プロファイルと履歴を組織間で統合利用するプライバシ保護推薦システム

山口高康
電気通信大学大学院 情報理工学研究科
博士 (工学) の学位申請論文

2019 年 3 月

プロファイルと履歴を組織間で統合利用する
プライバシ保護推薦システム

博士論文審査委員会

主査　吉浦　裕　　教授
委員　太田　和夫　教授
委員　﨑山　一男　教授
委員　坂本　真樹　教授
委員　大坐畠　智　准教授

Abstract

Privacy-Preserving Recommender System Joining Profiles and Records Among Organizations

Statistical recommendation of goods and services becomes important by making use of such data as purchase records. However it is only used by large organizations that hold all the necessary data such as purchase records, data about goods/services, and user profiles. Customers' privacy may be infringed because these large organizations hold all information about them. Our research objective is to establish an Inter-Organization Privacy-Preserving Recommender System Based on Secure and Efficient Use of Profiles and Purchase Records (Inter PPR) that enables recommendation by using data from cooperating organizations while protecting privacy of the organizations and customers.

Our model of Inter PPR consists of an ID manager, retailers, and customers. The ID manager provides IDs and profiles of users while a retailer provides purchase records. Content-based recommendation with the profiles enables accurate recommendations from small amount of data. The user IDs are used to join databases from organizations to generate a cross-tabulation table. Credit-card and telephone companies are candidates for the ID manager. Private set product and inner product protocols, lightweight cryptographic tools, are used to generate the cross-tabulation table and recommendation values respectively while hiding data of each player, i.e. ID manager, retailer, and customer. Smoothing is used for precise recommendation from small amount of data. Differential privacy (DP) is applied to the cross-tabulation table so that the retailer cannot infer the profiles from it.

Because DP heavily degrades data such as a cross tabulation table that has many attributes, we develop two DP methods. One method uses only part of the attributes of the table to reduce degradation of the table. It then complements information of other attributes from the used attributes. The other method normalizes data in the table to maintain S/N ratio by shrinking the signal (i.e. original table values) while more shrinking noises from DP.

We have implemented core parts of Inter PPR and evaluated them to show the system's achievement of privacy, recommendation preciseness, processing efficiency, and usability.

概要

購買履歴等のデータを活用して顧客に商品やサービスを推薦する統計的推薦が盛んに利用されている．しかし，統計的推薦は購買履歴の他，商品情報やユーザプロファイルを必要とするため，これらの情報を全て利用可能な大組織しか実施できなかった．その結果，少数の大組織によるビジネスの寡占および中小組織の凋落が懸念されている．また，大組織が，ユーザプロファイルと購買履歴を全て管理するため，ユーザのプライバシが損なわれる懸念がある．そこで，複数の組織が連携して，各組織およびユーザのプライバシを確保しながら推薦するシステム Inter PPR の実現を研究目的とする．

中小組織は保有するデータが少ないので，データ量への依存が小さいコンテンツベース推薦を採用する．コンテンツベース推薦を高精度化するにはユーザプロファイル (以下プロファイル) が必要であるが，プロファイルは個人情報であるため，中小組織には管理負担が大きい．また，複数の組織のデータを結合するには共通のユーザ ID が必要だが，その管理負担も大きい．そこで，ユーザ ID とプロファイルの管理を日常的に行っているカード会社や携帯電話会社を ID 管理組織とし，購買履歴を保有する商店と連携する．ID 管理組織と商店が互いのデータを秘匿しながら，クロス集計表を生成し，これを用いて商店がユーザに推薦する．クロス集計表の生成には，効率の良い暗号技術である秘匿積集合プロトコルを用いる．また，プロファイルや購買履歴の情報がクロス集計表から漏洩することを防止するために差分プライバシを用いて匿名化する．クロス集計表のような多属性データは差分プライバシにより大幅に劣化するため，劣化の少ない多属性対応差分プライバシを新たに提案する．推薦精度を向上するためにスムージングを用いるが，従来のスムージング方式はプロファイルと履歴情報の両方を必要とするので，大組織しか利用できない．そのため，クロス集計表に適用可能な分散環境対応スムージングを新たに提案する．商店とユーザが互いのデータを秘匿しながら推薦するために，効率の良い暗号技術である秘匿内積プロトコルを用いる．さらに，ID 管理組織が複数の商店との間でクロス集計表を各々生成し，これらを統合した後，差分プライバシで匿名化して各商店に配布することで，3 つ以上の組織の連携を可能とし，推薦精度をさらに向上する．

Inter PPR の部品のうち従来存在しない技術として，分散環境対応スムージングと多属性対応差分プライバシを提案する．従来のスムージングは，ID 管理組織のプロファイルと商店の履歴情報を両方必要とするので，Inter PPR には利用できない．また，パラメータ数が，プロファイルの属性数 × 商品数となり多いため，データが少ない場合にパラメータを最適化できず，精度向上の効果が発揮できない．そこで，クロス集計表に適用可能

で，パラメータ数が商品数となるスムージング方式を提案した．

差分プライバシは匿名化のためにデータにノイズを付加するが，そのノイズの大きさはデータの属性数に比例する．クロス集計表は多属性であるためノイズが大きくなり，集計表の有用性が失われる．そこで，一部の属性のみを考慮し，その結果生じるデータの劣化を属性間の関係によって補正する．属性間の関係は元データの情報を含むので，これも匿名化が必要となるが，属性間の関係に加えるノイズが小さければ全体として有用性が維持される．また，クロス集計表の中の情報を個人単位で正規化して，突出した個人情報の漏洩可能性を低減し，匿名化に必要なノイズを低減する．クロス集計表は統計であり，個人単位に切り分けることはできないが，秘匿積集合プロトコルによるクロス集計表の生成過程に立ち入ることで，個人単位の正規化を可能にする．

提案した Inter PPR を実装し，そのプライバシ，推薦精度，処理性能，社会実装容易性を評価した．プライバシについては，秘匿積集合プロトコル，秘匿内積プロトコル，差分プライバシの組合せ効果を評価し，シームレスなプライバシ保護を確認した．推薦精度については，提案した分散環境対応スムージングおよび多属性対応差分プライバシの有効性とその組合せ効果を明らかにした．処理性能については，ID 管理組織が 1,000 万顧客の 57 属性，商店が 10 万顧客と 1,000 商品の購買履歴，ユーザが 57 属性を用いる場合に，1 か月に 1 回の頻度でクロス集計表を更新し，ユーザに 5 秒で推薦できることを明らかにした．社会実装容易性については，ID 管理組織の導入による共通ユーザ ID およびプロファイルの管理容易化，多組織間への拡張の可能性を明らかにした．

以上を通じて，シームレスなプライバシ保護と実用性を備えた組織間連携型推薦システムの構成法，分散環境対応スムージングおよび多属性対応差分プライバシの方式を明らかにした．

目次

第 1 章

序論

1.1 研究の背景

ネットワークや計算機の処理能力の向上によって，私たちの身近な物や事に関する情報が容易に集められるようになり，収集された巨大なデータ (ビッグデータ) を分析して私たちの生活に役立てる取り組みが広まっている．たとえば，2017 年には 1 分間あたり，Amazon が 258,751USD を売り上げ，Google が 3,607,080 回の検索結果を返し，Instagram が 46,740 枚の写真を掲載し，Netflix が 69,444 本の映画を流し，Twitter が 456,000 件の呟きを広め，Uber が 45,787 人の乗車を手配している [1]．ビッグデータには，ソーシャルメディアデータ，マルチメディアデータ，カスタマーデータ，センサーデータ，オフィスデータ，ログデータ，オペレーションデータ，ウェブサイトデータといったものがあり，利用者のニーズに合ったサービスを提供したり，サービス提供者の事業を効率化できるようになると期待されている [2]．ビッグデータの一つであるウェブサイトデータには，電子商取引 (EC: electronic commerce) サイトなどで蓄積される商品の購入履歴などが含まれる．このようなデータを分析して利用者の気を引き，商品やサービスの購入に結びつける統計的推薦 (以後，推薦と略す) は，ビッグデータの代表的なアプリケーションの 1 つとなっている．たとえば，推薦を活用して EC サイト最大手となった Amazon は，2011 年から 2015 年にかけて売上が倍増し，1,000 億ドルの大台を超えた [3]．このように，近年の商取引における，顧客の購入履歴などの統計的性質に基づいた商品の推薦は，企業の売上拡大のための有力な手段となっている．

現在の推薦は，推薦を行う組織が必要な情報を全て保有していることを前提としている．蓄積されたデータの量が推薦精度を左右するため，ユーザが大規模なデータを保有する組織 (Amazon などの大組織) の会員になれば，会員番号という個人識別子 (ID:

identification) に紐づけられて情報が蓄積され続け，その蓄積されたデータに基づく的確な推薦を得られるようになる．また，組織は，会員番号と共に，性別や年代などのプロファイルを保有し，推薦に利用することができる．しかし，ユーザは大規模な組織からしか推薦を得られず，商品の選択肢が狭くなる．また，商品を購入する組織の選択肢が減ることは，ベンダロックオンに繋がる危険性がある．ユーザがその組織のサービス無しでは生活できない状況になってしまったら，その組織の思いのままに商品の価格をコントロールされてしまう可能性もある．さらに，大組織がユーザのプロファイルや購入履歴を全て保有することから，ユーザのプライバシが侵害される懸念が出てくる．

一方，中小組織では，推薦に用いることができる情報が限られるが，セッション ID(来店毎に都度割り振られる ID) に紐づけて情報を蓄積し，その蓄積したデータに基づく推薦が行なわれることがある．たとえば，来訪客による一度の買い物 (同じレシートの中) で，一緒に買われる商品を統計化し，紙オムツとビールが一緒に買われているならば，店内の近くに両者を配置するといった工夫が可能である．これは，紙オムツを買う来訪客にビールを推薦していることを意味する．しかし，このような工夫にも関わらず，データ数の少なさ，およびユーザ ID に基づく継続的な購入履歴の蓄積が困難であることから，大規模なデータを保有する組織に比べて推薦の精度が大きく劣る．また，個人情報保護が叫ばれる中，個人のユーザ ID やそれに紐づくプロファイルなどの個人情報の管理は，中小組織にとって，大きな負担である．

1.2 先行研究

推薦の代表的なアルゴリズムに協調フィルタリング方式とコンテンツベース方式がある [4]．協調フィルタリング方式は，多数のユーザの購入履歴の中から類似するユーザやアイテムの情報を探して，ユーザ毎の嗜好に合わせて商品を推薦する [5]．しかし，履歴が少ない場合は，この類似度を正しく求められなくなるため，推薦精度が低下してしまう．Amazon は，自社で保有する膨大な履歴から類似度を算出することで，高い推薦精度を確保している [6]．協調フィルタリング方式を用いる場合は，多くの履歴データを保有している大組織が推薦精度の面で有利である．情報検索に端を発するコンテンツベース方式は，協調フィルタリングに比べて少ないデータで推薦することができ，推薦対象の商品の内容を上手く特徴づけられれば推薦精度を高めることができる [7]．コンテンツベース方式は，書評やレビューなどから商品の内容を特徴づけやすいため，書籍や映画などの推薦で用いられることが多い．コンテンツベース方式で用いる特徴は，推薦対象の商品の内容から得られるものだけでなく，ユーザのプロファイル (性別や年代など) なども用いるこ

とで精度を向上することができる [8]．ユーザのプロファイルを利用するには，ユーザに ID を振り，プロファイルと共に管理すればよい．しかし，ユーザのプロファイルを収集・更新することは，個人情報保護の面から，中小組織には負担が大きい．コンテンツベース方式を用いる場合も，ユーザの ID やプロファイルなどの個人情報を保有している大組織が推薦精度の面で有利である．

プライバシを保護しながらデータを解析する技術は，暗号応用による技術 [9] と匿名加工による技術 [10] に大別される．暗号応用によるものは，データを秘匿したまま演算を行うことにより，プライバシを保護しながら分析を行えるようにする．しかし，複雑な演算を行うには多くの処理時間がかかるため，推薦のように多くのデータを複雑な演算で処理しなくてはならない場合は，大きな遅延が生じる．また，鍵の管理も容易ではない．匿名加工によるものには，たとえば，データにランダム性を加えて，それぞれのデータを真の値から異なる値へと変えることにより，プライバシを保護するものがある．そして，プライバシ保護後のデータを集約して，それぞれのデータに加えられたランダム性を相殺させて分析を行う．しかし，どれだけランダムにすれば安全なのかを判断することが難しい．安全性を優先して強く匿名加工すると，プライバシ保護後のデータの値が元のデータの値とかけ離れた値になってしまうために，精度が著しく低下する．

スムージングとは，データの特異な値やノイズを平滑化する処理である．スムージングを推薦に適用すると，限られた少ないデータをスムーズにして，推薦精度を高める効果が得られる．スムージングは協調フィルタリング方式とコンテンツベース方式のどちらにも適用できる．協調フィルタリング方式は，ユーザや商品の種類が多くなると類似度の行列の要素数が急激に増加し，その行列を埋めるデータが足りなくなるので，それを補うスムージングの手法が研究されている [11, 12, 13]．情報検索においては，古くからスムージングの適用がなされており，その流れを受け継ぐコンテンツベース方式への適用も容易である．しかし，いずれの場合でも高度なスムージングの適用には，推薦に用いる商品の購入履歴やユーザのプロファイルなどの情報が必要である．これらの情報を 1 つの組織が保有していれば，問題なくスムージングを利用できるが，情報が複数の組織に分散している場合には，一方の組織からもう一方の組織へプライバシが漏洩してしまう問題が生じる．

1.3　本研究の目的

先に述べた背景および先行研究を踏まえると，複数の組織が連携し，各組織およびユーザのプライベートな情報を保護しながら，大組織と同等あるいは大組織に近い精度でユーザに推薦を行うシステムが社会的に強く求められる．このシステムは，プライバシと推薦

精度だけでなく，処理性能，信頼性，設備や管理コストの点でも実用性が求められる．また，中小組織による利用を想定すると，ユーザ ID とプロファイルの管理，および購入履歴との統合を無理なく実現できることが重要である．しかし，著者の知る限り，このようなシステムは従来提案されたことはない．そこで，このシステムを Inter-Organization Privacy-Preserving Recommender System Based on Secure and Efficient Use of Profiles and Purchase Records (Inter PPR) と名付け，本論文において提案する．

本論文の目的は，Inter PPR のコンセプトと実現方法を確立することである．そのために Inter PPR に関して以下の目標を設定する．

1. 技術要件を明らかにする．
2. ユーザ ID とプロファイルを無理なく管理する方法を明らかにする．
3. システム構成を設計する．その際，暗号応用と匿名加工の適切な組み合わせによって，プライバシ，推薦精度，処理性能を実用レベルで確保する．従来技術から適切な要素技術を選定するとともに，新たに創出すべき要素技術を明らかにする．
4. 新たに創出すべき要素技術として，分散環境対応スムージングおよび多属性対応差分プライバシを提案し，有用性を明らかにする．
5. 多組織間への拡張性等の社会実装容易性を確保する．

上記の目標を達成することにより，ユーザメリットの確保と産業の発展に資することが期待できる．

1.4 本論文の構成

本論文では，2 章で，Inter PPR 実現の観点から，先行研究を概観・分析し，有用な要素技術を明らかにすると共に，従来技術の限界を明らかにする．3 章では先行研究の分析を踏まえ，Inter PPR のシステム構成を設計し，従来技術から最適な要素技術を選定すると共に，新たに創出すべき要素技術を明らかにする．4 章では，設計したシステム構成に沿って，データ表現と処理フローを設計し，選定した要素技術を適用することにより，Inter PPR を具体化する．5 章では新たに創出すべき要素技術のうち分散環境対応スムージングを提案し，定式化する．6 章では，新たに創出すべき要素技術のうち多属性対応差分プライバシを提案，定式化し，その推薦精度への影響を評価する．また，有用性について，分散環境対応スムージングと合わせて評価する．7 章では，システムの中核部分を実装し，ユースケースに沿って，Inter PPR のシステムとしてのプライバシ，推薦精度，処理性能，社会実装容易性を評価する．8 章で結論と今後の課題を述べる．

第 2 章

先行研究

2.1 はじめに

前章では，ビッグデータの活用が期待されている中，統計的推薦が重要になっているが，多くの履歴データやユーザの個人情報を保有している大組織しか統計的推薦を利用できないことを述べた．中小組織は，大規模な履歴データを収集できない上，個人情報保護の面から，ユーザの個人情報の管理が困難である．その結果，大組織による情報とビジネスの独占により様々な弊害が生じている．そのため，複数の組織が連携して各々のプライベートな情報を秘匿しながら推薦するシステム Inter PPR を提案した．本章では，Inter PPR の技術を確立する観点から，推薦，暗号応用，匿名加工，スムージングについて先行研究を概観する．Inter PPR に利用可能な要素技術を明らかにすると共に，従来技術の限界を明らかにする．

2.2 推薦技術

推薦の手法には，ルールベース方式 [14]，アソシエーション分析 [15]，協調フィルタリング方式 [5, 16]，コンテンツベース方式 [17, 18]，ベイジアンネットワーク [19, 20] などがある．

ルールベース方式は人手により記述されたヒューリスティックなルールを用いて推薦を行う．ベテランの人間の経験をルール化するが，人によって解釈の仕方が多様であること，ビッグデータとそれに含まれる知識は日々増え続けることから，あらゆるルールを人間が記述し続けることは困難であるため，ビッグデータの良さを活かすことはできない．

アソシエーション分析は，蓄積したデータから有益なパターンや組み合わせを発見する

分析であり，代表的なものにマーケットバスケット分析がある．マーケットバスケット分析は，紙オムツとビールの併売傾向などの相関性を探り出す分析手法であり，販売時点情報管理 (POS: point-of-sale) データや EC サイトの分析に用いられる．しかし，アソシエーション分析は主にマーケティングで用いられ，ユーザ毎の嗜好に合わせた推薦までを行うものではない．

協調フィルタリング方式は，多数のユーザの購入履歴の中から類似するユーザやアイテムの情報を探して，各々のユーザの嗜好に合わせた商品を推薦する．協調フィルタリング方式は，ユーザやアイテムの種類が多くなると類似度の行列の要素数が急激に増加し，その行列を埋めるデータが足りなくなって類似度を正しく求められなくなり，推薦精度が低下してしまうという問題 (コールドスタート問題) がある．協調フィルタリング方式には，似た商品を評価したユーザを類似ユーザとし，類似ユーザが評価していて対象ユーザがまだ評価していない商品を推薦するユーザ間型と，似たユーザに評価された商品を類似商品とし，対象ユーザが評価した商品の類似商品を推薦するアイテム間型がある．履歴を記憶しておいて推薦候補を予測する方法をメモリベース法と呼び，履歴から類似度行列などを算出しておいて推薦候補を予測する方法をモデルベース法と呼ぶ．メモリベース法はいろいろな尺度で類似度を測れるという利点があるが，データの量に応じて推薦にかかる時間が長くなるという欠点がある．モデルベース法は尺度を変える度に類似度を測り直さなくてはならないという欠点があるが，類似度を事前計算しておけば推薦にかかる時間が短くて済むという利点がある．高い推薦精度を得るためには多くのデータが必要であるため，多くのデータを利用できるモデルベース法が有利である．Amazon は，アイテム間型モデルベース法協調フィルタリング方式を用いて「この商品を購入した方は，他にこれらの商品も購入してます」という推薦を提供しているが，自社で保有している膨大なデータによって有効な推薦ができている．協調フィルタリング方式を用いる場合は，多くの履歴データを保有している大組織が推薦精度の面で有利である．

コンテンツベース方式は，情報検索から発展した推薦の手法であり，商品などの推薦対象を上手く特徴づけることによって推薦の精度を高めることができる．コンテンツベース方式は，協調フィルタリングに比べて原理的にデータ量への依存度が小さいため，協調フィルタリングが陥りやすいコールドスタート問題，すなわち，新商品などを推薦する場合にデータ量が少ないため精度が低くなる問題を緩和しやすい．その代償として，推薦対象へその特徴を付与する手間が避けられないため，コンテンツベース方式は商品の特徴を自動で抽出しやすい書籍や映画などの推薦で用いられることが多いが，推薦対象への特徴の付与にユーザのプロファイルなどを利用することもできる [4, 8]．たとえば，ある商品を良く購入しているユーザの年代と性別を抽出して，「30 代の男性に好まれている商品」

という特徴を付与することもできる．だが，ユーザのプロファイルなどを含む個人情報の管理は中小組織にとって負担が大きいため，コンテンツベース方式を用いる場合も，ユーザのプロファイルを保有している大組織が推薦精度の面で有利である．

ベイジアンネットワークは，因果関係を確率分布のネットワークでモデル化するものである．たとえば，商品の購入履歴 (因果の果に相当) にあるユーザのプロファイル (因果の因に相当) の因果関係を学習しておけば，ユーザのプロファイルを入力として有り得そうな (因果が妥当な) 商品をベイジアンネットワークで出力することができるため，ユーザのプロファイルに応じた商品を推薦することができる．しかし，ベイジアンネットワークで取り扱うような複雑なモデルの学習には多くの学習データが必要であるため，ベイジアンネットワークを用いる場合も，多くの履歴データやユーザの個人情報を保有している大組織が推薦精度の面で有利である．

2.3 ID 管理技術

Inter PPR で複数の組織のデータを結合するためには，ユーザに共通の ID を付与する必要がある．また，2.4.2 節で述べる組織間の暗号プロトコルは，組織間に共通のユーザ ID が存在することを前提としている．共通 ID の付与と管理は様々な分野で必要とされる普遍的な課題であり，分野毎に手法が提案され，標準化と実用化が重ねられてきた [21]．ID はユーザの識別子を指すが，一般に ID 管理技術は識別子だけでなくユーザの属性情報や認証に関わる情報も含めたアイデンティティの管理を意味する．また，ユーザの性別などの属性情報をプロファイルと呼び，その認証において正当なユーザであることを示す情報をクレデンシャルと呼び，ID とプロファイルとクレデンシャルを合わせてアイデンティティと呼ぶ [22].

大規模な ID としては，国が国民に固有の番号を振って特定個人を管理しやすくするために付与した国民識別番号があり，1936 年にアメリカ合衆国で社会保障番号が開始され，1948 年にはイギリスで国民保険番号とシンガポールで国民登録番号が開始された．

IT 分野においては，1960 年代に複数のユーザで同じ計算機を利用するメインフレームの Time Sharing System のユーザへ ID が付与されるようになり，ID とクレデンシャルとの組み合わせでユーザ認証できるようになったが，プロファイルはまだ用いられていなかった．1980 年代になると，パーソナルコンピュータとローカルエリアネットワークが普及して多数の計算機が接続されるようになり，ITU-T(国際電気通信連合 電気通信標準化部門) はアイデンティティにプロファイルも記述できる X.500 の規格を策定し，マサチューセッツ工科大学はアイデンティティを集中管理できる共通認証プラットフォームで

ある Kerberos を実現した.

1990 年代にはインターネットが普及し始めて世界中の計算機が接続されるようになり，Internet Engineering Task Force の Request for Comments でディレクトリサービスの Lightweight Directory Access Protocol (LDAP) が策定され，Open Group は Kerberos と LDAP を組み合わせて Distributed Computing Environment を実現し，後に Windows サーバ OS の基にもなる，同一ドメイン内 (同一組織内) での Single Sign-On (SSO) を実現した.

2000 年代には Web テクノロジーが進化して，ID 連携と呼ばれる組織間を横断した Web SSO が行われるようになり，the Organization for the Advancement of Structured Information Standards はアイデンティティに関する情報を eXtensible Markup Language で記述する規格である Security Assertion Markup Language (SAML) を策定した．しかし，SAML は各ドメインで独自の ID 体系を構築していることを前提とし，ドメイン間の ID を連携する技術である．Inter PPR では，中小組織にとって独自のユーザ ID を付与して管理することが負担であるため，SAML の適用は困難である.

OpenID Foundation は SAML のような事前の信頼関係が不要でアイデンティティに関する情報を自由に交換できる OpenID を提案した．OpenID 自体は，Web サービスにおいてユーザのリソースへのアクセスを他の Web サービス提供者に認可する機能を持たないため，OAuth に依存していた．2009 年に OAuth 1.0 に重大なセキュリティホールが見つかり，パッチで対応されたものの，2012 年には後方互換性を諦めて OAuth 2.0 にバージョンアップされた．OAuth 2.0 は Facebook のソーシャルグラフへアクセスする Graph API の実装に用いられたことで知られ，Google や Microsoft でも試験的に利用された．しかし，OAuth 2.0 は曖昧であるため，実装次第でセキュリティホールが生じてしまうという問題がある．2014 年の調査では，Google Play や Apple マーケットといった公式アプリサイトで提供されていた OAuth 2.0 を使用するモバイルアプリの 6 割 (89 種類) に脆弱性が発見された [23]．2016 年の調査でも条件は異なるが 4 割 (85 種類) のモバイルアプリに脆弱性が発見されている [24]．OAuth 2.0 の利用の是非については，金融機関のように安全性を最優先すべき組織でも足並みが揃っていない．2017 年に全国銀行協会は OAuth 2.0 の利用を推奨しているが [25]，2018 年に日本銀行は OAuth 2.0 の脅威について指摘している [26]．現在検討されている ID 管理技術には OAuth の後継となる OpenID Connect も検討されているが，これまでの Web サービス提供者を起点とした技術ではなく，よりユーザ起点でのデータの扱いを重視した User-Managed Access や openPDS といったパーソナルデータサービスが注目を集め始めている．しかし，これらはまだ発展途上の段階であり，社会に受け入れられて共通 ID が広く普及するまでには，まだかなりの時間を要す

ると考えられる．

以上のように，Inter PPR におけるユーザ ID の付与と管理に適した技術は現時点では存在しない．また，この種の技術の普及には標準化が必要となるため，技術の開発と実用化には長期間を要する．

2.4　組織間連携のためのプライバシ保護技術

2.4.1　保護対象と保護技術の概要

組織間連携における保護対象は，参画する主体が各々保有している情報であり，これを他の主体に対して秘匿する必要がある．ここでは，図 2.1 に示すように，組織 1 と組織 2 がそれぞれ保有している元データからユーザ ID をキーとして片方の組織 2 に合成データを生成し，その組織を訪れた訪問者のプロファイルに応じて推薦を行う場合を考える．守るべきものは，組織 1 の元データ 1 と，組織 2 の元データ 2 と，訪問者のプロファイルである．たとえば，元データ 1 はユーザ ID と年代と性別からなる属性の情報であり，元データ 2 はユーザ ID と商品の情報であり，訪問者のプロファイルは性別と年代からなる属性の情報であり，他者に知られたくない．

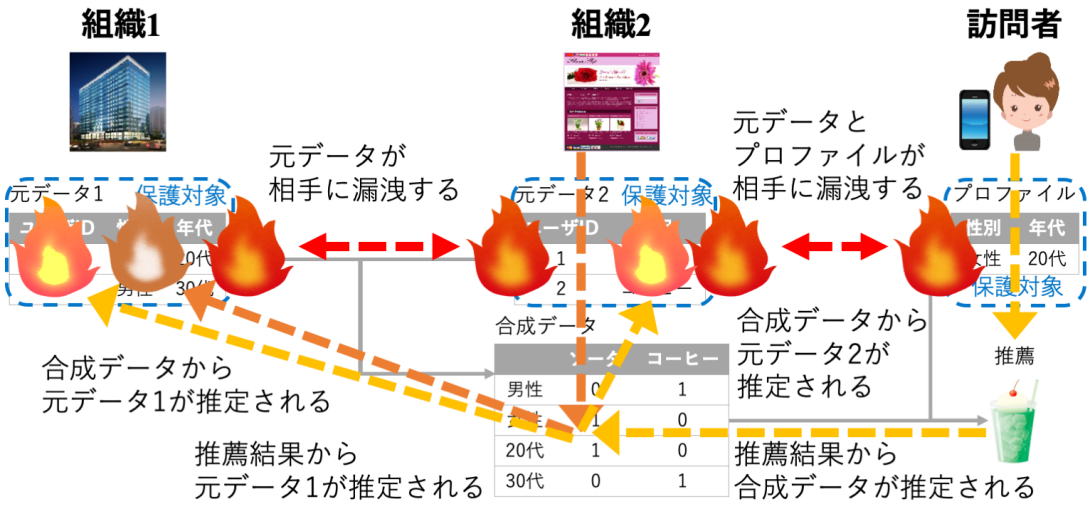


図 2.1　組織間連携におけるプライバシの漏洩可能性

しかし，組織 1 から組織 2 へ合成データを提供する過程において，組織 1 と組織 2 の情報のやりとりから，組織 1 の元データ 1 が組織 2 へ，組織 2 の元データ 2 が組織 1 へ互いに漏洩する可能性がある．また，組織 2 から訪問者へ推薦を提供する過程において，組

織 2 と訪問者の情報のやりとりから，組織 2 の合成データが訪問者へ，訪問者のプロファイルが組織 2 へ互いに漏洩する可能性がある．

さらに，提供された情報から，提供された情報の元となった情報が漏洩する可能性がある．すなわち，組織 1 から組織 2 へ提供された合成データから，組織 2 が元データ 1 を推定することで，元データ 1 が組織 2 へ漏洩する可能性がある．また，組織 2 から訪問者へ提供された推薦の結果から，訪問者が合成データを推定する可能性がある．合成データは保護対象ではないが元データ 1 と元データ 2 から生成されたものであるため，訪問者が合成データからさらに元データ 1 と元データ 2 を推定することで，元データ 1 と元データ 2 が訪問者へ漏洩する可能性がある．

一般に，組織間の情報提供の過程における漏洩は，情報のやりとりを秘匿できる暗号応用 [9] で防止できる．また，組織間で提供された情報からの漏洩は，その情報の元となった情報を保護できる匿名加工 [10] で防止できる．これらの技術を組織間のプライバシ保護に適用すると図 2.2 のように情報の漏洩を防止できる．それぞれ，暗号応用と匿名加工の先行研究について以下に述べる．

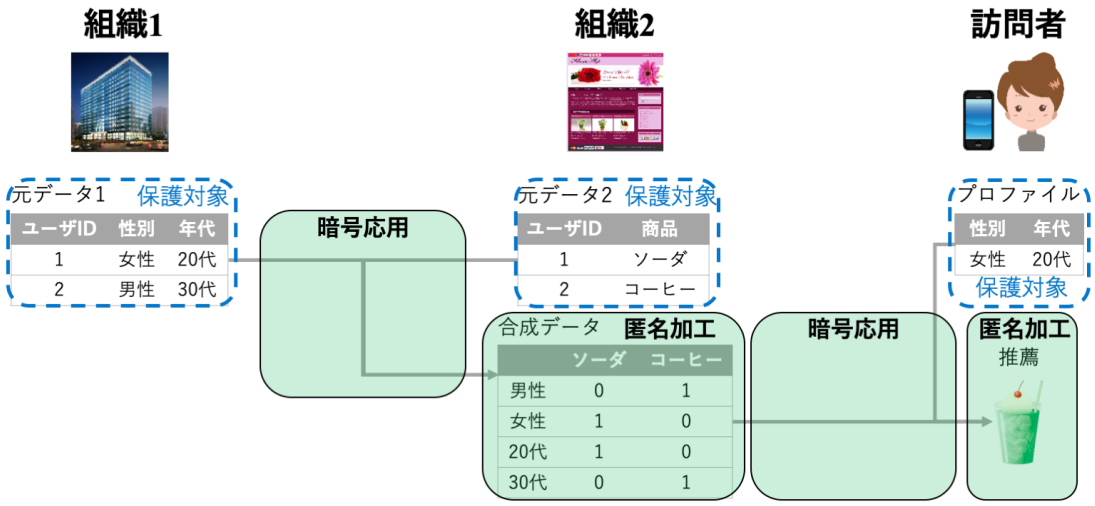


図 2.2　暗号応用と匿名加工の利用によるプライバシ保護

2.4.2　暗号応用

組織間の情報のやりとりを秘匿して計算ができる暗号応用については，1982 年の Yao による多項式評価を用いた 2 者間での秘密回路 [27] が始まりだと言われている．Goldreich らは，1987 年に多項式評価を多者間に拡張した [28, 29].

1988 年に Ben-Or らは多者間で任意の計算ができる Multi Party Computation[30] を考案した．Multi Party Computation はその適用範囲の広さから，その後も多くの研究がなされている [31, 32, 33, 34, 35, 36]．任意の計算ができる暗号応用としては，Multi Party Computation の他，2004 年の Malkhi による万能関数計算機 [37] や，2009 年の Gentry による完全準同型暗号 [38, 39] がある．しかし，これらの暗号応用は，任意の計算ができるという柔軟性を持つ反面，処理コストが大きいため計算が遅い．また，鍵管理や Trusted Third Party 設定など，運用における問題も複雑である．たとえば，Gentry の完全準同型暗号は，鍵や暗号文が数百メガバイトから数ギガバイトにも及ぶため，実用に向けて研究がなされている途中である [40, 41].

