi† Masayuki Terada‡ Kazuhiko Ishii*
Kimihiko Sekino*

†Graduate School of Engineering, Tokai University,
4-1-1 Kitakaname, Hiratsuka, Kanagawa, 259-1292, Japan.
ringo@cs.dm.u-tokai.ac.jp, kikn@tokai.ac.jp

‡Research Labs, NTT DOCOMO, Inc.,
3-6 Hikarinooka, Yokosuka-shi, Kanagawa,239-8536, Japan.
teradama@nttdocomo.co.jp

* Service & Solution Development Department, NTT DOCOMO, Inc.,
3-6 Hikarinooka, Yokosuka-shi, Kanagawa,239-8536, Japan.
{ishiikaz,sekino}@nttdocomo.co.jp

Abstract This paper proposes a new communication efficient distributed concordance evaluation in dataset vertical partition multiple parties. In conventional method, compute transitivity similarity between n users. We must send n^2 ciphertexts, which is very inefficient in communication. To solve the problem, we use a transitivity of concordance measure in order to reduce by $O(n)$. We evaluate our scheme in terms of quantity of disclosed private data.

1 はじめに

情報推薦が広く利用されている．n 人のユーザ間の類以度を計算する時，従来手法では n の二乗個の暗号文を通信する必要があり非常に効率が悪い．また，類以度はベクトル同士の計算であるため，欠損値が存在すると Default Voting[3] や Prediction Voting[1, 2] などで補完したり，二つの集合間の積集合（Intersection）を取得する必要がある．そこで，Somers' d，一致度（Concordant）という類以度の推移性に注目し，この n^2 のコストを，n に削減する近似方法を提案する．

本提案方式は三者間以上の垂直分割データベースへにおいて，多者間の秘匿協調フィルタリングへの応用が可能である．

通常，表 2 の未評価値を全て予測するためには，各ユーザ間全ての類以度を計算する必要がある．このユーザ間類以度の組み合わせはユーザ数を n としたとき，$O(n^2)$ に比例し大きくなる．

2 要素技術

2.1 準同型暗号

平文を秘匿したまま，平文同士を加算 (乗算) した暗号文を計算することができる性質を加法 (乗法) に関して準同型性を持つ，という．RSA 暗号や，ElGamal 暗号は乗法準同型性（Multiplicative Homomorphic）を満たす暗号として有名である．

本研究では加法に関して準同型性を持つ Paillier 暗号を用いた．g を生成元，r を乱数，n を大きな二つの素数 p, q の合成数 pq とした時，平文 m の 暗号文は，

$$g^m r^n \pmod{n}^2 \tag{1}$$

で計算される．このとき，g, n は公開鍵，p, q は秘密鍵となる．

入力を平文 m，暗号化関数を $E(\cdot)$ とした時，Paillier 暗号は次の三つの特徴を持つ．

1. 異なる暗号文
   同じ平文から暗号文を生成しても，乱数 r が異なるため生成された暗号文は一致しない．

2. 平文の加算
   暗号文同士を乗算することによって，平文の加算を行うことが出来る．

$$E(m_1) * E(m_2) = E(m_1 + m_2).$$

3. 平文の乗算
   暗号文に平文をべき乗する事によって，平文の乗算を行うことが出来る．

$$E(m_1)^{m_2} = E(m_1 * m_2).$$

2.2 Somers' d 類以度

Somers' d 類以度 [4] は，複雑な計算を必要とせず，平均に対して正か負かだけの定性的な値に基づいて計算することができる．

Somers' d 類以度は，C(Concordant), D(Discordant), T(Tied), N(Number of item), を数えることによって計算される．表 1 の値を例に説明していく．

表 1: Somers' d 類以度のサンプルデータ

	i_1	i_2	i_3	i_4	i_5	i_6	$\overline{r}$
u	4	2	5	1	5		3.4
v	1	4	3		4		3
$f_{u,*}$	0.6	−1.4	1.6	−2.4	1.6		
$f_{v,*}$	−2	1	0		1		
	D	D	T	T	C	T	

まず，ユーザ u のアイテム i の正規化評価値 $f_{u,i}$ を計算する．

$$f_{u,i} = r_{u,i} - \overline{r}_u \tag{2}$$

次に，ユーザ u の $f_{u,i}$ と v の $f_{v,i}$ の一致度を C, D, T, N の四値で表す．

1. $C_{u,v}$ **（Concordance）**

$$f_{u,i} > 0 \wedge f_{v,i} > 0, \text{ or}$$
$$f_{u,i} < 0 \wedge f_{v,i} < 0.$$

を満たすアイテムの数.

