情報処理学会論文誌 Vol.56 No.9 1754–1769 (Sep. 2015)

セキュアスムージング手法による組織間プライバシ保護リコメンドシステム

山口 高康[1,a)] 寺田 雅之[1,b)]

受付日 2014年12月4日, 採録日 2015年6月5日

概要：近年，顧客の属性や購買行動の統計的性質に基づいたリコメンドが注目されている．しかし，多くの場合，これらの情報は組織間で分散してしまっている．組織間の壁を越えてリコメンドを実現するためには，従来の PPDM 技術では不十分であり，分散形態の運用，分散形態でのプライバシ保護，処理コスト，計算精度に関する課題を解決する必要がある．本稿では組織間プライバシ保護リコメンドシステムの構成法と，PPDM 環境に適した新たなスムージング手法を提案することで，これらの課題を解決する．これにより，必要な情報をすべて保有できる大きな組織だけでなく，小さな組織もリコメンドを行うことが可能となる．

キーワード：プライバシ保護データマイニング，リコメンデーション，セキュアマッチング，秘匿内積

A Privacy-preserving Inter-organizations Recommendation System Based on A Secure-somoothing Method

Takayasu Yamaguchi[1,a)] Masayuki Terada[1,b)]

Received: December 4, 2014, Accepted: June 5, 2015

Abstract: Recommendation systems based on the statistical property of some personal profile and his purchase history receive a great amount of attention in the trade and the industry these days. Nevertheless, the information indespensable for recommendations is most commonly scattered or separated among organizations, and besides, conventional PPDM techniques are insufficient. Therefore, we make four breakthroughs, 1) working it in the distributed or cross-organization situation, 2) ensuring the privacy protection for each person in the situation, 3) suppressing the computational complexity required for the protection, 4) enhancing the precision of the recommendations quickly and securely. These breakthroughs could be able to provide regular shops and customers to the opportunity that they obtain the effective recommendations.

Keywords: privacy preserving data mining, recommendation, secure matching, secure scalar product

1. はじめに

近年の商取引において，顧客の属性や購買行動の統計的性質に基づいた商品の推薦（統計的リコメンド）は，企業の売上拡大のための有力な手段となっている．しかし，現在の統計的リコメンドは，推薦を行う組織が必要な情報をすべて保有していることを前提としている．この前提は，有力組織が商取引全体をコントロールする場合には成立する．しかし，複数の組織が対等な関係で連携する場合には，推薦に必要な情報が複数の組織に分散し，現在の統計的リコメンドの前提が成立しないことがある．また，推薦に必要な情報をすべて保有することが可能であっても，個人情報の管理などの理由から，分散保有が望ましい場合もある．たとえば，会員の属性情報のような機密性の高い個人情報については，個別の店舗で管理することは望ましくなく，クレジット会社などのプロバイダに管理を求める場合が多い．

[1] 株式会社 NTT ドコモ先進技術研究所
Research Laboratories, NTT DOCOMO, Inc., Yokosuka, Kanagawa 239–8536, Japan
[a)] yamaguchitaka@nttdocomo.com
[b)] teradam@nttdocomo.com

以上の背景から，複数の組織に分散した情報を用い，各組織が他の組織にプライベートな情報を開示することなく，1 組織がすべての情報を保有している場合と同等あるいはそれに近い精度の推薦を行う技術が望まれる．本稿では，このような技術を組織間プライバシ保護リコメンド技術と呼ぶことにする．組織間プライバシ保護リコメンド技術が実現すれば，一部の有力組織だけでなく，多くの組織が統計的リコメンドを実施可能となると期待される．また，有力組織にとっても，統計的リコメンドの適用範囲を拡大することが可能と考えられる．

しかしながら，組織間プライバシ保護リコメンド技術を実現するためには，従来の PPDM（Privacy Preserving Data Mining）技術だけでは不十分であり，下記に示す，分散形態の運用，分散形態でのプライバシ保護，処理コスト，計算精度に関する課題を解決する必要がある．

（課題 1）分散形態の運用

従来のプライバシ保護型リコメンド技術は，推薦する側が推薦に必要な情報をすべて保有していることを前提としていた．しかし，組織間リコメンドでは，推薦に必要な情報を複数の組織が分散保有している場合が多い．たとえば，2 章で述べるプロバイダ，店舗，来訪客の三者モデルの場合，来訪客への推薦に必要な情報をプロバイダと店舗が分散保有している．そのため，組織間リコメンドに適した三者間のシステムおよびプロトコルを設計する必要がある．

（課題 2）プライバシ保護

上記の三者システムにおいて，各々の者は，他の二者に対して，自分のプライベートな情報を秘匿可能とする必要がある．

（課題 3）処理性能

実運用時のデータ数を想定したうえで，そのデータを所定の許容時間内で処理できる方式とする必要がある．特に，三者モデルの場合，各々の二者間での処理量および許容時間が大きく異なる可能性があり，それらの特性を考慮した方式が必要である．

（課題 4）計算精度

非 PPDM 環境では一般的であるスムージング最適化などの推薦精度向上のための方法が，プライバシ保護のための PPDM 環境では適用困難であるため，PPDM 環境に適した推薦精度向上の手法が必要である．

本稿では，上記の課題 1 から課題 4 に沿って実用的な組織間リコメンドの要件を整理し，それらの要件を満たすシステムと技術を提案する．本稿の内容の進歩性は，下記のシステム構成および要素技術にあると考える．

（進歩性 1）組織間プライバシ保護リコメンドシステム

上記の課題 1，課題 2，課題 3 を同時に満たすシステム構成法を提案し，2 種類の標準的なベンチマークである Book-Crossing および Play Tennis データセットを用いて，その実用性を明らかにした．

（進歩性 2）セキュアスムージング手法

上記の課題 4 を満たす PPDM 環境に適した新たなスムージング手法を提案し，上記の標準ベンチマークを用いて，その精度と処理時間を明らかにした．

本稿では，まず 2 章で組織間プライバシ保護リコメンドのユースケースを設定し，これに沿って要件を整理する．次に，3 章で，従来方式とその問題点について述べ，要件に照らして課題を抽出する．4 章では組織間プライバシ保護リコメンドシステムを構成し，5 章ではセキュアスムージング手法を提案する．6 章では 2 種類の標準ベンチマークを用いた実験評価で処理性能と計算精度を明らかにし，7 章でそれらを要件に照らして考察する．8 章では提案方式の安全性をプライバシ保護の要件に照らして考察し，最後に 9 章で PPDM 組織間リコメンド研究における本稿の進歩性をまとめる．

2. 組織間プライバシ保護リコメンドのユースケースと要件

2.1 ユースケース

会員情報を管理しているプロバイダと，小売業を営んでいる店舗と，店舗へ買い物にやってきた来訪客の三者での運用を想定する．プロバイダは会員情報として，それぞれの会員について会員番号と属性情報（年代，性別，住所など）の組を管理している（**表 1**）．店舗は，プロバイダと提携しており，売上記録として過去に来店した会員の会員番号と売り上げた商品情報を管理している（**表 2**）．来訪客は，自身の属性情報を（携帯電話やスマートフォンなどの）計算能力を持つ携帯型電子機器に保持している（**表 3**）．ただし，来訪客はプロバイダの会員とは限らない．

上記をふまえて，店舗は来訪客に対し，来訪客の属性をふまえたリコメンドを行いたい．安全性を考慮しなければ，これは以下のようにして達成できる．

(1) プロバイダと店舗は，プロバイダの会員情報と店舗の売上記録のそれぞれのデータベースを会員番号を介して結合する．

(2) 店舗は，結合したデータベースから，属性と購入傾向

表 1 プロバイダの会員情報
Table 1 Provider's member information.

会員番号	属性	
	年代	性別
1	20	女性
3	30	男性
4	30	女性
5	40	男性
6	20	女性
7	40	女性
8	40	男性

表 2 店舗の売上記録
Table 2 Shop's sales record.

会員番号	商品
1	本 A
2	本 A，本 B
3	本 A
4	本 B
6	本 A
7	本 B

表 3 来訪客の属性情報
Table 3 Customer's attributes.

属性	
年代	性別
30	男性

の統計的な関係を推定する．

(3) 来訪客は，来店時に店舗に対して属性を提示する．

(4) 店舗は，上記で推定した属性と購入傾向の統計的関係を用い，来訪客から提示された属性を用い，来訪客が購入する可能性が高い商品を推定し，来訪客に対して推薦する．

(5) 来訪客がプロバイダの会員であった場合には，購入時に会員番号（会員証など）を提示し，店舗は売上記録に購入情報を反映させる．

しかし，プロバイダと店舗と来訪客の三者は，それぞれの立場から安全性に対して次の要求を持つ．プロバイダは会員から預かっている会員情報を漏洩したくない．店舗は店舗の売上記録を漏洩したくない．来訪客は，初めて立ち寄った店舗でも的確なリコメンドを受けたいが，自身の属性などのプライバシ情報は漏洩させたくない．

続いて，利便性への要求に目を向けると，来訪客は来店時にただちにリコメンドを受けたい．店舗は，定期的に新商品を取り入れたり売れ行きの良い商品を選別するなどして，魅力のあるリコメンドを保ちたい．

以下では，プロバイダや店舗が取り扱うデータの規模や内容を，代表的な電子マネーの 1 つである nanaco[*1] にならって想定する．

まず，プロバイダの会員情報について．2014 年 5 月末時点の会員数は 3,108 万人である．これより，会員情報の会員数（以後，N で表す）は 10^7～10^8 人を想定する．会員情報の属性については，年代と性別と住所を想定し，属性の種類数（以後，W で表す）は 3 種類を想定する．また，年代は 8 区分，性別は 2 区分，住所は 47 区分（都道府県）とし，属性値すなわち属性のとりうる値の種類数（以後，V で表す）は 57 種類を想定する．

次に，店舗の売上記録について．全国 14.5 万カ所で 1 カ月に 115 百万件の電子マネー決済が行われていることから，1 カ月 1 カ所あたりの売上情報の記録数の平均は 793 件である．この値は店舗の規模，所在地，業種，来訪客のリピート率などにより大きく異なると考えられるため，売上記録に含まれる人数（以後，M で表す）は 10^2～10^5 人を想定する．商品の種類数については，飲食店のメニューはたかだか 100 種類を想定すれば十分と考えられるが，コンビニエンスストアの商品は 2,800 種類もある[*2]．店舗の規模，所在地，業種などにより大きく異なると考えられるため，商品の種類数（以後，L で表す）は 10^1～10^4 種類を想定する．また，全国で 1 回あたりの決済金額は 907 円/回であるので[*3]，一度の来店で購入する商品は数種類であると考えられる．売上記録を月単位に締めるとすると，来訪客が毎週来店したとしても 1 カ月にユーザが購入する商品は十数種類であると考えられるので，売上記録に含まれる 1 人あたりの商品の種類数（以後，G で表す）は 10 種類を想定する．

表 4 データの規模や内容
Table 4 The parameters to estimate the computational complexity.