任意の計算は加法と乗法の計算ができれば可能となるが，Paillier は 1999 年に加法の計算に特化して高速に組織間の情報のやりとりを秘匿して計算する手法を提案した [42]. Paillier の手法は，準同型性を持つ Paillier 暗号を用いて暗号化したまま元のメッセージの足し算を行う．Paillier 暗号のように加法準同型性を有するものには modified-ElGamal 暗号がある．一方，暗号化したまま元のメッセージの掛け算を行える乗法準同型性を有するものには RSA 暗号や ElGamal 暗号がある．Cramer は 2000 年に，RSA 暗号を用いて高速に組織間の情報のやりとりを秘匿して計算する手法を提案している [43]．しかし，これらの暗号応用は，処理コストは抑えられるが計算方法が限定されるため，組織間で行えるタスクも限定されてしまう．

組織間暗号プロトコルの共通部品を対象とした，効率的な暗号プロトコルが提案されている．秘匿和集合は 2004 年に Kantarcioglu によって提案された [44]．秘匿積集合は，上記で述べた 1982 年の Yao の [27] に端を発するが，効率面では，2003 年に Agrawal らによってハッシュタグの照合を用いて 2 者間で高速に秘匿積集合を行う方法が提案された [45]．多者間における効率的な秘匿積集合も検討されている [46, 47]．安全面では，プロトコルに入力されるデータの要素数を秘匿できる秘匿積集合 [48] や，相手に悪意があっても秘匿積集合を行える方法 [49] も提案されている．秘匿内積は，ベクトルの要素の積とそれらの和を求める組み合わせ計算であり，2004 年以降に提案されている [50, 51]．秘匿内積は応用範囲が広い反面，さらなる高速化が望まれている．2014 年にはベクトルの次元数を削減する高速化手法 [52] が提案されているが，高速化による精度低下は避けられず，誤差が生じる．

組織間のデータ統合のための暗号プロトコルについて述べる．互いのデータを秘匿しながら，その等結合を算出する秘匿等結合プロトコルは応用範囲が広く，秘匿積集合に基づく手法 [53]，秘密分散とマルチパーティ計算に基づく手法 [54] が提案されている．互いのデータを秘匿しながらクロス集計表を生成する秘匿クロス集計プロトコルも応用範囲が

広く，秘匿内積を直接用いる手法 [55] [*1]，冪乗余の可換性に基づく手法 [45, 56] がある．[56] は，a 台のデータベースを跨いでクロス集計表を生成するタスクにおいて，Agrawal らの秘匿積集合 [45] を繰り返し行うよりも，1/a 倍の計算量と通信量でクロス集計表を生成することができる．

コンテンツベース推薦の代表的な方法にナイーブベイズ [57] がある．Vaidya らは互いのデータを秘匿してナイーブベイズ推薦を可能にする方法を提案した [55]．Vaidya らの方法は秘匿内積を用い，組織間でプライバシを保護しながらナイーブベイズの推薦確率を求める．ただし，Vaidya らが用いた秘匿内積は [58, 59] の方式をベースにしており，組織間で同じ長さのベクトルの積和を求めるため，組織間でのデータの同期 (データのサイズと順序の一致) が前提となっている．非同期のデータを同期させると処理コストが大きくなる．

組織間のデータ統合のための暗号プロトコルの中には，組織間のデータを統合して推薦するための暗号プロトコルもあり，その代表例はプライバシ保護協調フィルタリングと呼ばれる [60, 61, 62, 63, 64, 65]．プライバシ保護協調フィルタリングの基本形は，各個人が自身の購入履歴 (あるいは各商品に対するレーティング情報) を保有する状況で，加法準同型暗号等の高効率な暗号を用いて，自身の購買履歴を暗号化する．サーバが各個人の暗号化された購買履歴を集約し，準同型性等を利用して，推薦のための類似度行列を生成する．しかし，これらの手法は，多数の個人の連携が前提となっているため，安全性や信頼性の観点からビジネスへの利用は困難である．たとえば，個人が意図的に誤った購買履歴を用いて，誤った類似度行列を生成させることが可能である．また，類似度行列を更新する毎に，多数の個人が参加する必要があるが，その保証は困難である．計算量の問題もある．たとえば，[62] の方式は [60] の効率を向上しているが，商品数がボトルネックとなる．具体的には，13 種類の商品の類似度行列を生成するのに通常の PC を用いて 5 分間かかり，仮にコンビニエンスストアで扱っている程度の 1 万種類の商品になると 50 年以上かかる．さらに，類似度行列から，その元になった各個人の購買履歴の情報が漏洩することが知られている [66].

Zhan らは，加法準同型暗号の代わりに，乱数ベースの秘匿内積プロトコルを用いて計算量を削減した [63]．また，Nikolaenko らは，Yao の garbled circuit[27] を用いた行列分解による推薦手法を提案している [67]．しかし，これらの手法は，上記のうち計算量以外の問題は解決していない．[66] では，類似度行列からの情報漏洩を防止するために，類似

[*1] Vaidya らが行なっている条件付き確率の計算が，クロス集計表の各セルの値を秘匿内積で繰り返し算出しているとみなせる．

度行列への匿名加工の手法を提案している．しかし，類似度行列の匿名加工は行列の情報を大きく歪めてしまい，推薦精度が大幅に低下することが判明している [68].

プライバシ保護協調フィルタリングにおいて，個人の代わりに組織が連携するように改良することは可能である．その場合，各組織として商店等を想定し，商店毎に複数ユーザの購買履歴を保有する状況で購買履歴を暗号化する．サーバが，各商店から暗号化された購買履歴を集約し，推薦のための類似度行列を生成する．個人ではなく組織が連携する形にすると，契約等によって安全性や信頼性を向上させることができる．Jeckmans らは，加法準同型暗号に加え，比較，絶対値算出および割り算の 2 者間暗号プロトコルを用いて，このアプローチを実現している [65]．しかし，このアプローチでは，各組織が複数ユーザの購買履歴を管理する必要がある．ユーザに番号を付与し，ユーザ毎の購買履歴を管理することは，個人情報保護の観点から中小組織には負担が大きい．

推薦用のデータを集約する時だけでなく，推薦を実行する時のプライバシも保護する必要がある．Basu らの方式 [69] では，ユーザが自分のレイティング情報を加法準同型暗号によって暗号化した後，推薦サーバに送ると，推薦サーバは秘匿内積プロトコルを用いて暗号化された推薦情報を生成し，ユーザに返す．商店とユーザの間に介在者を配置することで情報を分散し，商店とユーザの情報を秘匿する方式 [70] もある．Cissée らは，ユーザプロファイルに加え，推薦結果 (誰に何が推薦されたか) もセンシティブな情報とみなし，これらが推薦の時だけ一時的に生成され，推薦終了時に消去されるべきとしている [71]．そのために，マルチエージェントモデルに基づいて，販売者エージェントと推薦者エージェントを分離し，さらに，永続的な推薦者エージェントからテンポラリな推薦者エージェントを生成して 1 回の推薦を行い，推薦後にテンポラリな推薦者エージェントを消去する手法を提案している．

2.4.3 匿名加工

組織間で統合された情報から，その統合情報の元となった情報を推定されないように保護する匿名加工について述べる．プライバシの保護に用いられる匿名加工の適用先は，個人情報のレコードの集合である個票と，個票から算出した統計量に大別される．個票から統計量を計算できるが，統計量から個票は計算できない．個票は情報量が多いので強い匿名加工が必要だが，統計量は情報量が少ないので適度な匿名加工で十分である．個票に対する匿名加工の役割は，個票の各レコードと個人との対応 (以後，リンケージと呼ぶ) がつかないようにすることなどがある．また，統計量に対する匿名加工の役割は，統計量から元の個票の推定 (以後，インファレンスと呼ぶ) をできないようにすることなどがある．そ

して，これらのプライバシ保護に関する達成度の定義を匿名加工基準と呼ぶ．以下では，個票と統計量のそれぞれに対して，匿名加工の基準と，基準を満たすための加工方法について述べる．

個票の匿名加工のための基準には，k-匿名性 [72, 73]，l-多様性 [74, 75]，t-近接性 [76, 77]，δ-存在性 [78] がある．たとえば，$k = 10$ の k-匿名性を満たすとは，10 人未満に個人を特定できないように，同一のプロファイルのユーザが 10 レコード以上ある場合を指す．k の値はセキュリティパラメータであり，k の値が大きいほど，よりプライバシが保護されていることを表す．

個票を匿名加工する方法には，上位概念統合，置換，外れ値無視，ノイズ重畳がある．上位概念統合には，データをより上位の概念で表して個別のレコードに絞り込めないようにする k-匿名化 [72] がある．たとえば，1 才刻みの年齢ではなく，20 代，30 代というように大まかな年代で個票のデータを書き換えてプライバシを保護する．置換には，個票の値を一定の確率でランダムに別の値へ置き換える PRAM(post randomization method)[79, 80] がある．外れ値無視には，特異な値を含むレコードを削除する方法があり，特に大きな値を削除する方法はトップコーディングと呼び，逆に小さな値の場合はボトムコーディングと呼ぶ．ノイズ重畳には，個票のレコードに平均値が 0 のノイズを加える方法がある．平均値が 0 のノイズは個々のレコードの値をランダムに変化させるが，多くのレコードを用いれば全体としてノイズがキャンセルされるため，データの全体的な傾向を維持し，精度の低下を抑制する効果が期待される．組織間の連携においては，自組織のデータに平均値が 0 の正規分布や一様分布に従うノイズを加えてから他の組織へ送り，他の組織がデータの正確な値を分からないようにする方法が提案されている [81]．

組織間プライバシ保護推薦の匿名加工からのアプローチとしては，各組織が個票データ (たとえば，ユーザ毎の購買履歴) を匿名加工した後，これらを代表組織で集約する方法がある [10, 82, 83]．しかし，上述したように，個票データの匿名加工はデータを大きく歪め，推薦精度を低下させる．Inter PPR の想定利用者である中小組織の場合，保有するデータ量が限られるため，匿名加工がなくても推薦精度が低い．そのため，個票の匿名加工を Inter PPR に適用することは困難である．

統計量の匿名加工のための基準には，individual risk [84]，n-k dominance rule [85]，prior-posterior rule [85]，差分プライバシ [86, 87] などがある．individual risk は母集団一意性に基づいて統計量から個人が特定される程度を示す．n-k dominance rule と prior-posterior rule は，統計量のうち特に集計表のセルの値の公開または非公開を判断する基準である．n-k dominance rule は，特定のセルに集計される個人のうちの上位 n 人がセルのすべての人数の k% を占める場合に，当該セルを非公開とする．prior-posterior rule は，セルの真

の値を特定のしきい値内の誤差で推定できてしまう場合に，当該セルを非公開とする．差分プライバシ (Differential Privacy) は，統計量から漏洩する可能性のある情報量を数学的に定式化した基準である．

統計量を匿名加工する方法には，上記で述べた n-k dominance rule や prior-posterior rule に従って開示する統計量を制御する統計的開示制御 [88, 89] がある．また，個票を匿名加工する方法として述べた上位概念統合，置換，外れ値無視などを施した個票から統計量を求めても構わない．差分プライバシのための代表的な加工方法は，ラプラスノイズの重畳である．一般に，統計量は個票に比べて情報が少ないので，統計量に対する匿名加工は個票に対する匿名加工に比べて弱い加工で充分である．

なお，何れの匿名加工であっても，データはオリジナルに比べて必ず歪んでしまう．匿名加工の安全性は，本節の冒頭で説明したリンケージやインファレンスなどの攻撃を匿名加工によってどれだけ防げるかに関係し，データの有用性は，匿名加工後のデータにどれだけ多くの情報を残せるかに関係する．加工が弱すぎると元のデータの値が推定されて安全性が低下してしまい，加工が強すぎると元のデータの値からかけ離れてデータの有用性が低下してしまう．プライバシとデータの有用性の間にはこのようなトレードオフの関係があるため，一定のデータの有用性を保ちつつプライバシをより高めること，もしくは，一定のプライバシを保ちつつデータの有用性をより高めること (言い換えれば，情報の劣化を抑えること) が課題となる [90]．差分プライバシを満たす匿名加工であれば，安全性と有用性のトレードオフを定式化することができる．

2.5 スムージング

スムージングは，データから特異な値やノイズを平滑化する処理であり，推薦での精度向上のために用いることができる．訪問客のプロファイルを踏まえて商品を推薦するには，過去に観測したプロファイル毎の商品の販売数に基づいて売れるであろう商品を予測するが，データが少ない場合はこの販売数の分布が真の分布とずれてしまい，推薦精度が低下する．たとえば，商品の販売直後の場合には，商品がまったく売れていないプロファイルがありうる．その場合，観測した値はゼロとなり，その商品はそのプロファイルの人に絶対に売れないという誤った予測を引き起こすため [91]，推薦精度が著しく低下してしまう．しかし，スムージングは，このようなゼロの値を埋めることができ，あたかも多くのデータがあるような効果を引き出して，推薦精度を向上させることができる．

ゼロの値を埋める平滑化の程度などによって異なる各種のスムージングについては 5 章で詳しく述べるが，中でも Minka は，ディリクレ分布と呼ばれる汎用性の高い多次元の

離散分布を前提として，その最適なスムージングの方法を明らかにしている [92]．できる限り滑らかにするには，1 レコードずつ入力してディガンマ関数を計算するか，全体の積集合から 1 レコードずつ取り除いて leave-one-out 尤度 (以後，LOO 尤度と呼ぶ) を計算し，スムージングの程度 (スムージングのパラメータの尤もらしい値) を推定する．そのため，Minka のスムージングを行うためには各レコードの情報，すなわち個票が必要である．また，Minka のスムージングで実際に平滑化する際には，プロファイルの属性毎の取りうる値数を全属性について積算した値 × 商品数と同数 (集計表ではセルの数と同じ数) のスムージングのパラメータを推定する必要があり，多くのデータが必要である．そのため，データ量が限られる中小組織では精度向上が難しい．

2.6 まとめ

本章では，Inter PPR を実現する観点から，推薦，暗号応用，匿名加工，スムージングについて先行研究を概観・分析し，下記の点を明らかにした．

1. 共通ユーザ ID の付番とプロファイルの管理が中小企業には大きな負担だが，解決する技術がない．
2. 推薦技術のうち小データに適するのはコンテンツベース推薦である．
3. 暗号応用の処理コストは膨大だが，特定タスク向けの効率的な手法もある．
4. 個票を対象とする匿名加工は，小データで推薦精度を維持する本システムには不適である．統計量を対象にする手法が良い．
5. スムージングが分散環境で利用できない．

第 3 章

Inter PPR の設計

3.1 はじめに

前章では，推薦，暗号応用，匿名加工，スムージングについて先行研究を概観し，Inter PPR に利用可能な要素技術を明らかにするとともに，従来技術の限界を明らかにした．本章では，先行研究を踏まえて Inter PPR の構成を設計する．その中で，従来技術のうち利用可能な要素技術を選定するとともに，新たに創出するべき要素技術を明らかにする．

3.2 推薦および ID 管理

推薦手法について，中小企業は保有するデータが少ないため，推薦精度が低いという課題がある．そこで，データ量への依存が小さいコンテンツベース推薦を選択する．

ユーザ ID とプロファイルの管理については，複数組織のデータを結合するために必要な共通ユーザ ID，および，コンテンツベース推薦を高精度化するためのプロファイルの管理コストが中小組織には大きな負担である．

しかし，共通ユーザ ID およびプロファイルの管理は，Inter PPR の実現において必須である．以下に，この点を説明する．複数間で推薦に必要な情報を結合する方法は大きく 2 つある．1 つ目は，図 3.1 に示すような，垂直統合と呼ばれる方法である．共通のユーザ ID を利用して，ユーザのプロファイルと，購入された商品の履歴データを結合する．たとえば，中小組織 1 において ID1 と ID2 のユーザが共に男性 30 代であり，中小組織 2 において ID1 のユーザが商品 1 を 20 個購入し，ID2 のユーザが商品 1 を 50 個購入したというデータを統合して，男性 30 代は商品 1 を 70 個購入したというデータを得ることができる．垂直統合では，共通 ID とプロファイルの管理が必須であることは明らかである．

2 つ目は，図 3.2 に示すような，水平統合と呼ばれる方法である．たとえば，中小組織 1

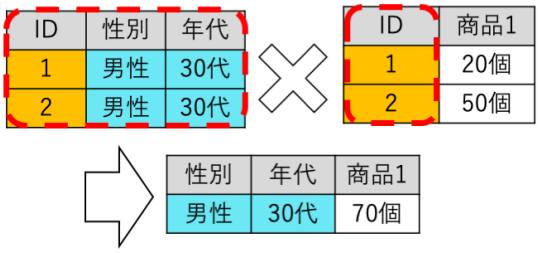


図 3.1　推薦に必要な情報の複数組織間での垂直統合

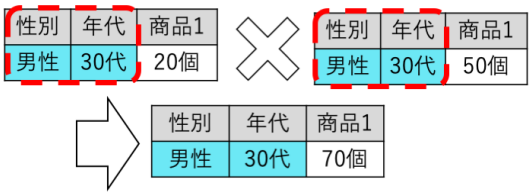


図 3.2　推薦に必要な情報の複数組織間での水平統合

において男性 30 代が商品 1 を 20 個購入し，中小組織 2 において男性 30 代が商品 1 を 50 個購入したというデータを統合して，男性 30 代は商品 1 を 70 個購入したというデータを得ることができる．水平統合では共通のユーザ ID の管理は一見不要であるが，両組織でユーザのプロファイルを保有することが避けられない．また，それぞれの組織において，ユーザのプロファイルとユーザを対応づけるために別途ユーザ ID を発行して，ユーザの登録および更新も行わなくてはならない．以上を踏まえると，組織間連携において共通ユーザ ID とプロファイルは不可避である．

Inter PPR では，図 3.3 に示すように，ユーザ ID とプロファイルの組 (以後，個人情報と呼ぶ) を日常的に管理している ID 管理組織を導入する．ID 管理組織の候補としては，カード会社，携帯電話会社，電子マネー会社などが挙げられる．これらの組織は，既に多数のユーザの ID およびプロファイルを管理しているので，ID およびプロファイルのための新たな負担は必要ない．商店は，ID 管理組織と共通のユーザ ID を用いて，購入された商品の履歴データを作成することにする．ここでの商店は，物理的な商店でも，インターネット上にあるようなバーチャルな商店でも構わない．訪問者は，携帯しているスマートフォンや自宅 PC などに自身のプロファイルを格納しているものとする．

個人情報を保有する ID 管理組織と，履歴データを保有する商店が ID 管理組織-商店間プロトコルでやりとりを行い，商店はプロファイルと商品に関する合成データを手に入れ

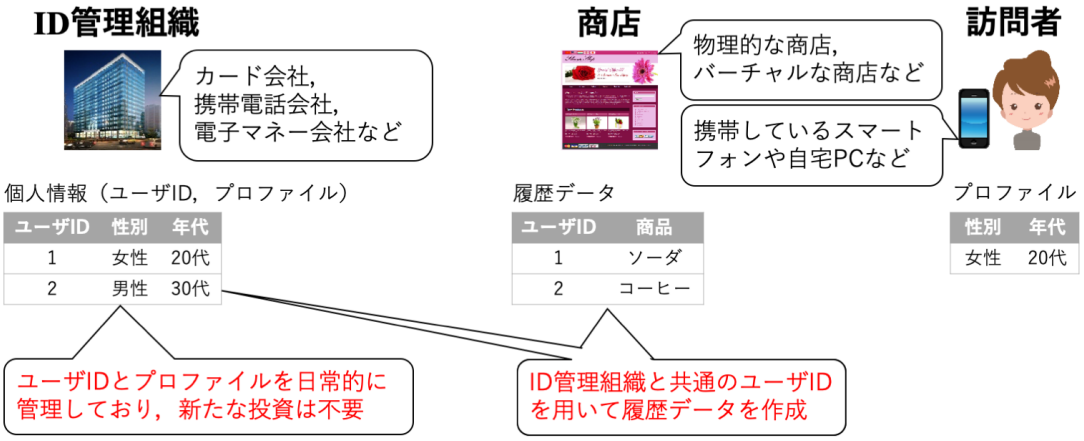


図 3.3　Inter PPR での主体

る．合成データを保有する商店とプロファイルを保有する訪問者が商店-訪問者間プロトコルでやりとりを行い，訪問者は訪問者のプロファイルに応じた商品の推薦を受ける．訪問者に対する推薦手法は，この節の冒頭で述べたように，少ないデータでも推薦精度を維持しやすいコンテンツベース推薦を用いることにしたので，図 3.4 のようにコンテンツベース推薦を商店-訪問者間プロトコルに組み込めるようにする．

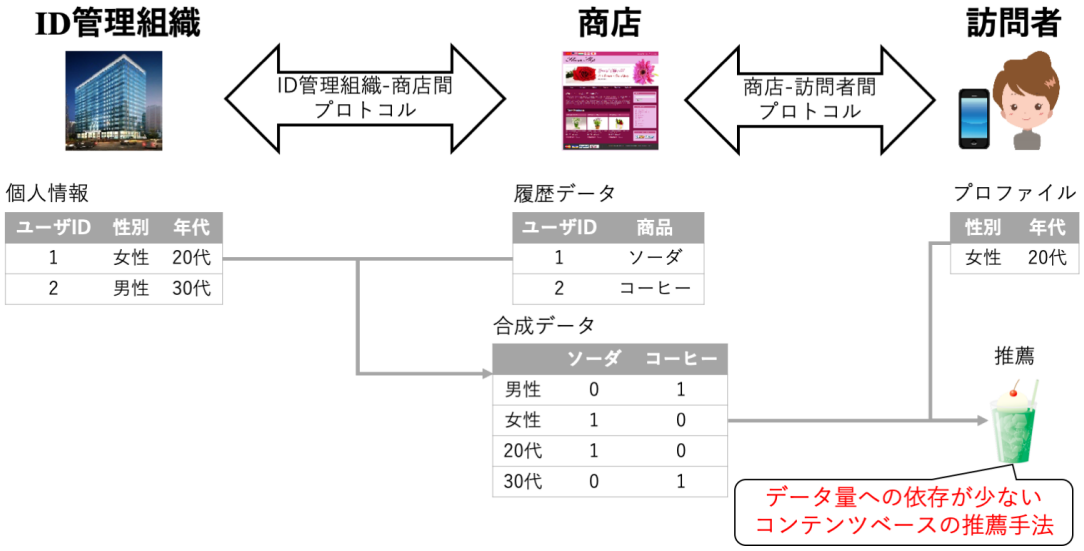


図 3.4　Inter PPR での主体間のプロトコルとコンテンツベースの適用

3.3 プライバシ保護

3.3.1 課題

中小組織が統計的推薦を行う場合，保有するデータ数が少ないことから精度が低くなるため複数の中小組織が連携し，データを統合して推薦精度を高めたい．しかし，各々の組織の保有するデータは組織の財産であるため，他組織への開示は経営上のリスクを伴う．また，個人情報保護の観点からも，他組織へのデータの開示は困難である．そこで，暗号処理と匿名加工によって，互いのデータを秘匿しながら統合利用する必要があるが，暗号応用は処理時間が膨大になりがちで，匿名加工はデータの劣化と推薦精度の低下を招いてしまう．Inter PPR におけるプライバシ保護の課題は，ID 管理組織，商店，訪問者が保有している情報 (個人情報，履歴データ，訪問者のプロファイル) を，各々他の 2 者に対して秘匿できることである．

Inter PPR におけるプライバシ保護の課題を，図 3.5 に示すように，プロトコル自体を安全にする暗号応用と，プロトコルの出力を安全にする暗号応用で解決する．プロトコル自体の安全性は，プロトコルの利用主体が，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入手できないようにすることで確保する．プロトコルの出力の安全性は，プロトコルの利用主体が，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できないことで確保する．

3.3.2 組織間の暗号プロトコル

組織間の暗号プロトコルにおいて，暗号応用は，処理時間が膨大になりがちである．一方で，カード会社，携帯電話会社，電子マネー会社などを想定した ID 管理組織は数千万人に及ぶ会員の個人情報を保有しており，中小規模の商店であっても多くの履歴データが発生すると考えられ，これらのような大規模なデータから妥当な期間内に合成データを算出しなくてはならない．また，訪問者が商店に来店したら速やかに推薦を提供しなくてはならない．そこで，組織間の暗号プロトコルでは，暗号応用のうち，2.4.2 節で述べた効率的な特定タスク向け技術を利用することにする．具体的には，ID 管理組織-商店間プロトコルについては，組織間のデータベース結合は集合の積演算に帰着するため，効率的な秘匿積集合プロトコルを利用する．商店-訪問者間プロトコルについては，ナイーブベイズでの推薦の確率を計算できる，効率的な秘匿内積プロトコルを利用する．

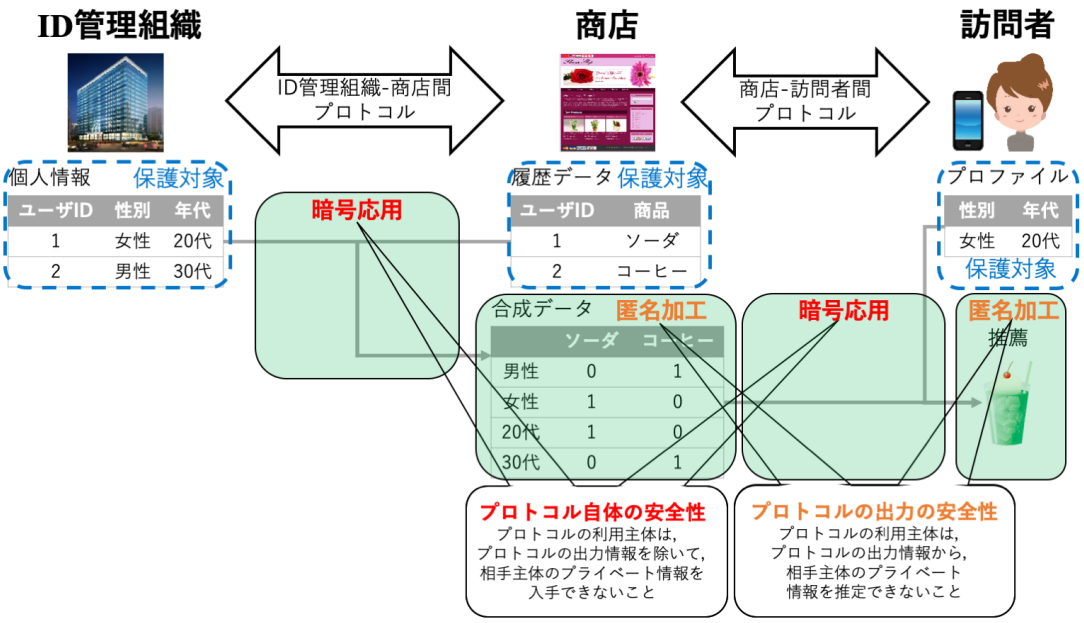


図 3.5　プライバシ保護の要件

3.3.3　匿名加工

匿名加工は，プロトコルの出力からの情報漏洩を防ぐために必要である．Inter PPR に匿名加工を組み込まなければ，次の 2 つのようなプライバシ漏洩の可能性が生じる (図 3.6).

(攻撃例 1) ソーダを買ったのは 20 代の女性である

ID 管理組織-商店間プロトコルが商店に出力した合成データから，ID 管理組織が保有する元データ 1 が商店に漏れる．

(攻撃例 2) 自分以外の 20 代の女性がソーダを買っている

商店-訪問者間プロトコルが訪問者に出力した推薦結果から，商店の合成データが訪問者に漏れる．さらに，合成データから，ID 管理組織が保有する元データ 1 と商店が保有する元データ 2 も訪問者に漏れる．

Inter PPR でのプライバシ保護において，プロトコルの出力の安全性を守る匿名加工は欠かせないが，データの劣化と推薦精度の低下をまねくため，データ劣化の大きい個票向け匿名加工ではなく統計量向け匿名加工を用いる．これに伴い，図 3.6 の合成データは個票ではなく統計量とし，具体的には，個人の属性毎に商品の購入数を集計したクロス集計

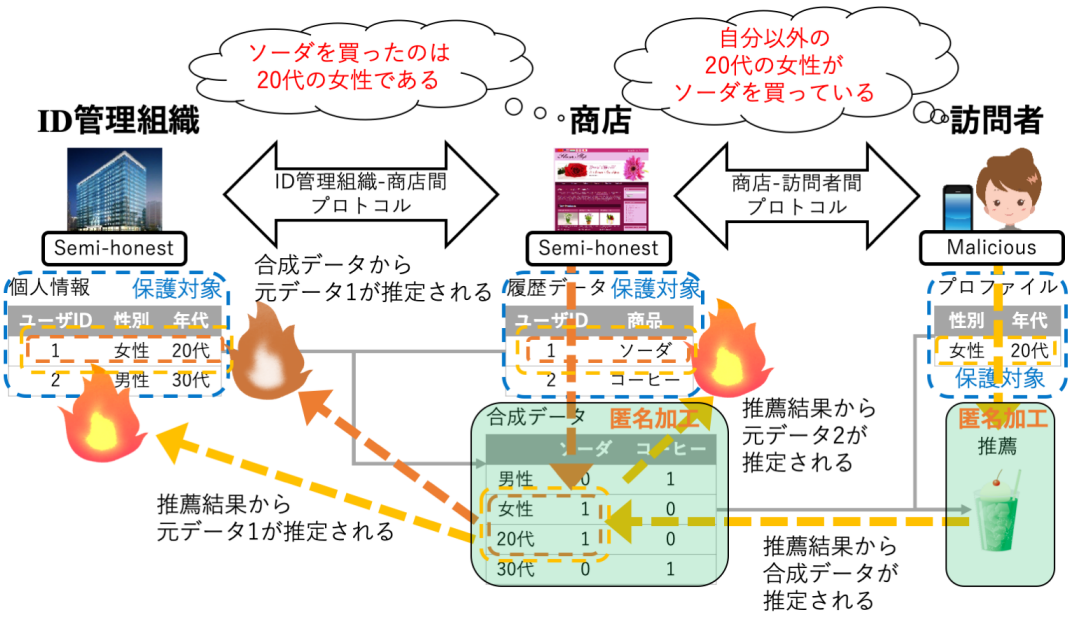


図 3.6　プライバシ漏洩の具体例

表とする．また，2.4.3 節で述べたように，統計量向けの匿名加工方式の中でも差分プライバシは安全性と有用性を数学的に定式化できるので，これを利用する．ID 管理組織-商店間プロトコルで，ID 管理組織が保有する個人情報と商店が保有する履歴データを合成して，商店にクロス集計表 (統計量) を出力する．しかし，クロス集計表のような多属性統計量の場合は差分プライバシによる劣化が大きいという問題があるため，多属性データの劣化を抑止する差分プライバシを新規提案する (図 3.7)．この差分プライバシについては 6 章で詳しく述べる．

商店がクロス集計表を差分プライバシで匿名加工すると，上記の攻撃例 2 は防げるようになる．しかし，攻撃例 1 は防ぐことができないため，Inter PPR への匿名加工の組み込み方には工夫が必要である．商店がクロス集計表を差分プライバシで匿名加工すると，商店-訪問者間プロトコルへの入力が保護されるため，訪問者による (攻撃例 2 を含む) いかなる攻撃も防ぐことができる．しかし，ID 管理組織-商店間プロトコルから出力される匿名加工前のクロス集計表は保護されないため，商店による (攻撃例 1 などの) 攻撃を防ぐことができない (図 3.8)．

そこで，ID 管理組織-商店間プロトコルを ID 管理組織へ一旦出力し，ID 管理組織がクロス集計表を差分プライバシで匿名加工してから商店へ出力するように Inter PPR を設計することで，商店および訪問者による (攻撃例 1 と攻撃例 2 を含む) いかなる攻撃も防げ

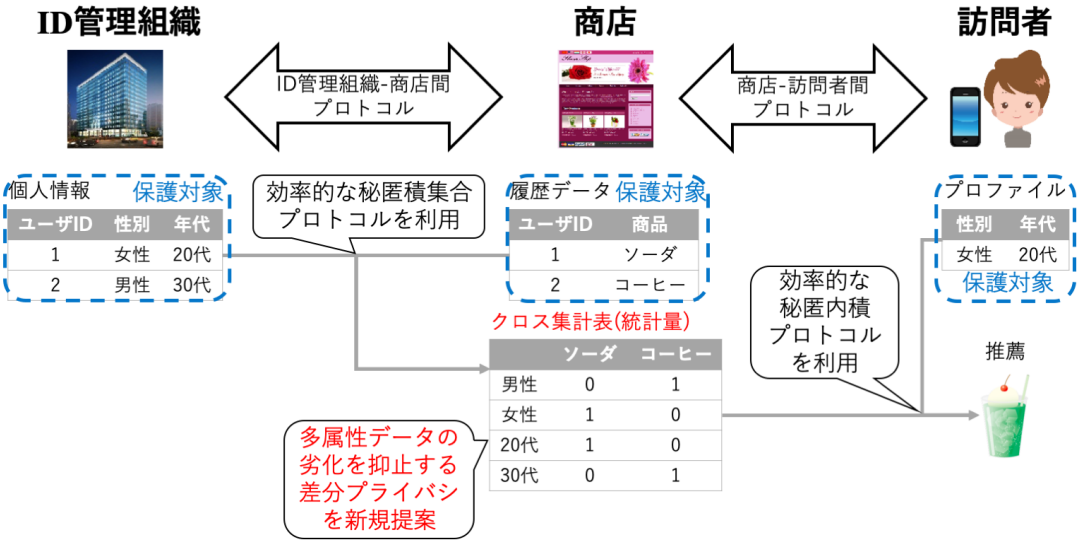


図 3.7　多属性対応差分プライバシの新規提案

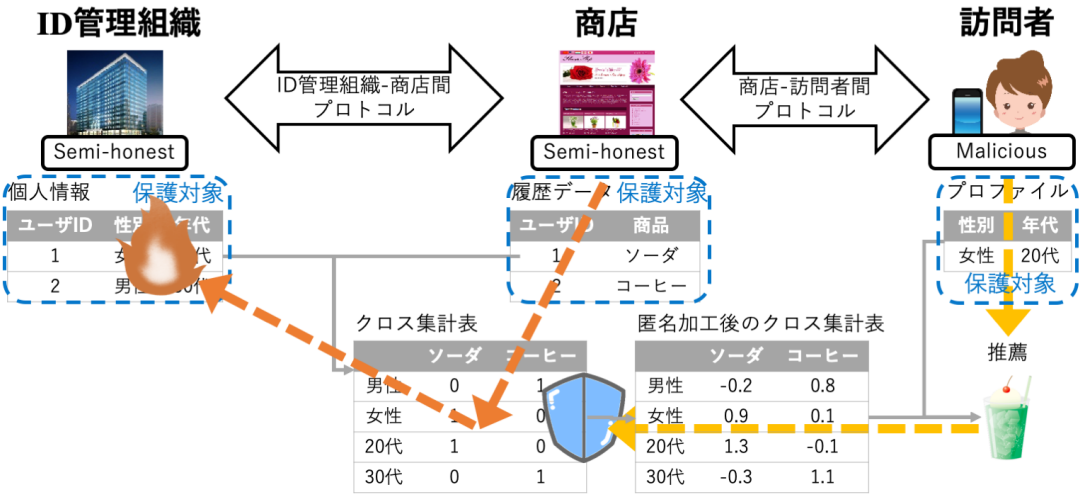


図 3.8　商店から ID 管理組織の個人情報への攻撃

るようにする．ID 管理組織は，カード会社，携帯電話会社，電子マネー会社などの事業者であり，商店との契約，従業員教育，計算機の操作に関する監視・監査により，商店に対して malicious な攻撃を行わないようにできると考えられる (図 3.9).