2. $D_{u,v}$（**Discordance**）

$$f_{u,i} > 0 \wedge f_{v,i} < 0, \text{ or}$$
$$f_{u,i} < 0 \wedge f_{v,i} > 0.$$

を満たすアイテムの数.

3. $T_{u,v}$（**Tied**）

$$f_{u,i} = 0 \vee f_{v,i} = 0, \text{ or}$$
$$r_{u,i} = \phi \vee r_{v,i} = \phi.$$

を満たすアイテムの数.

4. **N（Number of item）**
N は $N = C + D + T$ アイテムの総数を満たす.

表 1 に f を計算例を示す.
u と v の Somers' d 類以度は,

$$d_{u,v} = \frac{C - D}{N - T}$$

出計算される．ただし，$N = T$ のとき $d_{u,v} = 0$ とする.

表 1 の例では，$C = 1, D = 2, T = 3, N = 6$ となり，類以度は $d_{u,v} = (1-2)/(6-3) = -0.3$ と計算される.

Somers' d 類以度は −1 から 1 の値を取り，1 に近いほど一致度は高い,

Lathia らによると，ピアソン相関係数を用いた場合と同等の精度を得ることが出来る [5]．その MAE（Mean Abusolute Error）はピアソン相関係数が 0.826，Somers' d 類以度が 0.824 となっている.

また，推移性（Transitivity）を持っている．推移性とは，$A \subset B, C \subset A$ である時，$C \subset B$ が成り立つことを言う．この類以度においての推移性とは，f_u, f_v, f_c について図 1 に示す関係が成り立つ.

2.3 一致度（Concordant）値の区間推定 [5]

Lathia らはこの推移性を利用した水平分割環境における間の秘匿協調フィルタリング方式を提案している [5].

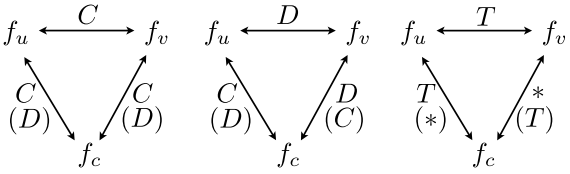

図 1: 左:C の例, 中:D の例, 右:T の例

評価値をランダムに与えた擬似ユーザを作成し，そのユーザとの類以度のみ公開することによって，プライバシを保護したまま対象ユーザ間の Somers' d 類以度を予測する．但し，この擬似ユーザ c は全てのアイテム i について次の条件を満たす.

1. 評価値に欠損値は存在しない．$r_{c,i} \neq \phi$.
2. 評価値は評価値の平均と一致しない．$r_{c,i} \neq \overline{r}_c$.

ユーザ u, v と c 間を，C_{uc}, C_{vc} とした時，C_{uv} の取り得る値を $\underline{C_{uv}} \leq C_{uv} \leq \overline{C_{uv}}$ を満たす区間 $[\underline{C_{uv}}, \overline{C_{uv}}]$ で表す．$[\underline{D_{uv}}, \overline{D_{uv}}]$ $[\underline{T_{uv}}, \overline{T_{uv}}]$ も同様にする.

T_{uv} のとりうる範囲 $[\underline{T_{uv}}, \overline{T_{uv}}]$ は,

$$\max(T_{uc}, T_{vc}) \leq T_{uv} \leq \min(T_{uc} + T_{vc}, N) \quad (3)$$

である.

C_{uv} のとりうる範囲 $[\underline{C_{uv}}, \overline{C_{uv}}]$ は,

$$\max(C_{uc} + C_{vc} - N, 0) + \max(D_{uc} + D_{vc} - N, 0)$$
$$\leq C_{uv} \leq \min(C_{uc}, C_{vc}) + \min(D_{uc}, D_{vc}) \quad (4)$$

となる[1].

この二つの性質と，$N = C + D + T$ の性質を利用して，D のとりうる範囲 $[\underline{D_{uv}}, \overline{D_{uv}}]$ を推定すると,

$$\max(N - (\overline{C_{uv}} + \overline{T_{uv}}), 0) \leq D_{uv} \leq N - (\underline{C_{uv}} + \underline{T_{uv}}) \quad (5)$$

となる.