データの種類	データの規模のパラメータ	値の範囲
プロバイダの会員情報	会員数 (N)	10^7～10^8 人
	属性の種類数 (W)	3 項目
	属性値の種類数 (V)	57 種類
店舗の売上記録 (1 カ月分)	売上記録に含まれる人数 (M)	10^2～10^5 人
	商品の種類数 (L)	10^1～10^4 種類
	1 人あたりの商品の種類数 (G)	10 種類

以上，プロバイダや店舗が実際に取り扱う必要のあるデータの規模や内容を整理すると**表 4** となる．

2.2 データの表現

前節のユースケースで示したデータを PPDM で処理するためのデータ表現の一例を示す．4 章以降で述べる提案手法は，この表現を前提としている．なお，以下の説明では，属性と属性値という用語を用いる．属性は年代，性別などである．属性値は属性のとりうる値であり，20 代，30 代，40 代，男性，女性などである．

表 1，表 2，表 3 の属性情報と商品情報は多値であるが，これらを**表 5**，**表 6**，**表 7** のように二値で表現する．さらに，これらの二値情報をベクトルとマトリクスにより表現する．

プロバイダが保有している表 1 の会員情報は，表 5 のように二値化され，さらに**図 1** のように，会員番号を表すベクトル $\boldsymbol{t}$ と，属性情報を表すマトリクス X により表現される．会員番号を表すベクトル $\boldsymbol{t}$ は長さ N である（N はプ

[*1] nanaco ホームページ，‘電子マネー nanaco とは’，http://www.nanaco-net.jp/corporate/company/

[*2] セブンイレブンホームページ，‘セブンイレブンまるわかり豆知識 今日も寄りたくなっちゃう！そのワケは？’，http://www.sej.co.jp/products/trivia/trivia_09.html

[*3] 日本銀行，“決済システムリポート”，2013 年 10 月

表 5 プロバイダの会員情報の二値表現
Table 5 The binarization of provider's dataset.

$\boldsymbol{t}$	X				
会員番号	年代：20	年代：30	年代：40	性別：男性	性別：女性
1	1	0	0	0	1
3	0	1	0	1	0
4	0	1	0	0	1
5	0	0	1	1	0
6	1	0	0	0	1
7	0	0	1	0	1
8	0	0	1	1	0

表 6 店舗の売上記録の二値表現
Table 6 The binarization of shop's dataset.

$\boldsymbol{u}$	Y	
会員番号	商品：本 A	商品：本 B
1	1	0
2	1	1
3	1	0
4	0	1
6	1	0
7	0	1

表 7 来訪客の属性情報の二値表現
Table 7 The binarization of customer's data.

$\acute{\boldsymbol{x}}$				
年代：20	年代：30	年代：40	性別：男性	性別：女性
0	1	0	1	0

$$(1,3,4,5,6,7,8)$$

(a) 会員番号のベクトル表現 $\boldsymbol{t}$

1	0	0	0	1
0	1	0	1	0
0	1	0	0	1
0	0	1	1	0
1	0	0	0	1
0	0	1	0	1
0	0	1	1	0

(b) 会員番号のマトリクス表現 X

図 1 プロバイダの会員情報のベクトルとマトリクスによる表現
Fig. 1 The vector and matrix expression of provider's dataset.

ロバイダに登録された会員の数）．$\boldsymbol{t}$ の n 番目の要素を t_n とする．表 5 の例では，$\boldsymbol{t}=(1,3,4,5,6,7,8)$，$N=7$ であり，$t_1=1$，$t_2=3$，...，$t_N=8$ である．

プロバイダが保有する属性情報は，図 1 (b) のように，N 行 V 列のマトリクス X により表現される．V は属性値の種類数（属性のとりうる値の総数）であり，図 1 (b) の場合は 5 である．v は属性値の識別子である．たとえば，$v=1$ は 20 代，$v=3$ は男性を表す．属性の種類数を W とする．表 5 の場合，属性は 2 種類（年代と性別）であるため，$W=2$ である．属性の識別子を w とする．w は 1 または 2 をとりうる．

マトリクス X の n 行, v 列の値を $x_{n,v}$ で表す．$x_{n,v}=1/0$ は，会員 t_n が v 番目の属性値を有する/有しないことを表す．図 1 (b) の場合，$x_{1,1}=1$，$x_{2,1}=0$，...，$x_{N,V}=0$ である．

店舗が保有する表 2 の売上記録は，表 6 のように二値化され，さらに，会員番号を表すベクトル $\boldsymbol{u}$ と商品情報を表すマトリクス Y により表現される．会員番号を表すベクトル $\boldsymbol{u}$ は長さ M である（M は売上記録に含まれる会員の数）．$\boldsymbol{u}$ の m 番目の要素を u_m とする．表 6 の例では，$\boldsymbol{u}=(1,2,3,4,6,7)$，$M=6$ であり，$u_1=1$，$u_2=2$，...，$u_M=7$ である．

店舗が保有する商品情報は，M 行 L 列のマトリクス Y により表現される．マトリクス Y の m 行, l 列の値を $y_{m,l}$ で表す．l は商品の識別子である．$y_{m,l}=1/0$ は，会員 u_m が l 番目の商品を購入した/していないことを表す．表 6 の場合，$y_{1,1}=1$，$y_{2,1}=1$，...，$y_{M,L}=1$ である．購入数量を直接属性値として扱うことも考えられるが，外れ値による精度の低下や，ある商品を流行させるために仲間内で多数購入するチーティングの懸念を考慮し，本稿では購入した/しないの二値のみで扱うとした．なお，購入数量を扱う方法としては，1 個購入，2 個購入，...，10 個以上購入を別属性とし，各々を 0/1 で表現することなどが考えられる．

最後に，来訪客が保有している表 3 の属性情報を，表 7 のように二値化し，さらに，この情報を長さ V のベクトルで表す．表 7 の例では，$\acute{\boldsymbol{x}}=(0,1,0,1,0)$ である．$\acute{x}_v=1/0$ は，来訪客が v 番目の属性値を有する/しないことを表す．

2.3 要件

これまで述べたユースケースをふまえて，組織間プライバシ保護リコメンドに対する要件をまとめると以下のようになる．

（要件 1）プライバシ要件
プロバイダが保有する顧客の属性情報，店舗が保有する売上情報，来訪客の属性情報を，各々他の二者に対して秘匿可能であること．

（要件 2）処理性能要件
来訪客が店舗に来たときの推薦プロトコルはリアルタイム（あるいは準リアルタイム）の必要があるので，その許容時間（T_1 とする）は小さく，たとえば 5 秒とする．表 4 に示した既存の電子マネーのデータの規模や内容で，この T_1 の制約を守ること．

（要件 3）計算精度要件
非 PPDM 環境と同等あるいはそれに近い推薦精度であること．

本稿では，後述する組織間プライバシ保護リコメンドシステムの構成法と，PPDM 環境に適した新たなスムージン

グ手法により，これらの要件を適える．

3. 従来方式とその問題点

組織間プライバシ保護データマイニングの研究は Yao らから始まる．組織間プライバシ保護データマイニングの基本演算の 1 つに，二者が各々保有する集合の間で積集合を求める演算がある．これは組織間クロス集計を行うために用いられる．Yao らは，多項式評価によって，二者間で集合を秘匿したまま積集合を求める手法を提案したが，計算量が多いという問題がある [1]．Goldreich の方式は，この多項式評価を三者間に拡張しているが，計算量はさらに増えてしまう [2]．Agrawal らはハッシュタグの照合により積集合を算出する方式を考案した [3]．千田らはこれを高速化し，効率的に行える方式を考案した [4]．

積集合を用いた組織間プライバシ保護データマイニング手法の 1 つにナイーブベイズがある．Vaidya らは秘匿内積で求めた積集合を用いるナイーブベイズの手法を提案した [5]．しかし，秘匿内積 [6], [7] は組織間で同じ長さのベクトルの積和を求めるため，組織間でのデータが同期している必要があり，非同期の場合は計算コストや安全性に問題が生じてしまう．そこで菊池らは，非同期でも少ない計算コストで積集合を求められる Agrawal らのタグ照合方式を用いてこの問題を回避した．そして，タグ照合方式で積集合の算出を繰り返すうちに他の組織に情報が漏れる差分攻撃の問題を指摘し，チャフを混在させてこの問題を回避する，チャフ付き照合タグ方式を提案した [8]．

しかし，菊池らの方式は組織間プライバシ保護データマイニングにフォーカスしており，組織間プライバシ保護リコメンドについては言及していなかった．すなわち，複数の組織のデータを統合マイニングする部分にフォーカスしており，複数の組織のデータを統合利用してリコメンドを行う問題は取り上げていない．たとえば，2 組織の場合，菊池らの方式は 2 組織間のデータマイニングプロトコルにフォーカスしており，2 組織のデータを利用してエンドユーザにリコメンドを行う三者間プロトコルは提案していない．

また，Vaidya や菊池らが用いているナイーブベイズの手法は，少ない計算量で高い計算精度が得られる手法の 1 つであるが，非 PPDM 環境ではスムージングと呼ばれるさらに計算精度を高める工夫もある．たとえば，Minka の LOO（leave-one-out）尤度を用いる手法により，適切にスムージングを行える [9]．しかし，LOO 尤度を算出するためには，積集合から 1 件ずつすべてのデータを取り出して，それぞれの尤度を算出しなければならない．すなわち，PPDM 環境で安全に組織間の積集合を求められたとしても，LOO を行うたびに 1 件ずつすべてのデータが相手に漏れてしまうという問題が生じる．これらの処理を安全に行うためには，いかなる計算も安全に行うことができる万能関数計算機 [10] や，マルチパーティプロトコルによるパラメータ推定 [11] など，非常に計算コストが高い方式が必要となってしまうため，PPDM 環境のセキュアなスムージングはこれまで実現できなかった．

4. 組織間プライバシ保護リコメンドシステム

4.1 設計方針

プロバイダが取り扱うデータの規模はプロバイダの会員数に等しく $N = 10^7 \sim 10^8$ と非常に大きいため，来訪客が来店したときの推薦プロトコルにプロバイダを含めることは現実的ではない．そこで，プロバイダと店舗の間で定期的に二者間プロトコルを実行することで，店舗が当該店舗での推薦に必要なデータを保有しておき，これを用いて来訪客が来店したときの推薦プロトコルを動作させる．このデータの規模は，商品ごとのデータとなるので，店舗で扱っている商品数に等しく $L = 10^1 \sim 10^4$ で済む．

プロバイダ–店舗プロトコルにより，このデータを作成するまでの許容時間（以後，T_2 で表す）は，その実行間隔から決定する必要がある．ユースケースで述べたようにこの実行間隔を 1 カ月とすると，プロバイダを含むプロトコルを 1 カ月以内で実行可能となるようにシステムを設計する必要がある．店舗が取り扱うデータの規模は店舗の売上記録に含まれる人数に等しく $M = 10^2 \sim 10^5$ である．プロバイダのデータと店舗のデータは 1 対 1 で対応しておらず非同期なので，プロバイダ–店舗プロトコルには効率的な方式を採用する必要があり，千田らの高効率なセキュアマッチングプロトコル [4] を用いる．

ただし，タグ照合方式は，両者の持つデータベースを同期させる際に差分攻撃が可能となるため，1) セキュアマッチングプロトコルの結果を店舗のみに開示するようにして，プロバイダから店舗に対する差分攻撃を防ぐ．2) プロバイダでサブサンプリングを行うようにして，店舗からプロバイダへの差分攻撃を防ぐ．サブサンプリングによる計算精度への影響については，サブサンプリングがランダムであるので，集計表から得られる確率の値の変化は小さく，その影響を無視できる．

来訪客が来店したときの推薦プロトコル（店舗–来訪客プロトコル）では，従来の標準的な PPDM 手法である秘匿内積プロトコル [6], [7], [12] を採用する．以上のように，プロバイダ–店舗プロトコルではセキュアマッチングプロトコル，店舗–来訪客プロトコルでは秘匿内積プロトコルを用いることで，基本的なプライバシ要件を満たす．

プロバイダ–店舗プロトコルにより，店舗がデータを生成する際にスムージングを行い，計算精度要件の達成を図る．ベイズの手法は，主観で設定する事前確率と観測データで算出する尤度とを用いて事後確率を算出する手法であり，事前確率と尤度をどのようにバランシングさせるかによって，識別の精度が左右されるという難しさがある．そ

表 8 プロバイダ–店舗プロトコルで店舗が手に入れる集計表（Φ）
Table 8 Cross-tabulation table.