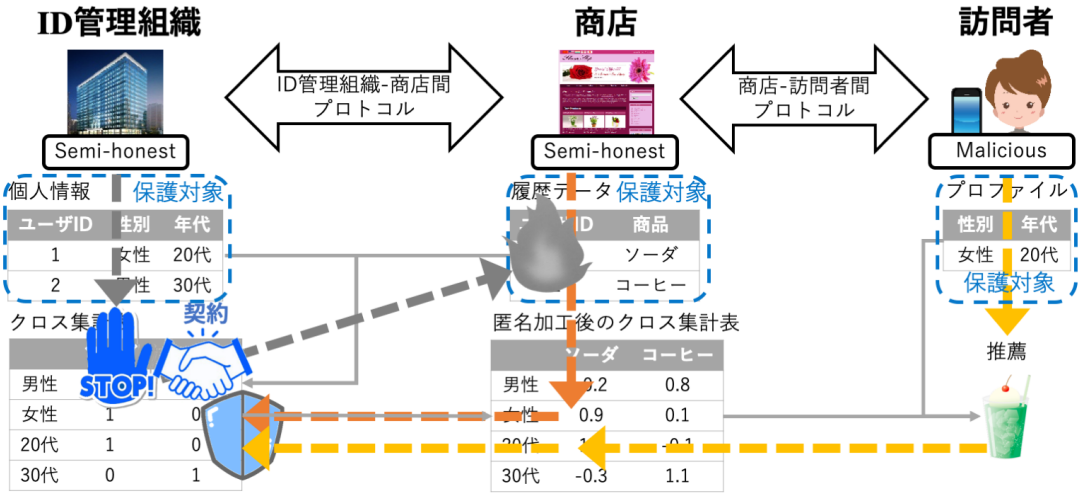


図 3.9　ID 管理組織から商店の履歴データへの攻撃の防止

3.4　スムージング

推薦の精度向上にスムージングが用いられるが，従来のスムージングは個票を必要とするため，分散環境で実行すると相手の組織にプライバシが漏洩する．たとえば，個票のレコードに性別と年代の属性がある状況において組織間でスムージングすると，相手の組織に性別と年代の組み合わせ，すなわち個人情報に含まれるユーザのプロファイルが漏洩してしまう．

一方，クロス集計表であれば，性別と年代毎に該当するレコードを集計する統計化によって，プロファイルを構成する属性間の組み合わせの情報が失われており，性別と年代のそれぞれに独立して該当するレコードの数の情報しか残っていない．そのため，クロス集計表をどのように処理しても，個人情報に含まれるユーザのプロファイルは漏洩しない．そこで，クロス集計表に対して適用可能なスムージングの手法を新規提案し，ID 管理組織と商店に情報が分散した環境において，プライバシを保護したままでスムージングによる推薦精度の向上を可能にする．この分散環境対応スムージングについては 5 章で詳しく述べる．

ところで，Minka のスムージングでは，プロファイルの属性毎の取りうる値数を全属性について積算した値 × 商品数と同数のスムージングのパラメータを推定する必要があり，多くのデータが必要になるためにデータ量が限られる中小組織では精度向上が難しい．しかし，提案する分散環境対応スムージングは，プロファイルを構成する属性間の組み合わ

せを用いないため，商品数と同数のスムージングのパラメータを推定すれば良く，データ量が限られる中小組織でも精度向上が期待できる．また，分散環境対応スムージングはクロス集計表であれば適用可能であるが，匿名加工のためにクロス集計表に付加されたノイズを特異値とみなして平滑化することで，推薦精度を向上できる可能性がある．そのため，Inter PPR においては，ID 管理組織-商店間プロトコルで出力される差分プライバシのノイズを重畳した匿名加工後のクロス集計表に分散環境対応スムージングを適用する設計とする．

組織間に適用可能な分散環境対応スムージングを新規提案し，適切に Inter PPR へ組み込むことにより，スムージングによる精度向上と，匿名加工による推薦精度の低下の抑止の効果を発揮させて，安全性と有用性の両立を図る．

3.5 多組織間への拡張性

中小企業は保有するデータが少ないため，推薦精度が低いという問題がある．中小組織のデータは大組織に比べて桁違いに少ないため，多組織間の連携により，データ量をスケールアップする必要がある．そこで，ID 管理組織が多数の商店との間で，各々，クロス集計表を生成し，これらを統合し，匿名加工して，各商店に配布する．

3.6 Inter PPR のシステム構成と創出すべき要素技術

以上のように設計した Inter PPR のシステム構成を図 3.10 に示す．ID 管理組織は，ユーザ ID とプロファイルを日常的に管理しており，新たな投資は不要である．商店は，ID 管理組織と共通のユーザ ID を用いて履歴データを作成する．ID 管理組織と商店は，効率的な秘匿積集合プロトコルを利用して，妥当な期間内に個人情報と履歴データを統計化したクロス集計表を生成する．プロトコルの出力からの情報漏洩を防ぐため，クロス集計表を ID 管理組織に一旦出力し，差分プライバシで匿名加工してから商店へ出力する．また，推薦精度の向上と匿名加工による精度低下の抑止のためにスムージングを適用してクロス集計表を平滑化する．ID 管理組織は，多数の商店との間で，各々，クロス集計表を生成し，これらを統合し，匿名加工およびスムージングして，各商店に配布することで，推薦精度を向上させる多組織間の連携を可能とする．商店と訪問者は，効率的な秘匿内積プロトコルを利用して，速やかに訪問者へ推薦を行う．少ないデータでも推薦精度を維持できるように，データ量への依存が少ないコンテンツベースの推薦処理を行う．

上記のシステム構成において，多属性統計量であるクロス集計表に従来の差分プライバ

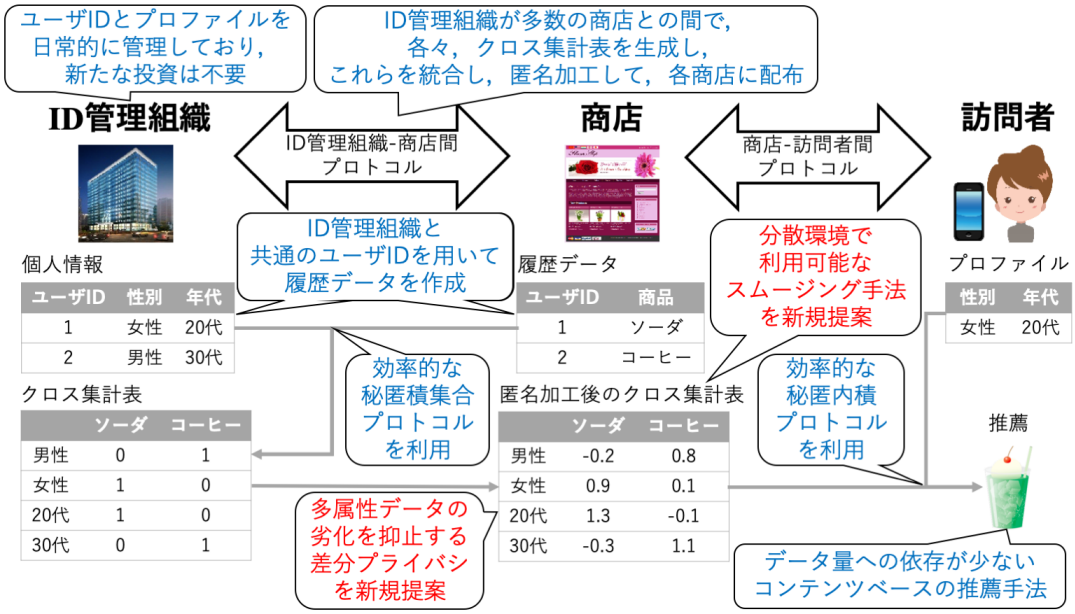


図 3.10　Inter PPR のシステム構成と創出すべき要素技術

シを適用するとデータの劣化が大きいため，多属性対応差分プライバシを新たに提案する．また，従来のスムージング手法は個票を対象とし，個人情報と履歴データの両方を保有する組織しか利用できないので，クロス集計表に適用可能な分散環境対応スムージングを新たに提案する．これらの多属性対応差分プライバシを 6 章，分散環境対応スムージングを 5 章で詳述する．

以上の方針から，目的のシステムは図 3.10 の構成となる．「組織間で情報を分散し協調することで，巨大企業大組織でなくても推薦できるシステム」というコンセプト，このコンセプトに基づく，上記の機能および図 3.10 の構成を持つ技術は，著者らの独自の提案である．

図 3.11 と図 3.12 に示す Inter PPR の社会実装が実現すれば，一部の大組織だけでなく，多くの組織が統計的推薦を実施できるようになると期待される．また，大組織にとっても，地方に特化した中小組織などとの連携が可能となり，統計的推薦の適用範囲を拡大できるようになると考えられる．

図 3.11　Inter PPR の利用イメージ

図 3.12　Inter PPR が実現する社会

3.7　まとめ

本章では，従来技術の分析を踏まえて，Inter PPR のシステム構成を設計すると共に，新たに開発すべき要素技術を明らかにした．

- データが少なくても推薦精度を維持するためにコンテンツベースの推薦手法を採用すると共に，スムージングを適用する．そのために，分散環境で利用可能なスムージング手法を新たに提案する必要がある．
- ユーザ ID とプロファイルの管理コストを低減させるために，ユーザ ID とプロファイルを日頃から管理している組織を ID 管理組織として活用する．

- プロトコル自体の安全性を確保するために，暗号応用のうち，効率的な秘匿積集合と秘匿内積プロトコルを利用する．
- プロトコル自体の安全性を確保するために，クロス集計表を匿名加工して商店に出力する．多属性のクロス集計表に差分プライバシを適用するとノイズが大きいため，多属性データの劣化を抑止する匿名加工を新たに提案する必要がある．
- 複数商店の履歴を統合するために ID 管理組織が多数の商店との間で，各々，クロス集計表を生成し，これを統合し，匿名加工して，各商店に配布する．

第 4 章

Inter PPR の実現

4.1 はじめに

前章では，Inter PPR のシステム構成を設計した．本章では，前章の設計に基づいて，Inter PPR の実現方法を明らかにする．まず，Inter PPR のユースケースを明らかにし，ユースケースに沿って Inter PPR の技術要件を定義する．この技術要件に沿って，データ表現と処理フロー，組織間の暗号プロトコルを設計する．

4.2 ユースケース

ユーザの個人情報を管理している ID 管理組織と，小売業を営んでいる商店と，商店を訪れた訪問者の 3 者での運用を想定する．ID 管理組織は，個人情報としてユーザの ID とプロファイルを管理している．商店は，ID 管理組織と提携しており，履歴データとして商品を販売したユーザの ID と商品情報を管理している．訪問者は，自身のプロファイルを (携帯電話やスマートフォンなどの) 計算能力を持つ携帯端末や PC，クラウドサービスの仮想端末やアカウント，もしくはブラウザなどに保持している．ただし，訪問者は ID 管理組織のユーザとは限らず，ID を持たずに商店を訪問することもある．

以上の状況において，商店は ID 管理組織の支援を受けて，商品の購入傾向を定期的に手に入れたい．また，商店は，訪問者の端末上で，購入に繋がりやすい商品を推薦したい．プライバシを考慮しなければ，これは以下のようにして達成できる．

1. ID 管理組織と商店は，ID 管理組織の個人情報と商店の履歴データのそれぞれのデータベースをユーザ ID を介して結合する．
2. 商店は，結合したデータベースから，プロファイルと商品の購入傾向の統計的な関

係を推定する．

3. 商店は，上記で推定したプロファイルと商品の購入傾向の統計的関係を用い，訪問者のプロファイルに応じて購入される可能性が高い商品を訪問者に推薦する．
4. 訪問者が ID 管理組織のユーザであった場合には，購入時にユーザ ID を提示し，商店はユーザ ID と商品情報を履歴データに追加する．

4.3 要件

4.2 節で述べたユースケースを踏まえて，Inter PPR に対する要件を述べる．まず，各組織に必要なプライバシ保護の要件について述べ，次に，プライバシ保護によって低下が懸念される，処理性能と推薦精度の要件について述べる．

4.3.1 プライバシ保護の要件

ID 管理組織と商店と訪問者の 3 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．ID 管理組織はユーザから預かっている個人情報を漏洩したくない．商店は商店が記録した履歴データを漏洩したくない．訪問者は自身のプロファイルに沿った適切な推薦を受けたいが，プロファイルなどのプライバシ情報は漏洩したくない．

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 ID 管理組織，商店，訪問者) と，外部からの攻撃者が想定される．本論文の主旨は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．

攻撃者のモデルとしては, 他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．ID 管理組織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 な攻撃を行わないようにできると考えられる. 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できるので，semi-honest を想定する．商店も事業者であり，ID 管理組織との提携における契約および ID 管理組織と同等の対策によって，同等の安全性を担保できるため，semi-honest を想定する．訪問者は，任意の個人が含まれるため，実装プログラムの認証および改ざん防止といった対策を講じたとしても，malicious を想定する必要がある．

以上のことから，プライバシ保護の要件は下記のように具体化される．

(要件 1) プライバシ保護の要件

ID 管理組織と商店が semi-honest で，訪問者が malicious の想定において，ID 管理組織が保有する個人情報と，商店が保有する履歴データと，訪問者のプロファイルを，各々他の 2 者に対して秘匿できること．

“(要件 1) 各々他の 2 者に対して秘匿する” をプロトコルの内と外で分割して，プライバシ保護の要件を詳細化する．

(要件 1a) プロトコル自体の安全性

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入手できないこと．

(要件 1b) プロトコルの出力の安全性

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できないこと．

4.3.2 推薦精度の要件

個人情報と履歴データと訪問者のプロファイルの全てを保有している大組織と同等あるいはそれに近い推薦精度であることが望ましい．大組織のみで処理が完結する場合は，スムージングなどの一般的な推薦精度向上の方法を適用できるため，組織間においてもスムージングによって推薦精度を向上させたい．また，大組織のみで処理が完結する場合は，プライバシ保護による精度低下は生じないため，組織間で新たに必要とされるプライバシ保護によって推薦精度を低下させたくない．

以上のことから，組織間での推薦における精度の要件は下記のように具体化される．

(要件 2a) スムージングに対する推薦精度の要件

個人情報と履歴データを直接扱えなくても，スムージングにより推薦精度を向上できること

(要件 2b) プライバシ保護に対する推薦精度の要件

匿名加工による推薦精度の低下を抑止できること

4.3.3 処理性能の要件

ID 管理組織と商店と訪問者の 3 者は，それぞれの立場から処理性能に対して次の要求を持つ．商店は，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，魅力のある推薦を保ちたい．訪問者は商店を訪れた際に直ちに推薦を受けたい．

以下では，ID 管理組織や商店が取り扱うデータの規模や内容を，日本の電子マネーで最大の決済件数を誇る nanaco[93] に倣って想定する．

まず，ID 管理組織の個人情報について．2017 年 2 月末時点のユーザ数は 5,350 万人である [94]．これより，個人情報のユーザ数 (以後，N と表す) は $10^7 \sim 10^8$ 人を想定する．個人情報のプロファイルについては，年代と性別と住所の項目を想定し，プロファイルの項目数 (以後，W と表す) は 3 項目を想定する．また，各プロファイルの項目が取りうる値は，年代は 8 区分 (10 代から 10 才刻み．80 才以上は区別しない)，性別は 2 区分，住所は 47 区分 (都道府県) とし，プロファイルの値の種類数 (以後，V と表す) は 57 種類を想定する．

次に，商店の履歴データについて．全国 23 万箇所で 1 ヶ月に 1.68 億件の電子マネー決済が行われていることから [94]，1 ヶ月 1 箇所あたりの決済件数は 730 件である．この値は商店の規模，所在地，業種，訪問者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，履歴データに含まれる人数 (以後，M と表す) は $10^2 \sim 10^5$ 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高々 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 [95]．商店の規模，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 と表す) は $10 \sim 10^4$ 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 [96]，一度の決済で購入される商品は数種類であると考えられる．履歴データを月単位に締めるとすると，訪問者が毎週商店を訪れたとしても 1 ヶ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，履歴データに含まれる 1 人あたりの商品の種類数 (以後，G と表す) は 10 種類を想定する．ID 管理組織や商店が実際に取り扱うデータの規模と内容を整理すると表 4.1 となる．

ID 管理組織-商店間プロトコルの処理に許容される時間は，このプロトコルの実行間隔 (以後，T_1 と表す) から決定する必要がある．T_1 が長すぎると推薦内容が流行から遅れてしまう．一方，短すぎると，T_1 の間に履歴データがほとんど変化しないので，ID 管理組織-商店間プロトコルを再実行しても推薦結果は変わらない．また，T_1 が短いとプロトコルを短時間に実行する必要があり，高速の計算設備が必要となる．そのため，T_1 が短すぎ

表 4.1　データの規模と内容

データの種類	データの規模のパラメータ	値の範囲
ID 管理組織の個人情報	個人情報に含まれるユーザ数 (N)	$10^7 \sim 10^8$ 人
	プロファイルの項目数 (W)	3 項目
	プロファイルの値の種類数 (V)	57 種類
商店の履歴データ (1 ヶ月分)	履歴データに含まれるユーザ数 (M)	$10^2 \sim 10^5$ 人
	商品の種類数 (L)	$10 \sim 10^4$ 種類
	1 人あたりの商品の種類数 (G)	10 種類

るとコストパフォーマンスが低くなる．最適な実行間隔 T_1 は業界等によって異なる．たとえば，洋服の推薦であれば季節ごとに商品を入れ替えるために 3 ヶ月であったり，ゲームのように日々新たなタイトルが登場するのであれば 1 日であったりする．そのため，T_1 の具体的な値は文献等でも殆ど言及されていないが，[97] では 1 ヶ月となっている．Inter PPR ではプライバシを保護するため，クロス集計表に匿名加工のノイズを加える．そのため，T_1 が短すぎると，ノイズによるクロス集計表の変化が，ユーザの嗜好の変化による履歴データの真の変化を上回ってしまい，ユーザの嗜好の変化が推薦結果に反映されなくなってしまう．以上から，今回は一例として T_1 を 1 ヶ月に設定することとし，T_1 の増減による影響については 7.4 節で考察を行うこととする．

以上から，表 4.1 に示したデータの規模や内容で，ID 管理組織と連携した購入傾向の生成を 1 ヶ月以内に実行可能となるようにシステムを設計する必要がある．また，商店-訪問者間プロトコルにより，訪問者が商店を訪れたときの商品の推薦はリアルタイム (あるいは準リアルタイム) である必要がある．その許容時間 (T_2 とする) は小さいため，たとえば 5 秒と想定すると，表 4.1 に示したデータの規模や内容で，訪問者への推薦を 5 秒以内に実行可能となるようにシステムを設計する必要がある．

以上のことから，処理性能の要件は下記のように具体化される．

(要件 3a) ID 管理組織-商店間の処理性能の要件

ID 管理組織-商店間プロトコルは，表 4.1 のデータの規模と内容において，許容時間 T_1 の制約を満たすこと．

(要件 3b) 商店-訪問者間の処理性能の要件

商店-訪問者間プロトコルは，表 4.1 のデータの規模と内容において，許容時間 T_2 の制約を満たすこと．

4.3.4　社会実装容易性の要件

中小組織の間で推薦に必要な情報を統合利用したいが，2.3 節で述べたように，ユーザ ID とプロファイルの管理が避けられず，中小組織には負担が重い．また，中小組織は保有するデータ数が少ないことから，多組織間の連携に拡張可能なシステムであることが望ましい．

以上のことから，社会実装容易性の要件は下記のように具体化される．

(要件 4a) ユーザ ID およびプロファイルの安全な利用と負担抑制の要件

ユーザ ID およびプロファイルを安全に管理し利用するにあたって，新たな負担が少ないこと．

(要件 4b) 多組織間連携への拡張性の要件

ID 管理組織と商店の 2 組織間の連携だけでなく，ID 管理組織と複数の商店の間の多組織間連携に拡張可能であること．

4.4　実装方法の概要

Inter PPR の構成を図 4.1 に示す．ID 管理組織，商店，訪問者の 3 者モデルにおいて，ID 管理組織はユーザの ID とプロファイルを紐づけて，ID 管理組織のサーバ上で管理している．商店は，ユーザの ID とユーザに販売した商品の情報 (以後，商品情報と呼ぶ) を紐づけて，商店のサーバ上で管理している．商店の用いるユーザ ID は ID 管理組織と共通のものである．訪問者は，自身のプロファイルを携帯端末に保持している．ID 管理組織，商店，訪問者が保有している情報 (個人情報，履歴データ，訪問者のプロファイル) を，各々他の 2 者に対して秘匿できるようにする．

ID 管理組織と商店は，ユーザ ID に基づいて互いが保有する情報 (ID 管理組織が保有する個人情報と商店が保有する履歴データ) を結合してクロス集計表を生成し，商店へ開示する．このクロス集計表は，ID 管理組織が保有するユーザのプロファイルを用いて，商店が保有する商品情報を集計したものであり，各商品のプロファイルの値ごとの購入傾向を示す．クロス集計表を生成する処理において，ID 管理組織は商店に対して個々のユーザのプロファイルを秘匿し，商店は ID 管理組織に対して個々のユーザの商品情報を秘匿する．互いの情報を秘匿しながらデータを結合する方法には，秘密回路，マルチパーティプロトコル，万能計算機などがあるが，2.4.2 節で述べたようにこれらは遅いので，ID 管理

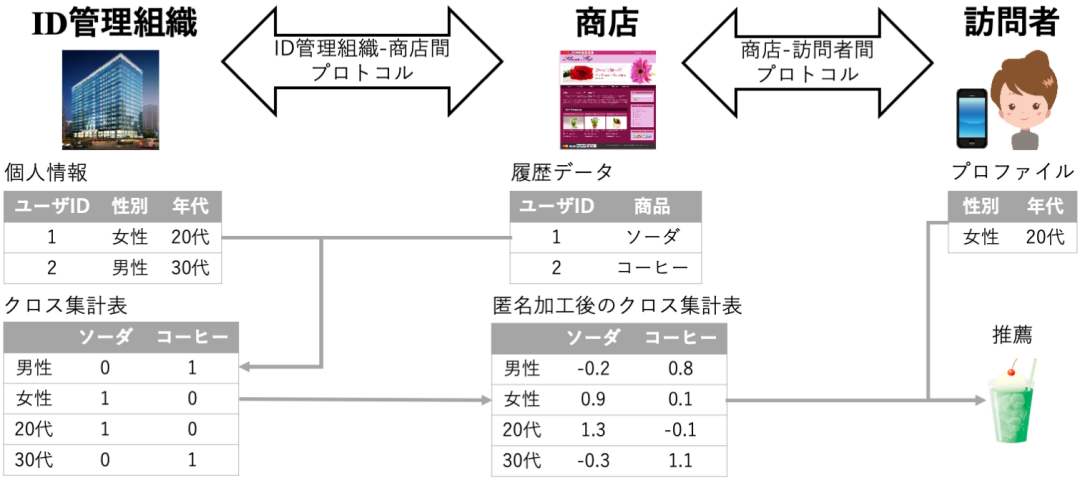


図 4.1　Inter PPR の構成

組織と商店は秘匿積集合プロトコルを利用して，互いがプライベートな情報を相手に対して秘匿しながらクロス集計表を生成する．この詳細については 4.6.1 節で詳しく述べる．クロス集計表に対して，多属性対応差分プライバシを適用して匿名加工を行い，さらに分散環境対応スムージングを適用する．訪問者に対して，クロス集計表を利用したコンテンツベースの推薦を行う．プロファイル毎のクロス集計表を用いることで，コンテンツベースでの商品の推薦に必要となる商品の特徴を得る．訪問者への推薦の処理において，訪問者は商店に対して訪問者のプロファイルを秘匿し，商店は訪問者に対して商店のノウハウである購入傾向を秘匿する．中小組織である商店が手に入れられる購入傾向は疎になりやすいため，コンテンツベースの推薦手法であるナイーブベイズを用いる．ナイーブベイズを秘匿して実行するために秘匿内積プロトコルを利用して，互いがプライベートな情報を相手に対して秘匿しながら推薦を行う．この詳細については，4.6.2 節で詳しく述べる．

4.5　データ表現と処理フロー

4.5.1　データの表現

ID 管理組織は個人情報として，ユーザの ID とプロファイルを管理している (表 4.2)．商店は，ID 管理組織と提携しており，履歴データとして商品を販売したユーザの ID と商品情報を管理している (表 4.3)．訪問者は，自身のプロファイルを携帯端末に保持している (表 4.4)．　表 4.2，表 4.3，表 4.4 のプロファイルと商品情報は多値であるが，これらを

表 4.2　ID 管理組織の個人情報

ユーザ ID	プロファイル	
	性別	年代
1	男性	20 代
2	女性	30 代
4	男性	30 代
5	女性	40 代
6	男性	20 代
7	男性	40 代
8	女性	40 代

表 4.3　商店の履歴データ

ユーザ ID	商品
1	本 A
2	本 A
3	本 A，本 B
4	本 B
6	本 A
7	本 B

表 4.4　訪問者のプロファイル

プロファイル	
性別	年代
男性	30 代

表 4.5，表 4.6，表 4.7 のように二値で表現する．

ID 管理組織が保有している表 4.2 の個人情報を表 4.5 のように表し，これを図 4.2 のように，ユーザ ID を表すベクトル $\boldsymbol{t}$ とプロファイルを表すマトリクス X で表す．$\boldsymbol{t}$ は図 4.2(a) のように，長さ N のベクトルである．ベクトル $\boldsymbol{t}$ の長さは ID 管理組織の個人情報に含まれるユーザ数に等しく，図 4.2(a) の場合は $N = 7$ である．$\boldsymbol{t}$ の n 番目の要素を t_n

表 4.5　ID 管理組織の個人情報の二値表現

$\boldsymbol{t}$	X				
ユーザ ID	性別:男性	性別:女性	年代:20 代	年代:30 代	年代:40 代
1	1	0	1	0	0
2	0	1	0	1	0
4	1	0	0	1	0
5	0	1	0	0	1
6	1	0	1	0	0
7	1	0	0	0	1
8	0	1	0	0	1

表 4.6　商店の履歴データの二値表現

$\boldsymbol{u}$	Y	
ユーザ ID	商品:本 A	商品:本 B
1	1	0
2	1	0
3	1	1
4	0	1
6	1	0
7	0	1

と表す．t_n は n 番目のユーザの ID を表し，図 4.2(a) の場合は $t_1 = 1, t_2 = 2, ..., t_N = 8$ である．X は図 4.2(b) のように N 行 V 列のマトリクスである．マトリクス X の列数は，プロファイルの値の種類数に等しく，図 4.2(b) の場合は $V = 5$ である．v はプロファイルの値の識別子であり，$v = 1$ は男性，$v = 3$ は 20 代を表す．プロファイルの項目数を W とする．表 4.5 の場合，プロファイルの項目数は 2 種類 (性別と年代) であるため，$W = 2$ である．プロファイルの項目の識別子を w とする．w は 1 または 2 を取りうる．マトリクス X の n 行，v 列の要素を $x_{n,v}$ と表す．$x_{n,v} = 1/0$ は，ユーザ t_n が v 番目のプロファイルの値を有する/有しないことを表す．図 4.2(b) の場合，$x_{1,1} = 1, x_{2,1} = 0, ..., x_{N,V} = 1$ である．

商店が保有している表 4.3 の履歴データを表 4.6 のように表し，これをユーザ ID を表

表 4.7　訪問者のプロファイルの二値表現

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

$$(1, 2, 4, 5, 6, 7, 8)$$

(a) ユーザの ID のベクトル表現 $\boldsymbol{t}$

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

(b) ユーザのプロファイルのマトリクス表現 X

図 4.2　ID 管理組織の個人情報のベクトルとマトリクスによる表現

すベクトル $\boldsymbol{u}$ と商品情報を表すマトリクス Y で表す．ベクトル $\boldsymbol{u}$ の長さは履歴データに含まれるユーザ数 M に等しく，表 4.6 の場合は $\boldsymbol{u} = (1, 2, 3, 4, 6, 7)$,$M = 6$ である．$\boldsymbol{u}$ の m 番目の要素を u_m と表す．u_m は m 番目のユーザの ID を表し，表 4.6 の場合は $u_1 = 1, u_2 = 2, ..., u_M = 7$ である．マトリクス Y は，M 行 L 列のマトリクスである．マトリクス Y の列数は，商品の種類数に等しく，表 4.6 の場合は $L = 2$ である．l は商品の識別子であり，$l = 1$ は本 A，$l = 2$ は本 B を表す．マトリクス Y の m 行，l 列の要素を $y_{m,l}$ と表す．$y_{m,l} = 1/0$ は，ユーザ 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 4.6 の場合は，$y_{1,1} = 1, y_{2,1} = 1, ..., y_{M,L} = 1$ である．購入数量の値を直接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，ここでは購入した/しないの二値のみで扱うことにする．なお，購入数量を扱う方法としては，1 個購入，2 個購入，...，10 個以上購入を別の商品として扱い，各々を 0/1 で表すことなどが考えられる．

訪問者が保有している表 4.4 の訪問者のプロファイルを表 4.7 のように表し，これを訪問者のプロファイルを表すベクトル $\acute{\boldsymbol{x}}$ で表す．ベクトル $\acute{\boldsymbol{x}}$ の長さはプロファイルの値の種類数に等しく，表 4.7 の場合は $\acute{\boldsymbol{x}} = (1, 0, 0, 1, 0)$，$V = 5$ である．$\acute{\boldsymbol{x}}$ の v 番目の要素を $\acute{x_v}$ と表す．$\acute{x_v} = 1/0$ は，訪問者が v 番目のプロファイルの値を有する/しないことを表し，表 4.7 の場合は $\acute{x_1} = 1, \acute{x_2} = 0, ..., \acute{x_V} = 0$ である．

4.5.2 処理フロー

ID 管理組織，商店，訪問者の 3 者を連携させるシステムを構成し，Inter PPR を実現するための，3 者間での処理の流れについて述べる．提案システムを，ID 管理組織のサーバ，商店のサーバ，訪問者の端末により構成する．ID 管理組織のサーバには，ID 管理組織の個人情報として，ユーザ ID ベクトル $\boldsymbol{t}$ およびプロファイルマトリクス X が格納されている．商店のサーバには，商店の履歴データとして，ユーザ ID ベクトル $\boldsymbol{u}$ および商品情報マトリクス Y が格納されている．訪問者の端末には，訪問者のプロファイルとして，プロファイルベクトル $\acute{\boldsymbol{x}}$ が格納されている．

商店は，秘匿積集合を応用した効率的なクロス集計 [56] を用いた ID 管理組織-商店間プロトコルにより，訪問者への推薦に必要な情報をクロス集計表 (マトリクス Φ) の形で手に入れる (表 4.8)．Φ は，V 種類のプロファイルの値と L 種類の商品ごとに，ID 管理

表 4.8　ID 管理組織-商店間プロトコルで商店が手に入れるクロス集計表 (Φ)

プロファイル	商品	
	本 A	本 B
男性	2	2
女性	1	0
20 代	2	0
30 代	1	1
40 代	0	1

組織と商店の間でマッチングできたユーザ数を集計したものである．このプロトコルについては，4.6.1 節で詳しく述べる．

商店は，ID 管理組織-商店間プロトコルで手に入れたクロス集計表 Φ (表 4.8) から，訪問者のプロファイルに応じて購入される可能性が高い商品から訪問者に推薦する．まず，それぞれの商品がどの程度購入されやすいかを示す確率 (以後，P(商品) と表す) を求め

る．P(商品) はそれぞれの商品を購入したユーザ数の割合とみなせるので，商店は Φ の商品ごとの値の割合から P(商品 = 本 A) = $\frac{6}{10}$ と P(商品 = 本 B) = $\frac{4}{10}$ を算出する．次に，それぞれの商品がどのようなプロファイルのユーザに購入されているかを示す，プロファイルの条件付き確率 P(プロファイル | 商品) (以後，Θ と表す) を求める．Θ はそれぞれの商品を購入したユーザのプロファイルの値の割合とみなせるので，商店は Φ の商品ごとのプロファイルの値の割合から，Θ の要素である $\theta_{\text{男性}}^{(\text{本 } A)} = P$(性別 = 男性 | 商品 = 本 A) = $\frac{2}{6}$ や $\theta_{30\text{ 代}}^{(\text{本 } A)} = P$(年代 = 30 代 | 商品 = 本 A) = $\frac{1}{6}$ を算出できる．以下同様に，全てのプロファイルの条件付き確率を求めると表 4.9 となる．

表 4.9　プロファイルの条件付き確率 (Θ)

プロファイル	商品	
	本 A	本 B
男性	$\frac{2}{6}$	$\frac{2}{4}$
女性	$\frac{1}{6}$	$\frac{0}{4}$
20 代	$\frac{2}{6}$	$\frac{0}{4}$
30 代	$\frac{1}{6}$	$\frac{1}{4}$
40 代	$\frac{0}{6}$	$\frac{1}{4}$

それぞれのプロファイルのユーザにどのような商品が購入されやすいかを示す，商品の条件付き確率 (P(商品 | プロファイル) と表す) は，ベイズのモデルによって P(商品 | プロファイル) = $\frac{P(\text{プロファイル} \mid \text{商品})P(\text{商品})}{P(\text{プロファイル})}$ と算出できる．しかし，このモデルでは，プロファイルの値の組み合わせ数に応じてクロス集計表が大きくなってしまうため，4.4 節で述べたようにクロス集計表が疎になって推薦の精度が低下してしまう．そこで，プロファイルの独立性を仮定してこの問題を回避するナイーブベイズを用いる．ナイーブベイズのモデルでは，P(商品 | 性別, 年代) $\propto P$(性別 | 商品)P(年代 | 商品)P(商品) を計算することによって，それぞれのプロファイルに購入されやすい商品の大小関係 (以後，推薦度と呼ぶ) を算出できる．このナイーブベイズのモデルに，表 4.9 で示した Θ の値と，上で述べた P(商品) の値を代入して，推薦度を計算するための値 (以後，パラメータと呼び，$\acute{\theta}$ と表す) を算出する．たとえば，男性が本 A を購入するパラメータは $\acute{\theta}_{\text{男性}}^{(\text{本 } A)} = \left(\frac{2}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$ であり，30 代が本 A を購入するパラメータは $\acute{\theta}_{30\text{ 代}}^{(\text{本 } A)} = \left(\frac{1}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$ である．なお，ユーザはプロファイルの項目数 W(今回の例では W=2) だけのプロファイルの値を持つので，P(商品) の値を W 乗根しておき，ナイーブベイズのモデルで W 個のパラメータを掛け合わせ