これらの性質を利用して算出した範囲の平均値 $\hat{C} = (\underline{C}_{uv} + \overline{C}_{uv})/2, \hat{D} = (\underline{D}_{uv} + \overline{D}_{uv})/2, \hat{T} =$

[1] [5] では，$C_{uc} + C_{vc} - N$ のみによる下限が示されているが，式 (4) の方がよりタイトな下限である.

$(\underline{T}_{uv} + \overline{T}_{uv})/2$ を推測値として，予測類以度 $\hat{d}$ を次式で計算する．

$$\hat{d}_{uv} = \frac{\hat{C}_{uv} - 0.5\hat{D}_{uv}}{N - \hat{T}_{uv}} \tag{6}$$

3 提案方式

3.1 モデル定義

$\mathcal{A}$,$\mathcal{B}$,$\mathcal{C}$ の三組織間で，データベースが垂直分割されているときの例を 2 に示す．ここで，i_j はアイテム，u_k はユーザを示す．例えば，ユーザ 4 がアイテム 2 に与えた評価は 4 となる．c は前述した擬似ユーザを表す．

表 2: データベースの例

	$\mathcal{A}$			$\mathcal{B}$			$\mathcal{C}$		
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
u_1	4	3	5	1	1		3	1	1
u_2	1	4			4	1	1		4
u_3	1		1	2	2	2	4	3	3
u_4	4	3		5	3	5	2	3	2
c_1	+	−	−	+	−	−	+	+	+
c_2	+	+	−	−	+	+	+	−	+

3.2 提案方式：一致度の推定

k 人の類似ユーザを作り，そのユーザとの nk 組の一致度を求める．

次に，一致度の推移性を利用し，任意の n^2 組の一致度を推定する．必要な通信コストは $nk = O(n)$ であり，効率が良いが，推定される一致度は誤差を含む．

Lathia らの提案した擬似ユーザを利用して $O(n^2)$ を n に削減する方法を提案する．まず，Algorithm**??**をつかって，三組織間で擬似ユーザ c との全域的一致度 C, D, T, N を共有する．

ここで，擬似ユーザとの各区間 $[\underline{C}_{uv,c_1}, \overline{C}_{uv,c_1}]$, $[\underline{D}_{uv,c_1}, \overline{D}_{uv,c_1}]$, $[\underline{T}_{uv,c_1}, \overline{T}_{uv,c_1}]$ が得られる．さらに，擬似ユーザを一人だけではなく，c_2, c_3 と増やしていくことによって，区間を狭めていくことが出来る．

Algorithm 1 一致度の予測方法

入力: $\mathcal{A}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{A}})$，$\mathcal{B}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{B}})$，$\mathcal{C}$ の評価値 $r_{u,i}$ ，$(i \in I_{\mathcal{C}})$.
出力: 擬似ユーザと各ユーザの C,T,D,N $\hat{C}_{u,c}, \hat{D}_{u,c}, \hat{T}_{u,c}, N$.

1. $\mathcal{A}, \mathcal{B}, \mathcal{C}$ は条件を満たす擬似ユーザ $c_1, \ldots, c_k$ の評価値を各々のアイテムについてランダムに決める．
2. $\mathcal{A}, \mathcal{B}, \mathcal{C}$ は全ての n 人のユーザ u について，式 2 で正規化し，擬似ユーザとの間の一致度 C, D, T, N を計算する．
3. $\mathcal{A}, \mathcal{B}, \mathcal{C}$ は Alg. 2 を用いて，擬似ユーザ $c_1, \ldots, c_k$ との一致度（C, D, T, N）を総和して，共有する．
4. 式 (3, 4, 5) の区間推定を行い，全アイテム間 u, v について $[\underline{C_{uv,c_1}}, \overline{C_{uv,c_1}}], \ldots, [\underline{C_{uv,c_k}}, \overline{C_{uv,c_k}}]$ を求め，その集約区間 $\underline{C_{uv,*}} = \vee_k \underline{C_{uv,c_k}}$, $\overline{C_{uv,*}} = \vee_k \overline{C_{uv,c_k}}$ を求める．D と T についても同様に算出する．

Algorithm 2 秘匿総和

入力: $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ がそれぞれ持つ値 a, b, c. $\mathcal{A}$ の公開鍵で暗号化，復号する関数 $E(\cdot), D(\cdot)$.
出力: $s = a + b + c$.