属性	商品	
	本 A	本 B
20 代	2	0
30 代	1	1
40 代	0	1
男性	1	0
女性	2	2

こで増村らは，事前分布にディレクレ分布を用いて MAP 推定を行うスムージングを行い，最尤推定で発生する零確率問題を回避している [13]．事前分布を適切に定めることは，事前分布として用いるディレクレ分布のパラメータを適切に設定することである．ディレクレ分布のパラメータの設定方法と計算精度については，Minka によって解析されており，LOO 尤度を用いたシンプルなパラメータ設定により，高い計算精度が得られるとされている [9]．しかし，この LOO 尤度を用いたスムージングの手法は，データを秘匿しない非 PPDM 環境を前提としており，プライバシ保護を必要とする PPDM 環境では利用できない．そこで，PPDM 向けのスムージングを新たに提案する．本提案手法については，5 章で詳しく述べる．

以上に述べた方針から，処理性能要件は下記のように具体化される．

（要件 2a）プロバイダ–店舗の処理性能要件

プロバイダ–店舗プロトコルは，プロバイダの会員数 $N = 10^7 \sim 10^8$ と，店舗の売上記録に含まれる人数 $M = 10^2 \sim 10^5$ という条件下で，許容時間 T_2 の制約を満たすこと．

（要件 2b）店舗–来訪客の処理性能要件

店舗–来訪客プロトコルは，許容時間 T_1 の制約を満たすこと．

4.2 システム構成法と処理概要

プロバイダ，店舗，来訪客の三者を連携させるシステムを構成し，組織間プライバシ保護リコメンドを実現するための，三者間での処理の流れについて述べる．提案システムは，プロバイダのサーバ，店舗の PC，来訪客の端末により構成される．サーバにはプロバイダの会員情報として，会員番号ベクトル $\boldsymbol{t}$ および属性情報マトリクス X が格納されている．PC には店舗の売上記録として，会員番号ベクトル $\boldsymbol{u}$ および商品情報マトリクス Y，端末には来訪客の属性として $\hat{\boldsymbol{x}}$ が格納されている．

店舗は，セキュアマッチングを用いたプロバイダ–店舗プロトコルにより，当該店舗での推薦に必要なデータをマトリクス Φ の形で手に入れる．Φ は，V 種類の属性と L 種類の商品ごとに，プロバイダと店舗の間でマッチングできた会員の数を集計したマトリクス（**表 8**）である．

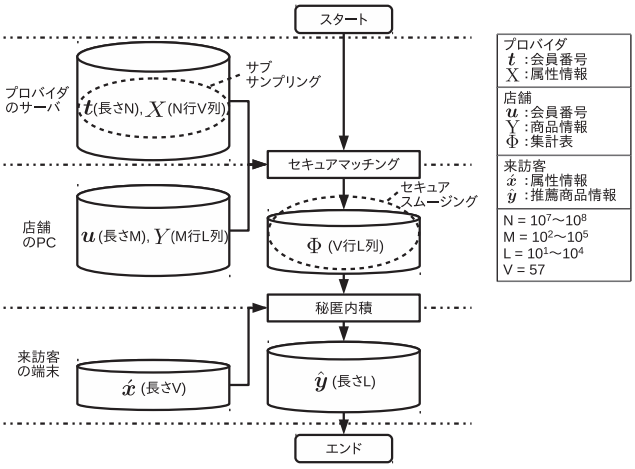


図 2 システム構成と処理概要
Fig. 2 System configuration diagram.

このマトリクスは，どの商品がどの属性の人にどれだけ売れたのかという店舗のノウハウに相当し，店舗はこれをふまえて，これからどの属性の人にどの商品を推薦するのかを決める．なお，店舗は月次で Φ を更新するので，差分開示により更新のたびにプロバイダの会員情報が店舗へ漏れる恐れがある．そこでプロバイダは，店舗へ $\boldsymbol{t}$ および X を送る際にランダムに会員を抜き（以後，サブサンプリングと表す），プロバイダの会員情報を守る．また，店舗がセキュアマッチングの安全性と処理性能を保ったままで推薦の精度も高めるべく，後述のセキュアスムージングを提案し，適用する．

来訪客は，秘匿内積を用いた店舗–来訪客プロトコルにより，店舗の集計表と来訪客の属性情報を互いに秘匿したまま，推薦情報 $\hat{\boldsymbol{y}}$ を手に入れる．$\hat{\boldsymbol{y}}$ は長さ L のベクトルであり，L 種類の商品ごとに推薦の度合いを数値の大小で表現している．来訪客は属性情報だけを秘匿内積すれば推薦商品情報が得られるので，たとえ来訪客がプロバイダの会員でなくても，推薦を得られる．

4.3 プロバイダ–店舗プロトコル

本節では，店舗が集計表を手に入れるまでの処理を例にあげて，セキュアマッチング [4] を用いたプロバイダ–店舗プロトコルについて説明する．

位数 q の巡回群 $\mathcal{G}$ と，その巡回群 $\mathcal{G}$ を値域とするハッシュ関数 $\mathbf{H}$ と乱数 $R \in Z_q$ を考える．プロバイダの乱数と店舗の乱数で互いに冪乗余すると，指数部における互いの乱数の和になるため，冪乗余の順序にかかわらず同じ値（以後，タグで表す）が得られる．この原理を用いて，プロバイダの会員番号のハッシュ乱数乗をさらに店舗の乱数乗したタグと，店舗の会員番号のハッシュ乱数乗をさらにプロバイダの乱数乗したタグの合致する個数を数えて，プロバイダと店舗で共通する会員の集計値を算出する．

本稿のユースケースにあてはめると，以下のようになる．

(1) プロバイダは属性ごとに 5 個の乱数 $R_p^{(v=1)}$, ..., $R_p^{(v=5)}$ を生成し，表 5 のマトリクス X の要素が 1 となっている 14 カ所について $\mathbf{H}(1)^{R_p^{(v=1)}}$, $\mathbf{H}(1)^{R_p^{(v=5)}}$, $\mathbf{H}(3)^{R_p^{(v=2)}}$, $\mathbf{H}(3)^{R_p^{(v=4)}}$, $\mathbf{H}(4)^{R_p^{(v=2)}}$, $\mathbf{H}(4)^{R_p^{(v=5)}}$, $\mathbf{H}(5)^{R_p^{(v=3)}}$, $\mathbf{H}(5)^{R_p^{(v=4)}}$, $\mathbf{H}(6)^{R_p^{(v=1)}}$, $\mathbf{H}(6)^{R_p^{(v=5)}}$, $\mathbf{H}(7)^{R_p^{(v=3)}}$, $\mathbf{H}(7)^{R_p^{(v=5)}}$, $\mathbf{H}(8)^{R_p^{(v=3)}}$, $\mathbf{H}(8)^{R_p^{(v=4)}}$ の 14 個のタグを生成し，これを属性ごとに分けてシャッフル[*4]して店舗に送る．

(2) 店舗は商品ごとに 2 個の乱数 $R_s^{(l=1)}$, $R_s^{(l=2)}$ を生成し，表 6 のマトリクス Y の要素が 1 となっている 7 カ所について $\mathbf{H}(1)^{R_s^{(l=1)}}$, $\mathbf{H}(2)^{R_s^{(l=1)}}$, $\mathbf{H}(2)^{R_s^{(l=2)}}$, $\mathbf{H}(3)^{R_s^{(l=1)}}$, $\mathbf{H}(4)^{R_s^{(l=2)}}$, $\mathbf{H}(6)^{R_s^{(l=1)}}$, $\mathbf{H}(7)^{R_s^{(l=2)}}$ の 7 個のタグを生成し，これをプロバイダへ送る．

(3) プロバイダは，(2) の 7 個のタグを 5 個の乱数でそれぞれ冪乗して 35 個のタグを生成し，これをすべてシャッフルして店舗へ送り返す．

(4) 店舗は，(1) の属性ごとに分けられた 14 個のタグを 2 個の乱数で商品ごとに冪乗し，属性と商品ごとに 28 個のタグを生成する．

(5) 店舗は，(3) のすべてシャッフルされた 35 個のタグと，(4) の属性と商品ごとに分けられた 28 個のタグを照合し，$\mathbf{H}(1)^{R_p^{(v=1)}R_s^{(l=1)}}$, $\mathbf{H}(6)^{R_p^{(v=1)}R_s^{(l=1)}}$, $\mathbf{H}(3)^{R_p^{(v=2)}R_s^{(l=1)}}$, $\mathbf{H}(4)^{R_p^{(v=2)}R_s^{(l=2)}}$, $\mathbf{H}(7)^{R_p^{(v=3)}R_s^{(l=2)}}$, $\mathbf{H}(3)^{R_p^{(v=4)}R_s^{(l=1)}}$, $\mathbf{H}(1)^{R_p^{(v=5)}R_s^{(l=1)}}$, $\mathbf{H}(6)^{R_p^{(v=5)}R_s^{(l=1)}}$, $\mathbf{H}(4)^{R_p^{(v=5)}R_s^{(l=2)}}$, $\mathbf{H}(7)^{R_p^{(v=5)}R_s^{(l=2)}}$ の 10 個のタグを合致させて表 8 の集計表を手に入れる[*5]．

以上のプロバイダ–店舗プロトコルにかかる冪乗余計算は (1) で $NW = 14$ 回，(2) で $MG = 7$ 回，(3) で $MGV = 35$ 回，(4) で $NWL = 28$ 回である．これを主体ごとに整理すると，プロバイダは (1) と (3) を行うので $NW + MGV$ 回であり，店舗は (2) と (4) を行うので $MG + NWL$ 回である．ただし，実際の想定では，(1) で $NW = 10^8 \times 3$ 回，(2) で $MG = 10^5 \times 10$ 回，(3) で $MGV = 10^5 \times 10 \times 57$ 回，(4) で $NWL = 10^8 \times 3 \times 10^4$ 回であり，(4) の NWL 回が支配的である．