ると P(商品) の値が復元されるようにする．商店は，訪問者への推薦に備えて，表 4.10 に示す，5 種類のプロファイルと 2 種類の商品に対応した 10 個 ($= 5 \times 2$) のパラメータからなるマトリクス $\acute{\Theta}$ を算出しておく．

表 4.10　推薦度の算出に用いるパラメータ ($\acute{\Theta}$)

プロファイル	商品	
	本 A	本 B
男性	$\left(\frac{2}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$	$\left(\frac{2}{4}\right)\left(\frac{4}{10}\right)^{\frac{1}{2}}$
女性	$\left(\frac{1}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$	$\left(\frac{0}{4}\right)\left(\frac{4}{10}\right)^{\frac{1}{2}}$
20 代	$\left(\frac{2}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$	$\left(\frac{0}{4}\right)\left(\frac{4}{10}\right)^{\frac{1}{2}}$
30 代	$\left(\frac{1}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$	$\left(\frac{1}{4}\right)\left(\frac{4}{10}\right)^{\frac{1}{2}}$
40 代	$\left(\frac{0}{6}\right)\left(\frac{6}{10}\right)^{\frac{1}{2}}$	$\left(\frac{1}{4}\right)\left(\frac{4}{10}\right)^{\frac{1}{2}}$

訪問者は，以上のナイーブベイズの処理を，商店のパラメータ $\acute{\Theta}$ と訪問者のプロファイル $\acute{\boldsymbol{x}}$ を互いに秘匿しながら計算するために，商店と訪問者の間で秘匿内積プロトコルを実行する．その結果，訪問者は各商品の推薦値を入手し，推薦度の高い商品から順に推薦を受ける．訪問者が，たとえ ID 管理組織のユーザでなくても，もしくは初めて訪れた商店でも，推薦を受けられるように，訪問者によるプロトコルへの入力はプロファイルのみとする．訪問者は，推薦度をベクトル $\acute{\boldsymbol{y}}$ の形で手に入れる．$\acute{\boldsymbol{y}}$ は商品の種類数 L(今回の例では L=2) の長さのベクトルであり，ベクトルの要素の値の大小がそれぞれの商品の推薦の度合いを表す．$\acute{\boldsymbol{y}}$ は，表 4.10 に表すパラメータのマトリクス $\acute{\Theta}$ から訪問者のプロファイルに該当するパラメータの値を取り出して，ナイーブベイズのモデルで算出する．たとえば，表 4.7 に示すように，訪問者が男性の 30 代の場合は，本 A の推薦度は $\acute{y}^{(\text{本 } A)} = \acute{\theta}^{(\text{本 } A)}_{\text{男性}} \acute{\theta}^{(\text{本 } A)}_{\text{30 代}} = \left(\frac{2}{6}\right)\left(\frac{1}{6}\right)\left(\frac{6}{10}\right) = 0.03$ であり，本 B の推薦度は $\acute{y}^{(\text{本 } B)} = \acute{\theta}^{(\text{本 } B)}_{\text{男性}} \acute{\theta}^{(\text{本 } B)}_{\text{30 代}} = \left(\frac{2}{4}\right)\left(\frac{1}{4}\right)\left(\frac{4}{10}\right) = 0.05$ であるため，訪問者へ本 B，本 A の順に推薦する．

4.6 組織間の暗号プロトコル

4.6.1 ID 管理組織-商店間プロトコル

ID 管理組織の個人情報と商店の履歴データを互いに秘匿しつつ，商店がクロス集計表を手に入れるまでの処理を例にあげて，ID 管理組織-商店間プロトコルのアルゴリズム [56] を詳しく述べる．また，このプロトコルの処理性能と安全性を述べる．

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．同じハッシュ値を ID 管理組織の乱数と商店の乱数で冪乗余すると，指数部が 2 つの乱数の和になるため，冪乗余の順序に関わらず同じ値が得られる．この原理を用いて，ID 管理組織のユーザ ID のハッシュ乱数乗をさらに商店の乱数乗した値と，商店のユーザ ID のハッシュ乱数乗をさらに ID 管理組織の乱数乗した値の合致する個数を数えて，ID 管理組織と商店で共通するユーザの集計値を算出する．

ID 管理組織-商店間プロトコルのアルゴリズムは以下の通りである．

1. ID 管理組織はハッシュ関数の計算に用いるシードをランダムに生成し，商店へ送る．
2. ID 管理組織はプロファイルの値ごとに乱数 $R_a^{\text{プロファイル値 }ID}(1 \leq \text{プロファイル値 } ID \leq V)$ を 5 個生成し，表 4.5 のマトリクス X の要素が 1 になっている 14 個所について，$\mathbf{H}(\text{ユーザ } ID)^{R_a^{\text{プロファイル値 }ID}}$ を算出する．各々の冪乗余値に，対応するプロファイル値 ID を付与し，これらのプロファイル値 ID 付き冪乗余値をシャッフルした後，商店に送る[*1]．
3. 商店は商品ごとに乱数 $R_b^{\text{商品 }ID}(1 \leq \text{商品 } ID \leq L)$ を 2 個生成し，表 4.6 のマトリクス Y の要素が 1 になっている 7 個所について，$\mathbf{H}(\text{ユーザ } ID)^{R_b^{\text{商品 }ID}}$ を算出する．各々の冪乗余値に，対応する商品 ID を付与し，これらの商品 ID 付き冪乗余値をシャッフルした後，ID 管理組織に送る．
4. ID 管理組織は，ステップ 3 の商品 ID を付与された 7 個の値を 5 個の乱数でプロファイルの値ごとに冪乗余し，対応するプロファイル値 ID を付与して 35 個の値を算出する．
5. 商店は，ステップ 2 のプロファイル値 ID を付与された 14 個の値を 2 個の乱数で

[*1] ステップ 2 のシャッフルにより，プロファイルの値ごとにランダムにユーザ ID を入れ替えてユーザのプロファイルを保護する．それぞれのプロファイルの値ごとに該当するユーザは複数必要であるため，ユーザ数が少ないプロファイルのデータは予め削除しておく．

商品ごとに冪乗余して 28 個の値を算出し，プロファイル値 ID を削除し，シャッフルして ID 管理組織へ送り返す．

6. ID 管理組織は，ステップ 4 で算出した 35 個のプロファイル値 ID および商品 ID 付き冪乗余値 $\mathbf{H}(\text{ユーザ } ID)^{R_b^{\text{商品 } ID} R_a^{\text{プロファイル値 } ID}}$ と，ステップ 5 で送られてきた 28 個の冪乗余値 $\mathbf{H}(\text{ユーザ } ID)^{R_a^{\text{プロファイル値 } ID} R_b^{\text{商品 } ID}}$ を照合し，等しければ，当該プロファイル値のユーザが当該商品を購入したとして，クロス集計表の当該プロファイルの値と商品の欄にポイント 1 を加算する．このポイント加算を等しいペア毎に行う[*2]．

ID 管理組織-商店間プロトコルの処理性能を見積もるため，処理性能を左右する冪乗余の回数を明らかにする．上記のアルゴリズムにかかる冪乗余の回数は，ステップ 2 で $NW = 14$ 回，ステップ 3 で $MG = 7$ 回，ステップ 4 で $MGV = 35$ 回，ステップ 5 で $NWL = 28$ 回である．これを主体ごとに整理すると，ID 管理組織はステップ 2 とステップ 4 を行うので $NW + MGV$ 回であり，商店はステップ 3 とステップ 5 を行うので $MG + NWL$ 回である．一方，Vaidya らの提案した基本的な秘匿内積の方式は，ID 管理組織と商店でそれぞれ NVL 回の冪乗余が必要となる．両者の冪乗余の回数を表 4.11 にまとめる．

表 4.11　ID 管理組織-商店間プロトコルの冪乗余の回数

利用主体	Vaidya らの方式 [55]	採用方式 [56]
ID 管理組織	NVL	$NW + MGV$
商店	NVL	$MG + NWL$

表 4.1 のデータの規模と内容に照らすと，採用方式はステップ 2 で $NW = 10^8 \times 3$ 回，ステップ 3 で $MG = 10^5 \times 10$ 回，ステップ 4 で $MGV = 10^5 \times 10 \times 57$ 回，ステップ 5 で $NWL = 10^8 \times 3 \times 10^4$ 回となる．つまり，採用方式は商店でのステップ 5 の NWL 回の冪乗余が支配的である．たとえば，長さ V のベクトルの積集合を求める場合，Vaidya らの方式では，ベクトルの全ての要素 (V 個) を冪乗余する必要があるが，採用方式はベクトルに値が入っている要素 (W 個) のみを冪乗余すれば良いので，冪乗余の回数を $O(NVL)$ から $O(NWL)$ に減らせる．プロファイルの項目は少なくとも 2 種類以上の

[*2] この例では 10 個の値が等しいペアとなり，ID 管理組織は表 4.8 のクロス集計表を手に入れる．ステップ 2 のシャッフルにより，ステップ 6 の時点では，それぞれの商品を購入したユーザ同士のプロファイルの値が入れ替わっているので安全となる．つまり，このプロトコルが安全であるためには，それぞれの商品を販売したユーザが少なくとも 2 人以上必要である．

値を取る (1 種類の値しかないプロファイルの項目には意味が無い) ので，プロファイルの値の種類数 V はプロファイルの項目数 W の少なくとも 2 倍以上となり，採用方式の処理性能は Vaidya らの方式より 2 倍以上優れると見積もれる．

以下では, プライバシ保護の “(要件 1a) プロトコル自体の安全性” および “(要件 1b) プロトコルの出力の安全性” について論じる．プロトコル自体の安全性について，採用したプロトコル [56] の安全性は，semi-honest モデルにおいて，DDH (Decisional Diffie-Hellman) 仮定の安全性に帰着する [45]．一方，プロトコル [56] は出力の安全性を保証しない．そこで，出力すなわちクロス集計表の安全性は，4.4 節で述べたように匿名加工によって保証する．

4.6.2 商店-訪問者間プロトコル

商店の履歴データと訪問者のプロファイルを互いに秘匿しつつ，訪問者が商品の推薦度を手に入れるまでの処理を例にあげて，商店-訪問者間プロトコルのアルゴリズム [55] を詳しく述べる．また，このプロトコルの処理性能と安全性を述べる．

商店のクロス集計表と訪問者のプロファイルを守るために，商店-訪問者間プロトコルに秘匿内積 [55] 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．平文に掛ける数が a の場合は，暗号文を a 乗することで積を計算でき，2 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 2 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している [55]．ナイーブベイズのモデルで推薦度を求めて，訪問者の端末画面の上から順に推薦度の高い商品を並べて推薦する．推薦する商品の順序関係が保たれれば十分であるので，この確率モデルに単調増加関数である対数を適用し，推薦する商品の順序関係 (推薦度の大小関係) は変えずに $\acute{\boldsymbol{x}}$ と $log\acute{\boldsymbol{\theta}}^{(l)}$ の秘匿内積で推薦度を算出できるようにして，商店の購入傾向と訪問者のプロファイルを守る．4.1 式の確率の対数 $log\acute{\theta}_v^{(l)}$ 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，して正の整数に変換する必要がある．そこで，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ることとする．また，冪乗余を積に置き換えることで処理性能も向上させる．すなわち，商店-訪問者間プロトコルは，商店と訪問者の間で 4.1 式の秘匿内積を商品の種類数 L だけ繰り返し，秘匿内積で得られる商品の推薦度 $\acute{\boldsymbol{y}}$ で商品を順位付ける．

$$\acute{y}^{(l)} = \acute{\boldsymbol{x}} log\acute{\boldsymbol{\theta}}^{(l)}. \tag{4.1}$$

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 [42] が知られている．後者を用いた商店-訪問者間プロトコルを設計する．

1. 訪問者は大きな素数 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. 訪問者は，訪問者のプロファイルの暗号文として 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. 訪問者は商店を訪れた際に公開情報 N_c，g と訪問者のプロファイルの暗号文 $E(x_1), ..., E(x_V)$ を商店へ送る．
4. 商店は，購入傾向を正の整数に変換した値 $\log \acute{\theta}_v^{(l)}$ およびそれぞれのプロファイルの値ごとの訪問者のプロファイルの値の暗号文 $E(x_v)$ 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 $E(x_v \log \acute{\theta}_v^{(l)}) = E(x_v)^{\log \acute{\theta}_v^{(l)}} mod N_c^2$ を計算する．
5. 商店は VL 個の $E(x_v \log \acute{\theta}_v^{(l)})$ を，それぞれの商品ごとに全てのプロファイルの値の暗号文を掛けあわせ，加法準同型暗号の性質を利用して全てのプロファイルの値の和をとり，L 個の推薦値の暗号文 $E(\acute{y}^{(l)})$ を計算する．
6. 商店は推薦値の暗号文 $E(\acute{y}^{(1)}), ..., E(\acute{y}^{(L)})$ を訪問者へ送る．
7. 訪問者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．

具体的には，商店は表 4.10 に表すパラメータのマトリクス $\acute{\Theta}$ から本 A に関するパラメータを抜き出して $\log \acute{\boldsymbol{\theta}}^{(l=1)} = \left(\log \left(\left(\frac{2}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{1}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{2}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{1}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{0}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right) \right)$ を求めておき，同様に本 B に関するパラメータを抜き出して $\log \acute{\boldsymbol{\theta}}^{(l=2)} = \left(\log \left(\left(\frac{2}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{0}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{0}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{1}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right), \log \left(\left(\frac{1}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right) \right)$ を求めておく．訪問者が表 4.7 に示す男性の 30 代というプロファイルベクトル $\acute{\boldsymbol{x}} = (1, 0, 0, 1, 0)$ を商店-訪問者間プロトコルに入力すると，秘匿内積によって本 A の推薦度は $\acute{y}^{(l=1)} = \log \left(\left(\frac{2}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right) + \log \left(\left(\frac{1}{6}\right) \left(\frac{6}{10}\right)^{\frac{1}{2}} \right) = -3.4$ と算出でき，同様に本 B の推薦度は $\acute{y}^{(l=2)} = \log \left(\left(\frac{2}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right) + \log \left(\left(\frac{1}{4}\right) \left(\frac{4}{10}\right)^{\frac{1}{2}} \right) = -3.0$ と算出できる．両者の推薦度の大小関係を比較して，訪問者へ本 B，本 A の順に推薦する．

商店-訪問者間プロトコルの処理性能を見積もるため，処理性能を左右する冪乗余の回数を明らかにする．上記のアルゴリズムにかかる冪乗余の回数は，ステップ 1 で 1 回，ス

テップ 2 で $2V$ 回，ステップ 4 で VL 回，ステップ 7 で L 回である．すなわち，商店で VL 回，訪問者で $1 + 2V + L$ 回の冪乗余計算となる．実際の想定では，訪問者はステップ 1 とステップ 2 を商店を訪れる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．Vaidya らと採用方式は同じ秘匿内積を用いているため，両者の冪乗の回数は同じである．両者の冪乗余の回数を表 4.12 にまとめる．

表 4.12　商店-訪問者間プロトコルの冪乗余の回数

利用主体	Vaidya らの方式 [55], 採用方式 [56]
商店	VL
訪問者	$1 + 2V + L$

以下では，“(要件 1a) プロトコル自体の安全性” および “(要件 1b) プロトコルの出力の安全性” について論じる．プロトコル自体の安全性について，Vaidya らのプロトコル [55] の安全性は，semi-honest モデルにおいて，利用する加法準同型暗号の安全性に帰着する．たとえば，Paillier 暗号の場合は，DCR(Decisional Composite Residuosity) 仮定の安全性に帰着する [42]．Vaidya らのプロトコルは，2 つの暗号化されたベクトルから，暗号化された内積値を求めるものであり，入力となるベクトルに制約はない．そのため，攻撃者がどのようなベクトルを入力しても，攻撃には相当しない．また，処理の途中で暗号が復号されることはないので，攻撃者が処理の途中に介入できるか否かは，利用する暗号の安全性に帰着する．そこで，Vaidya らのプロトコルの安全性は，malicios モデルにおいて，利用する加法準同型暗号の安全性に帰着する．Vaidya らのプロトコルは出力の安全性は保証しない．そのため，商店の持つクロス集計表の情報を，訪問者は推定することができる．推薦システムの目的上，この情報漏洩を完全に防止することは困難である[*3]．しかし，クロス集計表は差分プライバシによって匿名加工されているので，訪問者は，クロス集計表のもとになった個人情報および履歴データを知ることはできない．

4.7　まとめ

本章では，ユースケースに基づいて，Inter PPR の技術課題とシステム構成を明らかにした．技術課題では，プライバシについて，ID 管理組織と商店は semi-honest，訪問者は malicious の状況下で，ID 管理組織と商店と訪問者が，自己の保有する情報を他の 2 者に

[*3] たとえば，訪問者側のシステムを難読化しておき，各商品の推薦値を直接開示するのではなく，推薦度トップ 10 の商品のリストのみを開示する方法が考えられる．

秘匿する必要がある．その際，プロトコル自体の安全性とプロトコルの出力の安全性を満たす必要がある．推薦精度について，ID 管理組織と商店の両方の情報を保有する大組織が訪問者のプロファイルを知って推薦する場合と同等の推薦精度を得る必要がある．さらに，ID 管理組織が $10^7 \sim 10^8$ 規模のユーザの 57 種類のプロファイルを保有し，商店が $10^2 \sim 10^5$ 規模のユーザの $10 \sim 10^4$ 種類の商品に対する購入履歴を保有する状況で，2 者間において，ユーザのプロファイル毎の商品の購入傾向に関するクロス集計表を実用的な時間内に計算する必要がある．また，商店が，57 種類のプロファイル毎の $10 \sim 10^4$ 種類の商品の購入傾向を保有し，訪問者が 57 種類のプロファイルを保有する状況で，2 者間において，訪問者への商品の推薦を実用的な時間内に算出する必要がある．以上の技術課題を踏まえて，ID 管理組織と商店の間で，共通のユーザ ID に基づき秘匿積集合プロトコルを用いてクロス集計を行うこと，商店と訪問者の間で，ナイーブベイズに基づいて秘匿内積プロトコルを用いて推薦を算出することにした．また，クロス集計に安全に適用可能な分散環境対応スムージング手法の必要性を明らかにし，5 章で詳述することにした．さらに，訪問者が malicious である場合の対策として，クロス集計表を差分プライバシにより匿名加工することとした．その際，ユーザのプロファイルの値の種類数や推薦する商品の種類数に応じて，クロス集計表が大きくなり，差分プライバシを満たすために付加するノイズが大きくなるため，クロス集計に適した差分プライバシの実現方式について 6 章で詳述することにした．

第 5 章

分散環境対応スムージング

5.1 はじめに

スムージングはデータの特異値およびノイズを平滑化することにより，推薦精度を向上させる技術である．従来のスムージング手法は，組織間で分散していない個票データを対象としていた．しかし，Inter PPR では，個票の情報は，ID 管理組織の保有する個人情報と商店の保有する履歴情報に分散している．そのため，ID 管理組織と商店が互いの情報を秘匿しながら従来のスムージングを実行することはできない．また，従来のスムージング手法は，平滑化のパラメータが多いため，データ数が少ない場合には，パラメータを正しくチューニングできず，推薦精度を向上する効果が発揮できなかった．そこで，本章では，クロス集計表に適用可能で，パラメータの少ない分散環境対応スムージングを提案する．分散環境対応スムージングは，統計化によって個票の情報を失ったクロス集計表にスムージングを施し，データが少ない場合でも効果を発揮する．商品ごとに LOO 尤度を最大にするようにスムージングのパラメータをチューニングする具体的な方法について述べる．

5.2 スムージングとは

スムージングとは，少ないデータで確率分布のパラメータを求めることである．スムージングの起源は，1812 年に Laplace が書いたエッセイだと言われており，ある試行における成功の確率を求める際に，成功確率 = 成功回数/試行回数 ではなく，成功確率 = (成功回数 + 1)/(試行回数 + 2) として，分子と分母に定数を加算するスムージングが述べられている [91]．成功か失敗かの離散かつ二値の試行により観測される回数の分布は二項分

布に従い，この二項分布の共役事前分布はベータ分布であることが知られている．ラプラスのスムージングは，このベータ分布の期待値を求める際に成功回数と失敗回数にそれぞれ 1 回を加えたもの，すなわち，成功確率 $=$ (成功回数 $+1$)/[(成功回数 $+1$) $+$ (失敗回数 $+1$)] であるため，このようなスムージングは Add one スムージングと呼ばれる．機械学習の分野では，事前分布のパラメータをハイパーパラメータと呼ぶため，スムージングは，ハイパーパラメータを適切に定めようとする手法の一つであるとも言える．

1951 年にシャノンは，自然言語の生起をマルコフ過程とみなして単語列の分布でモデル化したが [98]，そのモデルのパラメータ (次の単語が出現する条件付き確率) は文の長さに応じて指数関数的に増加してしまうものであった．たとえば，たった 4 つの英単語列で構成する分布でも，2 万語の 4 乗で 1.6×10^{17} にもパラメータ数が増大するため，これらのパラメータの学習に膨大なデータと計算時間を要してしまう．また，仮に一つでも学習データに含まれない単語があれば，その単語を含む単語列の出現確率が全てゼロとなってしまうゼロ頻度問題 [91] が起こり，パラメータを推定することができなくなってしまう．たとえば，学習データが Saga, Kumamoto, rain という単語を含むが，偶然 Nagasaki という単語を含まないとき，$P(rain|Saga)$ や $P(rain|Kumamoto)$ は値を持つが，$P(rain|Nagasaki)$ は値を持たないため，自然言語として存在しないことになってしまう．このような場合に，少なくとも 1 回は全ての単語を観測したとみなす Add one スムージングを適用すれば，ゼロ頻度問題を回避でき，長崎にも雨が降る．しかし，Add one スムージングによって与えられる確率はパラメータ数が増えてデータが疎になると大きくなり過ぎるため，より小さな確率を与える Lidstone's Laws スムージング [99] や，一度確率を推定してから大き過ぎる確率を減らしていく Good-Turing スムージング [100, 101]，減らした確率を未知の事象に適切に分配する Kneser-Ney スムージング [102, 103] などが提案されている．また，本論文で対象とするディリクレ分布のスムージングはディリクレスムージングと呼ばれている [104, 105, 106, 107, 108].

正田らは，ディリクレスムージングとディリクレ分布のハイパーパラメータ推定が同じであることに着目した [107]．ディリクレ分布のハイパーパラメータの推定手法は Minka によって詳しく解析されている [92]．正田らは，Minka の手法を用いてディリクレ分布のハイパーパラメータ推定を行い，教師ありの文書分類の実験を通じて，ナイーブベイズに対する識別性能の優位性を定量的に示している．しかし，この Minka の手法を組織間で行うには，ID 管理組織が保有するユーザのプロファイルを商店へ開示しなければならないという問題がある．

5.3 分散環境における問題点

Minka の手法は個票データを対象としている．Inter PPR では，個票の情報は，ID 管理組織の保有する個人情報と商店の保有する履歴データに分散している．そのため，ID 管理組織と商店が互いの情報を秘匿しながら Minka の手法を実行することはできない．また，Minka の手法は，平滑化のパラメータ数が，プロファイルの属性毎の取りうる値数を全属性について積算した値 × 商品数となり多いため，データ数が少ない場合はパラメータを正しくチューニングできず，推薦精度を向上する効果が発揮できない．

Minka の手法によるスムージングを具体的な例で説明する．スムージングの計算には，ID 管理組織が保有するプロファイルと商店が保有する商品情報の両者が必要であることから，Minka の手法を組織間でプライバシを保護しながら実行することが困難であることを述べる．

説明を簡単にするために，シンプルなデータベースの例を用いる (表 5.1)．シンプ

表 5.1　シンプルなデータベースの例

ユーザ ID	ID 管理組織が保有するプロファイル			商店が保有する商品情報
(i)	20 代 $(v=1)$	30 代 $(v=2)$	40 代 $(v=3)$	コーヒー $(l=1)$
1	$x_{i=1,v=1}$	$x_{i=1,v=2}$	$x_{i=1,v=3}$	$y_{i=1}^{(l=1)}$
2	$x_{i=2,v=1}$	$x_{i=2,v=2}$	$x_{i=2,v=3}$	$y_{i=2}^{(l=1)}$
3	$x_{i=3,v=1}$	$x_{i=3,v=2}$	$x_{i=3,v=3}$	$y_{i=3}^{(l=1)}$

ルなデータベースでは，ID 管理組織が保有するプロファイルを年代のみ $(W=1)$ とし，そのプロファイルの値が 20 代，30 代，40 代の 3 種類 $(V=3)$ であるとする．商店が扱う商品情報はコーヒーのみの 1 種類 $(L=1)$ であるとする．ID 管理組織のプロファイルと商店の商品情報を結合したユーザの総数を $I=3$ であるとする．ユーザ ID を i，プロファイルの値のインデックスを v，商品のインデックスを l と表す．ID 管理組織が保有するプロファイルを $x_{i,v}$，商店が保有する商品情報を $y_i^{(l)}$ と表す．コーヒーを購入するユーザのプロファイル (20 代，30 代，40 代) の真の分布を $\boldsymbol{\alpha}^{(l=1)}=(1,5,10)$ のディリクレ分布とする．この分布から 3 つのデータをランダムサンプリングすると $\boldsymbol{x}_{i=1}=(x_{i=1,v=1},x_{i=1,v=2},x_{i=1,v=3})=(0.035,0.286,0.679)$ と $\boldsymbol{x}_{i=2}=(0.060,0.407,0.533)$ と $\boldsymbol{x}_{i=3}=(0.057,0.457,0.487)$ が得られる (図 5.1)．なお，本論文のユースケースの想定では x_i は整数であるが，ここでは少ないデータで分かりや

すく説明するために，x_i の値を敢えて実数としている (乱数のシードが 0 の条件において，Numeric python の random クラスの dirichlet メソッドにより発生させたランダムデータを用いている)．生成したデータを表 5.2 にまとめる．

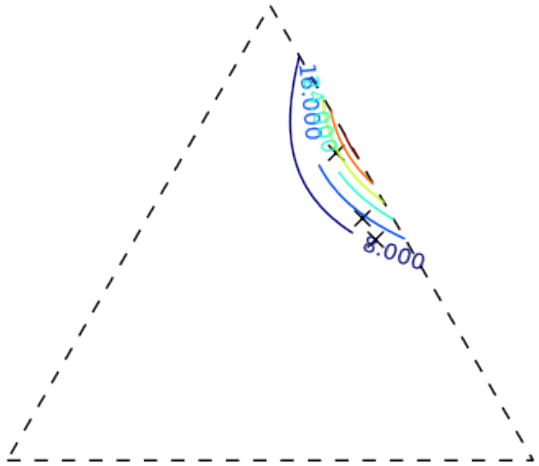


図 5.1　$\boldsymbol{\alpha}^{(l=1)} = (1, 5, 10)$ のディリクレ分布から $\boldsymbol{x}_{i=1} = (0.035, 0.286, 0.679)$, $\boldsymbol{x}_{i=2} = (0.060, 0.407, 0.533), \boldsymbol{x}_{i=3} = (0.057, 0.457, 0.487)$ のデータをサンプリング

表 5.2　Minka の手法へ入力するデータの一例

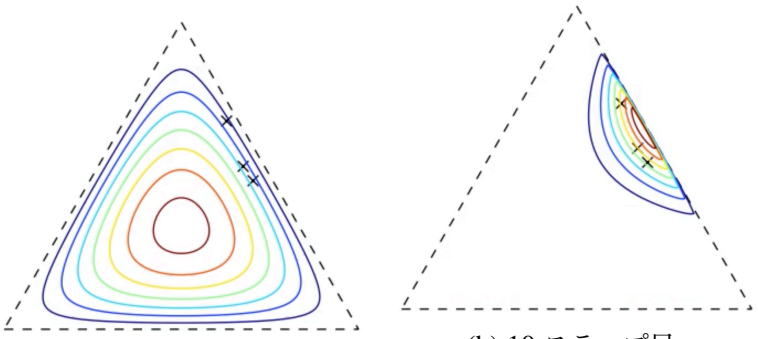
ユーザ ID	ID 管理組織が保有するプロファイル			商店が保有する商品情報
(i)	20 代	30 代	40 代	コーヒー
1	$x_{i=1,v=1} = 0.035$	$x_{i=1,v=2} = 0.286$	$x_{i=1,v=3} = 0.679$	$y_{i=1}^{(l=1)} = 1$
2	$x_{i=2,v=1} = 0.060$	$x_{i=2,v=2} = 0.407$	$x_{i=2,v=3} = 0.533$	$y_{i=2}^{(l=1)} = 1$
3	$x_{i=3,v=1} = 0.057$	$x_{i=3,v=2} = 0.457$	$x_{i=3,v=3} = 0.487$	$y_{i=3}^{(l=1)} = 1$

次に，5.1 式に示す Minka の更新式で，データから真の分布のパラメータを推定する．

$$\alpha_v^{(l,s+1)} = \alpha_v^{(l,s)} \frac{\sum_{i=1}^{I} \left[\Psi\left(x_{i,v} y_i^{(l)} + \alpha_v^{(l,s)} \right) - \Psi\left(\alpha_v^{(l,s)} \right) \right]}{\sum_{i=1}^{I} \left[\Psi\left(\sum_{v=1}^{V} x_{i,v} y_i^{(l)} + \sum_{v=1}^{V} \alpha_v^{(l,s)} \right) - \Psi\left(\sum_{v=1}^{V} \alpha_v^{(l,s)} \right) \right]}. \quad (5.1)$$

パラメータの初期値は，Add one スムージングに相当する $\boldsymbol{\alpha}^{(l=1,s=0)} = (\alpha_{v=1}^{(l=1,s=0)}, \alpha_{v=2}^{(l=1,s=0)}, \alpha_{v=3}^{(l=1,s=0)}) = (2, 2, 2)$ とする．s はパラメータ更新のステップ数である．まず，20 代のユーザがコーヒーを購入する分布のパラメータを計算すると，$\alpha_{v=1}^{(l=1,s=1)} = 2 \times \{ [\Psi(0.035 \times 1 + 2) - \Psi(2)] + [\Psi(0.060 \times 1 + 2) - \Psi(2)] + [\Psi(0.057 \times 1 + 2) -$

$\Psi(2)]\big\}/\big\{[\Psi(0.035\times1+0.286\times1+0.679\times1+2+2+2)-\Psi(2+2+2)]+[\Psi(0.060\times1+0.407\times1+0.533\times1+2+2+2)-\Psi(2+2+2)]+[\Psi(0.057\times1+0.457\times1+0.487\times1+2+2+2)-\Psi(2+2+2)]\big\}=0.39$ となる．同様に計算すると，30 代のユーザがコーヒーを購入する分布のパラメータは $\alpha_{v=2}^{(l=1,s=1)}=2.65$ となり，40 代は $\alpha_{v=3}^{(l=1,s=1)}=3.74$ となるので，1 ステップ目の計算によって，パラメータは $\boldsymbol{\alpha}^{(l=1,s=1)}=(0.39,2.65,3.74)$ と推定できる．



(a) 初期値: $\boldsymbol{\alpha}^{(l=1,s=0)}=(2,2,2)$, 尤度 $=1.94$

(b) 10 ステップ目: $\boldsymbol{\alpha}^{(l=1,s=10)}=(1.11,5.71,8.20)$, 尤度 $=7.16\times10^3$

図 5.2　Minka の手法で推定したディリクレ分布のパラメータとデータの尤度

パラメータの推定結果が妥当であるかを，尤度の上昇によって確認してみる．0 ステップ目の 1 サンプル目の確率は $P\big(\boldsymbol{\alpha}^{(l=1,s=0)}=(2,2,2);\boldsymbol{x}_{i=1}=(0.035,0.286,0.679)\big)=\big\{\Gamma(2+2+2)/[\Gamma(2)\Gamma(2)\Gamma(2)]\big\}\times0.035^{2-1}\times0.286^{2-1}\times0.679^{2-1}=0.82$ である．他のデータも同様に計算すると，2 サンプル目の確率は $P\big(\boldsymbol{\alpha}^{(l=1,s=0)}=(2,2,2);\boldsymbol{x}_{i=2}=(0.060,0.407,0.533)\big)=1.56$ であり，3 サンプル目の確率は $P\big(\boldsymbol{\alpha}^{(l=1,s=0)}=(2,2,2);\boldsymbol{x}_{i=2}=(0.057,0.457,0.487)\big)=1.52$ である．尤度はこれらの確率の積で表せるので，0 ステップ目の尤度は $0.82\times1.56\times1.52=1.94$ である．1 ステップ目も同様に計算すると，尤度は $P\big(\boldsymbol{\alpha}^{(l=1,s=1)}=(0.39,2.65,3.74);\boldsymbol{x}_{i=1}=(0.035,0.286,0.679)\big)\times P\big(\boldsymbol{\alpha}^{(l=1,s=1)}=(0.39,2.65,3.74);\boldsymbol{x}_{i=2}=(0.060,0.407,0.533)\big)\times P\big(\boldsymbol{\alpha}^{(l=1,s=1)}=(0.39,2.65,3.74);\boldsymbol{x}_{i=2}=(0.057,0.457,0.487)\big)=10.9\times7.23\times7.05=5.55\times10^2$ であり，尤度が 1.94 から 5.55×10^2 に上昇したことから，妥当なパラメータを推定できていることが分かる．Minka の更新式を繰り返していくと，10 ステップ目の計算によって，パラメータは $\boldsymbol{\alpha}^{(l=1,s=10)}=(1.11,5.71,8.20)$ に更新され，尤度は $20.3\times19.9\times17.7=7.16\times10^3$ に上がる (図 5.2)．ステップ数に応じた尤度を図 5.3 に

まとめる．Minka の更新式は，パラメータ推定に用いるデータに対する尤度を最大化することが保証されている [92].