1. $\mathcal{B}$ は $E(b)$ を $\mathcal{C}$ へ送る．
2. $\mathcal{C}$ は $x = E(b)E(c)$ を $\mathcal{A}$ へ送る．
3. $\mathcal{A}$ は $s = D(x) + a$ を出力する．

例えば，三つの擬似ユーザ c_1, c_2, c_3 から，$[\underline{C}_{uv,c_1}, \overline{C}_{uv,c_1}]$, $[\underline{C}_{uv,c_2}, \overline{C}_{uv,c_2}]$, $[\underline{C}_{uv,c_3}, \overline{C}_{uv,c_3}]$, が得られた時，**集約区間**は

$$[max(\underline{C}_{uv,c_1}, \underline{C}_{uv,c_2}, \underline{C}_{uv,c_3},), \\ min(\overline{C}_{uv,c_1}, \overline{C}_{uv,c_2}, \overline{C}_{uv,c_3})] \quad (7)$$

となり，より狭められた区間で算出される．

区間を狭めることによって，正確に一致度を予測することが出来る．

3.3 類以度推定

類以度を求める時は，このあらかじめ計算された擬似ユーザ c_1 との一致度を利用して式 (6) から予測類以度 $\hat{d}$ を求める．この方法を使うことによって，以降，暗号通信をせずにユーザ間類以度を計算することが出来る．

4 評価

ユーザ数6, アイテム数17, レーティング数102のランダムな評価データを作成し，提案方式の性能を評価する．

4.1 区間の変化

$k = 5$ の擬似ユーザを利用して予測した C の区間と，真の C の値を図 3 に示す．異なる 5 組のユーザ間の一致度を示している．エラーバーで示されているのが，各擬似ユーザを利用して予測した時の C のとりうる範囲である．各擬似ユーザの評価値に依って範囲がばらついていることが分かる．

これらの集約区間を図 3 に示す．

擬似ユーザ数 k を増やした時の集約区間の幅，$\overline{C_{uv} - \underline{C_{uv}}}$ の変化を図 4 に示す．$k = 1$ のときは C 範囲が 18 となっているのに比べ，$k = 5$ の時は C の範囲が 7 と半分以下に狭められている．

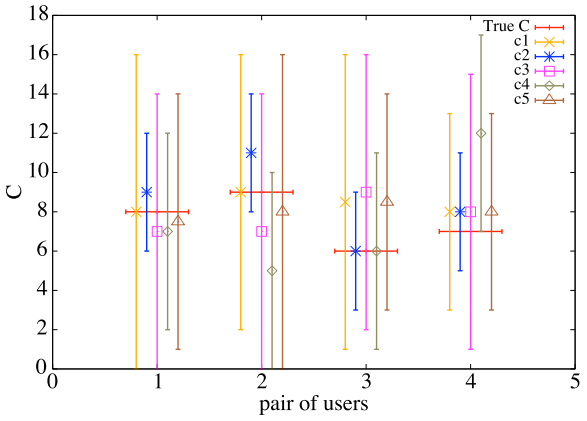

図 2: 各擬似ユーザから予測された C の区間範囲と，真の C

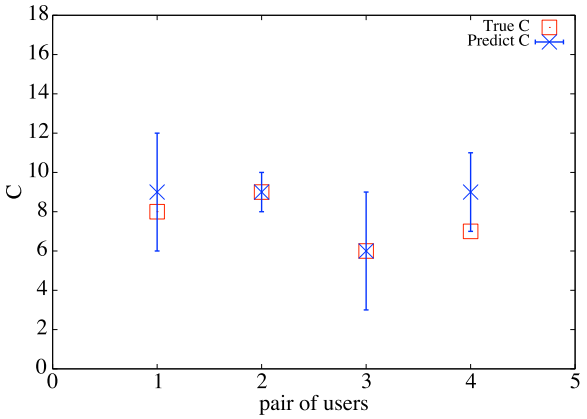

図 3: 擬似ユーザ数 $k = 5$ の集約区間

4.2 予測類以度と真の類以度の比較

集約をしない全域的な Somers'd 類以度とピアソン相関係数の散布図を図 5 に示す．この二つの類以度には正の相関が見える．

Somers'd 類以度と擬似ユーザ数 5 のとき予測された集約区間から算出された Somers'$\hat{d}$ 類以度の散布図を図 6 に示す．この散布図の相関は小さい．