なお，上記の説明では，プロバイダは (1) で会員番号をすべてタグ化して店舗へ送ったが，もしも店舗が (2) で会員番号の一部を取り除いてしまうと，その取り除いた会員の属性が (5) で集計表の差分として見えてしまう差分攻撃が可能となる．そのため，プロバイダは (1) で一部の会員番号をサブサンプリングしてこれを回避する．実際の想定では，プロバイダの会員番号の数は 10^7〜10^8 であるが，そのうち 10%（10^6〜10^7）程度をサブサンプリングする．

[*4] それぞれの属性ごとにランダムに会員番号を入れ替えて会員の属性を保護する．それぞれの属性に該当する会員が複数必要なので，会員数が少ない属性のデータはあらかじめ削除しておく．

[*5] (1) での属性ごとのシャッフルにより，(5) で属性と商品ごとに集計する時点では，それぞれの商品を購入した会員同士の属性が入れ替わるので安全となる．つまり，セキュアマッチングが安全であるためには，それぞれの商品を購入した会員が少なくとも 2 人以上必要である．

4.4 店舗–来訪客プロトコル

本節では，来訪客が推薦商品情報を手に入れるまでの処理を例にあげて，秘匿内積を用いた店舗–来訪客プロトコルについて概説する．

本稿では，店舗の集計表と来訪客の属性情報を守るために，店舗–来訪客プロトコルに秘匿内積を用いる．秘匿内積は，暗号文と暗号文の掛け合わせが両者の和の暗号文となる加法準同型暗号の原理を利用している．掛け合わせる回数を増やせば両者の積の暗号文にもなるので，二者間でベクトルの積和を安全に計算することができる．Vaidya らはこの秘匿内積で二者の確率ベクトルの積和を求めて，ナイーブベイズ識別器を構成 [5] している．

推薦精度の高いベイズの手法を用いるためには，確率での計算が必要である．しかし，本稿のユースケースの想定では，来訪客は V 種類の属性がそれぞれ出現したことを表す頻度のベクトル $\boldsymbol{x}$ しか持ち合わせていない．そこで本稿では，これと同じ長さ V の確率ベクトル $\boldsymbol{\theta}$ を用いた確率モデル $P(\boldsymbol{x}) \propto \prod_v^V \theta_v^{x_v}$ を導入し，確率で扱えるようにする．このモデルは多項分布モデルと呼ばれる一般的なモデルであり，それぞれの属性 v について，出現確率 θ_v を出現回数 x_v だけ掛け合わせたものに確率が比例するモデルである．

推薦商品を来訪客の端末画面の上から順に並べて表示するには，この確率モデルで各商品ごとの確率を求めて，その確率が高い商品から順に推薦すればよい．このことから，この確率モデルで算出する確率の計算精度は，推薦商品の順序関係が同じであれば十分である．そこで，この確率モデルに単調増加関数である対数を適用し，商品の順序関係（この確率モデルで計算した確率の大小関係）は変えずに $\boldsymbol{x}$ と $\log \boldsymbol{\theta}^{(l)}$ の秘匿内積で確率を計算できるようにして[*6]，店舗の売上記録と来訪客の属性情報を守る．また，冪乗を積に置き換えることで処理性能を向上する．

以上をふまえて本稿では，店舗と来訪客の間で式 (1) の秘匿内積を L 種類の商品について行い，この店舗–来訪客プロトコルで得られる推薦商品情報 $\hat{\boldsymbol{y}}$ を用いて推薦商品を順位付ける．

$$\hat{y}^{(l)} = \boldsymbol{x} \log \boldsymbol{\theta}^{(l)} \tag{1}$$

本稿のユースケースにあてはめると，店舗は表 8 から $\log \boldsymbol{\theta}^{(l=1)} = (\log(\frac{2}{6}), \log(\frac{1}{6}), \log(\frac{0}{6}), \log(\frac{1}{6}), \log(\frac{2}{6}))$ と $\log \boldsymbol{\theta}^{(l=2)} = (\log(\frac{0}{4}), \log(\frac{1}{4}), \log(\frac{1}{4}), \log(\frac{0}{4}), \log(\frac{2}{4}))$ を作成しておく．ここで，来訪客が $\boldsymbol{x} = (0, 1, 0, 1, 0)$ を提

[*6] 実装においては，式 (1) の確率の対数 $\log \theta_v^{(l)}$ は 0 か負の実数値となるので，これに秘匿内積を適用するためには，1) 小数を整数に変換，2) 負の値を正の値に変換，する必要がある．これらに対しては，確率の対数に負の大きな定数をかけてから秘匿内積を行い，秘匿内積で得られる確率の大小関係の判定を逆にすればよい．

示すると，秘匿内積で得られる商品推薦情報は本 A が $\hat{y}^{(l=1)} = \log(\frac{1}{6}) + \log(\frac{1}{6}) = -1.1$ で，本 B が $\hat{y}^{(l=2)} = \log(\frac{1}{4}) + \log(\frac{0}{4}) = -1.4$ である．よって，来訪客には本 A，本 B の順に推薦する．

秘匿内積に用いることができる加法準同型暗号には，Modified-ElGamal 暗号や Paillier 暗号 [14] が知られている．後者を用いた店舗–来訪客プロトコルを設計する．

(1) 来訪客は大きな素数 p，q を生成し，公開情報として $N_c = pq$ と $g \in \mathbf{Z}^*_{N_c^2}$ を，秘密情報として $\lambda = lcm(p-1, q-1)$ と $g^\lambda mod N_c^2$ を計算する．

(2) 来訪客は，来訪客の属性の暗号文として V 個の $E(x_v) = g^{x_v} r_c^{N_c} mod N_c^2$ を計算する．ここで，$r_c \in \mathbf{Z}^*_{N_c^2}$ は暗号文ごとに異なる乱数である．

(3) 来訪客は来店時に公開情報 N_c，g と来訪客の属性の暗号文 $E(x_1)$，...，$E(x_V)$ を店舗へ送る．

(4) 店舗は，店舗のノウハウの（注釈 6 の方法などにより整数化した）値 $\log \theta_v^{(l)}$ およびそれぞれの属性ごとの来訪客の属性の暗号文 $E(x_v)$ をもとに，加法準同型暗号の性質を利用して両者の積をとり，VL 個の $E(x_v \log \theta_v^{(l)}) = E(x_v)^{\log \theta_v^{(l)}} mod N_c^2$ を計算する．

(5) 店舗は VL 個の $E(x_v \log \theta_v^{(l)})$ を，それぞれの商品ごとに全属性の暗号文を掛け合わせ，加法準同型暗号の性質を利用して全属性の和をとり，L 個の推薦値の暗号文 $E(\acute{y}^{(l)})$ を計算する．

(6) 店舗は推薦値の暗号文 $E(\acute{y}^{(1)})$，...，$E(\acute{y}^{(L)})$ を来訪客へ送る．

(7) 来訪客は推薦値の暗号文を秘密情報の λ と $g^\lambda mod N_c^2$ で復号し，L 個の推薦値 $\acute{y}^{(l)} = \frac{\frac{(E(\acute{y}^{(l)})^\lambda mod N_c^2) - 1}{N_c}}{\frac{(g^\lambda mod N_c^2) - 1}{N_c}} mod N_c$ を得る．

以上の店舗–来訪客プロトコルにかかる冪乗余計算は，(1) で 1 回，(2) で $2V$ 回，(4) で VL 回，(7) で L 回である．すなわち，店舗で VL 回，来訪客で $1 + 2V + L$ 回の冪乗余計算となる．実際の想定では，来訪客は (1) と (2) を来店前に事前計算しておくことができるため L 回で済ませることができる．

5. セキュアスムージング手法

本章では，PPDM 環境で利用可能なセキュアスムージング手法を提案する．まず最初に，ナイーブベイズについて詳しく述べ，スムージングの必要性を述べる．非 PPDM 環境における Minka のスムージング手法とその問題点を述べた後，提案法を説明する．

5.1 ナイーブベイズ

本節では，Vaidya らのナイーブベイズの手法を，本稿のユースケースの例に照らして概説する．

ユースケースでは，プロバイダと店舗がともに会員番号 1 番，3 番，4 番，6 番，7 番の情報を保有している．1 番と 3 番と 6 番の会員は本 A を，4 番と 7 番の会員は本 B を購入していることから，店舗は本 A が $\frac{3}{5}$ の確率で，本 B が $\frac{2}{5}$ の確率で売れると考える．これを事前確率と呼び，$P(\boldsymbol{y}) = \left(\frac{3}{5}, \frac{2}{5}\right)$ で表す．次に店舗は，これまでどの商品がどの属性の人にどれだけ売れたかという表 8 の集計表の値を観測する．この観測値を商品ごとに足して 1 に正規化したものを商品ごとの条件付き確率と呼び，$P(\boldsymbol{x}|\boldsymbol{y}) = \begin{pmatrix} 2/6 & 0 \\ 1/6 & 1/4 \\ 0 & 1/4 \\ 1/6 & 0 \\ 2/6 & 2/4 \end{pmatrix}$ で表す．ナイーブベイズの手法は，事前確率にこの条件付き確率を掛け合わせて補正を行い，より正確な事後確率を算出して推薦を行う．ユースケースの来訪客は 30 代の男性であるので，この来訪客が本 A を購入する確率は $\frac{3}{5} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{60}$，本 B を購入する確率は $\frac{2}{5} \times \frac{1}{4} \times 0 = 0$ である．よって，店舗は来訪客に本 A，本 B の順に推薦する．

しかし，来訪客が本 B を購入する確率が 0（絶対に売れない）という予測は極端である（零確率問題と呼ばれる [13]）．そこで次の節では，これを緩和するスムージングについて述べる．

5.2 スムージング

ベイズの手法は，事前確率を観測値で補正し，より正しい事後確率を算出する方法である．前節ではこの補正を条件付き確率で行ったが，この補正の度合いは観測値に対する尤もらしさの度合いであることから，尤度とも呼ばれる．スムージングは，たとえば表 8 の集計表にすべて 1 を足してから（少なくとも 1 回は観測したと見なして）尤度を算出すれば実現できる[*7]．スムージングは，このようにして零確率問題を回避し，計算精度を高める．次節では，このスムージングの効果をさらに引き出して計算精度を高めるために，具体的にどのような値にスムージングを設定するかについて述べる．

5.3 Minka のスムージング

ベイズの手法には事前確率の設定に難しさがある．それを回避する方法の 1 つに，事前確率に分布を設定して，その分布を適切にパラメータ学習することで，計算精度を高める方法がある．増村らは，頻度分布の観測値に対してディレクレ分布の事前分布を設定し，その分布を LOO 尤度を用いて適切に学習するディレクレスムージングを提案している [13]．また，Minka は，このディレクレ分布のパラメータ学習の手法のバリエーションの説明と，それぞれの手法による計算精度を解析しており，LOO 尤度を用いて高い計算精度が得られると述べている [9]．しかし，この LOO 尤度を用いたスムージングの手法は，データを秘

[*7] 観測していないものを観測したと見なすのは，あくまで工学上の工夫にすぎない．未観測のものは事前確率で扱うべきであるし，もっと多くのデータを観測すべきともいえる．

匿しない非 PPDM 環境を前提としており，プライバシ保護を必要とする PPDM 環境にこのまま適用してしまうと，LOO を行うたびに 1 件ずつすべてのデータが相手に漏れてしまうという問題が生じるため，利用できない．