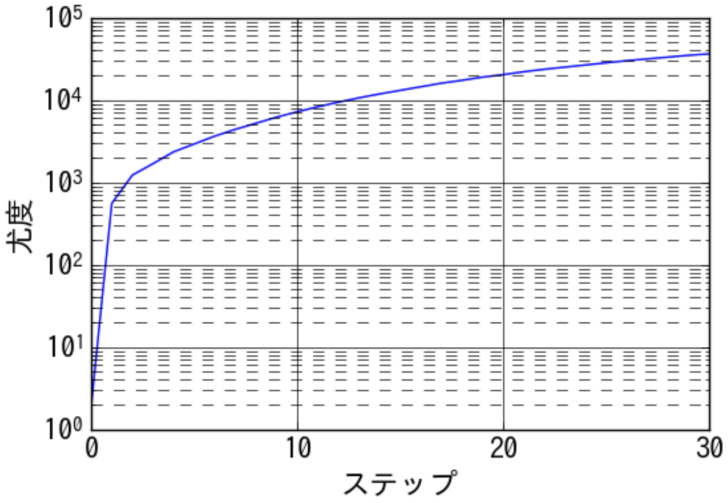


図 5.3　Minka の手法で推定したデータの尤度の推移

しかし，Minka のスムージングの手法は，互いのプライバシ保護を必要としない環境を前提としており，プライバシ保護を必要とする環境にこのまま適用してしまうと，$x_{i,v}y_i^{(l)}$ を用いる度に 1 件ずつ，合計 I 件のデータが相手に漏れてしまうという問題が生じる．たとえ，ID 管理組織と商店が暗号化して $x_{i,v}y_i^{(l)}$ を演算しても，商店は商品情報の $y_i^{(l)}$ を知っているので，演算結果からプロファイル $x_{i,v}$ を把握できてしまう．また，4.6.1 節で述べた ID 管理組織-商店間プロトコルで商店が受け取れるデータはプロファイルの値 (たとえば，男性，20 代，30 代) ごとにシャッフルされているため，プロファイルの値の組み合わせ (たとえば，男性 20 代，男性 30 代) でパラメータを推定する Minka のスムージングの手法を適用しても，パラメータの推定を誤ってしまう．そこで，次節では，プライバシを保護しながら組織間リコメンドを実現できる，分散環境対応スムージングの手法を新たに提案する．

5.4　分散環境対応スムージング

Minka のスムージングを行うにはレコードが残っている状態の個票が必要であり，スムージングをそのまま適用すると相手の組織にプライバシが漏洩してしまう．また，Minka のスムージングでは，プロファイルと商品数の積に応じた (集計表ではセルの数と同じ) 数のスムージングのパラメータを推定する必要があり，推薦精度を向上するために

多くのデータが必要になってしまうという問題があった．そこで，新規提案する分散環境対応スムージングでは，ID 管理組織または商店の一方だけの情報でスムージングを実行できるようにする．また，パラメータ数を削減し，データが少ない場合でも推薦精度を向上できるようにするといった課題に取り組む.

これらの課題を解決するため，プロファイルを構成する属性間の組み合わせの情報が失われていてどのように処理してもユーザのプライバシが漏洩しない，クロス集計表に対してスムージングを適用できるようにする．また，プロファイルを構成する属性間の組み合わせを用いないようにして，プロファイルの属性毎の取りうる値数を全属性について積算した値 × 商品数だけあったスムージングパラメータの数を，商品数まで減らすという方針で解決する．具体化には，クロス集計表から仮想のレコードをサンプリングして，商品ごとに異なるスムージングパラメータを，その LOO 尤度が最大となるようにチューニングできるようにする.

5.4.1 方式概要

提案手法を実現する更新式を導出し，この更新式が，Minka の手法と同じく，最尤推定であることを示す.

商店は，ID 管理組織-商店間プロトコルによって，表 5.2 を集計した表 5.3 のクロス集計表を手に入れる．提案手法は LOO 尤度を用いるため，クロス集計表から仮想のサンプルを leave-one-out したクロス集計表 (表 5.4) を生成しておく．なお，今は説明を簡略化するために年代のみで説明しているので 30 代と 40 代を leave-one-out している (20 代は 1 サンプル未満なので leave-one-out できない)．属性が複数の場合，たとえば，プロファイル項目が性別と年代で，プロファイル値が男性，女性，20 代である場合には，男性 20 代と女性 20 代のように leave-one-out する．男性女性や 20 代 20 代のようにはしない.

5.1 式に替わって，5.2 式の更新式で，データから真の分布のパラメータを推定する.

$$\alpha^{(l,s+1)} = \frac{(J^{(l)}-1)W}{V} \frac{\sum_v \phi_v^{(l)} \frac{\alpha^{(l,s)}-1}{\phi_v^{(l,-i)}+\alpha^{(l,s)}-1}}{\sum_v \phi_v^{(l)} \frac{\phi_v^{(l,-i)}}{\phi_v^{(l,-i)}+\alpha^{(l,s)}-1}} + 1. \tag{5.2}$$

コーヒーを購入した会員数は $J^{(l=1)} = \frac{\sum_v^V \phi_v^{(l=1)}}{W} = \frac{0.152+1.150+1.698}{1} = 3$ であり，パラメータの初期値は，Add one スムージングに相当する $\boldsymbol{\alpha}^{(l=1,s=0)} = (2,2,2)$，すなわち $\alpha^{(l=1,s=0)} = 2$ とする．ユーザがコーヒーを購入する分布のパラメータを計算すると，$\alpha^{(l=1,s=0)} = \big\{[(3-1)\times 1/3] \times [0.152 \times (2-1)/(0.152+2-1) + 1.150 \times (2-1)/(0.150+2-1) + 1.698 \times (2-1)/(1.698+2-1)]/[0.152 \times 0.152/(0.152+2-1) +$

表 5.3　提案手法へ入力するデータの一例

	コーヒー
20 代	$\phi_{v=1}^{(l=1)} = 0.152$
30 代	$\phi_{v=2}^{(l=1)} = 1.150$
40 代	$\phi_{v=3}^{(l=1)} = 1.698$

表 5.4　提案手法へ入力するデータの生成の一例

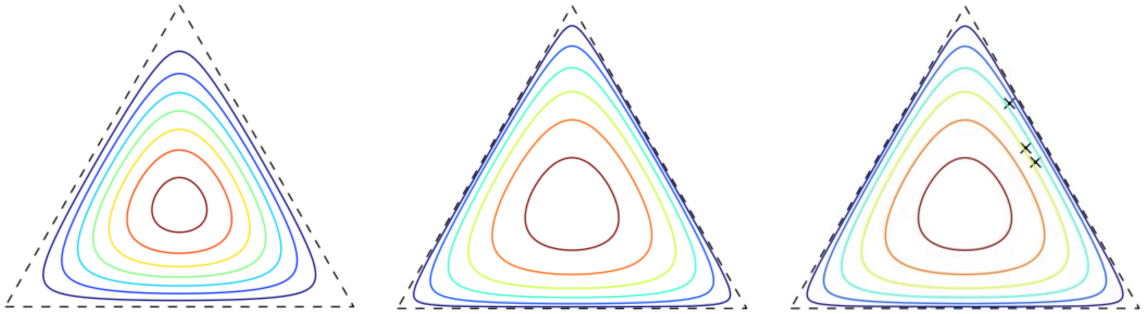
(a) 仮想の 1 サンプル目を leave-one-out したクロス集計表 $\left(\boldsymbol{\phi}^{(l=1,s \bmod 2=0)}\right)$

	コーヒー
20 代	$\phi_{v=1}^{(l=1,s \bmod 2=0)} = 0.152$
30 代	$\phi_{v=2}^{(l=1,s \bmod 2=0)} = 0.150$
40 代	$\phi_{v=3}^{(l=1,s \bmod 2=0)} = 1.698$

(b) 仮想の 2 サンプル目を leave-one-out したクロス集計表 $\left(\boldsymbol{\phi}^{(l=1,s \bmod 2=1)}\right)$

	コーヒー
20 代	$\phi_{v=1}^{(l=1,s \bmod 2=1)} = 0.152$
30 代	$\phi_{v=2}^{(l=1,s \bmod 2=1)} = 1.150$
40 代	$\phi_{v=3}^{(l=1,s \bmod 2=1)} = 0.698$

$1.150 \times 0.150/(0.150+2-1) + 1.698 \times 1.698/(1.698+2-1)]\}+1 = 1.95$ となる．よって，パラメータは $\boldsymbol{\alpha}^{(l=1,s=1)} = (1.95, 1.95, 1.95)$ と推定できる．



(a) 初期値:
$\boldsymbol{\alpha}^{(l=1,s=0)} = (2, 2, 2)$

(b)10 ステップ目:
$\boldsymbol{\alpha}^{(l=1,s=10)} = (1.52, 1.52, 1.52)$

(c)10 ステップ目:
元のデータの尤度 = 4.78

図 5.4　提案手法で推定したディリクレ分布のパラメータ

以上のように，提案手法はクロス集計表の値からパラメータを推定するため，ID 管理組織から商店へプライバシが漏れない．すなわち，安全なスムージングを実現できる．提案手法はプロファイルの個々の値ごとに推定を行うため，ID 管理組織-商店間プロトコルで商店が受け取るデータからパラメータを推定できる．

パラメータの推定結果が妥当であるかを，尤度の上昇によって確認してみる．0 ステップ目の尤度は Minka の説明の際と同じく 1.94 である．1 ステップ目の尤度は $P\big(\boldsymbol{\alpha}^{(l=1,s=1)} = (1.95, 1.95, 1.95); \boldsymbol{x}_{i=1} = (0.035, 0.286, 0.679)\big) \times P\big(\boldsymbol{\alpha}^{(l=1,s=1)} = (1.95, 1.95, 1.95); \boldsymbol{x}_{i=2} = (0.060, 0.407, 0.533)\big) \times P\big(\boldsymbol{\alpha}^{(l=1,s=1)} = (1.95, 1.95, 1.95); \boldsymbol{x}_{i=2} = (0.057, 0.457, 0.487)\big) = 0.87 \times 1.60 \times 1.56 = 2.17$ であり，尤度が 1.94 から 2, 17 に上昇したことから，妥当なパラメータを推定できていることが分かる．提案手法の更新式を繰り返していくと，10 ステップ目の計算によって，パラメータは $\boldsymbol{\alpha}^{(l=1,s=10)} = (1.52, 1.52, 1.52)$ に更新され，尤度は $1.35 \times 1.90 \times 1.86 = 4.78$ に上がる (図 5.4)．ステップ数に応じた尤度を図 5.5 にまとめる．

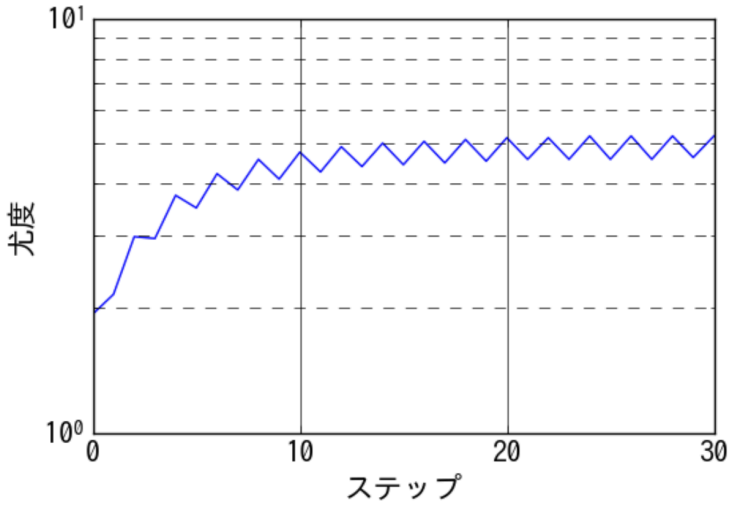


図 5.5　提案手法で推定した元のデータの尤度の推移

提案方式の更新式も，Minka の更新式 (5.1 式) と同じく，パラメータ推定に用いるデータに対する尤度を最大化することを保証する．以下，提案方式の更新式 (5.2) の導出について述べ，提案方式の更新式が，パラメータ推定に用いるデータに対する尤度を最大化することを明らかにする．

5.4.2　確率モデルの設定

提案手法で用いる確率モデルを，プロファイル ($\boldsymbol{x}$) と商品 ($y^{(l)}$) とディリクレ分布のパラメータ ($\boldsymbol{\theta}^{(l)}$ と $\boldsymbol{\alpha}^{(l)}$) を用いて設定する．まず，プロファイル ($\boldsymbol{x}$) と商品 ($y^{(l)}$) の組みで表される I 件のデータがそれぞれ独立であり，その頻度が多項分布 $Multi\{I, \boldsymbol{x}_i, y_i^{(l)}, \boldsymbol{\theta}^{(l)}\} =$

$\frac{I!}{\prod_v (x_{i,v} y_i^{(l)})!} (\theta_v^{(l)})^{x_{i,v} y_i^{(l)}}$ に従うと仮定すると，その確率は 5.3 式のように表せる．

$$P(X, \boldsymbol{y}^{(l)} | \boldsymbol{\theta}^{(l)}) = \prod_i P(\boldsymbol{x}_i, y_i^{(l)} | \boldsymbol{\theta}^{(l)}) \propto \prod_i \prod_v (\theta_v^{(l)})^{x_{i,v} y_i^{(l)}}. \tag{5.3}$$

一方，ディリクレ分布 $Dirichlet\{\boldsymbol{\theta}^{(l)}, \boldsymbol{\alpha}^{(l)}\} = \frac{\Gamma\left(\sum_v \alpha_v^{(l)}\right)}{\prod_v \Gamma\left(\alpha_v^{(l)}\right)} \prod_v (\theta_v)^{\alpha_v^{(l)} - 1}$ の確率は 5.4 式のように表せる．

$$P(\boldsymbol{\theta}^{(l)}; \boldsymbol{\alpha}^{(l)}) \propto \prod_v (\theta_v^{(l)})^{\alpha_v^{(l)} - 1}. \tag{5.4}$$

多項分布とディリクレ分布の共役性を利用して，両者の積に比例する 5.5 式の確率モデルを設定する．

$$\begin{aligned} P(\boldsymbol{\theta}^{(l)} | X, \boldsymbol{y}^{(l)}; \boldsymbol{\alpha}^{(l)}) &\propto P(X, \boldsymbol{y}^{(l)} | \boldsymbol{\theta}^{(l)}) P(\boldsymbol{\theta}^{(l)}; \boldsymbol{\alpha}^{(l)}) \\ &\propto \left[\prod_i \prod_v (\theta_v^{(l)})^{x_{i,v} y_i^{(l)}} \right] \left[\prod_v (\theta_v^{(l)})^{\alpha_v^{(l)} - 1} \right]. \end{aligned} \tag{5.5}$$

5.4.3 パラメータ ($\hat{\boldsymbol{\theta}}^{(l)}$) の学習

プロファイル (X) と商品情報 ($\boldsymbol{y}^{(l)}$) を踏まえて，尤もらしいディリクレ分布のパラメータ ($\hat{\boldsymbol{\theta}}^{(l)}$) を推定する．尤もらしいパラメータとは，5.5 式の提案手法の確率モデルを最大化するパラメータであり，ハットを付与して表す．5.5 式の対数をとって最大化を行うと 5.6 式となる．

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{(l)} &= \underset{\boldsymbol{\theta}^{(l)}}{\operatorname{argmax}} \left[P(X, \boldsymbol{y}^{(l)} | \boldsymbol{\theta}^{(l)}; \boldsymbol{\alpha}^{(l)}) \right] \\ &= \underset{\boldsymbol{\theta}^{(l)}}{\operatorname{argmax}} \left[\sum_{i,v} x_{i,v} y_i^{(l)} \log \theta_v^{(l)} + \sum_v (\alpha_v^{(l)} - 1) \log \theta_v^{(l)} \right]. \end{aligned} \tag{5.6}$$

5.6 式の大括弧の中を F_1 とおき，ラグランジュの未定係数法により，$\sum_v \theta_v^{(l)} = 1$ のディリクレ分布の制約のもとで F_1 を最大にするパラメータを求める．C はラグランジュの未定係数である．

$$F_1 = \sum_{i,v} x_{i,v} y_i^{(l)} \log \theta_v^{(l)} + \sum_v (\alpha_v^{(l)} - 1) \log \theta_v^{(l)} + C \left(\sum_v \theta_v^{(l)} - 1 \right). \tag{5.7}$$

$$\frac{\partial F_1}{\partial \theta_v^{(l)}} = \frac{1}{\hat{\theta}_v^{(l)}} \sum_i x_{i,v} y_i^{(l)} + (\alpha_v^{(l)} - 1) \frac{1}{\hat{\theta}_v^{(l)}} + C = 0. \tag{5.8}$$

$$\hat{\theta}_v^{(l)} = \frac{\sum_i x_{i,v} y_i^{(l)} + \alpha_v^{(l)} - 1}{-C}. \tag{5.9}$$

ディリクレ分布の制約から $C = -\sum_{i,v} x_{i,v} y_i^{(l)} - V(\alpha_v^{(l)} - 1)$ であるので,

$$\hat{\theta}_v^{(l)} = \frac{\sum_i x_{i,v} y_i^{(l)} + \alpha_v^{(l)} - 1}{\sum_{i,v} x_{i,v} y_i^{(l)} + V(\alpha_v^{(l)} - 1)} = \frac{\phi_v^{(l)} + \alpha_v^{(l)} - 1}{\sum_v \phi_v^{(l)} + V(\alpha_v^{(l)} - 1)} \tag{5.10}$$

のように推定できる．$\phi_v^{(l)}$ はクロス集計表の集計値である．

5.4.4 パラメータ ($\hat{\boldsymbol{\alpha}}^{(l)}$) の学習

同一ユーザのプロファイルに含まれる情報を切り離してプライバシ保護するため，全てのプロファイルの値 (v) に対応するパラメータの値 ($\alpha_v^{(l)}$) を等しく扱う．5.10 式は，$\alpha_v^{(l)} = 1$ のときに最尤推定になり，$\alpha_v^{(l)} = 2$ のときに Add one スムージングになる，そこで，$\beta^{(l)} = \alpha_v^{(l)} - 1$ ($\beta^{(l)} \geq 0$) とおいて，Add $\beta^{(l)}$ スムージングの問題を解くことで，ディリクレ分布のパラメータ ($\hat{\boldsymbol{\alpha}}^{(l)}$) を推定することにする．

$$\hat{\theta}_v^{(l)} = \frac{\phi_v^{(l)} + \beta^{(l)}}{\sum_v \phi_v^{(l)} + V\beta^{(l)}}. \tag{5.11}$$

5.11 式は商品 (l) ごとに独立であるため，以下，特に指示しない限り商品の添え字を省略し，部分分数で展開する．

$$\hat{\theta}_v = \frac{\phi_v + \beta}{\sum_v \phi_v + V\beta} = \frac{\sum_v \phi_v}{\sum_v \phi_v + V\beta} \left(\frac{\phi_v}{\sum_v \phi_v} \right) + \frac{V\beta}{\sum_v \phi_v + V\beta} \left(\frac{1}{V} \right). \tag{5.12}$$

ここで，プロファイルの項目数 (W) と，今着目している商品を購入した人数 J から，$\sum_v \phi_v = JW$ より，

$$\hat{\theta}_v = \frac{JW}{JW + V\beta} \left(\frac{\phi_v}{JW} \right) + \frac{V\beta}{JW + V\beta} \left(\frac{1}{V} \right) \tag{5.13}$$

である．ここで，$\omega_1 = \frac{JW}{JW+V\beta}$，$1 - \omega_1 = \frac{V\beta}{JW+V\beta}$ とすると

$$\hat{\theta}_v = \omega_1 \left(\frac{\phi_v}{JW} \right) + (1 - \omega_1) \left(\frac{1}{V} \right) \tag{5.14}$$

となる．

次に，学習データから i 番目の仮想サンプルを抜いて LOO 尤度を求める際のパラメータの推定値を $\hat{\theta}_v^{(-i)}$ と表す.

$$\hat{\theta}_v^{(-i)} = \omega_1^{(-i)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1 - \omega_1^{(-i)}\right) \left(\frac{1}{V} \right). \tag{5.15}$$

ここで，$\omega_{2(\beta)} = \omega_1^{(-i)} = \frac{(J-1)W}{(J-1)W+V\beta}$ ，$1 - \omega_{2(\beta)} = 1 - \omega_1^{(-i)} = \frac{V\beta}{(J-1)W+V\beta}$ とすると，

$$\hat{\theta}_v^{(-i)} = \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1 - \omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \tag{5.16}$$

となる．LOO 尤度を用いて尤もらしい $\hat{\beta}$ を推定する．尤もらしい $\hat{\beta}$ とは，5.3 式の多項分布の確率モデルを最大化するパラメータでありハットを付与して表す.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left(\sum_v \phi_v \log \hat{\theta}_v^{(-i)} \right). \tag{5.17}$$

括弧の中を F_2 とおいて 5.16 式を代入して LOO 尤度を最大化する.

$$F_2 = \sum_v \phi_v \log \hat{\theta}_v^{(-i)} = \sum_v \phi_v \log \left[\omega_{2(\beta)} \frac{\phi_v^{(-i)}}{(J-1)W} + \left(1 - \omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \right]. \tag{5.18}$$

ここで，$\omega_{3(\beta)} = \frac{\omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right)}{\omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1-\omega_{2(\beta)}\right)\left(\frac{1}{V}\right)}$，$\omega_{4(\beta)} = \frac{\left(1-\omega_{2(\beta)}\right)\left(\frac{1}{V}\right)}{\omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1-\omega_{2(\beta)}\right)\left(\frac{1}{V}\right)}$ とすると，

$$\begin{aligned} F_2 \;=\; & \sum_v \phi_v \left[\; \omega_{3(\beta^{(s)})} \log \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \omega_{4(\beta^{(s)})} \log \left(1 - \omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \; \right] \\ & - \sum_v \phi_v \left[\omega_{3(\beta^{(s)})} \log \omega_{3(\beta)} + \omega_{4(\beta^{(s)})} \log \omega_{4(\beta)} \right] \\ \overset{def}{=} \; & F_3(\beta|\beta^{(s)}) - F_4(\beta|\beta^{(s)}). \end{aligned} \tag{5.19}$$

ここで F_4 は，イェンゼンの不等式を用いた 5.20 式より，β の更新によって必ず減少するので，F_3 のみを最大化すれば F_2 を最大化できる．

$$
\begin{aligned}
F_4(\beta|\beta^{(s)}) - F_4(\beta^{(s)}|\beta^{(s)}) &= \sum_v \phi_v \left[\omega_{3(\beta^{(s)})} \log \frac{\omega_{3(\beta)}}{\omega_{3(\beta^{(s)})}} + \omega_{4(\beta^{(s)})} \log \frac{\omega_{4(\beta)}}{\omega_{4(\beta^{(s)})}} \right] \\
&\leq \sum_v \phi_v \left[\log \omega_{3(\beta)} + \log \omega_{4(\beta)} \right] = 0. \qquad (5.20)
\end{aligned}
$$

LOO 尤度を最大にする β を求めるために，F_3 を最大にする β を求める．

$$
\begin{aligned}
\frac{\partial F_3(\beta|\beta^{(s)})}{\partial \beta} &= \sum_v \phi_v \left[\omega_{3(\beta^{(s)})} \frac{1}{\omega_{2(\beta)}} \frac{\partial \omega_{2(\beta)}}{\partial \beta} + \omega_{4(\beta^{(s)})} \frac{1}{1-\omega_{2(\beta)}} \frac{\partial (1-\omega_{2(\beta)})}{\partial \beta} \right] \\
&= \frac{1}{\beta((J-1)W + V\beta)} \left[-\sum_v \phi_v \omega_{3(\beta^{(s)})} V\beta + \sum_v \phi_v \omega_{4(\beta^{(s)})} ((J-1)W) \right] \\
&= 0. \qquad (5.21)
\end{aligned}
$$

(5.19) 式の展開.

$$
\begin{aligned}
&\sum_v \phi_v \log \left[\ \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \ \right] \\
&= \sum_v \phi_v \left[\omega_{3(\beta^{(s)})} + \omega_{4(\beta^{(s)})} \right] \log \left[\ \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \ \right] \\
&= \sum_v \phi_v \omega_{3(\beta^{(s)})} \log \frac{\omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right)}{\omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right)} \left[\ \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \ \right] \\
&\quad + \sum_v \phi_v \omega_{4(\beta^{(s)})} \log \frac{\left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right)}{\left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right)} \left[\ \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \ \right] \\
&= \sum_v \phi_v \omega_{3(\beta^{(s)})} \log \frac{\omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right)}{\omega_{3(\beta)}} + \sum_v \phi_v \omega_{4(\beta^{(s)})} \log \frac{\left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right)}{\omega_{4(\beta)}} \\
&= \sum_v \phi_v \left[\omega_{3(\beta^{(s)})} \log \omega_{2(\beta)} \left(\frac{\phi_v^{(-i)}}{(J-1)W} \right) + \omega_{4(\beta^{(s)})} \log \left(1-\omega_{2(\beta)}\right) \left(\frac{1}{V} \right) \right] \\
&\quad - \sum_v \phi_v \left[\omega_{3(\beta^{(s)})} \log \omega_{3(\beta)} + \omega_{4(\beta^{(s)})} \log \omega_{4(\beta)} \right].
\end{aligned}
$$

上式を満たす β は $\beta^{(s)}$ の条件下で F_3 を最大にするので，β を $\beta^{(s+1)}$ とおいて更新式を導出する．

$$\beta^{(s+1)} = \frac{(J-1)W}{V}\frac{\sum_v \phi_v \omega_{4(\beta^{(s)})}}{\sum_v \phi_v \omega_{3(\beta^{(s)})}} = \frac{(J-1)W}{V}\frac{\sum_v \phi_v \frac{\beta^{(s)}}{\phi_v^{(-i)}+\beta^{(s)}}}{\sum_v \phi_v \frac{\phi_v^{(-i)}}{\phi_v^{(-i)}+\beta^{(s)}}}. \tag{5.22}$$

(5.21) 式の展開.

$$\sum_v \phi_v \left[\omega_{3(\beta^{(s)})} \frac{1}{\omega_{2(\beta)}} \frac{\partial \omega_{2(\beta)}}{\partial \beta} + \omega_{4(\beta^{(s)})} \frac{1}{1-\omega_{2(\beta)}} \frac{\partial (1-\omega_{2(\beta)})}{\partial \beta} \right]$$

$$\left[\begin{array}{l} \text{ここで} \\ \frac{\partial \omega_{2(\beta)}}{\partial \beta} = \frac{\partial}{\partial \beta}\frac{((J-1)W)}{((J-1)W)+V\beta} = ((J-1)W) \times (-1)\{((J-1)W)+V\beta\}^{-2} \times V \\ \qquad = -\frac{V((J-1)W)}{\{((J-1)W)+V\beta\}^2}, \\ \frac{\partial (1-\omega_{2(\beta)})}{\partial \beta} = \frac{\partial}{\partial \beta}\frac{V\beta}{((J-1)W)+V\beta} = \frac{\partial}{\partial \beta}\frac{V}{\frac{(J-1)W}{\beta}+V} \\ \qquad\qquad = V \times \{\frac{(J-1)W}{\beta}+V\}^{-2} \times (-1) \times ((J-1)W) \times \beta^{-2} \times (-1) \\ \qquad\qquad = \frac{V((J-1)W)}{\left(\frac{(J-1)W}{\beta}+V\right)^2 \beta^2} = \frac{V((J-1)W)}{\{((J-1)W)+V\beta\}^2} \\ \text{より} \end{array} \right.$$

$$= \sum_v \phi_v \left[\begin{array}{l} \omega_{3(\beta^{(s)})} \left(\frac{((J-1)W)+V\beta}{((J-1)W)}\right)\left(-\frac{V((J-1)W)}{\{((J-1)W)+V\beta\}^2}\right) \\ +\omega_{4(\beta^{(s)})} \left(\frac{((J-1)W)+V\beta}{V\beta}\right)\left(\frac{V((J-1)W)}{\{((J-1)W)+V\beta\}^2}\right) \end{array} \right]$$

$$= \frac{1}{\beta((J-1)W+V\beta)} \left[\ -\sum_v \phi_v \omega_{3(\beta^{(s)})} V\beta + \sum_v \phi_v \omega_{4(\beta^{(s)})}((J-1)W) \ \right].$$

よって，商品 l のハイパーパラメータ $\hat{\alpha}^{(l)}$ は以下の 5.23 式の収束値として求められる．以上が 5.2 式で述べた提案手法の更新式の導出である．

$$\alpha^{(l,s+1)} = \frac{(J^{(l)}-1)W}{V} \frac{\sum_v \phi_v^{(l)} \frac{\alpha^{(l,s)}-1}{\phi_v^{(l,-i)}+\alpha^{(l,s)}-1}}{\sum_v \phi_v^{(l)} \frac{\phi_v^{(-i)}}{\phi_v^{(l,-i)}+\alpha^{(l,s)}-1}} + 1. \tag{5.23}$$

なお，β の更新によって F_4 は必ず増加し，$\beta^{(s+1)}$ は $\beta^{(s)}$ の時点において F_3 を最大にするように求まるため，LOO 尤度は常に増大する．ただし，5.23 式の更新式において，$\frac{(J^{(l)}-1)W}{V}$ は更新の幅の大きさの係数を，$\frac{\sum_v \phi_v^{(l)} \frac{\alpha^{(l,s)}-1}{\phi_v^{(l,-i)}+\alpha^{(l,s)}-1}}{\sum_v \phi_v^{(l)} \frac{\phi_v^{(l,-i)}}{\phi_v^{(l,-i)}+\alpha^{(l,s)}-1}}$ はパラメータの値をどれだけ大きな値に更新するか (または，どれだけ小さな値に更新するか) を決めるため，$\frac{(J^{(l)}-1)W}{V} = 0$ の場合は値が更新されない．よって，この更新式を用いるには $J^{(l)}-1 > 0$ である必要があり，商品 l を購入した人数 $J^{(l)}$ が 2 人以上いる事が必要条件である．

(5.22) 式の展開.

$$\frac{(J-1)W}{V} \frac{\sum_v \phi_v \omega_{4(\beta^{(s)})}}{\sum_v \phi_v \omega_{3(\beta^{(s)})}}$$

ここで

$$\omega_{3(\beta^{(s)})} = \frac{\omega_{2(\beta^{(s)})}\left(\frac{\phi_v^{(-i)}}{(J-1)W}\right)}{\omega_{2(\beta^{(s)})}\left(\frac{\phi_v^{(-i)}}{(J-1)W}\right)+\left(1-\omega_{2(\beta^{(s)})}\right)\left(\frac{1}{V}\right)} = \frac{\left(\phi_v^{(-i)}\right)}{\left(\phi_v^{(-i)}\right)+\frac{(J-1)W}{\omega_{2(\beta^{(s)})}}\left(1-\omega_{2(\beta^{(s)})}\right)\left(\frac{1}{V}\right)}$$
$$= \frac{\left(\phi_v^{(-i)}\right)}{\left(\phi_v^{(-i)}\right)+\frac{(J-1)W}{V}\left(\frac{1}{\omega_{2(\beta^{(s)})}}-1\right)} = \frac{\left(\phi_v^{(-i)}\right)}{\left(\phi_v^{(-i)}\right)+\frac{(J-1)W}{V}\left(\frac{(J-1)W+V\beta^{(s)}}{(J-1)W}-1\right)}$$
$$= \frac{\left(\phi_v^{(-i)}\right)}{\left(\phi_v^{(-i)}\right)+\beta^{(s)}},$$

$$\omega_{4(\beta^{(s)})} = \frac{\left(1-\omega_{2(\beta^{(s)})}\right)\left(\frac{1}{V}\right)}{\omega_{2(\beta^{(s)})}\left(\frac{\phi_v^{(-i)}}{(J-1)W}\right)+\left(1-\omega_{2(\beta^{(s)})}\right)\left(\frac{1}{V}\right)} = \frac{\frac{(J-1)W}{V}\left(\frac{1}{\omega_{2(\beta^{(s)})}}-1\right)}{\left(\phi_v^{(-i)}\right)+\frac{(J-1)W}{V}\left(\frac{1}{\omega_{2(\beta^{(s)})}}-1\right)}$$
$$= \frac{\frac{(J-1)W}{V}\left(\frac{(J-1)W+V\beta^{(s)}}{(J-1)W}-1\right)}{\left(\phi_v^{(-i)}\right)+\frac{(J-1)W}{V}\left(\frac{(J-1)W+V\beta^{(s)}}{(J-1)W}-1\right)} = \frac{\beta^{(s)}}{\left(\phi_v^{(-i)}\right)+\beta^{(s)}}$$

より

$$= \frac{(J-1)W}{V} \frac{\sum_v \phi_v \frac{\beta^{(s)}}{\phi_v^{(-i)}+\beta^{(s)}}}{\sum_v \phi_v \frac{\phi_v^{(-i)}}{\phi_v^{(-i)}+\beta^{(s)}}}.$$

5.5 まとめ

スムージングは，特異な値やノイズを平滑化する技術である．Minka の手法は，データの生起確率がディリクレ分布に従うことを前提として，最適性を保証可能である．しかし，Minka の手法は，ID 管理組織と商店の両者にまたがる個票の情報を対象とするため，Inter PPR では利用できない．また，パラメータの個数は，プロファイルの属性毎の取りうる値数を全属性について積算した値 × 商品数となり多いため，データ数が少ない場合はパラメータを正しくチューニングできず，推薦精度を向上する効果が発揮できない．本章ではクロス集計表の情報だけを用いて実施可能なスムージング手法を提案した．提案手法は，仮想的な訪問者を想定し，クロス集計表に仮想的な leave-one-out を適用するため，個人情報を利用する必要がない．また，提案手法のパラメータ数は商品数となり，Minka の手法に比べると少ないため，データが少ない場合にも推薦精度を向上させる効果があると期待される．提案手法の推薦精度に対する効果は，次章で差分プライバシの精度とまとめて評価する．

第 6 章

多属性対応差分プライバシ

6.1 はじめに

Inter PPR では，ID 管理組織と商店が ID 管理組織-商店間プロトコルを実行することで，ID 管理組織の側にクロス集計表を生成する．ID 管理組織はクロス集計表を匿名加工した後，商店へ開示する．この匿名加工により，ID 管理組織の保有する個人情報が商店に漏洩すること，および ID 管理組織の個人情報と商店の履歴データが訪問者に漏洩することを防ぐ．匿名加工にはさまざまな手法が提案されているが，攻撃者の背景知識に依存せず，プライバシとデータの有用性のトレードオフを数学的に定式化できる差分プライバシに基づく手法 [86, 109, 110, 111] を取り上げる．しかし，差分プライバシはデータにノイズを付加する手法であり，このノイズの大きさ (標準偏差) は基本的にデータの属性数に比例する．Inter PPR の場合，4.3.3 節の処理性能の要件で示したように，個人情報であるプロファイルの属性数が 57，履歴データにある商品数が最大 10^4 であり，クロス集計表の属性数は両者の属性数の積になるため，付加するノイズが非常に大きくなる．その結果，データの有用性が失われ，推薦精度が大幅に低下する．そこで，多属性データにおいて，データ値とノイズ値の比率 (S/N 比) の低下を抑制する差分プライバシ手法を検討する．