5 まとめ

三者間以上で垂直分割されたデータベース環境において，互の値を秘匿したまま通信効率の良い一致度の秘匿分散計算方式を提案した．一致度という類以度の推移性を利用し，この n^2 のコストを，n に削減する近似方法を提案し，擬

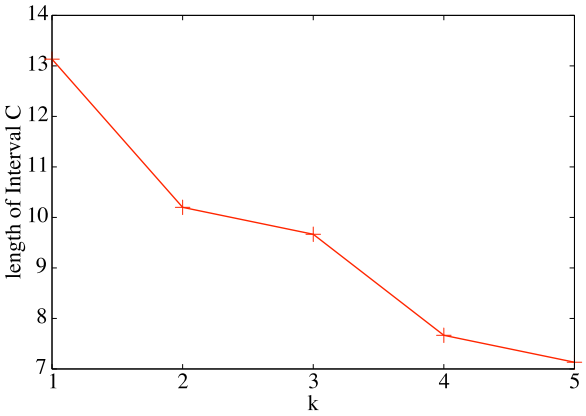

図 4: 擬似ユーザ数 k についての集約区間の変化

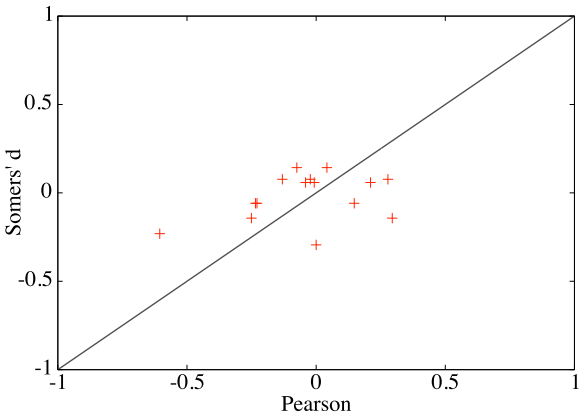

図 5: Somers'd 類以度とピアソン相関係数の散布図

似ユーザを増やすことによって予測値の区間を狭めることをシミュレーションにより確認した．

今後の課題として，擬似ユーザ数 $k = 5$ のとき予測された Somers'$\hat{d}$ 類以度の間が低相関である原因の調査と改善を行う．

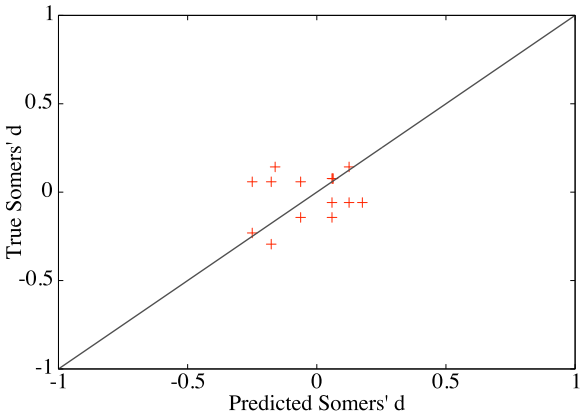

図 6: Somers'd 類以度と予測された Somers'$\hat{d}$ 類以度の散布図

参考文献

[1] 高島 秀佳, 山岸 英貴, 平澤 茂一, “欠損値推定による協調フィルタリング手法” 情報科学技術フォーラム一般講演論文集, Vol. 4, No. 1, pp. 15-16, 2005. 人工知能学会誌, Vol. 23 No. 2, pp. 248-263, 2008.

[2] 平山 巧馬, 小柳 滋, “協調フィルタリングにおける相関係数法の予測性能向上”, 電子情報通信学会論文誌. D, 情報・システム, Vol J90-D(2), 223-232, 2007.

[3] John S. Breese, David Heckerman, and Carl Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, Proceedings of the 14th conference on Uncertainty in Artificial Intelligence, pp. 43-52, 1998.

[4] A. Agresti, “Analysis of Ordinal Categorical Data”, 1984.

[5] Neal Lathia, Stephen Hailes, and Licia Capra, “Private Distributed Collaborative Filtering Using Estimated Concordance Measures”, In ACM 2007 Conference on Recommender Systems (RecSys). Minneapolis, Minnesota, pp. 1-8, USA. October 19-20, 2007.