また，これらの方法をそのまま適用しても，組織間リコメンドにおいて問題が生じる．増村らや Minka の手法はディレクレ分布の V 個の属性値のパラメータをそれぞれ学習するが，この学習には少なくともすべての属性値について複数の観測データが必要である．保有するデータの規模が異なるプロバイダと店舗のセキュアマッチングによって，店舗が十分なデータを手に入れられるとは限らない．本稿のユースケースにあてはめると，店舗が月の初めにある商品を投入したとして，その商品を 20 代，30 代，40 代，男性，女性の来訪客が少なくとも 2 人以上ずつ購入してからでなければ，スムージングを行えないので，商品を推薦できない．Minka の更新式を式 (2) に示す．s は更新のステップ数を，ξ はディレクレ分布のハイパーパラメータ（以後，ξ で表す）である．

$$\xi_v^{(l,s+1)} = \xi_v^{(l,s)} \frac{\sum_i \frac{\left(x_v y^{(l)}\right)_i}{\left(x_v y^{(l)}\right)_i - 1 + \xi_v^{(l,s)}}}{\sum_i \frac{\left(\sum_v x_v y^{(l)}\right)_i}{\left(\sum_v x_v y^{(l)}\right)_i - 1 + \sum_v \xi_v^{(l,s)}}} \tag{2}$$

Minka の更新式を動作させるために必要なデータの量は，セキュアマッチングの安全性の要件（商品ごとに 2 人以上の会員をマッチングできること）よりも多い．Minka の方法は属性ごとにスムージングを行うが，このまま適用してしまうと属性の種類だけデータが疎になってしまうので，それを補うだけのデータが必要となる．すなわち，ある商品を購入した会員の属性が 1 種類しかなければ（たとえば，店舗である商品を買った会員に 20 代が 1 人しかいなければ），1 巡目で $\xi_{v=age:20}^{(l,s+1)} = 0$，2 巡目で $\xi_{v=age:20}^{(l,s+1)} = \inf$ となってしまい，スムージングできない．通常，商品を購入した会員の属性には偏りがある（この偏りを利用してリコメンドを行う）ので，属性の種類（$V = 57$）倍以上のデータがあったとしても十分ではない．

そこで，次節では，PPDM 環境で組織間リコメンドを実現できるスムージングを新たに提案する．

5.4 セキュアスムージング

4.3 節で述べたプロバイダ–店舗プロトコルにおいて，店舗は，すべてシャッフルされたタグと，属性と商品ごとに分けられたタグを照合して集計表を作成する．このプロトコルはセキュアマッチングと同等に安全であり，店舗がプロバイダの会員情報を手に入れようとしても，会員番号が秘匿されているうえ，ある商品を購入した会員同士で属性がランダムに入れ替わっているので，個々の会員の属性を正しく復元できない．

ところが，本稿で用いるナイーブベイズは属性を独立に扱うものであり，たとえ個々の会員の属性を正しく復元できなくても，それぞれの商品を購入した会員の属性の頻度の合計は変わらないので，このシャッフルされた暗号文から復元した仮想のデータで LOO 尤度を算出しても，スムージングを行える可能性がある．そこで，実際のデータを模倣して，W 個ずつの暗号文を属性が重複しないように抜き出して仮想データを生成することとする[*8]．

さらに組織間リコメンドに適するよう，属性と商品ごとではなく，商品ごとにスムージングのパラメータ（以後，$\gamma^{(l)}$ で表す）を設定する[*9]．このスムージングは集計表 Φ の商品 l の列にすべて $\gamma^{(l)}$ を足してから尤度を算出するのと同じであり，式 (3) で算出できる．

$$\theta_v^{(l)} = \frac{\phi_v^{(l)} + \gamma^{(l)}}{\sum_v \left(\phi_v^{(l)} + \gamma^{(l)}\right)} \tag{3}$$

式 (3) を式 (1) へ代入し，Minka の方法のように事後確率を最大にできる $\gamma^{(l)}$ の更新式を導出すると式 (4) となる．

$$\gamma^{(l,s+1)} = \frac{(J^{(l)} - 1)W}{V} \frac{\sum_v \phi_v^{(l)} \frac{\gamma^{(l,s)}}{\phi_v^{(l,-i)} + \gamma^{(l,s)}}}{\sum_v \phi_v^{(l)} \frac{\phi_v^{(l,-i)}}{\phi_v^{(l,-i)} + \gamma^{(l,s)}}} \tag{4}$$

ここで，$J^{(l)}$ は商品 l に該当する会員数であり，$\phi_v^{(l,-i)}$ は i 番目の仮想データを抜いた集計表の値である．この更新式を用いれば，ある商品を購入した来訪客が 2 人以上いれば動作でき，店舗は安全にスムージングを行うことができる[*10]．

次章では，このセキュアスムージングに実際のデータセットを適用し，PPDM の Vaidya らの手法および非 PPDM の LOO による Minka のスムージングとの計算精度の違いを明らかにする．

6. 実験評価

6.1 Play Tennis データセット

Play Tennis データセット [15]（**表 9**）は，14 日間分の天候条件と，それぞれの天候条件の日にテニスと休息のどちらを選択したかを記したものであり，データマイニングの分野では例題として取り上げられる，標準的なベンチマークデータセットである．

それぞれの日を会員に，天候の条件を属性情報に，テニスか休息かを商品情報に，それぞれ見立ててリコメンドを

[*8] たとえば，属性が 20 代，20 代，男性，女性であるとき，20 代，男性と 20 代，女性のように抜き出して仮想サンプルを作る．20 代，20 代と男性，女性のようにはしない．

[*9] 一般に，多項分布の事前分布として用いられるディレクレ分布はハイパーパラメータ（本稿では ξ で表す）を持つが，本稿では説明を平易にするために，集計表に具体的に足し込む値 γ で説明する．両者は $\xi = \gamma + 1$ の関係となる．

[*10] 商品ごとに 2 人以上の会員をマッチングできれば，$J^{(l)} \geq 2$ となり，式 (4) の更新式が動作する．すなわち，これを動作させるために必要なデータの量は，セキュアマッチングの安全性の要件（商品ごとに 2 人以上の会員をマッチングできること）と一致する．

表 9 Play Tennis データセット
Table 9 Play Tennis data set.

Day	Outlook	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	Rest
2	Sunny	Hot	High	Strong	Rest
3	Overcast	Hot	High	Weak	Tennis
4	Rain	Mild	High	Weak	Tennis
5	Rain	Cool	Normal	Weak	Tennis
6	Rain	Cool	Normal	Strong	Rest
7	Overcast	Cool	Normal	Strong	Tennis
8	Sunny	Mild	High	Weak	Rest
9	Sunny	Cool	Normal	Weak	Tennis
10	Rain	Mild	Normal	Weak	Tennis
11	Sunny	Mild	Normal	Strong	Tennis
12	Overcast	Mild	High	Strong	Tennis
13	Overcast	Hot	Normal	Weak	Tennis
14	Rain	Mild	High	Strong	Rest

表 10 Play Tennis データセットに対する従来方式と提案方式による識別結果
Table 10 Results of recommendations on Play Tennis data set.

方式	True Positive	True Negative	False Positive	False Negative	Accuracy
VKC08	7	1	3	3	57%
M00	5	3	1	5	57%
提案方式	7	4	1	2	79%

行う．プロバイダは $N = 14$ 人，$W = 4$ 項目からなる会員情報を保有している．会員番号 $\boldsymbol{t}$ は長さ $N = 14$ のベクトルであり，属性情報 X は $N = 14$ 行，$V = 10$ 列のマトリクスである[*11]．店舗は売上 1 人あたり $G = 1$ 種類の商品を含む $M = 13$ 件[*12]の売上記録を保有している．$N \neq M$ より，プロバイダと店舗のデータは非同期である．店舗が保有する会員番号 $\boldsymbol{u}$ は長さ $M = 13$ のベクトルである．店舗が保有する商品情報 Y は $M = 13$ 行，$L = 2$ 列のマトリクスである．来訪客は長さ $V = 14$ の属性情報 $\boldsymbol{x}$ のベクトルを保有している．Play Tennis データセットは件数が少ないので，実験の精度を高めるために LOO の交差検定で評価する．すなわち，1 日ずつ順に抜き出したデータをテストデータとし，残りを学習データとして，計 14 回の試行を行う．なお，この LOO は，5.3 節で述べた LOO 尤度とは無関係である．

実験結果を**表 10** に示す．リコメンドが Tennis で事実も Tennis で当てた場合を True Positive，リコメンドが Rest で事実も Rest で当てた場合を True Negative，リコメンドは Tennis だが事実は Rest で外した場合を False Positive，リコメンドは Rest だが事実は Tennis で外した場合を False Negative と数え，正解率（以後，*Accuracy* で表す）を $Accuracy = \frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$ として求めた．提案方式は Vaidya らの手法 [5]（以後，VKC08 で表す）や Minka の手法 [9]（以後，M00 で表す）よりも 22%高い *Accuracy* が得られた．Minka の手法は非 PPDM の手法だが，参考のために比較した．また，菊池らの手法の計算精度は，原理的に Vaidya らの手法と同等であるので割愛する．

なお，非 PPDM の LOO 尤度と PPDM の提案手法の仮想 LOO 尤度によるスムージングの収束速度について，表 10 の精度に達するまでに要した更新式のステップ数は，M00 が 7 回で，提案手法は 11 回であった．

6.2 Book-Crossing データセット

Book-Crossing データセット[*13]を用いて，スムージングのない VKC08 と，属性と商品ごとにスムージングを行う M00 と，商品ごとにスムージングを行う提案手法の *Accuracy* を比較し，スムージングの設定方法による計算精度を明かにする．

Book-Crossing データセットは，Book-Crossing community における本の評価のデータセットである．これは Cai-Nicolas Ziegler によって 2004 年 8 月から 9 月のうちの 4 週間クローリングされたものであり，278,858 人のユーザの 271,379 冊の本に対する 1,149,780 件の評価値を含む．ユーザは匿名化されているが，年齢と居住地の情報は有している．

年齢については 80 歳を超えるデータは取り除き[*14]，10 歳刻みの年代とした．また，居住地については，国際標準化機構（ISO）によって公表されている 249 国の国名を用い，ISO 3166-1 における英語名（アメリカ合衆国の場合であれば ‘United States’）を含むデータに国名を割り当て，もしも ISO 3166-1 における英語名を含まない場合は，ISO 3166-1 alpha-3（アメリカ合衆国の場合であれば ‘USA’）を含むデータに国名を割り当てた[*15]．

Book-Crossing データセットの評価値は 10 点満点で付与されている．本実験では，ユーザが 5 点を超える評価値を与えた本を，ユーザが満足した本であると見なす．それぞれの本ごとに満足したユーザの属性を学習しておき，未知のユーザの属性をもとに，そのユーザが満足できる本を

*11 会員の属性（天候条件）は 4 種類だが，Outlook の属性値は Sunny，Overcast，Rain の 3 種類，Temparature の属性値は Hot, Mild, Cool の 3 種類，Humidity の属性値は High, Normal の 2 種類，Wind の属性値は Weak，Strong の 2 種類なので，属性値は合計で $V = 10$ となる．

*12 それぞれの試行で，順に 1 件ずつを来訪客で用い，残りの 13 件ずつを店舗で用いる．

*13 Institut für Informatik, Universität Freibur, Book-Crossing Dataset, http://www.informatik.uni-freiburg.de/~cziegler/BX/

*14 Book-Crossing データセットには 200 歳を超えるデータも含まれていたが，これらの異常データは取り除いた．

*15 ‘space, space, somewherein space’ のような意味不明のデータや，複数の国名が記載されたデータなど，国名を 1 つに特定できないデータは取り除いた．また，10 人未満の国も取り除いた．これらのクレンジングにより 110 国を実験に用いた．

表 11 Book-Crossing データセットに対する従来方式と提案方式による識別結果

Table 11 Results of recommendations on Book-Crossing Dataset.