6.2 差分プライバシとラプラスノイズ

6.2.1 匿名加工

2.4.3 節で述べたように，匿名加工の対象には，個人情報のレコードを集めた個票と，個票から算出された統計量の 2 種類があり，各々に対する匿名加工の手法がある．個票を対象とする匿名加工の目的は，個票中の各レコードと個人との対応がつかないようにする

ことである．一方，統計値を対象とする匿名加工の目的は，統計値から元の個票を推定できないようにすることである．個票を対象にする手法は元データを大きく劣化させ，推薦精度を低下させる．Inter PPR では少ないデータを用いて推薦するので，推薦精度の低下は致命的である．そのため，3.3.3 節で述べたように．統計値を対象とする手法を用いる．中でも，プライバシとデータ劣化のトレードオフを数学的に定式化できる差分プライバシを用いて，クロス集計表の匿名加工を行うことにする．

6.2.2 差分プライバシの概要

差分プライバシは，匿名加工の安全性を数学的に定式化できる基準であり，ラプラスメカニズムと呼ばれる匿名加工の方法とともに提案された [86]．ラプラスノイズを統計量に重畳することによって，統計量の元になった個人情報の推定を防止できる．プライバシの基準とラプラスメカニズムの詳細については次節で述べる．

Inter PPR では，ID 管理組織の個人情報と商店の履歴データを保護するにあたり，両者から算出した統計量であるクロス集計表にノイズを重畳することで互いの組織からのプライバシの漏洩を防止する．プライバシ保護のために重畳するノイズが大きいほど安全性を高められるが，データの有用性は低下するため，安全性と有用性のトレードオフが課題となる．安全性については，次節で述べる確率を用いた定式化によって定量的な議論が可能であるが，直感的に言うと，元データ中の任意の 1 ユーザの情報を変えた場合に，匿名加工後の統計量に違いが見られなければ，任意のユーザの情報は守られているという考えに基づく．データの有用性については応用に依存するが，本論文の対象とする推薦においては，元データによる推薦と，差分プライバシを適用した匿名加工後のデータによる推薦の差が小さいほど有用性が高い，すなわち，推薦する商品の変化が小さいほど有用性が高いと定義する．

6.2.3 プライバシ基準とラプラスメカニズム

差分プライバシは匿名加工の安全性を統一的に評価可能な基準である．また，数学的な裏付けがあり，プライバシの安全性を定量的に議論することができる．6.1 式の不等式が成り立つ場合に，差分プライバシは保証される [86].

$$\forall D_1, \forall D_2 \in D, \forall S \subseteq Range(F)$$
$$e^{-\epsilon} \leq \frac{Pr[F(D_1) \in S]}{Pr[F(D_2) \in S]} \leq e^{\epsilon}. \tag{6.1}$$

D はメカニズムのドメイン，すなわち，メカニズムが処理する個票が取りうる値の集合を表す．D_x は個票を表し，D_1 と D_2 は 1 レコードだけ異なる個票である．F はメカニズムと呼ばれ，個票から統計量を算出し，さらに匿名加工の処理を施す．S はメカニズムの出力の部分集合，すなわち，メカニズムが処理した匿名加工後のデータが取りうる値の部分集合を表す．メカニズム F は個票 D_x を匿名加工する．ϵ はセキュリティのパラメータである．6.1 式の不等式が満たされる場合は，あるレコードを含む個票 D_1 を匿名加工したデータが S 内に生じる確率と，そのレコードを含まない個票 D_2 を匿名加工したデータが同じ S 内に生じる確率の比が，エクスポネンシャル ϵ 以下に抑えられることを保証する．個票 D_1 と D_2 が 1 レコード (各レコードが個人を表す場合は 1 人) しか差分がないことから，個人のプライバシが漏洩される確率の比もエクスポネンシャル ϵ 以下に抑えられることを保証する．両辺の対数を取ると，匿名加工によって漏れる情報量が ϵ 以下となることから，差分プライバシは攻撃者が攻撃に利用できる情報量を ϵ 以下に抑えることを保証していると言える．

$\epsilon = 0$ の場合は，データが 1 人分変化しても匿名化後のデータが不変なので，プライバシは完全に保護できるが，データの有用性はゼロとなる．$\epsilon = \infty$ の場合は，プライバシはゼロだが，データの有用性は完全に維持できる．ϵ が小さいほど，プライバシは大きくなり，データの有用性は小さくなる．だが，ϵ の値は守るべき情報資産や想定する攻撃者の強さによって要求される安全性が異なるため，その値を適切に決めることが難しいという課題がある．そこで，先行研究 [112, 113, 114, 115] を鑑みて ϵ の値を変化させながら，安全性とデータの有用性のトレードオフを定量的に評価することにする．6.8 節で述べる匿名加工の推薦精度に対する影響評価では，具体的に $\epsilon = 0.1 \sim 2$ の値を用いる．

差分プライバシを満たす代表的な匿名加工の方法であるラプラスメカニズムは，個票を統計化し，その統計値にラプラスノイズを付与して匿名加工する．ラプラスノイズの大きさはスケールパラメータ λ に比例し (λ を $\sqrt{2}$ 倍するとラプラスノイズの標準偏差になる)，λ が大きくなると匿名加工後のデータの有用性は低下する．ラプラスメカニズムは，6.2 式の不等式を満たす場合に差分プライバシを保証できる．

$$\lambda \geq \frac{\Delta F}{\epsilon}. \tag{6.2}$$

ϵ は前述した差分プライバシのセキュリティパラメータであり，ΔF は個票を統計化した際の感度を表す．感度は個票の任意の 1 レコード (典型的には 1 人のデータ) を変化させたときに，統計値に現れる変化量の最大値である．感度が大きいほど個人情報が漏洩しやすいため，大きなラプラスノイズを重畳する必要がある．そのため，感度が小さければ匿名加工後のデータの有用性を維持できるが，感度が大きければデータの有用性は低下して

しまう．つまり，データの有用性を維持するために感度を抑えなくてはならないという課題がある．そこで，ユーザのレコードを変化させても統計値に現れる変化量を少なくする工夫を講じることにする．具体的には，Inter PPR で匿名加工するクロス集計表の感度を抑えることによって，クロス集計表に重畳するラプラスノイズの大きさを減らして推薦精度を維持する．

6.3 属性数の増加に伴う情報の劣化の問題

ID 管理組織と商店が，個人情報と履歴データを互いに秘匿しながらクロス集計表を生成した後，クロス集計表からプライバシの漏洩を防ぐために，差分プライバシの匿名加工を適用する．しかし，クロス集計表は，1 人分の変化がクロス集計表の全てのセルに反映される可能性があるため，Q 種類の属性と Q' 種類の属性をもつデータから生成したクロス集計表は，感度が $\Delta F = Q \times Q'$ とクロス集計表の大きさに比例し，匿名加工後のデータの有用性が著しく劣化してしまうという問題がある．たとえば，Inter PPR では個人情報の属性 (プロファイルの値の種類) 数は 57 で，履歴データの属性 (商品の種類) 数は $10 \sim 10^4$ なので，ノイズの大きさ $\lambda \geq \frac{\Delta F}{\epsilon}$ が大きくなり，クロス集計表の有用性が失われてしまう．このような属性数の増加に伴う情報の劣化の問題に対して，差分プライバシによる匿名加工後のデータの有用性を維持するために，感度を抑えるアプローチが試みられてきた．

第 1 のアプローチとして，$1/n$ に比例する統計量の組み合わせによって所望の統計量を算出して感度を抑えるアプローチがある [109, 116]．ラプラスメカニズムでの匿名加工による情報の劣化は，統計化の手法に応じて決まる感度を小さくすることによって，抑えることができる．たとえば，個票から平均値を求める場合には，個票を構成するレコード数 n に反比例させて，感度を $1/n$ に小さくできる．このように感度を減らせる理由は，個票を構成する 1 レコードを変化させても，平均をとった匿名加工後の統計値は $1/n$ 倍しか変化しないからである．同様に分散の場合も匿名加工後の統計値は $1/n$ 倍しか変化しないので，同じく感度を $1/n$ に抑えることができる．つまり，所望の統計量を平均や分散などの $1/n$ に比例する統計量の組み合わせによって算出することで，感度を抑えることができる．しかし，このアプローチには算出可能な統計量が限定されてしまうという限界がある．たとえば，最大や最小の計算を含む場合は，上記のようなレコード数に応じた $1/n$ 倍の効果が生まれないため，感度を小さくすることはできない．

第 2 のアプローチとして，1,000 種類以上の商品を購入するような例外的な個人を無視して感度を抑えるアプローチがある [117, 118, 119, 120]．感度の大きさは，個票を構成す

る 1 レコードのあらゆる変化を考慮して，匿名加工後の統計値に現れる変化量の最大値とするので [121, 122, 123]，例外的な個人に引きずられて感度が大きくなりやすい．そのため，このアプローチではこれらの個人を一定の確率で無視するが，例外状況ではプライバシを保護できないという限界がある．たとえば，滅多に生じないと考えていたレコードが万が一生じてしまうとその個人のプライバシは漏洩し，匿名加工の安全性が低下する．

第 3 のアプローチとして，統計化に用いる属性数を減らすことによって感度を抑えるアプローチがある [124]. たとえば，性別と身長，体重の 3 つの属性からなるレコードで構成されている個票から，背の高い男性 (180cm を超える男性) と体重が重い男性 (80kg を超える男性) の人数を匿名加工しつつ集計する場合を考える．この場合の感度には 2 を用いるべきである．なぜならば，匿名加工の処理で “男性 & 190cm & 90kg” のレコードを “女性 & 150cm & 50kg” というレコードに変えてしまうと，匿名加工後の集計表から背の高い男性が 1 人減り，体重が重い男性も 1 人減るので，合わせて 2 人の変化が生じるからである．そこで，体重の属性を用いないことにすると，感度を 1 に減らせる．なぜならば，匿名加工の処理で “男性 & 190cm” のレコードを “女性 & 150cm” というレコードに変えても，匿名加工後の集計表から背の高い男性が 1 人減るだけなので，1 人の変化しか生じないからである．しかし，属性の部分集計を用いるアプローチは，オリジナルのデータの情報を欠落させてしまう (上記の例では，体重の情報が欠落している) ため，匿名加工後のデータの有用性が低下してしまう [125].

表 6.1 に示すように，多くの属性を用いれば，統計が有する情報量も多くなるが，同時にプライバシ漏洩のリスクも大きくなってしまう．属性数が多くなると感度が大きくなる

表 6.1　匿名加工の要件

	属性数	
	多い	少ない
統計が有する情報量	多い	少ない → 多い
プライバシ漏洩のリスク 匿名加工に要するデータの歪み	大きい	小さい
匿名加工後のデータの有用性	低い	低い → 高い

ため，匿名加工に要するデータの歪みも大きくなって，匿名加工後のデータの有用性を低下させてしまう．属性数を少なくすれば，プライバシ漏洩のリスクと匿名加工に要するデータの歪みを小さくできるが，統計が有する情報量も少なくなるため，匿名加工後のデータの有用性を低下させてしまう．そこで，少ない属性数を用いつつも統計が有する情

報量の低下を抑止し，匿名加工後のデータの有用性を高める必要がある．

第 4 のアプローチとして，主成分分析などを用いて情報量をできるだけ保ちながら属性数を減らして感度を抑えるアプローチがある [66, 126, 127, 128]．しかし，主成分分析や相関分析後の値から各属性間の関係を把握できてしまうことから，属性数削減の処理によってプライバシが漏洩する可能性が生じるという限界がある．たとえば，McSherry は差分プライバシを適用した協調フィルタリングを提案しているが [66]，協調フィルタリングで用いる相関の情報からプライバシが漏洩することが指摘されている [68].

6.4 多属性データの劣化防止

本節では，多属性データの劣化を防止するための方針として，感度を抑えるデータの正規化と属性間の関係の利用，さらにスムージングについて述べる．

6.4.1 データの正規化

推薦の現場では，突出して多くの商品を購入するユーザ (以後，爆買いユーザと呼ぶ) に引きずられて推薦が偏らないようにするためなどの理由で，履歴データをユーザごとに正規化することが行われている．しかし，この正規化は現場のノウハウであるとして，差分プライバシの研究では無視されてきた．ところが，差分プライバシの感度すなわちノイズの大きさが属性数に比例する原因は，1 人分のレコードの変化の最大量が属性数に比例する．Inter PPR において差分プライバシの対象となるデータはクロス集計表であるが，そこでの属性には商品の種類がある．全種類の商品を購入する 1 人が加わると，クロス集計表に商品数だけ変化が生じるので，商品数 (属性数) に比例した感度となる．ここで，ユーザ毎の正規化の効果を考える．ユーザの購入した商品種類数の和を 1 に正規化するとする．1 商品のみ購入したユーザが加わった場合には，クロス集計表の該当箇所の値が 1 だけ大きくなる．5 商品購入したユーザが加わった場合には 5 カ所の値が $\frac{1}{5}$ ずつ大きくなり，全商品を購入したユーザが加わった場合には全個所の値が $\frac{1}{A}$ だけ大きくなる．ここで，A は商品の総種類数である．したがって，ユーザ毎の正規化により，ユーザの購入数に関係なく，感度を 1 に抑えることができ，差分プライバシのノイズを押さえることができる．

正規化によると，ノイズが小さくなる一方で，データの値も $\frac{1}{5}$ や $\frac{1}{A}$ のように小さくなるが，S/N 比は一般に向上する．この点について以下に説明する．最初に，商店の履歴データに含まれるユーザが，1 人当たり平均 A 種類の商品を購入していた場合を考える．

この場合，各ユーザの各購入情報は正規化によって $\frac{1}{A}$ に圧縮されるので，クロス集計表の A 個所に $\frac{1}{A}$ が加算される．ユーザが B 人とすると，上記の加算が B 人分発生する．これに対して，クロス集計表の全個所に $\frac{\text{感度}}{\epsilon} = \frac{1}{\epsilon}$ のノイズが重畳されるので，S/N 比は $\frac{BA}{A} / \frac{C}{\epsilon} = \frac{B\epsilon}{C}$ となる．ここで，C はクロス集計表のセルの総数である．一方，正規化しない場合には，1 人当たりクロス集計表の A 個所に 1 が加算される一方，C 個所に $\frac{\text{商品総種類数}}{\epsilon}$ だけのノイズが重畳されるので，S/N 比は $AB/\frac{C \cdot \text{商品総種類数}}{\epsilon} = \frac{B\epsilon}{C} \times \frac{A}{\text{商品総種類数}}$ となる．以上から，正規化 “する場合” は “しない場合” に比べて，S/N 比が $\frac{\text{商品総種類数}}{\text{1 人あたりの平均購入種類数}}$ だけ向上する．表 4.1 のデータの規模と内容にあてはめると，商品総種類数は高々 1 万，1 人あたりの平均購入種類数は高々 10 なので，正規化により S/N 比は 1,000 倍程度向上すると期待できる．具体的な分析は 6.5.3 節で述べる．

6.4.2 属性間の関係の利用

6.3 節で述べた第 3 のアプローチ，すなわち，個票のレコードを構成する属性の一部を部分集計して統計化するアプローチを発展させて，匿名加工後のデータの有用性を維持する．統計化に用いる属性数を減らすことによって感度を小さく抑えつつ，統計化に用いる属性とそれ以外の属性との間の関係を利用して，統計化した属性の統計値からそれ以外の属性の統計値を推定する．両者の関係は，プライベートな情報が漏れないようにするため，公的情報などの公開知識から導き出す．もしも，公的情報のみから両者の関係を導き出せない場合は，個票の情報のうち，平均や分散などの感度の低い情報を用いることで導き出す．

公開知識を利用して情報の劣化を抑える新たな匿名加工を，以下の 5 つのステップにより実現する．

1 公開知識を用いて，属性間の関係を導き出す

1'a 匿名加工において感度が小さくて済む統計値を算出する

1'b 算出した統計値からプライバシが漏洩しないように統計値を匿名加工する

1'c 匿名加工した統計値から，属性間の関係を導き出す

2 個票のレコードを構成する属性の一部を取り出して部分的に統計化する

3 部分的に統計化した統計値を小さなノイズで匿名加工する

4 導き出した関係を用いて，統計化に用いなかった属性の統計値を推定する

5 統計化に用いた属性とそれ以外の属性の統計値を用いて，所望の統計値を求める

上記の匿名加工の処理の流れを，図 6.1 に沿って具体的に説明する．図の左上に個票を示す．所望の統計値は背の高い男性 (180cm を超える男性) と体重の重い男性 (80kg を超える男性) の人数であり，集計表の形式で表す．従来の手法では，それぞれの属性に該当する人数を集計し，ノイズを付加して匿名加工することで，プライバシを保護した集計表を生成する．しかし，6.3 節で述べた通り，この場合の感度は 2 を用いなければならないため，匿名加工後のデータの有用性を劣化させてしまう恐れがある．そこで，提案手法は，図の右上に示す性別と身長だけの (体重を除いた) 個票を用いて部分的に集計を行う．このようにすると感度は 1 で済むため，匿名加工に要する歪みを小さくできる．体重の重い男性の人数は，図の右に示すような，性別と身長と体重の関係を用いて推定する．これらの関係を，理想的には国が公表しているような公開情報から，正確に導き出せることが望ましい．もしくは，そのような理想的な状況ではなくても，個票のレコード数に反比例して感度が低くなるような平均や分散などの統計値を用いて，これらの関係を導き出す．

以下では，6.4.1 節で述べた正規化と，本節で述べた少ない属性と属性間関係の利用との組み合わせについて述べる．正規化を行うと感度は正規化幅 (たとえば 1) に固定され，それに伴ってノイズの大きさも固定されるため，扱う属性数を減らしてもノイズの大きさは減らない．しかし，S/N 比は改善される．扱う属性数 (商品の種類) を減らすと，6.4.1 節の S/N 比の分析において，ユーザ 1 人あたりの商品の平均購入種類数 A がより小さな値に変わる．S/N 比の向上は，$\frac{\text{商品総種類数}}{\text{1 人あたりの平均購入種類数}}$ であるため，1 人あたりの平均購入種類数が小さくなることで，S/N 比がより向上する．

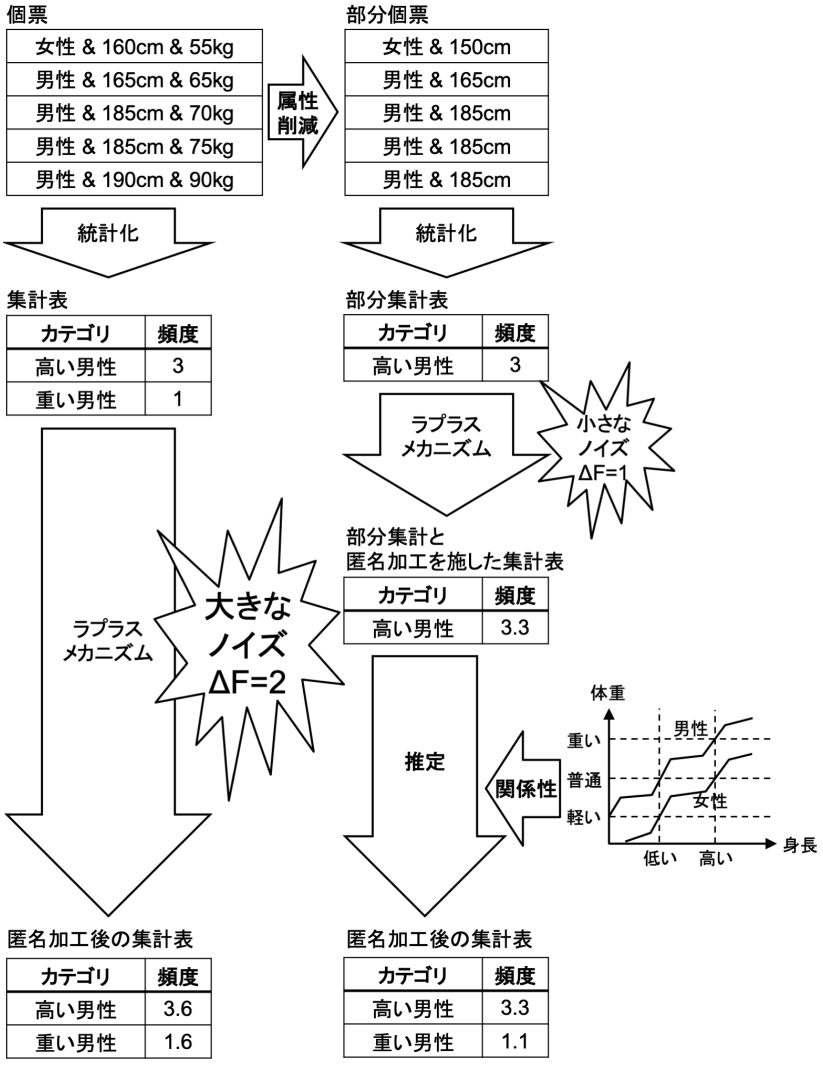


図 6.1　匿名加工の処理の一例 (従来手法は左側．提案手法は右側)

6.4.3　スムージングの利用

差分プライバシによって統計量に大きなラプラスノイズが付加された場合，その統計量は異常値とみなすことができるので，スムージングによって，その影響を低減できる可能性がある．

6.5　多属性対応差分プライバシの実現

クロス集計表は感度が大きく，感度に応じた大きなノイズを重畳して匿名加工すると，クロス集計表が劣化して推薦精度が低下してしまう．前節では，この推薦精度の低下を抑止する方針について述べた．本節では，現実のデータセットを用いて映画の推薦システムを想定し，前節の方針に沿って，データの正規化，属性間の関係の利用，スムージングの利用を Inter PPR で実現する方法について具体化する．

6.5.1　データセット

4.2 節で述べたユースケースに合致する評価用データセットが存在しないため，利用可能なデータセットのうち最もユースケースに近い MovieLens 1M データセット [129] を用いて，提案手法を具体化する．MovieLens 1M データセットは，6,040 人のユーザによる，3,952 本の映画に対する 1,000,209 レコードの評価値 (レーティングと呼ぶ) の個票のセットである．MovieLens 1M データセットの概要を表 6.2 に示す．それぞれの映画

表 6.2　MovieLens 1M データセットの概要

ユーザ ID	プロファイル	映画 (カテゴリ)	レーティング
1	Male, 18–24, programmer	Waterworld, (Action & Adventure)	5
1	Male, 18–24, programmer	Beverly Hills Cop, (Action & Comedy)	4
2	Female, 25–34, writer	Sabrina, (Comedy & Romance)	3
3	Male, 18–24, programmer	Star Trek, (Action & Adventure & Sci-Fi)	5
4	Female, 25–34, artist	Sound of Music, (Musical)	4

には Action や Adventure，Comedy などの 18 種類の基本カテゴリ，および基本カテゴリの組み合わせで “Action & Adventure” や “Comedy & Romance”，“Action & Adventure & Sci-Fi” などの複合カテゴリが付与されている．複合カテゴリは $2^{18} - 19$ 種類 (基本カテゴリーおよび Null カテゴリを除く) の可能性がある．各ユーザには性別 (2 種類)，年代 (7

種類)，職業 (21 種類) などのプロファルが付与されている．ユーザは映画に 1 点から 5 点までの 5 段階のレーティングをしており，5 点が最も良い評価値である．

6.5.2　想定システム

上記の MovieLens 1M データセットを用いて，できるだけ実用に近い推薦システムを想定し，その推薦システムの推薦精度を評価することにする．MovieLens 1M データセットのレーティングのデータを，商店が映画を販売した履歴データとみなす．具体的には，ユーザによって 4 点以上が付与されたレコードにある映画は，そのユーザによって購入されたとみなす．実験のシナリオを図 6.2 に示す．

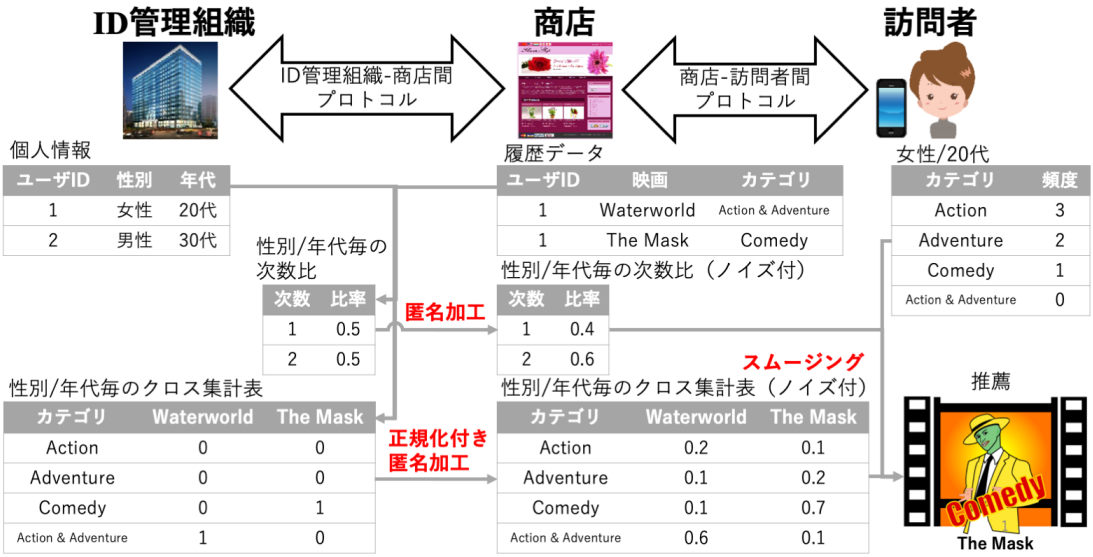


図 6.2　MovieLens を用いた Inter PPR の推薦精度の評価

商店は，販売した映画のカテゴリを，ID 管理組織の協力を得てユーザのプロファイルごとに集計し，プロファイルごとに人気のある (よく購入されている) 映画とそのカテゴリをクロス集計表として統計化する．たとえば，商店は 20 代の女性が購入した映画とそのカテゴリを集計して統計化し，訪問者のプロファイルと訪問者が過去に購入した映画のカテゴリに応じて，訪問者が好みそうな映画を推薦する．訪問者への推薦は，訪問者の携帯端末上で R 件をランキング形式で提示する．ユーザが購入した映画がランキングに含まれる割合 (推薦システムの精度評価で用いられる指標の一つであり，Precision at R と呼ばれる [130]．本論文では P@R と表す) で推薦の精度を評価する．

組み合わせられている基本カテゴリの数を次数と呼ぶ．たとえば, “Action & Adventure”

のように 2 種類の基本カテゴリを組み合わせたカテゴリを 2 次のカテゴリと呼び，3 種類を組み合わせたカテゴリを 3 次のカテゴリと呼び，i 種類を組み合わせたカテゴリを i 次のカテゴリと呼ぶ．基本カテゴリと複合カテゴリを合わせると，それぞれの映画には $2^{18} - 1$ 種類のカテゴリが付与される可能性がある．システムで扱う次数の上限を高くしていくと，商店が手に入れるクロス集計表や，訪問者が保有しているプロファイルの表のサイズが大きくなるため，情報量が増えて推薦の手がかりも増える．これは，より多くの属性を考慮することに相当する．ID が 1 のユーザが Waterworld と The Mask を購入し，ID が 2 のユーザが The Mask と Golden Child (カテゴリは”Action & Adventure & Comedy”) を購入したとする．そのとき，ID 管理組織が手に入れるクロス集計表を表 6.3 に示す．

表 6.3　クロス集計表の例

カテゴリ	Waterworld	The Mask	Golden Child
Action	0	0	0
Adventure	0	0	0
Comedy	0	2	0
Action & Adventure	1	0	0
Action & Comedy	0	0	0
Adventure & Comedy	0	0	0
Action & Adventure & Comedy	0	0	1

商品のカテゴリーが組合せ型である場合に，実用的な推薦システムでは，あるカテゴリーの商品をカウントする際，その要素となるカテゴリーもカウントすることが多い．MovieLens の場合，“Action & Adventure” の映画が購入された際，Action の映画および Adventure の映画も購入されたとみなして，Action および Adventure もカウントする．これは，“Action & Adventure” の映画を購入しそうな人は，Action の映画および Adventure の映画も購入しそうなので，Action の映画および Adventure の映画も推薦できるようにカウントしておくという経験則を反映している．本評価実験でも，できるだけ実用的な推薦システムを想定するために，この経験則を用いることにした．この経験則を以下では，「部分一致カウント」と呼ぶことにする．部分一致カウントを用いた場合に，表 6.3 のクロス集計表は，表 6.4 のようになる．このクロス集計表にラプラスノイズを加えて匿名加工を施し，クロス集計表から ID 管理組織の個人情報と商店の履歴データからのプライバシ漏洩を防止する．この匿名加工後のクロス集計表にスムージングを適用することで，プ

ライバシを保護したまま推薦精度を向上させる．なお，要素となるカテゴリーをカウントしない場合，完全一致カウントと呼ぶことにする．

表 6.4　部分一致カウントしたクロス集計表

カテゴリ	Waterworld	The Mask	Golden Child
Action	1	0	1
Adventure	1	0	1
Comedy	0	2	1
Action & Adventure	1	0	1
Action & Comedy	0	0	1
Adventure & Comedy	0	0	1
Action & Adventure & Comedy	0	0	1

6.5.3　データの正規化

表 6.4〜表 6.8 を用いて，本例題における正規化の流れと S/N 比の向上について説明する．ID が 1 のユーザは男性で Waterworld (カテゴリは “Action & Adventure”) と The Mask (Comedy) を購入していた．ID が 2 のユーザは女性で The Mask (Comedy) と Golden Child (“Action & Adventure & Comedy”) を購入した．その結果の集計が表 6.4 となる．部分一致カウントにより，ユーザ 1 が購入した Waterworld は “Action & Adventure”，Action，Adventure の 3 商品としてカウントされる．同様に，ユーザ 1 およびユーザ 2 が購入した The Mask は Comedy としてカウントされる．ユーザ 2 が購入した Golden Child は “Action & Adventure & Comedy”，“Action & Adventure”，“Action & Comedy”，“Adventure & Comedy”，Action，Adventure，Comedy の 7 商品としてカウントされる (表 6.4).

そこで，正規化により，ユーザ 1 の購入した各商品は 1/4 ずつのカウント，ユーザ 2 の購入した各商品は 1/8 ずつのカウントとなる (表 6.5).

今，簡単のために $\epsilon = 1$ とする (一般の ϵ 値でも同様の議論が成立する)．また，ノイズの大きさが λ のラプラスノイズを $\mathbf{L}(\lambda)$ で表すことにする．クロス集計表の感度は 1 に固定されるので，集計表の各セルに，$\mathbf{L}(\lambda = \frac{1}{\epsilon}) = \mathbf{L}(\lambda = 1)$ が重畳される (表 6.6).