方式	推薦した本の種類	推薦の成功率（Accuracy）	推薦に成功した場合での表示順位得点（Score）
VKC08	1,555 種類	12%	7 点
M00	89 種類	1%	6 点
提案方式	1,841 種類	4%	6 点

推薦できるかを計測する．データの件数が多いので三交差検定を用いて実験を行う．なお，交差検定での分割後の件数を確保するため，評価値が 10 件を超える本（4,518 種類）を対象とする．

Book-Crossing データセットは，Play Tennis データセットと異なり，正解となりうる商品が複数あるので，この実験における評価指標を改めて設定する．この実験では，来訪客の端末の画面に推薦商品を推薦度順に 10 個表示する試行を行い，その中の本をどれか 1 つでも来訪客が購入すれば，その試行を正解とする．*Accuracy* は正解数を試行数で除した値であり，すなわち推薦の成功率である．また，正解の場合での表示順位に対して得点（以後，*Score* で表す）を与える．その得点は 1 位を 10 点，2 位を 9 点，...，最下位の 10 位を 1 点とする．すなわち *Score* は，来訪客が購入した本の推薦順位が 1 番だった場合には 10 点，2 番だった場合は 9 点，...，10 番だった場合は 1 点となる．

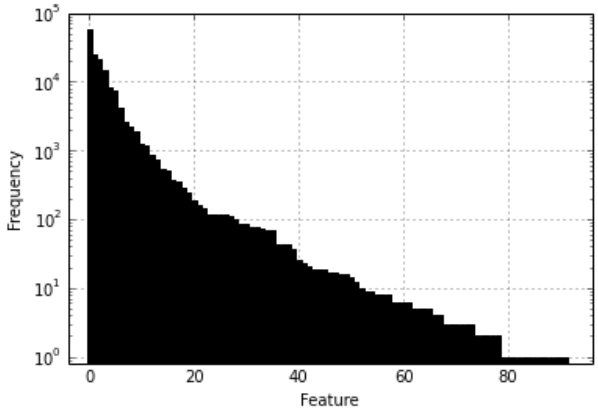
実験結果を**表 11** に示す．*Score* は表示順位得点の分布の中央値を示した．*Accuracy* と *Score* ともに VKC08 が最上位であった．M00 は推薦した本の種類と *Accuracy* が極端に低く，適切な推薦ができなかった．提案手法は推薦した本の種類こそ最も多かったが，*Accuracy* も *Score* も VKC08 に及ばなかった．

7. 計算精度と処理性能の考察

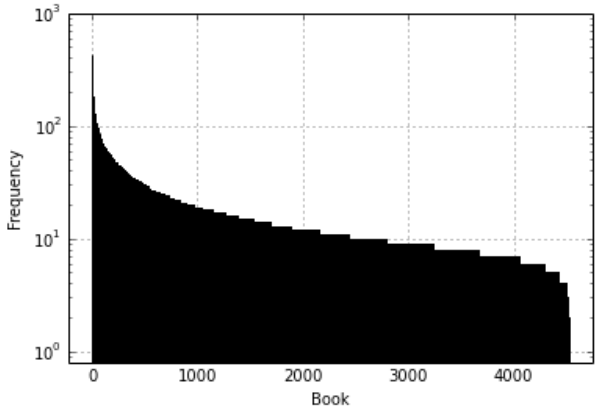
7.1 計算精度の考察

非 PPDM の既存手法の 1 つである M00 は，Book-Crossing データセットでの実験において，推薦した本の種類と推薦の成功率が極端に低かった．M00 は属性と商品それぞれの VL 個のスムージングのパラメータがあり，それらを設定するために十分なデータが必要である．このデータを把握するために，プロバイダ–店舗プロトコルでのセキュアマッチング後に店舗が手に入れたデータの属性ごとの頻度を**図 3** にプロットした．

図 3 (a) の横軸は会員の属性（年代，性別）で，縦軸はデータの量を表す頻度の対数である．頻度が多い順に属性を並べてあるので右肩下がりとなっており，ロングテールとなっている様子を観測できる．図 3 (b) の横軸は商品（本）で，縦軸は同じく頻度の対数である．商品も頻度が多い順に並べてあるので右肩下がりとなっておりロングテールではあるが，図 3 (a) と比べれば一部の商品だけが極端に多く（ロングテールを示す傾斜がきつく），その他の商品での偏りは少ない（ロングテールを示す傾斜が緩い）ことを観測できる．



(a) 属性（年代，性別）の頻度



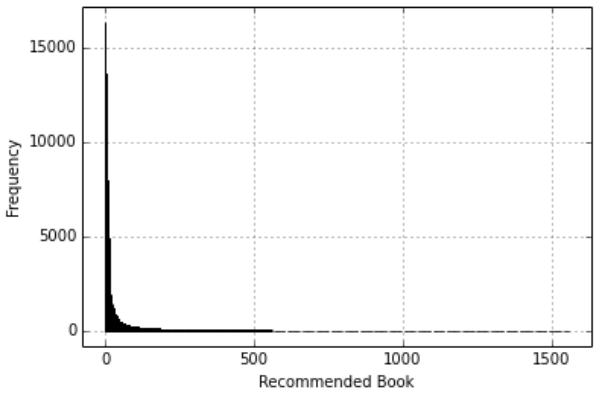
(b) 商品（本）の頻度

図 3 セキュアマッチング後のデータの特性

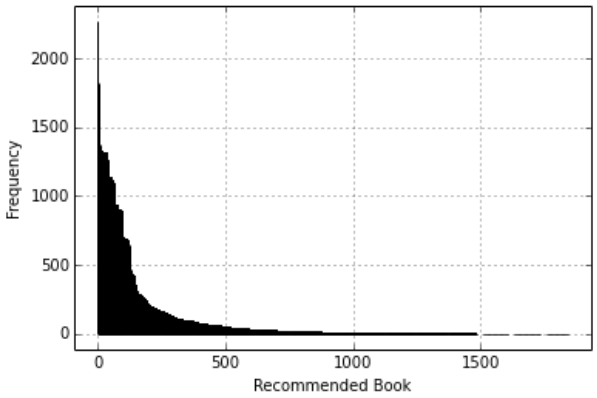
Fig. 3 The frequency of the user's feature. The frequency of the books.

Book-Crossing データセットの属性値は 119（年代：9 + 国：110）で，商品の種類は 4,520 であり，537,880 個のスムージングのパラメータを設定する必要がある．しかし，プロバイダ–店舗プロトコル後のデータに 0 または 1 件しか現れない属性が 2 割程度もあり，これが原因で計算精度が低下したと考えられる．もしも，プロバイダと店舗のデータが同期していてすべての会員の売上情報を把握できる組織であれば，このような問題は生じないので高い計算精度が得られたと考えられる．本稿で対象としている組織間プライバシ保護リコメンドでは，プロバイダと店舗のデータが同期しておらず，セキュアマッチング後のデータをスムージングのために属性と商品ごとに分割してしまうと少ないため，M00 は適していない．

一方，提案手法は商品それぞれの L 個のスムージングのパラメータを設定すればよい．店舗は店舗で販売した商品



(a) VKC08



(b) 提案方式

図 4 推薦された本の種類と頻度

Fig. 4 The frequency of the recommended books.

は分かるが，プロバイダに登録している会員の属性は分からないので，商品数に応じたパラメータ数のみでスムージングを行うのが確実である．セキュアマッチング後のデータ量は 10^2 種類の商品を取り扱う店舗と 10^5 種類の商品を取り扱う店舗では異なると考えられるが，提案手法はこれに応じてスムージングのパラメータ数を対応させる．この違いにより，M00 に比べて提案手法のスムージングの方がうまく動作し，計算精度が向上したのだと考えられる．

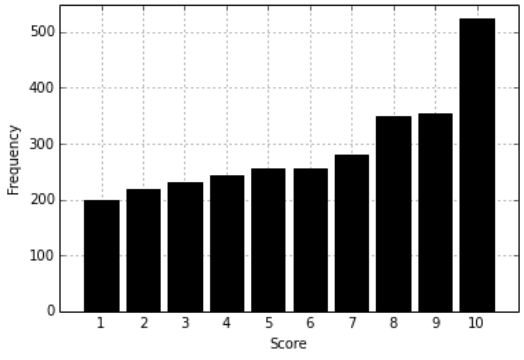
ところで，VKC08 は，なぜ最も高い正解率と表示順位得点を得られたのであろうか？ 提案手法との違いを明らかにするために，両者に推薦された本の種類と頻度を**図 4** にプロットした．

VKC08 はスムージングを行わないため，原理的に零確率などの問題が生じやすく，提案方式よりも推薦が偏りやすい．図 4 からも，その傾向を確認できる．

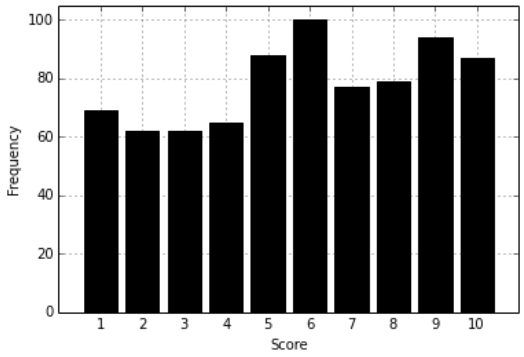
VKC08 の正解率とスコアが高い理由を引き続き探るために，推薦に成功した場合での表示順位得点（10 件表示．上位より 10 点〜1 点）の分布を**図 5** にプロットした．

これによると，VKC08 は 1 位で推薦した商品（図 5 (a) 右端のスコア 10 の商品）で獲得したスコアが抜きん出ており，高いスコアを得られたと考えられる．Book-Crossing データセットは，図 3 に示したように，よく売れている本とそうでない本には 100 倍程度の差がある．Book-Crossing データセットで最も多い売上を記録していたのは “The Lovely Bones”（Alice Sebold 著，売上 663 冊）であり，VKC08 が最もよく推薦していた本と一致した．そこで，“The Lovely Bones” を除いて再度実験を行ったところ，VKC08 のスコアは 6 点となり，この 1 冊を除いただけで提案手法と同点となった．VKC08 は多く売れている本を推薦するのでスコアが高いと考えられる．

次に，多く売れている本から順に除いていくと，推薦に成功する確率がどのように変化するかを**図 6** にプロット



(a) VKC08



(b) 提案方式

図 5 推薦に成功した場合での表示順位得点の分布

Fig. 5 The score of purchased books in the top ten recommend ranking.