表 6.6 の各値に，ユーザ毎の購入数 (4 と 8) の最小公倍数を掛けると，S/N 比を維持したまま，表 6.7 のようになる．

表 6.5　正規化したクロス集計表

カテゴリ	Waterworld	The Mask	Golden Child
Action	1/4	0	1/8
Adventure	1/4	0	1/8
Comedy	0	1/4 + 1/8	1/8
Action & Adventure	1/4	0	1/8
Action & Comedy	0	0	1/8
Adventure & Comedy	0	0	1/8
Action & Adventure & Comedy	0	0	1/8

表 6.6　正規化して匿名加工したクロス集計表

カテゴリ	Waterworld	The Mask	Golden Child
Action	1/4 + $\mathbf{L}(\lambda = 1)$	0	1/8 + $\mathbf{L}(\lambda = 1)$
Adventure	1/4 + $\mathbf{L}(\lambda = 1)$	0	1/8 + $\mathbf{L}(\lambda = 1)$
Comedy	0	1/4 + 1/8 + $\mathbf{L}(\lambda = 1)$	1/8 + $\mathbf{L}(\lambda = 1)$
Action & Adventure	1/4 + $\mathbf{L}(\lambda = 1)$	0	1/8 + $\mathbf{L}(\lambda = 1)$
Action & Comedy	0	0	1/8 + $\mathbf{L}(\lambda = 1)$
Adventure & Comedy	0	0	1/8 + $\mathbf{L}(\lambda = 1)$
Action & Adventure & Comedy	0	0	1/8 + $\mathbf{L}(\lambda = 1)$

表 6.7　感度を映画のユーザ毎の売上総数の最小公倍数にして匿名加工したクロス集計表

カテゴリ	Waterworld	The Mask	Golden Child
Action	2 + $\mathbf{L}(\lambda = 8)$	0	1 + $\mathbf{L}(\lambda = 8)$
Adventure	2 + $\mathbf{L}(\lambda = 8)$	0	1 + $\mathbf{L}(\lambda = 8)$
Comedy	0	3 + $\mathbf{L}(\lambda = 8)$	1 + $\mathbf{L}(\lambda = 8)$
Action & Adventure	2 + $\mathbf{L}(\lambda = 8)$	0	1 + $\mathbf{L}(\lambda = 8)$
Action & Comedy	0	0	1 + $\mathbf{L}(\lambda = 8)$
Adventure & Comedy	0	0	1 + $\mathbf{L}(\lambda = 8)$
Action & Adventure & Comedy	0	0	1 + $\mathbf{L}(\lambda = 8)$

一方で，正規化しない場合には，クロス集計表の感度は構造的ゼロを除く集計表のセル数 (11) なので[*1]，正規化前の表 6.4 に $\mathbf{L}(\lambda = 11)$ を加えることになり，表 6.8 のようになる．正規化した場合の表 6.7 と正規化しない表 6.8 を比較すると，明らかに正規化した場合の S/N 比が優れる．また，商品購入数が少ないユーザ (ここでは ID が 1 のユーザ) の情報ほど S/N 比が優れることが分かる．

表 6.8　感度をクロス集計表のセル総数にして匿名加工したクロス集計表

カテゴリ	Waterworld	The Mask	Golden Child
Action	$1 + \mathbf{L}(\lambda = 11)$	0	$1 + \mathbf{L}(\lambda = 11)$
Adventure	$1 + \mathbf{L}(\lambda = 11)$	0	$1 + \mathbf{L}(\lambda = 11)$
Comedy	0	$2 + \mathbf{L}(\lambda = 11)$	$1 + \mathbf{L}(\lambda = 11)$
Action & Adventure	$1 + \mathbf{L}(\lambda = 11)$	0	$1 + \mathbf{L}(\lambda = 11)$
Action & Comedy	0	0	$1 + \mathbf{L}(\lambda = 11)$
Adventure & Comedy	0	0	$1 + \mathbf{L}(\lambda = 11)$
Action & Adventure & Comedy	0	0	$1 + \mathbf{L}(\lambda = 11)$

以上の正規化の処理は，匿名加工を行う ID 管理組織がユーザ毎の購入商品を知っていることが前提となっている．しかし，Inter PPR 環境では，ユーザ毎の購入商品は商店の履歴データであり，ID 管理組織は直接知ることはできない．そこで，4.6.1 節で述べた ID 管理組織-商店間プロトコルを改変することで，Inter PPR 環境でも正規化を実施できるようにする．以下ではその工夫を説明する．

4.6.1 節の ID 管理組織-商店間プロトコルでは，商店は商品の種類毎の異なる乱数を用意し，ID 管理組織はプロファイルの属性値毎の異なる乱数を用意していた．ユーザ ID のハッシュ値を各々の乱数で冪乗余し，結果の値に商品名あるいはプロファイルの属性値をラベルとして付加して交換し (ステップ 2，3)，照合していた (ステップ 6)．本章の実施例の場合，商店はステップ 3 においてユーザ ID のハッシュ値を，購入した映画毎の異なる乱数で冪乗余し，結果の値に映画名をラベルとして付加して ID 管理組織に送ることになる．

[*1] クロス集計表における構造的ゼロとは，ゼロ以外の値が入り得ないセルを意味する．すなわち，Waterworld のカテゴリが “Action & Adventure” であると知られている場合は，Action, Adventure, “Action & Adventure” 以外のセルは構造的ゼロである．構造的ゼロのセルの値はゼロだと知られているので，ノイズを重畳する必要はない．一方，映画のカテゴリが非公開の場合は，クロス集計表の感度は集計表のセル数 (21) となる．

これを以下のように改変する．商店は映画とは関係なく D 種類の乱数を用意することにする．D は公開値である．ステップ 3 において，商店は各ユーザの ID を D 種類の乱数で冪乗余し，ユーザ毎に D 個の冪乗余値を得る．ID が 1 のユーザが D 種類の映画を購入していた場合には，D 個の冪乗余値に異なる映画名を一つずつ付与して ID 管理組織に送る．ID が 2 のユーザが 1 種類の映画のみ購入していた場合には，D 個の冪乗余値全てにその映画名を重複して付与し，ID 管理組織に送る．ID が 3 のユーザが 3 種類の映画を購入していた場合には，D 個の冪乗余値のうち $D/3$ 個ずつに，同じ映画名を重複付与して，ID 管理組織に送る．その結果，ステップ 6 において，クロス集計表には ID が 1 のユーザの情報として D 種類の映画のセルに値 1 が加算され，ID が 2 のユーザの情報として当該 1 映画のセルに値 D が加算され，ID が 3 のユーザの情報として当該 3 映画のセルに値 $D/3$ が加算される．ただし，元のプロトコルと同様に，ユーザ ID はハッシュ化された上に冪乗余されているので，ID 管理組織は加算される情報がどのユーザの情報であるのかを知ることはできない．ここで，生成したクロス集計表の各値を D で割ることで，各ユーザからの情報の合計を 1 に正規化することができる．

上記のステップ 3 の処理を一般化すると下記のようになる．商店は各ユーザの ID を D 種類の乱数で冪乗余し，ユーザ毎に D 個の冪乗余値を得る．

(**ケース** 1) ユーザの購入商品数 E が D 以下であり，E が D の約数である場合には，D 個の冪乗余値のうち D/E 個ずつに同じ商品名を付加する．

(**ケース** 2) $E < D$ で，E が D の約数でない場合には，D 個の冪乗余値のうち $\lfloor D/E \rfloor$ 個ずつに同じ商品名を付加する．さらに，E 個の商品から $D \bmod E$ 個をランダムに選択し，残りの冪乗余値に付加する．

(**ケース** 3) $E > D$ の場合には，E 個の商品から D 個をランダムに選択し，D 個の冪乗余値に各々付加する．商品名を付加した D 個の冪乗余値を ID 管理組織に送る．

D を全てのユーザの購入映画数の最小公倍数に設定すると，ケース 1 のみが生じ，クロス集計表に誤差は生じない．全ユーザの購入映画数の最小公倍数が大きいと D が大きくなり，ID 管理組織-商店間プロトコルの計算量および通信量が大きくなる．そこで，D を全ユーザの購入映画数の最小公倍数より小さくすると，ケース 2 およびケース 3 になるためクロス集計表に誤差が生じる．D が最小公倍数より小さくても，ユーザの購入商品数の最大値より十分に大きい場合にはケース 3 は生じず，ケース 2 の誤差も小さいので，誤差は小さい．以上から，D は推薦精度と処理性能のトレードオフを制御するパラメータと考えられる．

6.5.4 属性間の関係の利用

6.4.2 節の方針に沿って，直接扱う属性を限定する一方で，属性間の関係を用いて，他の属性の値を推定する．本実施例では，映画が属性に相当する．そのため，考慮する映画を 1 次〜i 次カテゴリ ($i < 18$) に限定することが，属性を限定することになる．また，次数の異なる映画数の比率を，属性間の関係として利用する．次数の異なる映画数の比率を以下では次数比と呼ぶことにする．次数比の情報自体に匿名加工を施す必要があるが，その感度は小さいことを後述する．

次数比を用いて，ランキングを補正する．高次のカテゴリを持つ映画は少なく，次数比は，1 次と 2 次，3 次だけで大凡 3 : 5 : 2 の割合になっている．各カテゴリ間の独立性のナイーブベイズ仮定が崩れやすい高次の (複合カテゴリの) ランキングを次数比で補正する．

ID 管理組織は，商店が販売した映画のカテゴリをユーザのプロファイルごとに集計して，プロファイルごとに人気のある (よく購入されている) 映画とそのカテゴリをクロス集計表として統計化する．i 次のカテゴリーまでを考慮してクロス集計表を生成することでノイズを低減する．また，属性間の関係についての情報として，感度の低い次数比の統計量を算出し，この次数比にもノイズを付与して商店に開示する．次数比の統計量は，各プロファイルのユーザが購入した映画の次数毎の購入数の比率を正規化したものである．商店は，匿名化されたプロファイル毎クロス集計表およびプロファイル毎次数比を，ID 管理組織から受け取る．そして，訪問者のプロファイルと訪問者が過去に購入した映画のカテゴリに応じて，訪問者が好みそうな映画を推薦する．

訪問者への推薦は，訪問者の携帯端末上で R 件をランキング形式で提示する．その際，次数比の情報を利用して，R 件のうち 1 次の映画数，2 次の映画数，... の個数を調整する．たとえば，次数比は 1 次が 1/3 で 2 次が 2/3 であるとする．3 件の推薦を行う場合に，補正前のランキングは 1 位が 1 次の映画 A，2 位が 2 次の映画 B，3 位が 1 次の映画 C，4 位が 2 次の映画 D，... とすると，順位の高い方から 1 次の映画 1 件 (A)，2 次の映画 2 件 (B，D) を取り出して A，B，D の順に推薦する．

プライバシ保護のために，クロス集計表と次数比の両方にラプラスメカニズムを適用する必要がある．クロス集計表の感度，すなわちサイズは図 6.2 に示したように映画の数 (図ではクロス集計表の横方向．MovieLens では 3,952 本) に比例するが，次数比のサイズ (すなわち感度) は基本カテゴリの数 (図では次数比の表の縦方向．MovieLens では 18 カテゴリ) と同じである．映画を分類するための基本カテゴリの数は，分類対象である映画の数を超えない (3,952 本の映画を分類するために，3,952 種類を超えるカテゴリを用意し

ない) ので，次数比の感度がクロス集計表の感度を超えることはない．システムの安全性を左右する差分プライバシのセキュリティパラメータ (ϵ) を，両者の匿名加工において $\rho\epsilon$ と $(1-\rho)\epsilon$ のように分配する．次数比はクロス集計表よりも感度を小さく抑えられるので，加えるノイズが少なくて済む．正規化を施せば両者とも感度を 1 にできるが，代わりに S/N 比がサイズに比例するので，やはり次数比はクロス集計表よりも加えるノイズが少なくて済む．以降の評価では $\rho = 0.1$ を用いることにする．

6.5.5 スムージングの利用

ラプラスノイズ付加後のクロス集計表に，分散環境対応スムージングを適用する．

6.6 提案方式のプライバシ評価

6.6.1 データの正規化

(1) 匿名加工のプライバシ

差分プライバシに基づく匿名加工のプライバシは，6.2 式のパラメータ ϵ によって決まる．そのため，差分プライバシに基づく従来の匿名加工法と提案法のパラメータ ϵ を同じ値に設定すれば，同じプライバシが達成される．

(2) 秘匿積集合の改変の影響

4.6.1 節で述べた秘匿積集合 [56] を提案法と連結するために，秘匿積集合を 6.5.3 節のように改変するので，その影響を考察する．秘匿積集合は以下のように改変される．

- 4.6.1 節ステップ 3 の修正

  改変前

  商店は商品ごとに乱数 $R_b^{\text{商品 }ID}(1 \leq \text{商品 } ID \leq L)$ を生成し，表 4.6 のマトリクス Y の要素が 1 になっている個所について，$\mathbf{H}(\text{ユーザ } ID)^{R_b^{\text{商品 }ID}}$ を算出する．各々の冪乗余値に，対応する商品 ID を付与し，これらの商品 ID 付き冪乗余値をシャッフルした後，ID 管理組織に送る．

  改変後

  商店は D 個の乱数 $R_b^d(1 \leq d \leq D)$ を生成し，$\mathbf{H}(\text{ユーザ } ID)^{R_b^d}$ を算出する．各々の冪乗余値に，6.5.3 節 (ケース 1〜3) のルールに沿って商品 ID を付与し，これらの商品 ID 付き冪乗余値をシャッフルした後，ID 管理組織に送る．

- 4.6.1 節ステップ 6 の修正

  改変前

  ID 管理組織は，ステップ 4 で算出した 35 個のプロファイル値 ID および商品 ID 付き冪乗余値 $\mathbf{H}(\text{ユーザ } ID)^{R_b^{\text{商品 } ID} R_a^{\text{プロファイル値 } ID}}$ と，ステップ 5 で送られてきた 28 個の冪乗余値 $\mathbf{H}(\text{ユーザ } ID)^{R_a^{\text{プロファイル値 } ID} R_b^{\text{商品 } ID}}$ を照合し，等しければ，当該プロファイル値のユーザが当該商品を購入したとして，クロス集計表の当該プロファイルの値と商品の欄にポイント 1 を加算する．このポイント加算を等しいペア毎に行う．

  改変後

  ID 管理組織は，ステップ 4 で算出した算出したプロファイル値 ID および商品 ID 付き冪乗余値 $\mathbf{H}(\text{ユーザ } ID)^{R_b^d R_a^{\text{プロファイル値 } ID}}$ と，ステップ 5 で送られてきた冪乗余値 $\mathbf{H}(\text{ユーザ } ID)^{R_a^{\text{プロファイル値 } ID} R_b^d}$ を照合し，等しければ，当該プロファイル値のユーザが当該商品を購入したとして，クロス集計表の当該プロファイルの値と商品の欄にポイント 1 を加算する．このポイント加算を等しいペア毎に行う．

4.6.1 節のステップ 6 では，あるプロファイル値のユーザがある商品を購入したとして，クロス集計表の該当欄にポイントを加算する．しかし，冪乗余によってユーザ ID は秘匿されるので，誰が商品を購入したかは秘匿される．また，クロス集計表の複数の欄に加算されたポイントが，同じユーザの購入によるものか，異なるユーザの購入によるものかが秘匿される．その結果，semi-honest な攻撃者は，秘匿積集合の実行を観察しても出力 (クロス集計表) 以外の情報を得ることがない．この秘匿の安全性は，DDH (Decisional Diffie-Hellman) 仮定の安全性に帰着する．上記の改変前と改変後を比較すると，改変後の秘匿積集合の安全性は改変前と同様に DDH 仮説の安全性に帰着することは明らかである．以上から，DDH 仮定の安全性の下で，攻撃者は改変後の秘匿積集合からクロス集計表以外の情報を得ることはできない．そのため，提案法 (正規化を用いた匿名加工法) の達成するプライバシは，秘匿積集合の改変の影響を受けず，従来の安全性と同等である．

6.6.2 属性間の関係の利用

提案法は，各次数の映画の推薦数の割合 (次数比) を用いて，推薦順位を補正する．この次数比は，商店の履歴データから算出するので，それ自体が個人情報を含んでおり，匿名加工を施す必要ある．そこで，提案法では，6.2 式のパラメータ ϵ のうち一定の割合を次

数比の匿名加工に割り当て，残りをクロス集計表の匿名加工に割り当てる．文献 [131] によれば，ϵ を複数の匿名加工に割り当て，加工結果の統計量から最終出力を合成した場合，その最終出力のプライバシは ϵ の合計値によって決まる．そのため，ϵ の合計値を従来法と同じ値に設定すれば，同じプライバシが達成される．

6.6.3 スムージングの利用

5 章で述べた分散環境対応スムージングは，クロス集計表に適用され，クロス集計表の情報だけを用いて実行される．そのため，匿名加工値のクロス集計表に分散環境対応スムージングを適用しても，新たな個人情報が混入することはない．ϵ を同じ値に設定すれば，スムージングを行う提案法は，行わない従来法と同等のプライバシを達成する．

6.7 スムージングの推薦精度に対する影響評価

精度評価には交差検定を用いる．データセットに含まれる各ユーザのデータを，学習データとテストデータに分けて，学習データで推薦に用いる統計量を算出しておく．そして，統計量を用いてテストデータにあるユーザに映画のランキングを提示する．ユーザに推薦する際に，そのユーザが購入した映画を 1 つずつ除いておき，それぞれ除いておいた映画がランキングに含まれる割合 (P@R) で推薦の精度を評価する [130]．交差数は 10 として，3 件以下しか売れていない映画を除いて 3,883 本の映画を実験に用いる．

性別と年代のプロファイルを用いて，完全一致カウントの場合の精度を表 6.9 に示す．表 6.9 において，カテゴリの次数は何次の映画までを考慮したかを示す．4 次以上の映画は殆どないので 3 次までを評価した．Minka のスムージング (以後，MK と表す) はプライバシ保護を無視して Minka のスムージングを適用した場合の推薦精度を，分散環境対応スムージング (以後，PS と表す) は提案する分散環境対応スムージングを適用した場合の推薦精度を，スムージングなし (以後，NS と表す) はスムージングを適用しなかった場合の推薦精度を示す．

推薦数が 10 件の 1 次に着目すると，MK の精度 (4%) が最も良い．次数を 1 次から 2 次に増やすと推薦の手がかりとなる情報量が増えるので，NS や提案手法である PS の精度が 3% から 4% に向上する．MK はデータがスパースになるため，精度が 4% から 1% に低下する[*2]．次数を 2 次から 3 次に増やすと，NS の精度は 4% のままだが PS は 5% に向

[*2] 一般に，識別に用いる分離超平面を学習する (分離超平面を構成する関数のパラメータを推定する) ためには，少なくとも超平面の容量 (次元数に 1 を加えた値の 2 倍) を超えるデータ数が必要とされる

表 6.9　性別と年代のプロファイルを用いて，完全一致カウントの場合の精度

推薦数	カテゴリの次数	スムージングなし (NS)	Minka のスムージング (MK)	分散環境対応スムージング (PS)
10	1 次	3%	4%	3%
	2 次	4%	1%	4%
	3 次	4%	1%	**5%**
100	1 次	14%	16%	14%
	2 次	20%	3%	22%
	3 次	24%	3%	**25%**
1,000	1 次	33%	29%	31%
	2 次	58%	25%	62%
	3 次	68%	25%	**77%**

上する．NS の精度が変わらない理由は，次数を増やすと情報量が増える一方でデータがスパースになっていくためだと考えられる．一方で PS はスパースなデータをスムーズにする効果があるため，精度が向上したのだと考えられる．MK もスムージングの効果は期待されるが，PS よりもパラメータが多いので，より多くのデータが必要であったと考えられる．推薦数が 10 件で推薦を行う場合は，3 次の PS を用いると最も高い精度 (5%) が得られる．

推薦数が 100 件の 1 次に着目すると，MK の精度 (16%) が最も良い．次数を 1 次から 2 次に増やすと NS は 14% から 20% に，PS は 14% から 22% に向上する．MK はデータがスパースになるため精度が 16% から 3% に低下する．次数を 2 次から 3 次に増やすと，NS は 20% から 24% に，PS は 22% から 25% に向上する．MK は 3% のままである．推薦数が 100 件で推薦を行う場合は，3 次の PS を用いると最も高い精度 (25%) が得られる．

推薦数が 1,000 件の 1 次に着目すると，NS の精度 (33%) が最も良い．次数を 1 次から 2 次に増やすと NS は 33% から 58% に，PS は 31% から 62% に向上する．MK はデータがスパースになるため精度が 29% から 25% に低下する．次数を 2 次から 3 次に増やす

[132, 133]．MK は次数が 1 次の場合でも映画毎に 18 個の (すなわち 18 次元空間で) パラメータを推定する必要がある．2 次になると ${}_{18}\mathrm{C}_1 + {}_{18}\mathrm{C}_2 = 171$ 個，3 次では ${}_{18}\mathrm{C}_1 + {}_{18}\mathrm{C}_2 + {}_{18}\mathrm{C}_3 = 987$ 個，18 次では $\sum_{i=1}^{18} {}_{18}\mathrm{C}_i = 2^{18} - 1 = 262,143$ 個ものパラメータを推定しなくてはならないため，データが足りなくなってしまう．一方 PS は映画毎に 1 個のパラメータを推定すれば良い．

と，NS は 58% から 68% に，PS は 62% から 77% に向上する．MK は 25% のままである．推薦数が 1,000 件で推薦を行う場合は，3 次の PS を用いると最も高い精度 (77%) が得られる．

性別と年代のプロファイルを用いて，部分一致カウントの場合の精度を表 6.10 に示す．

表 6.10　性別と年代のプロファイルを用いて，部分一致カウントの場合の精度

推薦数	カテゴリの次数	スムージングなし (NS)	Minka のスムージング (MK)	分散環境対応スムージング (PS)
10	1 次	4%	1%	**5%**
	2 次	4%	1%	**5%**
	3 次	4%	2%	**5%**
100	1 次	26%	17%	26%
	2 次	25%	3%	**27%**
	3 次	26%	3%	**27%**
1,000	1 次	83%	71%	84%
	2 次	81%	25%	**85%**
	3 次	76%	25%	**85%**

いずれの推薦数でも PS の精度が最も良い．NS の精度は PS を下回る．MK は精度が最も悪い．部分一致カウントは高次の情報を低次に集めて低次のデータを密にする．実際には観測していない高次の情報を低次で観測したとみなすことで，MK はパラメータの値が狂って精度が低下したと考えられる．一方，NS はデータが密になったことで精度が向上したと考えられる．NS にスムージングを加える PS は，MK のように次数に応じたパラメータを持たないので精度が低下せずに精度が向上したと考えられる．

性別と年代のプロファイルを用いて，部分一致カウント&正規化ありの場合の精度を表 6.11 に示す．

いずれの推薦数でも PS の精度が最も良い．NS の精度は PS を僅かに下回る．MK は精度が最も悪く，特に 1 次の精度が劣化する．正規化は爆買いユーザなどの特異なデータによる影響を抑制する．データが丸められることで，MK はパラメータの値が狂って精度が低下したと考えられる．シンプルなモデルである NS は正規化によって若干精度が向上し，PS は僅かだが NS から精度を向上させた．

以上の結果から，訪問者に提示する推薦数にかかわらず，クロス集計表を分散環境対応

表 6.11　性別と年代のプロファイルを用いて，部分一致カウント&正規化ありの場合の精度

推薦数	カテゴリの次数	スムージングなし (NS)	Minka のスムージング (MK)	分散環境対応スムージング (PS)
10	1 次	4%	0%	**5%**
	2 次	4%	1%	**5%**
	3 次	4%	1%	**5%**
100	1 次	26%	0%	**27%**
	2 次	26%	3%	**27%**
	3 次	26%	3%	**27%**
1,000	1 次	84%	25%	**85%**
	2 次	84%	25%	**85%**
	3 次	84%	25%	**85%**

スムージングで処理すると推薦精度を高められることがわかった．

6.8　匿名加工の推薦精度に対する影響評価

推薦精度 (すなわちデータの有用性) の評価基準として，6.7 節と同様に，映画がランキングに含まれる割合 (P@R) を用いる．性別と年代のプロファイルを用いて，部分一致カウントの場合の精度を表 6.12 に示す．安全のためにノイズを加えると全ての推薦数において精度が低下した．

性別と年代のプロファイルを用いて，部分一致カウント&正規化ありの場合の精度を表 6.13 に示す．正規化を行うことで感度を抑えてノイズを減らし，表 6.12 よりも精度が向上した．

性別と年代のプロファイルを用いて，部分一致カウント&正規化あり&次数比ありの場合の精度を表 6.14 に示す．次数比を用いることで，表 6.13 よりも次数の高い場合の精度が向上した．

性別と年代のプロファイルを用いて，部分一致カウント&正規化あり&次数比あり&分散環境対応スムージングありの場合の精度を表 6.15 に示す．部分一致カウントと次数比に加えて分散環境対応スムージングも用いることで，表 6.14 よりも精度が向上した．

以上の結果から，推薦数や要求される安全性に関わらず，次数比と部分一致カウントと分散環境対応スムージング (PS) を用いた手法で処理すると，推薦精度を高められるこ

表 6.12　性別と年代のプロファイルを用いて，部分一致カウントの場合の精度

推薦数	カテゴリの次数	要求される安全性			
		無 ($\epsilon = \infty$)	弱 ($\epsilon = 2.0$)	中 ($\epsilon = 1.0$)	強 ($\epsilon = 0.1$)
10	1 次	**4%**	0%	0%	0%
	2 次	**4%**	**1%**	**1%**	**1%**
	3 次	**4%**	**1%**	**1%**	**1%**
100	1 次	**26%**	**3%**	**3%**	**3%**
	2 次	**26%**	**3%**	**3%**	**3%**
	3 次	**26%**	**3%**	**3%**	**3%**
1,000	1 次	**84%**	24%	24%	24%
	2 次	**84%**	**25%**	**25%**	**25%**
	3 次	**84%**	**25%**	**25%**	**25%**

表 6.13　性別と年代のプロファイルを用いて，部分一致カウント&正規化ありの場合の精度

推薦数	カテゴリの次数	要求される安全性			
		無 ($\epsilon = \infty$)	弱 ($\epsilon = 2.0$)	中 ($\epsilon = 1.0$)	強 ($\epsilon = 0.1$)
10	1 次	**4%**	**4%**	**4%**	**3%**
	2 次	**4%**	**4%**	**4%**	1%
	3 次	**4%**	**4%**	**4%**	1%
100	1 次	**26%**	**26%**	**25%**	**16%**
	2 次	**26%**	25%	24%	8%
	3 次	**26%**	23%	20%	5%
1,000	1 次	**84%**	**83%**	**80%**	**53%**
	2 次	**84%**	78%	71%	40%
	3 次	**84%**	70%	60%	33%

とがわかった．以下では，この提案手法 (表 6.15) と従来手法 (表 6.12) の推薦精度をプロファイル毎に比較し，提案手法の特性や適用限界を探る．

要求される安全性が強 ($\epsilon = 0.1$) における，性別と年代のプロファイルを用いた場合の精度の詳細 (推薦数が 10 件の 1 次) を表 6.16 に示す．

14 種類の性別と年代の組み合わせのうち 12 種類において，提案手法の精度は従来手法よりも優れている．提案手法は幅広い性別や年代で推薦精度を高めることができたが，女性の 18 歳未満は 22 件の正解が 21 件に，女性の 56 歳以上は 33 件の正解が 17 件に低下してしまった．女性の 18 歳未満は購入数が 0.9%(574,689 件のうち 5,329 件) と 14 種類

表 6.14　性別と年代のプロファイルを用いて，部分一致カウント&正規化あり&次数比ありの場合の精度

推薦数	カテゴリの次数	要求される安全性			
		無 ($\epsilon = \infty$)	弱 ($\epsilon = 2.0$)	中 ($\epsilon = 1.0$)	強 ($\epsilon = 0.1$)
10	1 次	**4%**	**4%**	**4%**	**3%**
	2 次	**4%**	**4%**	**4%**	1%
	3 次	**4%**	**4%**	3%	1%
100	1 次	**25%**	**25%**	**25%**	**16%**
	2 次	**25%**	24%	23%	9%
	3 次	**25%**	23%	20%	6%
1,000	1 次	**84%**	**83%**	**80%**	**56%**
	2 次	**84%**	78%	72%	45%
	3 次	**84%**	71%	63%	41%

表 6.15　性別と年代のプロファイルを用いて，部分一致カウント&正規化あり&次数比あり&分散環境対応スムージングありの場合の精度

推薦数	カテゴリの次数	要求される安全性			
		無 ($\epsilon = \infty$)	弱 ($\epsilon = 2.0$)	中 ($\epsilon = 1.0$)	強 ($\epsilon = 0.1$)
10	1 次	**4%**	**4%**	**4%**	**3%**
	2 次	**4%**	**4%**	**4%**	2%
	3 次	**4%**	**4%**	**4%**	2%
100	1 次	**26%**	**26%**	**25%**	**17%**
	2 次	**26%**	**26%**	**25%**	15%
	3 次	**26%**	**26%**	**25%**	14%
1,000	1 次	**84%**	**83%**	**81%**	**59%**
	2 次	**84%**	82%	79%	56%
	3 次	**84%**	82%	79%	55%

中最も少なく，女性の 56 歳未満は購入数が 1.1%(574,689 件のうち 6,489 件) と 14 種類中二番目に少なく，推薦の手がかりとなる情報量が足りないことが考えられる．

要求される安全性が強 ($\epsilon = 0.1$) における，職業のプロファイルを用いた場合の精度の詳細 (推薦数が 10 件の 1 次) を表 6.17 に示す．

21 種類の職業の組み合わせのうち 19 種類において，提案手法の精度は従来手法よりも優れている．提案手法は幅広い職業で推薦精度を高めることができたが，職業が “farmer”

表 6.16　要求される安全性が強 ($\epsilon = 0.1$) における，性別と年代のプロファイルを用いた場合の精度の詳細 (推薦数が 10 件の 1 次)

プロファイル		購入数	従来手法 スムージングなし (NS)		提案手法 次数比あり&部分一致カウント& 分散環境対応スムージングあり (PS)	
男性	–18	10,265	33	0%	71	1%
	18–24	76,828	2,174	3%	2,986	4%
	25–34	168,861	4,150	2%	6,376	4%
	35–44	86,821	1,812	2%	2,504	3%
	45–49	34,756	126	0%	411	1%
	50–55	33,193	120	0%	337	1%
	56–	18,537	76	0%	103	1%
女性	–18	5,329	22	0%	21	0%
	18–24	23,888	84	0%	344	1%
	25–34	53,488	640	1%	1,603	3%
	35–44	29,753	107	0%	357	1%
	45–49	14,663	54	0%	81	1%
	50–55	11,818	52	0%	65	1%
	56–	6,489	33	1%	17	0%
	計	574,689	9,483	2%	15,276	3%

の 6 件の正解が 3 件に低下してしまった．“farmer” は購入数が 0.2%(574,689 件のうち 1,417 件) と 21 種類中最も少なく，推薦の手がかりとなる情報量が足りないことが考えられる．また，職業が “retiree” の 40 件の正解が 21 件に低下してしまった．“retiree” の購入数は全体の 1.5%(574,689 件のうち 8,886 件) と推薦の手がかりとなる情報量が決して多くないことに加えて，今はリタイアしているが以前は様々な職業に就いていたユーザが混在している可能性が高いことが考えられる．

表 6.17　要求される安全性が強 ($\epsilon = 0.1$) における，職業のプロファイルを用いた場合の精度の詳細 (推薦数が 10 件の 1 次)

プロファイル	購入数	従来手法 スムージングなし (NS)		提案手法 次数比あり&部分一致カウント& 分散環境対応スムージングあり (PS)	
academic / educator	49,143	378	1%	1,173	2%
artist	28,572	92	0%	246	1%
clerical / admin	19,045	61	0%	106	1%
college / grad student	73,427	1,990	3%	2,741	4%
customer service	11,774	43	0%	59	1%
doctor / health care	22,741	81	0%	172	1%
executive / managerial	61,270	596	1%	1,691	3%
farmer	1,417	6	0%	3	0%
homemaker	6,717	25	0%	29	0%
K–12 student	13,253	48	0%	88	1%
lawyer	12,353	57	0%	58	0%
programmer	34,543	102	0%	600	2%
retiree	8,886	40	0%	21	0%
sales / marketing	28,885	77	0%	375	1%
scientist	14,214	64	0%	67	0%
self-employed	26,586	85	0%	196	1%
technician / engineer	42,338	228	1%	863	2%
tradesman / craftsman	6,546	25	0%	26	0%
unemployed	7,650	19	0%	27	0%
writer	32,996	122	0%	395	1%
other / not specified	72,333	1,170	2%	1,855	3%
計	574,689	5,309	1%	10,791	2%

6.9 まとめ

差分プライバシによる匿名加工は，安全性とデータ劣化のトレードオフを数学的に定式化できるという特徴がある．しかし，この匿名加工は，データに重畳するノイズの大きさが，データの属性数に比例する．そのため，クロス集計表のような多属性の統計値に適用すると，データが大幅に劣化するという問題があった．

そこで，統計値のもとになった個票を想定し，個票の各レコードの値を正規化することで，ノイズの大きさを属性数への比例でなく 1 に抑える手法を提案した．この手法では，ノイズが $\frac{1}{\text{属性数}}$ に低減する一方，信号は平均的には $\frac{1}{\text{属性数}}$ より大きくなるため，S/N 比が向上する．また，一部の属性のみを直接利用することでノイズを低減する手法を提案した．一部の属性しか利用しないことによる情報の劣化を，属性間の関係を利用することで補正した．さらに，差分プライバシの大きなノイズは異常値とみなせることから，スムージングによってデータの劣化を抑止した．

Inter PPR に組み込む形で提案手法を実装し，MovieLens 1M データセットを用いた評価により，推薦精度の低下に関する抑止効果を明らかにした．

第 7 章

Inter PPR の評価

7.1 はじめに

4 章で ID 管理組織-商店間プロトコル，商店-訪問者間プロトコルを提案し，5 章で分散環境対応スムージング，6 章で多属性対応差分プライバシを提案し，これらの個々の部品のプライバシ保護と推薦精度を評価した．本章では，4.3 節で述べた Inter PPR の要件に基づき，これらの部品を結合したシステム全体のプライバシ保護と推薦精度を評価する．また，部品のうち実用時の性能への影響が大きいと考えられる ID 管理組織-商店間プロトコル，商店-訪問者間プロトコルについて，実装に基づいて処理性能を評価する．

7.2 プライバシ保護

4.3.1 節で述べたように，攻撃者は ID 管理組織，商店，訪問者であり，プライバシ要件は，各者の保有する情報を他の 2 者に対して秘匿することである．その際，ID 管理組織と商店は semi-honest，訪問者は malicious とする．また，安全性として，プロトコル自体の安全性とプロトコルの出力の安全性を考慮する．

守りたい情報，各者の行う処理，各者の間で実行するプロトコルを表 7.1 にまとめる．各々のプロトコルおよび処理の安全性については，4.6.1 節，4.6.2 節，5 章，6 章で述べた．表 7.1 に各部品の安全性をまとめる．

表 7.1 に示すように，Inter PPR 全体の安全性は，秘匿積集合プロトコルの安全性，秘匿内積プロトコルの安全性および差分プライバシにおける安全性と推薦精度のトレードオフに帰着する．ただし，以下の 2 か所のセキュリティホールがある．

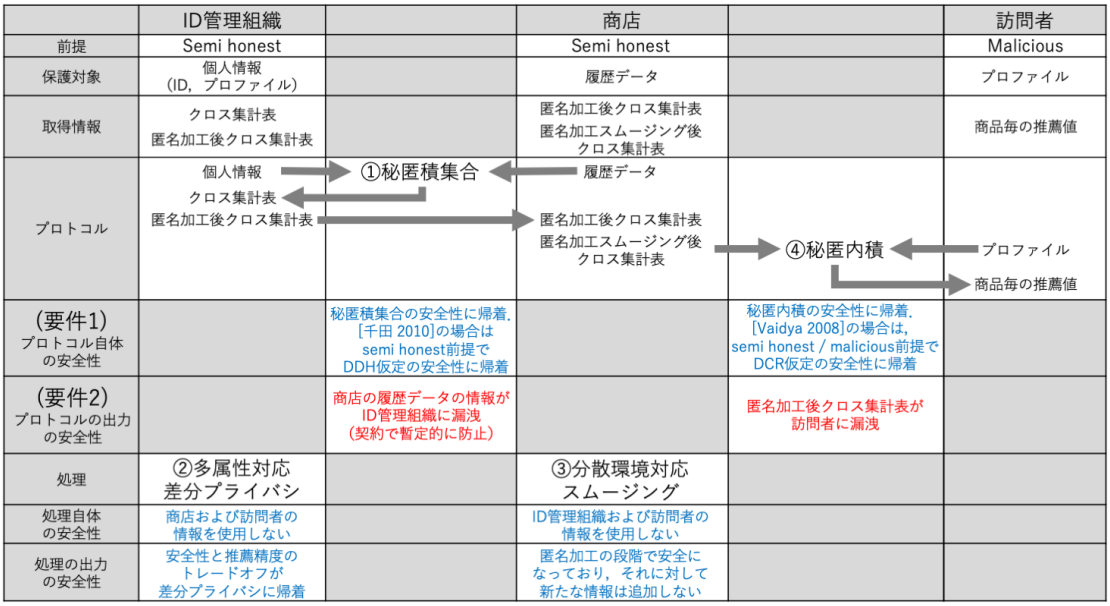


図 7.1　プライバシ保護の評価

(1) ID 管理組織が，秘匿積集合の出力 (クロス集計表) から，商店の保有する履歴情報を推定可能である．

(2) 訪問者が，秘匿内積の出力 (商品毎の推薦値) から，商店の保有するクロス集計表の一部を推定可能である．

4.4 節で述べたように，(1) のセキュリティホールについては実害に至る可能性は小さいが，厳密な安全性を満たすには，秘匿積集合から匿名加工までの一連の処理およびプロトコルを，マルチパーティ計算や準同型暗号の利用によって，秘匿状態で実行する必要がある．この秘匿計算の実現は今後の課題としたい．