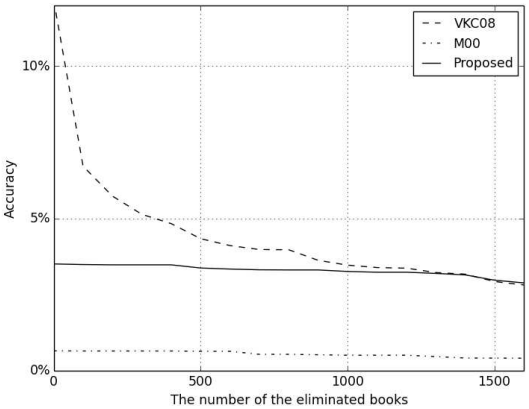


図 6 多く売れている本を除いた場合に，推薦に成功する確率

Fig. 6 The accuracy without the popular books.

表 12 冪乗の回数と処理時間
Table 12 The computational complexity and the processing time.

登場人物	VKC08	提案方式
プロバイダ	NVL （約 4×10^5 日）	$NW + MGV$ （約 6 日）
店舗	NVL （約 4×10^5 日）	$MG + NWL$ （約 2×10^4 日）
	-	VL （約 1 時間）
来訪客	-	$1 + 2V + L$ （約 60 秒）

した．

多く売れている本を除いていくと，VKC08 の正解率は大幅に低下したが，提案手法の正解率は変化が少ない傾向を確認できた．1,500 冊の本を除いた時点で提案手法の正解率は VKC08 を超えた．売上が上位の本を把握するのは，プロバイダの協力を得なくても店舗だけで，もしくは人間にも簡単に分かる．重要なのは，簡単には分からない下位の部分で適切な推薦を行うことである．その部分においては，提案方式が VKC08 よりも優れている[*16]．

以上の考察から，提案方式は従来方式に比べて，バリエーション豊富なリコメンドを来訪客の属性に応じて提供できる計算精度を有しているといえる．

7.2 処理性能の考察

提案方式による処理性能を明らかにするため，4.3 節および 4.4 節で述べた計算量の一般式を**表 12** に纏める．なお，括弧内は本稿でのユースケースの想定における処理時間である．プロバイダが会員情報の 10%をサブサンプリングする（VKC08 も 10%のサブサンプリングする）ものとし，1 回の冪剰余にかかる計算時間を 6 [ms] とし[*17]，並列計算は行っていない．

VKC08 は三者間での運用を想定していないので，プロバイダ–店舗間のみを考慮する[*18]．VKC08 はプロバイダと店舗でそれぞれ NVL 回の冪乗が必要となる[*19]．提案方式は 4.3 節で述べたように NWL 回の冪乗が支配的である．処理性能において提案方式が VKC08 よりも優れる理由は，冪乗の回数を $O(NVL)$ から $O(NWL)$ に減らせるところにある．属性の項目は 2 種類以上の値をとる（1 種類の値しかない属性の項目には意味がない）ので，属性値の種類数 V は属性の種類数 W の少なくとも 2 倍以上である．よって，提案方式の処理性能は VKC08 より 2 倍以上優れる．

次に，本稿のユースケースでの想定において，提案方式で要件を満たせるかを考察する．

プロバイダ–店舗プロトコルの処理性能について，提案方式は VKC08 よりも高速であるが，店舗での NWL の冪乗がボトルネックとなるため，このままでは T_2（1 カ月）以内の要件を満たすことができない．ただし，この問題は一般的になりつつあるクラウド環境の利用などにより解決できる．たとえば，Amazon などの外部のクラウドで 64 コアのインスタンスを 16 個用いて約 1,000 倍のスケールアウトを行うとすると，店舗での処理を約 2×10^4 日から約 20 日に短縮できるので，プロバイダでの処理の 6 日を加えても T_2（1 カ月）以内の要件を満たすことができる．なお，プロバイダの会員数 N が大きくなるほどサブサンプリングによる計算精度への影響は小さくなるので，サブサンプリングの量を減らすことで処理時間をさらに短縮できる可能性もある．

店舗–来訪客プロトコルの処理性能について，店舗と来訪客のいずれでも，このままでは T_1（5 秒）以内の要件を満たすことができない．ここでは素早いレスポンスが求められるため，外部のクラウドを利用することは難しい．この問題は，店舗における一般的な多コア CPU の利用と，ユースケースに則した計算上の工夫により解決できる．店舗の PC（8 コアの CPU を搭載していると想定）を用いて 8 並列で冪乗する．さらに，$L = 10^4$ 種類ある商品を 50 種類ずつに 200 分割して，ユーザが 50 種類の商品を眺めている間に次の 50 種類の商品の推薦を計算する．これらの工夫により，店舗での処理を 1,600 倍，来訪客での処理を 200 倍高速化する．すると，店舗での処理を約 1 時間から約 2 秒に，来訪客での処理を約 60 秒から 0.3 秒に短縮できるので，T_1（5 秒）以内の要件を満たすことができる．なお，ハードウェアの進歩により，一度に推薦できる商品の種類は今後増やしていけると考えられる．

8. 安全性の考察

8.1 攻撃者のモデル

提案システムに対する攻撃者としては，システムの利用主体（すなわちプロバイダ，店舗，来訪客）と，外部からの攻撃者が想定される．本稿の主旨は，1 章で述べたように，主体間のプライバシ保護であるため，ここでは攻撃者としてシステムの利用主体を前提とする．外部からの攻撃

*16 今回用いた Book-Crossing データセットは上位と下位の売上数の差が大きいが，リコメンドが適切に行われれば下位の本がより多く売れるようになると考えられる．下位の本が売れている状況を記録したデータセットで評価を行えば，VKC08 の正解率は低下し，提案手法の正解率は向上すると考えられる．

*17 通常で入手できるデスクトップ PC に搭載されている，Intel(R) Core(TM) i7 CPU 870 (2.93 GHz) を用い，The GNU Multiple Precision Arithmetic Library を用いた C 言語のプログラムで 2,048 ビットの冪剰余計算を行った測定結果より．

*18 VKC08 は秘匿内積を用いている．提案方式の店舗–来訪客プロトコルも秘匿内積を用いているので，仮に VKC08 を店舗–来訪客プロトコルに適用した場合の冪乗の回数は，提案方式と同等である．

*19 長さ V のベクトルの積集合を求める場合，提案方式がプロバイダ–店舗プロトコルで用いているタグ照合方式では，ベクトルに値が入っている要素（W 個）のみを冪乗すればよいが，VKC08 が用いている秘匿内積では，ベクトルのすべての要素（V 個）を冪乗する必要がある．

者については，SSL（Secure Sockets Layer），WPA（Wi-Fi Protected Access），ファイアーウォール，侵入検知などの別技術によって対応するものとする．

提案システムは，プロバイダ–店舗プロトコルと店舗–来訪客プロトコルを用いる．ここでの攻撃とは，これらのプロトコルにおいて，他の主体の有するプライベートな情報を取得することである．

攻撃者のモデルとしては，他との主体との通信において能動的な攻撃を仕掛ける malicious adversary と，そのような攻撃を行わない semi-honest adversary が考えられる．ここでは semi-honest adversary を前提とする．

なお，攻撃者が malicious adversary である場合については，8.4 節で今後の課題として考察する．

8.2 プライバシ要件

要件 1 をプロトコルの内と外で分割し，以下の 2 つの要件を目標とする．

（要件 1a）プロトコル自体のプライバシ要件

プロトコルの利用主体は，そのプロトコルの出力情報を除いて，相手主体のプライベート情報を推定することができない．

（要件 1b）プロトコルの出力のプライバシ要件

プロトコルの利用主体は，そのプロトコルの出力情報から．相手主体のプライベート情報を推定することができない．

8.3 安全性の考察

8.3.1 プロバイダ–店舗プロトコル

要件 1a について，本プロトコルのセキュアマッチングで用いるタグは暗号化されており，その安全性は DDH（Decisional Diffie-Hellman）仮定の安全性に帰着する [3]．店舗はプロバイダから受け取ったタグ $\mathbf{H}(t_n)^{R_p^{(v)}}$ の中身 $\mathbf{H}(t_n)$ を復元できない．同様にプロバイダもタグの中身を復元できない．よって，要件 1a は満たされる．

要件 1b について，プロバイダには表 8 に相当する出力情報は開示されないので，店舗にとって要件 1b は満たされる．店舗については，セキュアマッチングの 1 回目の実行時の売上情報と 2 回目の実行時の売上情報に差がある場合に，その差分に相当する会員情報が，1 回目の出力情報（たとえば表 8）と 2 回目の出力情報の差分として見えてしまう場合が考えられ，プロバイダの会員情報が漏れる懸念がある．提案方式はこの懸念をサブサンプリングによって解決する．以下，この懸念がサブサンプリングによって，どの程度防止できるかを考察する．

サブサンプリングがどの程度 SDC（Statistical Disclosure Control）における安全性に寄与するかについては，竹村によって議論されている [16]．母集団一意と呼ばれる安全性指標に基づき，サブサンプリングの比率と総セル数の積が 2 桁程度ならば個票データ（本稿のユースケースにおけるプロバイダの会員情報に相当）は安全であると述べている．提案方式はリコメンドに必要で，かつ個票よりもさらに安全な集計表に情報を絞っている．本稿のユースケースでは，総セル数は性別 2 と年代 8 と住所 47 の積で 752 セルであるので，約 13%未満のサブサンプリングを行えば，プロバイダの会員情報は安全であるといえる．よって，プロバイダにとっても要件 1b は満たされる．

8.3.2 店舗–来訪客プロトコル

要件 1a について，本プロトコルの秘匿内積で用いるメッセージは暗号化されており，その安全性は用いる準同型暗号の安全性に依存する．たとえば Paillier 暗号を用いた場合は DCR（Decisional Composite Residuosity）仮定の安全性に帰着する [14]．店舗は来訪客から受け取った暗号文 $E(x_v)$ の中身 x_v を復元できない．来訪客は推薦値のみしか受け取らない．よって，要件 1a は満たされる．

要件 1b について，来訪客が得ることができるのは推薦値のみである．推薦値は来訪客の属性のベクトルと店舗のノウハウのベクトルの内積である．内積から元のベクトルを一意に復元することは困難であるため，一般的には，要件 1b は満たされる．しかし，以下のような問題が残る．

本プロトコルにおいて，来訪客は推薦値 $\hat{y}^{(l)}$ の値そのものを得ることができるが，システムの機能に必要な最低限の情報は，それぞれの商品に対する推薦値 $\hat{y}^{(1)}, ..., \hat{y}^{(L)}$ の大小関係と考えられる．つまり，余分な情報が来訪客に伝わっている．このことが情報漏洩につながる場合がある．たとえば，本稿のユースケースでの想定では，来訪客の属性は年代と性別と住所であるため，属性ベクトルには 1 が 3 カ所含まれる．しかし，他のユースケースにおいては，属性ベクトルに 1 が 1 カ所だけ含まれることもありうる．その場合に，店舗のノウハウの $\frac{1}{V}$ が来訪客に漏れる，たとえば，来訪客の属性ベクトルが $\boldsymbol{x} = (1, 0, 0, 0, ...)$ である場合，店舗は店舗のノウハウの一部である $\log \theta_{v=1}^{(1)}, ..., \log \theta_{v=1}^{(L)}$ を推薦値 $\hat{y}^{(1)}, ..., \hat{y}^{(L)}$ として来訪客へ送り返すので，VL 個ある店舗のノウハウのうちの L 個が漏洩する．したがって，このような漏洩が起こらないように対策を講じる必要がある．

8.4 安全性に関する今後の課題

上述したように，店舗–来訪客プロトコルにおいて，来訪客は推薦値 $\hat{y}^{(l)}$ の値そのものを得ることができる．その結果，店舗のノウハウが漏洩する場合があるので，この点を改良する必要がある．たとえば，それぞれの商品に対する推薦値 $\hat{y}^{(1)}, ..., \hat{y}^{(L)}$ の大小関係のみを来訪客に伝えるように改良したい．また，これまでの安全性の考察は semi-honest adversary を前提としてきたが，実用の場面では malicious adversary も存在するため，その対応について考察する．プロバイダおよび店舗は事業者であるため，契約，従業員

教育，計算機の操作に関する監視・監査により，maliciousな攻撃を抑止することが考えられる．また，プロトコルの実装プログラムとして正規のプログラムのみ利用されるように，プログラムの認証および改ざん防止機能を設けることが考えられる．来訪客については，契約や実装プログラムの認証および改ざん防止によりmaliciousな攻撃を抑止することが考えられる．しかし，このような対策にもかかわらず，プロトコルの主体が強固な悪意を持ってmaliciousな攻撃を行う場合も考えられるので，提案プロトコルにおいてmaliciousな攻撃への安全性を向上させることが課題となる．以下では，malicious adversaryを前提として，プロトコルの改良に関する課題を考察する．

8.4.1 プロバイダ–店舗プロトコル

malicious adversaryを前提とする場合，プロバイダ自体の改良が必要になる．たとえば，攻撃者がプロバイダ–店舗プロトコルの手順(2)において，ある会員番号Iとある商品lに対応する乱数$R_s^{(l)}$から生成されるタグ$\mathbf{H}(I)^{R_s^{(l)}}$を2つに複製してプロバイダに送ることが考えられる．このとき，会員番号Iがある属性vを持っているとすると，プロバイダはまず手順(1)でその属性に対応する乱数$R_p^{(v)}$を生成したうえで「その属性に対応するタグ」の1つとして$\mathbf{H}(I)^{R_p^{(v)}}$を（他のタグと一緒に）店舗に送る．次に店舗は，手順(2)において2つの$\mathbf{H}(I)^{R_s^{(l)}}$を（他のタグと一緒に）プロバイダに送る．プロバイダは，手順(3)において，2つの$\mathbf{H}(I)^{R_p^{(v)}R_s^{(l)}}$を（他のタグと一緒に）店舗に送る．すると手順(4)において，店舗は「その属性に対応するタグ」であると分かっている$\mathbf{H}(I)^{R_p^{(v)}}$を$R_s^{(l)}$乗して得られるタグ$\mathbf{H}(I)^{R_p^{(v)}R_s^{(l)}}$が，手順(3)でプロバイダから送られたタグ一覧中に2回現れるという事実から，そのタグが会員番号Iに関連するタグであることを検知できてしまう．このような攻撃に対しては，手順(3)において，プロバイダが店舗から送られてきたタグに重複がないことを確認し，重複があればプロトコルを中止するか，重複を取り除いてからプロトコルを続行することで攻撃を回避するなどの対策を行う．

8.4.2 店舗–来訪客プロトコル

来訪客がmalicious adversaryである場合，次のような攻撃が考えられる．来訪客が$\log\theta_v^{(l)}$を正の整数に変換する定数の値の上限がN_dであることを知っており，$\boldsymbol{x} = (1, N_d, N_d^2, N_d^3, ...)$という不正な属性ベクトルを店舗へ送る．すると，来訪客は，店舗が来訪客へ送り返す推薦値$\hat{y}^{(l)}$をN_d進数展開するだけで，店舗のノウハウの確率ベクトルのすべての属性の要素$\log\theta_1^{(l)}, ..., \log\theta_V^{(l)}$を一度に手に入れてしまう．このような攻撃に対しては，推薦値がPaillier暗号のメッセージ空間の最大値N_cを意図的に超えるようにして，誤った計算結果にする対策が考えられる．

Paillier暗号のメッセージ空間はN_cであるので，本プロトコルが動作するために推薦値は$\hat{y}^{(l)} = \sum_v^V x_v \log\theta_v^{(l)} \leq N_c$を満たす必要がある．これを満たすため，店舗–来訪客プロトコルの手順(4)において$\log\theta_v^{(l)}$を正の整数に変換する際に，各要素の最大値が$N_d = \frac{N_c}{W}$となるようにする．すなわち，不正な属性ベクトルと掛け合わせた各要素の値を$\alpha_v \left(\frac{N_c}{W}\right)^{v-1}$とする（係数$\alpha_v$は$0 < \alpha_v \leq 1$の値域を持つ）．来訪客による不正な属性ベクトルの送付の結果，誤った計算結果となる条件を式(5)に示す．

$$\alpha_1 + \alpha_2\left(\frac{N_c}{W}\right) + \alpha_3\left(\frac{N_c}{W}\right)^2 + \alpha_4\left(\frac{N_c}{W}\right)^3 + ... + \alpha_V\left(\frac{N_c}{W}\right)^{V-1} > N_c \quad (5)$$

属性の種類数Wに比べてPaillier暗号のメッセージ空間Ncは十分に大きいと考えられるので，$Nc \gg W$より，式(5)の左辺の第3項以降はこの条件を満たす[*20]．よって，この攻撃が成功するのは属性値の種類数Vが2以下の場合であり，この属性の種類数が3以上に増えていくと攻撃が成功する確率は単調に減少する．今後は，上記の改良案について，攻撃の成功確率の見積りや正当な利用への影響の見積りなど，厳密な分析を行うとともに，他の改良案を検討していきたい．また，malicious adversaryを前提とする場合は他の攻撃もありうるので，さらなる検討が必要である．

9. おわりに

本稿では，複数の組織に分散した情報を用い，各組織が他の組織にプライベートな情報を開示することなく，1組織がすべての情報を保有している場合と同等あるいはそれに近い精度の推薦を行う技術の実現に取り組んだ．組織間リコメンドに適した三者間のシステムおよびプロトコルを設計し，分散形態の運用を可能とした．この三者システムにおいて，各々の者が，他の二者に対して，自分のプライベートな情報を秘匿可能であることを明らかにした．実運用時のデータ数を想定し，そのデータを所定の許容時間内で処理できる方式を新たに提案した．さらに，スムージング最適化による推薦精度向上のための方法をPPDM環境で具体化し，PPDM環境における推薦精度向上の一手法を新たに提案した．

この組織間プライバシ保護リコメンド技術により，多くの組織が統計的リコメンドを実施可能となり，これが社会全体における統計的リコメンドの適用範囲の拡大の一助となれば幸いである．

[*20] 式(5)が左辺の第3項のみで成り立つ場合のα_3を考えると，$\alpha_3 > \frac{W^2}{N_c} \sim 0$であり，$\alpha$の値が小さい場合でも式(5)を満たす．

参考文献

[1] Yao, A.C.-C.: How to generate and exchange secrets, *Proc. SFCS '86*, pp.162–167 (1986).
[2] Goldreich, O.: Foundations of Cryptography, *Basic*

Tools, Vol.1, Cambridge University Press (2001).
[3] Agrawal, R., Evfimievski, A. and Srikant, R.: Information sharing across private databases, *Proc. ACM SIGMOD 2003*, pp.86–97 (2003).
[4] 千田浩司，寺田雅之，山口高康，五十嵐大，濱田浩気：セキュアマッチングを用いた組織間クロス分析，情報処理学会コンピュータセキュリティシンポジウム 2010 (2010).
[5] Vaidya, J., Kantarcioglu, M. and Clifton, C.: Privacy-preserving Naive Bayes classification, *The VLDB Journal*, Vol.17, No.4, pp.879–898 (2008).
[6] Ioannidis, I., Grama, A. and Atallah, M.: A Secure Protocol for Computing Dot-Products in Clustered and Distributed Environments, *Proc. ICPP 2002*, pp.379–384, IEEE Computer Society (2002).
[7] Freedman, M.J., Nissim, K. and Pinkas, B.: Efficient private matching and set intersection, *Proc. Eurocrypt 2004*, pp.1–19 (2004).
[8] 菊池浩明，香川大介，石井一彦，寺田雅之，本郷節之：組織間プライバシー保護データマイニングの考察，電子情報通信学会暗号と情報セキュリティシンポジウム 2010，pp.1–5 (2010).
[9] Minka, T.P.: Estimating a Dirichlet distribution, Microsoft (online), available from ⟨http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf⟩ (accessed 2014-11-28).
[10] Malkhi, D., Nisan, N., Pinkas, B. and Sella, Y.: Fairplay – A Secure Two-party Computation System, *Proc. SSYM 2004*, p.20 (2004).
[11] Yang, B. and Nakagawa, H.: Computation of Ratios of Secure Summations in Multi-Party Privacy-Preserving Latent Dirichlet Allocation, *PAKDD 2010*, pp.21–24 (2010).
[12] Tran, D.H., Ng, W.K., Lim, H.W. and Nguyen, H.-L.: An Efficient Cacheable Secure Scalar Product Protocol for Privacy-Preserving Data Mining, *Data Warehousing and Knowledge Discovery, Lecture Notes in Computer Science*, Vol.6862, pp.354–366 (2011).
[13] 増村　亮，咸聖　俊，伊藤彰則：確率的言語モデルに基づく音声ドキュメント検索のための Web を利用したモデル拡張の検討，情報処理学会研究報告，Vol.SLP-84, No.20, pp.1–6 (2010).
[14] Paillier, P.: Public-Key Cryptosystems Based on Composite Degree Residuosity Classes, *Advances in cryptology – EUROCRYPT '99*, Vol.1592, pp.223–238 (1999).
[15] Mitchell, T.M.: *Machine Learning*, McGraw Hill (1997).
[16] 竹村彰通：個票開示問題の研究の現状と課題，統計数理，Vol.51, No.2, pp.241–260 (2003).

山口 高康 （正会員）

2001 年電気通信大学大学院電気通信学研究科博士前期課程修了．同年株式会社 NTT ドコモ入社．以後，携帯端末での，撮影対象判別技術，権利価値流通技術，コンテンツ検索技術，統計情報作成技術，プライバシ保護技術の研究開発に従事．

寺田 雅之 （正会員）

1995 年神戸大学大学院工学研究科修士課程修了，同年日本電信電話株式会社入社．同社情報通信研究所，情報流通プラットフォーム研究所を経て，2003 年株式会社 NTT ドコモへ転籍．2008 年電気通信大学大学院電気通信研究科博士後期課程修了．博士（工学）．情報セキュリティ技術，プライバシ保護技術，大規模データ処理技術の研究開発に従事．電子情報通信学会会員．