(2) のセキュリティホールは，訪問者が都合の良いプロファイルを用いると，匿名加工後のクロス集計表の一部あるいは全部を取得できてしまう．たとえば，女性 20 代の訪問者が自身のプロファイルを女性年齢不詳と偽って商店を訪問すると，匿名加工後のクロス集計表の女性の値が各商品毎の推薦値として訪問者に漏洩してしまう．訪問者が自身のプロファイルを性別不詳 20 代と偽って商店を再度訪問すると，今度は匿名加工後のクロス集計表の 20 代の値が各商品毎の推薦値として訪問者に漏洩してしまう．このような攻撃を訪問者が行っても，商店-訪問者間プロトコルは商店に対して秘匿内積で訪問者のプロファイルを秘匿するので，商店は攻撃を受けていることに気づけない．このような攻撃は

推薦システムの目的上，完全に防止することは困難であるが，漏洩を少なくすることは可能である．たとえば，訪問者側のシステムを難読化しておき，各商品の推薦値を直接開示するのではなく，推薦度トップ 10 の商品のリストのみを開示する方法が考えられる．また，秘匿内積から推薦商品のリスト生成までの一例の処理を秘匿計算で実行することが考えられる．このような手法の実現は今後の課題としたい．

7.3 推薦精度

4.2 節で述べたユースケースに合致するサンプルデータが存在しなかったため，MovieLens 1M データセットを用い，6.7 節，6.8 節にて推薦精度を評価した．その結果をまとめると表 7.1 のようになる．

表 7.1 推薦精度まとめ

アウトプットプライバシ	カウント	分散環境対応スムージング	ラプラスノイズ	正規化	属性間関係	推薦精度 P@10, P@100, P@1,000
-	完全一致	-	-	-	-	4%, 24%, 68%
	完全一致	あり	-	-	-	5%, 25%, 77%
	部分一致	-	-	-	-	4%, 26%, 83%
	部分一致	あり	-	-	-	5%, 27%, 84%
あり	部分一致	-	弱 ($\epsilon = 2.0$)	-	-	1%, 3%, 25%
	部分一致	-	弱 ($\epsilon = 2.0$)	あり	-	4%, 26%, 83%
	部分一致	-	弱 ($\epsilon = 2.0$)	あり	あり	4%, 25%, 83%
	部分一致	あり	弱 ($\epsilon = 2.0$)	あり	あり	4%, 26%, 83%
あり	部分一致	-	中 ($\epsilon = 1.0$)	-	-	1%, 3%, 25%
	部分一致	-	中 ($\epsilon = 1.0$)	あり	-	4%, 25%, 80%
	部分一致	-	中 ($\epsilon = 1.0$)	あり	あり	4%, 25%, 80%
	部分一致	あり	中 ($\epsilon = 1.0$)	あり	あり	4%, 25%, 81%
あり	部分一致	-	強 ($\epsilon = 0.1$)	-	-	1%, 3%, 25%
	部分一致	-	強 ($\epsilon = 0.1$)	あり	-	3%, 16%, 53%
	部分一致	-	強 ($\epsilon = 0.1$)	あり	あり	3%, 16%, 56%
	部分一致	あり	強 ($\epsilon = 0.1$)	あり	あり	3%, 17%, 59%

表 7.1 は，分散環境対応スムージングも差分プライバシも利用しない場合，分散環境対応スムージングのみ利用する場合，差分プライバシのみ利用する場合，分散環境対応スムージングと差分プライバシの両者を利用する場合の推薦精度を示している．差分プライ

バシについては，プライバシの度合いを弱，中，強の 3 段階にしている．また，学習に用いる正例のレコードの数え方として，完全一致と部分一致でのカウントを考慮した．推薦精度は，上位 10 位，100 位，1000 位までの適合率，すなわち Precision@10，100，1,000 によって測定し，百分率で数値化した．差分プライバシを利用しない場合において，分散環境対応スムージングによって推薦精度を向上させることができた．Minka の手法のように多くのパラメータを用いるスムージングの方が高い推薦精度を得られる場合 (表 6.9 に示したように，完全一致カウントで 1 次のカテゴリの P@10 の精度) もあったが，今回の MovieLens 1M データセットを用いた実験においては，提案手法のように少ないパラメータを用いる分散環境対応スムージングの方が高い推薦精度を安定して得ることができた．以上のことから，差分プライバシを利用しない場合に，提案システムは，個人情報と履歴データを直接扱えなくても，スムージングにより推薦精度を向上できることという “(要件 2a) スムージングに対する推薦精度の要件” を満たす．

少ないパラメータでスムージングを行う提案手法や，そもそもスムージングを行わないナイーブベイズでは，次数を高くする (統計が有する情報量を多くする) と推薦精度を高められる．しかし，匿名加工を行うと，有用性と安全性の間にトレードオフが生じる．同じ次数 (統計が有する情報量) において，安全性を高く (ϵ を小さく) すると匿名加工に要するデータの歪みが大きくなるために，推薦精度 (匿名加工後のデータの有用性) は低下する．また，同じ安全性 (同じ ϵ) において，次数 (統計が有する情報量) を高くすると，考慮する属性数が増加するため，匿名加工のラプラスノイズが大きくなるので，推薦精度 (匿名加工後のデータの有用性) は低下する．そこで，正規化によって感度を抑制しつつ，感度の小さい統計量である属性間関係を用いることで，小さなラプラスノイズで匿名加工可能な統計情報を導出して推薦精度を高めた (表 6.14)．以上のことから，提案システムは “(要件 2b) プライバシ保護に対する推薦精度の要件” すなわち匿名加工による推薦精度の低下を抑止できることを満たす．

匿名加工後のデータにはラプラスノイズが付加されているので，ラプラスノイズを異常値とみなしてスムージングを適用したところ，推薦精度の向上を確認できた (表 6.15)．以上のことから，アウトプットプライバシが求められる場合でも，提案システムは，個人情報と履歴データを直接扱えなくても，スムージングにより推薦精度を向上できることという “(要件 2a) スムージングに対する推薦精度の要件” を満たす．

Inter PPR を実用化する場合には，ID 管理組織および商店が有する現実のデータを用いた実証実験を行い，推薦精度を評価する必要がある．その点は今後の課題としたい．

7.4 処理性能

システム全体の処理性能に大きな影響を与えると考えられる ID 管理組織-商店間プロトコルおよび商店-訪問者間プロトコルを実装し，その処理性能を評価した．4.6.1 節で述べたように，ID 管理組織-商店間プロトコルの処理性能は冪乗余の計算にかかる時間が支配的である．たとえば，表 4.11 に示したように，冪乗余の回数は NWL 回にも及ぶため，仮に 1 回の冪剰余にかかる計算時間が数 [ms] であっても 100 年以上かかってしまう．分散環境対応スムージングと匿名加工の処理時間は，ID 管理組織または商店においてオフラインで処理することができ，両者を合わせても数時間であることから全体の処理性能に影響しない．そこで次節では，処理性能を明らかにするために行った実装と，これとは別に推薦精度を明らかにするために行った実装について述べる．

7.4.1 実装

実験は Intel Xeon E5-2697 v4 (2.30GHz, 18 cores) の CPU を 2 台と 512GByte のメモリを搭載した PC で行った．OS は Linux の Ubuntu 18.04 で，Python3 の Anaconda(科学技術計算用パッケージ) を用いた．

処理性能を明らかにするために，ID 管理組織-商店間プロトコルおよび商店-訪問者間プロトコルを以下の 6 つのプログラムで実装した．

worker **プログラム** (60 **行**)
: 引数を解析して実験を開始する．

eval_data **プログラム** (60 **行**)
: データを読み込んで，ID 管理組織-商店間プロトコル，商店での推薦に備えた学習，商店-訪問者間プロトコル，訪問者への推薦の順に実行する．

secure_matching **プログラム** (66 **行**)
: ID 管理組織-商店間プロトコルの自作クラスである．Crypto(暗号) ライブラリの SHA256 関数と randint 関数を利用している．また，gmpy2(任意精度演算) ライブラリの mpz(多倍長整数) 型と next_prime 関数を利用している．

provider **プログラム** (20 **行**)
: ID 管理組織が商店からのリクエストに応える．

shop **プログラム** (76 **行**)
: 商店が ID 管理組織からクロス集計表を手に入れ，推薦に備えて学習する．そして，商店が訪問者からのリクエストに応える．

guest **プログラム** (52 **行**)

訪問者が商店から推薦を受ける．phe(準同型案号) ライブラリの paillier クラスを利用している．

処理性能を左右する暗号関連のライブラリは個別のプログラムに対応させて記載したが，プログラムを簡潔に記述するために math, numpy, itertools などの一般的なライブラリも用いている．なお，ベンチマークは繰り返し計測を行う必要があるため，該当する箇所のみを timeit ライブラリで計測した．

推薦精度を明らかにするために，NS, MK, PS の手法を以下の 4 つのプログラムで実装した．

estimate **プログラム** (134 **行**)

下記の 3 つの手法を動作させるプログラムである．各手法へ渡すデータにラプラスノイズを付加するために，scipy ライブラリの laplace 関数を利用している．

naive_bayes **プログラム** (54 **行**)

ナイーブベイズの手法の実装である．

minka **プログラム** (110 **行**)

Minka の手法の実装である．scipy ライブラリの gamma 関数と digamma 関数を利用している．

secure_smoothing **プログラム** (124 **行**)

提案手法 (分散環境対応スムージング) の実装である．

推薦精度を左右する数学関連のライブラリは個別のプログラムに対応させて記載したが，プログラムを簡潔に記述するために一般的なライブラリも用いている．なお，次数や安全性などの条件を変えながら推薦精度を測定する必要があるため，上記のプログラムおよびデータの入出力を並列で処理できるようにして実験の効率化を図っているが，これについては本質ではないので割愛する．

7.4.2 処理時間

提案方式による処理性能を明らかにするため，計算量の一般式と本論文でのユースケースの想定における処理時間を表 7.2 にまとめる．1 回の冪剰余にかかる計算時間を 3.8[ms] とし[*1]，並列計算は行っていない．

[*1] 訪問者については，携帯端末を使わずに PC でエミュレートしたため，ID 管理組織-商店間プロトコルと商店-訪問者間プロトコルでの 1 回の冪剰余にかかる計算時間は同じとして見積もっている

表 7.2　冪乗余の回数と処理時間

登場人物	Vaidya らの方式 [55]	提案方式
ID 管理組織	NVL (約 2.5×10^6 日)	$NW + MGV$ (約 16 日)
商店	NVL (約 2.5×10^6 日)	$MG + NWL$ (約 1.3×10^5 日)
	-	VL (約 36 分)
訪問者	-	$1 + 2V + L$ (約 38 秒)

Vaidya らの方式 [55] は 3 者間での運用を想定していないので，ID 管理組織-商店間のみを考慮する[*2]．Vaidya らの方式は ID 管理組織と商店でそれぞれ NVL 回の冪乗余が必要となる[*3]．提案方式は NWL 回の冪乗余が支配的である．処理性能において提案方式が Vaidya らの方式よりも優れる理由は，冪乗余の回数を $O(NVL)$ から $O(NWL)$ に減らせるところにある．属性の項目は 2 種類以上の値を取る (1 種類の値しかない属性の項目には意味が無い) 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 Vaidya らの方式より 2 倍以上優れる．

ID 管理組織-商店プロトコルの処理性能について，提案方式は Vaidya らの方式よりも高速であるが, 商店での NWL の冪乗余がボトルネックとなるため[*4]，このままでは T_1(1 ヶ月) 以内の要件を満たすことができない (表 7.2)．ただし，この問題は一般的になりつつあるマルチコア CPU やクラウド環境の利用などにより解決できる．たとえば，商品数が 1 万種類に及ぶチェーン店や大規模な商店であっても，16 コアのマルチコア CPU を 6 個用いて 96 倍の高速化を行うとすると，100 万人の会員を有する ID 管理組織と T_1(1 ヶ月) 以内の要件を満たすことができる．商品数が 1,000 種類程度の中規模の商店であれば 1,000 万人の会員を有する ID 管理組織と，商品数が 100 種類程度の小規模の商店であれば 1 億

[*2] Vaidya らの方式は秘匿内積を用いている．提案方式の商店-訪問者プロトコルも秘匿内積を用いているので，仮に Vaidya らの方式を商店-訪問者プロトコルに適用した場合の冪乗余の回数は，提案方式と同等である．

[*3] 長さ V のベクトルの積集合を求める場合，提案方式の ID 管理組織-商店プロトコルでは，ベクトルに値が入っている要素 (W 個) のみを冪乗余すれば良いが，Vaidya らの方式が用いている秘匿内積では，ベクトルの全ての要素 (V 個) を冪乗余する必要がある．

[*4] ID 管理組織と商店での処理は並行できるので，商店の冪乗余がボトルネックとなる．

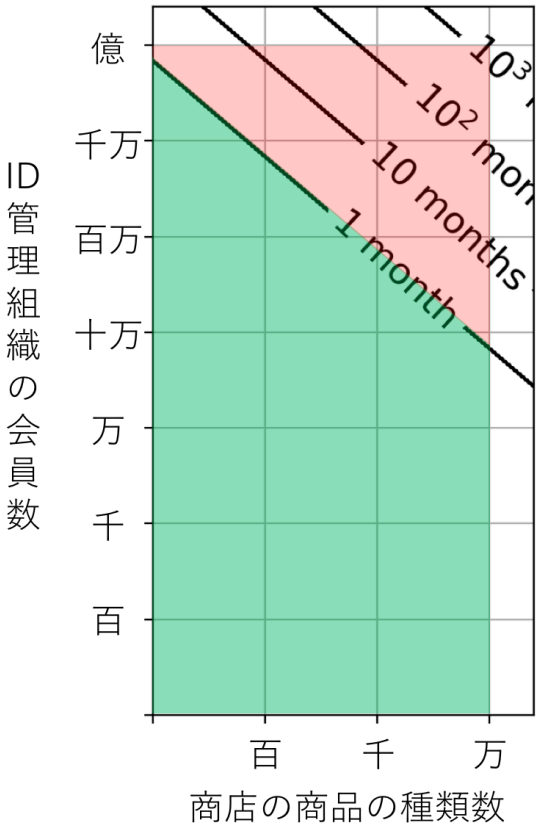


図 7.2　処理性能を満たす条件 (商店での $MG + NWL$ 回の冪乗余にかかる期間．1 人あたりの商品の種類数 $G = 10$, プロファイルの値の種類数 $W = 3$，シングルコアの場合)，緑の領域は条件を満たし，赤の領域は満たさない．

人の会員を有する ID 管理組織と T_1(1 ヶ月) 以内の要件を満たすことができる．以上のことから，提案システムは条件によって T_1(1 ヶ月) 以内という “(要件 3a) ID 管理組織-商店間の処理性能の要件” を満たす．

なお，T_1 を 1 ヶ月から 1 週間に変更する場合は，商店は費用をかけて約 4 倍の性能の CPU に買い換える必要が生じる．逆に T_1 を 1 ヶ月から 3 ヶ月に変更する場合は，商店は 1/3 倍の性能の CPU で足りるため，CPU の買い替えは不要で若干電気代が減少すると考えられる．

商店-訪問者プロトコルの処理性能について，商店と訪問者のいずれでも，このままでは T_2(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事は難しい．この問題は，商店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．商店の PC(16 コ

アの CPU を搭載していると想定) を用いて 16 並列で冪乗余する．さらに，$L = 10^4$ 種類ある商品を 100 種類ずつに 100 分割して，ユーザが 100 種類の商品を眺めている間に次の 100 種類の商品の推薦を計算する．これらの工夫により，商店での処理を 1,600 倍，訪問者での処理を 100 倍高速化する．すると，商店での処理を約 36 分から約 1.4 秒に，訪問者での処理を約 38 秒から約 0.4 秒に短縮でき，両者を合わせても約 1.8 秒なので T_2(5 秒) 以内の要件を満たすことができる．以上のことから，提案システムは T_2(5 秒) 以内という “(要件 3b) 商店-訪問者間の処理性能の要件” を満たす．

なお，ハードウェアの進歩により，ID 管理組織-商店間の月次処理や，商店-訪問者間で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．

7.5 社会実装容易性

ユーザ番号とプロファイルの管理については，ユーザの登録，付番，プロファイル管理を日頃から行っている ID 管理組織を導入し，商品の販売と履歴管理を行う商店と連携することにした．その結果，商店は ID 管理組織と共通のユーザ ID を用いて履歴データを管理できるようになる．ID 管理組織はカード会社や携帯電話会社等であり，Inter PPR に参加する前からユーザ ID およびプロファイルを日々管理している．そのため，Inter PPR の運用にあたって，ユーザ ID の付与やプロファイルの収集を新たに行う必要はない．ユーザ ID およびプロファイルの組織間利用は，秘匿積集合を用いた ID 管理組織-商店間プロトコルによって安全かつ全自動で行われる．ユーザ ID およびプロファイルの情報から生成されたクロス集計表は，匿名加工後に商店に開示される．商店が，クロス集計表から，ユーザ ID およびプロファイルを推定する攻撃については，差分プライバシを用いた匿名加工によって防止する．差分プライバシの安全性は，クロス集計表の有用性すなわち推薦精度とトレードオフの関係にあるが，このトレードオフは差分プライバシのセキュリティパラメータ ϵ によって制御することができる．以上から，Inter PPR は 4.3.4 節で述べた “(要件 4a) ユーザ ID およびプロファイルの安全な利用と負担抑制の要件” を満たす．

多組織間の情報統合への拡張性については，ID 管理組織が，各商店との間で結合処理を行って各々クロス集計表を生成し，これらのクロス集計表を統合．統合したクロス集計表を匿名加工して，各商店に配布することにより解決する．このようにすることで，ID 管理組織と各商店が各々のデータを秘匿しながら，ID 管理組織の個人情報と複数の商店の履歴データを統合したクロス集計表を利用することができる．以上から，Inter PPR は 4.3.4 節で述べた “(要件 4b) 多組織間連携への拡張性の要件” を満たす．

7.6 まとめ

プライバシ保護，推薦精度，処理性能，社会実装容易性の 4 点について，Inter PPR を評価した．Inter PPR のプライバシは，秘匿積集合プロトコルの安全性，秘匿内積プロトコルの安全性および差分プライバシにおける安全性と推薦精度のトレードオフに帰着する．ID 管理者がクロス集計表から商店の情報を推定可能である点，訪問者が商品毎の推薦値から商店の保有するクロス集計表の一部を推定可能である点を明らかにし，秘匿計算やプログラム難読化を用いる今後の課題を明らかにした．

推薦精度については，MovieLens 1M データセットを用い，Inter PPR の推薦と，ユーザの実際の評価値に基づく購入予測との一致度を推薦精度とした．分散環境対応スムージングを用いる場合と用いない場合の推薦精度を比較した結果，推薦数 10，100，1,000 において精度が 1% から 9% 向上した．このことから，MovieLens 1M データセットの場合は，Inter PPR 環境において分散環境対応スムージングの効果を維持できることを明らかにした．また，提案した匿名加工を用いた場合の推薦精度と，従来の匿名加工法を用いた場合の推薦精度を比較した結果，推薦数 10，100，1,000 において精度が 2% から 31% 向上した．さらに，匿名加工に加えて分散環境対応スムージングを適用した結果，匿名加工可能のみの場合に比べて，精度が 1% から 3% 向上した．これらのことから，提案した匿名加工法による精度低下の抑制効果を明らかにすると共に，分散環境対応スムージングが匿名加工との組み合わせ下においても有効であることを明らかにした．

処理性能については，ID 管理組織と商店の間の秘匿積集合プロトコルの計算量が，会員数 × プロファイル項目数 × 商品種類数に比例する点がボトルネックとなる．会員 100 万人 × プロファイル 3 項目 × 商品 1 万種類の場合や，会員 1 億人 × プロファイル 3 項目 × 商品 100 種類の場合は処理時間が 1 ヶ月以内となり実用的であるが，ユースケースの最大規模である会員 1 億人 × プロファイル 3 項目 × 商品 1 万種類の場合は実用的ではない．しかし，50 倍程度の性能不足はハードウェアの進歩や実装の工夫によって短期間に対応可能である．

社会実装容易性については，ユーザ ID の付与およびプロファイルの収集を新たに行う必要はない．また，ID 管理組織が，ID 管理組織-商店間プロトコルを複数の商店との間で各々実行し，複数商店の履歴データを統合したクロス集計表を生成して匿名加工することで，各商店に利用させることができる．

以上の評価から，Inter PPR の実用性を明らかにした．

第 8 章

結論

8.1 まとめ

本論文では，複数の組織が互いの情報を秘匿しながら，情報を統合利用して統計的推薦を行う Inter-Organization Privacy-Preserving Recommender System Based on Secure and Efficient Use of Profiles and Purchase Records (Inter PPR) を提案した．Inter PPR により，推薦に必要な情報を全て保有可能な大組織だけでなく，一部の情報しか保有できない中小組織も統計的推薦を行うことが可能になる．また，ユーザの個人情報 (ユーザ ID とプロファイルの組) や購入履歴を一つの組織が一元管理する必要がないので，プライバシ侵害の可能性を低減することができる．

本論文の 1 章では，ビッグデータを活用した統計的推薦が産業上重要になっていることを述べた．ところが，従来の推薦システムを中小組織が利用する場合，データ数が少ないため推薦精度が低い，個人情報の管理負担が大きいという問題がある．そのため，従来の推薦システムは，必要な情報を全て保有可能な大組織だけが利用可能であり，ユーザの選択肢が狭くなるという問題がある．また，ユーザの個人情報や購入履歴を大組織が一元管理することから，プライバシ侵害の懸念がある．そこで，複数の組織が連携し，各組織およびユーザの情報を秘匿しながら推薦を行うシステムの確立を研究目的とした．

2 章では，先行研究を分析した．従来の統計的推薦技術のうち，コンテンツベース推薦はデータ量への依存度が比較的小さく，ユーザのプロファイルと組み合わせれば中小組織でも高精度の推薦が可能となる．しかし，中小組織にとってプロファイルの管理は負担が大きい．また，中小組織にとってユーザ ID の管理は負担が大きいので，組織毎の ID 管理を前提とした従来の組織間 ID 管理技術は利用できない．プライバシ保護技術には暗号応用と匿名加工がある．一般に暗号応用は処理効率が問題になるが，秘匿積集合および秘匿

内積は，二つのデータベースを秘匿したまま結合する処理およびデータを秘匿したままコンテンツベース推薦を行う処理を効率的に実行可能である．組織間プライバシ保護推薦のための暗号応用も研究されているが，安全性，信頼性，処理性能，匿名加工との連携，IDおよびプロファイル管理の面で実用性が低い．匿名加工については，プライバシ保護とデータ劣化のトレードオフの問題がある．匿名加工技術のうち差分プライバシは，トレードオフを数学的に定式化できるが，個票に適用するとデータ劣化が大きいため，統計量に適用する方が良い．また，従来のスムージング技術のうち Minka の手法は最適性が保証されているが，これを推薦のデータに適用しようとすると，プロファイルと購買履歴の両方が必要になり，大組織しか利用できない．

3 章では，以上の社会的および技術的背景を踏まえ，Inter PPR のシステム構成を提案した．Inter PPR は，ユーザの ID およびプロファイルを管理する ID 管理組織と，小売業を営み購買履歴を管理する商店と，商店で購入する訪問者から成る構成とする．ID 管理組織として，日常的に共通ユーザ ID とプロファイルを管理しているカード会社や携帯電話会社を想定する．ID 管理組織と商店は秘匿内積プロトコルを用いて互いの情報を秘匿しながら統計量であるクロス集計表を生成する．ID 管理組織がクロス集計表に差分プライバシによる匿名化およびスムージングを加えた後，処理後のクロス集計表を商店に送る．商店は秘匿内積プロトコルにより，訪問者との間で互いの情報を秘匿したまま，コンテンツベースの推薦処理を実行する．ID 管理組織が複数の商店との間で各々クロス集計表を生成し，これらのクロス集計表を統合することにより，多組織間の連携を行う．以上の構成において，分散環境で利用可能なスムージングとクロス集計表等の多属性データの劣化を抑止する差分プライバシが存在しないので，これらを新たに開発する必要がある．

4 章では，3 章で述べたシステムの実現方法を検討した．ユースケースに沿って，プライバシ，推薦精度，処理性能の要件を明らかにした．プライバシでは，3 者の各々が，他の 2 者に対して自己の保有する情報を秘匿すること，その秘匿ではプロトコル自体の安全性とプロトコルの出力の安全性を考慮することとした．また，ID 管理組織と商店は semi honest，訪問者は malicious であるとした．推薦精度については，スムージングにより推薦精度を向上し，匿名加工による推薦精度の低下を抑止することで，全ての情報を利用可能な大組織と同等あるいはそれに近い推薦精度であることとした．処理性能については，ID 管理組織が 10^8 ユーザの 57 属性値を保有し，商店が 10^5 ユーザの 10^4 商品に関する購買履歴を保有し，訪問者が自己に関する 57 属性値を保有する状態を前提とし，ID 管理組織と商店の間のクロス集計生成および商店と訪問者の間の推薦値算出を実用上支障のない時間で実行できることとした．これらの要件を満たすデータ表現と処理フローを設計した．また，秘匿積集合および秘匿内積プロトコルを用いて，ID 管理組織-商店間プロトコ

ルおよび商店-訪問者間プロトコルの詳細を設計し，計算量の理論値を示した．

5 章では，従来のスムージング方式のうち最適性の保証されたディリクレスムージング (Minka の手法) を分析し，個人情報と購買履歴の両方が必要になるため，大組織しか利用できないことを明らかにした．また，データが少ない場合には，多数のパラメータを最適化できずスムージングの効果を発揮できない．この分析に基づき，ID 管理組織-商店間プロトコルの結果生成されるクロス集計表に直接適用可能でパラメータの少ないスムージング手法を提案した．提案手法は，クロス集計表以外の情報を必要としないので，クロス集計表を入手する ID 管理組織または商店が単独で実行可能であり，また，少ないデータにも有効である．

6 章では，差分プライバシによる匿名加工の Inter PPR への適用を検討した．従来の差分プライバシの重畳するノイズの大きさは，対象データの属性の総数に比例する．Inter PPR における匿名加工の対象はクロス集計表であるが，その属性数は商品の総種類数に比例するため，差分プライバシのノイズが非常に大きくなり，推薦精度が大幅に低下する．そこで，ユーザの購入した商品の種類数を正規化することで，ノイズの大きさを商品の総種類数ではなく 1 に抑えた．また，クロス集計する商品を一部の種類に限定し，他の種類の商品の集計値は商品間の関係から推定することで，ユーザの購入した商品の種類数すなわちノイズの大きさを抑えた．さらに，5 章で提案した分散環境対応スムージングを用いて，差分プライバシのノイズを平滑化した．MovieLens 1M データセットを用いて推薦精度を評価し，分散環境対応スムージングによる精度向上の効果，提案した匿名加工法による精度低下の抑止効果を確認した．

7 章では，プライバシ保護，推薦精度，処理性能，社会実装容易性の観点から，Inter PPR を評価した．Inter PPR のプライバシは，秘匿積集合プロトコルの安全性，秘匿内積プロトコルの安全性および差分プライバシにおける安全性と推薦精度のトレードオフに帰着する．ID 管理組織がクロス集計表から商店の情報を推定可能であり，訪問者が商品毎の推薦値から商店の保有するクロス集計表を推定可能であるため，その対策が今後の課題となる．推薦精度については 6 章の評価を総括し，提案した分散環境対応スムージング，多属性対応差分プライバシおよび両者の組合せの効果を明らかにした．処理性能については，ID 管理組織と商店の間の秘匿積集合プロトコルの計算量がボトルネックとなる．マルチコア CPU を用いて 96 倍の高速化を行うとすると，会員 1 億人，プロファイル 57 項目，商品 100 種類の場合や会員 100 万人，プロファイル 57 項目，商品 1 万種類の場合は処理時間が 1 ヶ月以内となり，実用的であるが，ユースケースの最大規模である会員 1 億人，プロファイル 57 項目，商品 1 万種類の場合は，さらに 50 倍の高速化が必要となるため実用的ではない．しかし，50 倍程度の性能不足はハードウェアの進歩や実装の工夫に

よって短期間に対応可能である．これらの評価から，Inter PPR の実用性を明らかにした．以上の検討を通じて，本論文では，推薦システムおよびプライバシ保護技術の発展に以下の貢献を達成した．

(1) 複数の組織が，各組織およびユーザの情報を秘匿しながら，情報を統合して推薦する課題に対して，ID 管理組織を導入し，推薦，暗号応用，匿名加工，スムージングの新たな組み合わせによるシステムを提案し，その実用性を明らかにした．本システムは，ユーザ ID とプロファイルの管理負担の軽減，シームレスなプライバシ保護，推薦精度の劣化防止，処理性能，多組織間への拡張性を同時に満たし，従来のプライバシ保護推薦技術よりも実用性が高い．
(2) 分散環境に適用可能で，データが少ない場合にも有効なスムージング手法を明らかにした．
(3) 多属性のデータにおいて有用性の低下を抑止可能な差分プライバシの手法を明らかにした．

8.2 今後の課題

Inter PPR の実用化に向けて，以下の課題があげられる．

(1) 訪問者の端末上で Inter PPR のクライアントシステムを実装し，エンドユーザにとっての使い勝手を評価する．
(2) 上記のクライアントシステムを難読化し，訪問者が推薦値からクロス集計表を推定する攻撃を防止する．
(3) ID 管理組織と商店の間の秘匿積集合プロトコルの計算時間を実装の工夫によって短縮する．
(4) 現実の ID 管理組織および商店を募り，現実のデータを用いて実証実験を行う．

また，プライバシ保護技術として，以下の課題があげられる．

(1) 6 章で述べた多属性対応差分プライバシ方式の有効性を，MovieLens 1M データセット以外のデータセットを用いて評価し，手法を一般化する．
(2) プレーヤ間の情報統合から匿名加工までの一連の処理を秘匿計算によって効率的に実施する方式を明らかにし，組織間の情報の統合利用において，シームレスなプライバシ保護を達成する．

謝辞

本研究を遂行し，学位論文としてまとめるにあたり，主任指導教員として終始多大なるご指導とご教示をいただいた吉浦裕教授，および副指導教員としてご指導とご教示をいただいた太田和夫教授，﨑山一男教授に心より感謝の意を表します．博士論文の審査委員として，ご指導いただいた坂本真樹教授，大坐畠智准教授に深く感謝申し上げます．様々なご指導ご助言をいただきました三木哲也特任教授，中嶋信生特任教授，市野将嗣准教授，吉浦研究室と市野研究室の皆様に深く感謝申し上げます．組織間でのプライバシ保護研究についてご指導ご助言をいただきました NTT ドコモ先進技術研究所 寺田雅之主任研究員，NTT セキュアプラットフォーム研究所 千田浩司主任研究員に深く感謝申し上げます．情報システムに関するご助言および学位取得を激励くださいましたドコモ・システムズ 西川清二代表取締役社長に深く感謝申し上げます．ご指導ご助言および社会人として働きながらの学位取得をご支援くださいました NTT ドコモ 中村寛取締役常務執行役員 (R&D イノベーション本部長)，NTT ドコモ先進技術研究所 滝田亘所長，梅田成視前所長 (現在，日本無線 研究所副所長)，浅井孝浩主幹研究員，池田大造主幹研究員，太田賢主幹研究員，岡島一郎主幹研究員，川上博主幹研究員，鈴木恭宜主幹研究員，檜山聡主幹研究員，山田暁主幹研究員，高畑実主任研究員，永田聡主任研究員，藤林暁主任研究員ならびに職場の皆様に深く感謝申し上げます．最後に，働きながらの学位取得を温かく見守り，辛抱強く支えてくれた妻 恵美，両親に深い感謝の意を表して謝辞といたします．

参考文献

[1] DOMO. Data never sleeps 5.0. *available from < https://www.domo.com/learn/data-never-sleeps-5*>, 2017.

[2] 森川博之. ビッグデータの活用に関するアドホックグループの検討状況. 情報通信審議会 ICT 基本戦略ボード, 2012.

[3] U.S. Securities and Exchange Commission. SEC filings & forms. *available from <https://www.sec.gov/edgar.shtml*>, accessed 2018-1-12.

[4] M. Balabanović and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, Vol. 40, No. 3, pp. 66–72, 1997.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. *CSCW*, pp. 175–186, 1994.

[6] G. Linden, B. Smith, and J. York. Amazon.com recommendations. *IEEE Internet Computing*, pp. 76–80, 2003.

[7] C. Buckley and G. Salton. Optimization of relevance feedback weights. *SIGIR*, pp. 351–357, 1995.

[8] 神嶌敏弘. 推薦システムのアルゴリズム. *available from <http://www.kamishima.net/archive/recsysdoc.pdf*>, accessed 2018-01-02.

[9] Y. Lindell and B. Pinkas. Privacy preserving datamining. *CRYPTO*, pp. 36–54, 2000.

[10] R. Agrawal and R. Srikant. Privacy-preserving data mining. *ACM SIGMOD*, pp. 439–450, 2000.

[11] G.-R. Xue, C. Lin, Q. Yang, W. Xi, H.-J. Zeng, Y. Yu, and Z. Chen. Scalable collaborative filtering using cluster-based smoothing. *SIGIR*, pp. 114–121, 2005.

[12] D. Zhang, J. Cao, J. Zhou, M. Guo, and V. Raychoudhury. An efficient collaborative filtering approach using smoothing and fusing. *ICPP*, pp. 558–565, 2009.

[13] I. Lee. Framework for smoothing-based collaborative filtering recommender system.

ACMSE, pp. 363–364, 2012.

[14] R. Zhang, Y. Zhou, L. Li, and C. Zou. A rule-based recommendation for personalization in social networks. *APSCC*, 2014.

[15] 藤沼貴士. 「紙オムツとビールが一緒に買われる」をどのように発見するか. *IS magazine*, Vol. 11, pp. 90–92, 2016.

[16] R. Zhang, Q. Liu, Chun-Gui, J.-X. Wei, and Huiyi-Ma. Collaborative filtering for recommender systems. *CBD*, pp. 301–308, 2014.

[17] M. J. Pazzani and D. Billsus. Content-based recommendation systems. *In P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.): The Adaptive Web*, Vol. LNCS 4321, pp. 325–341, 2007.

[18] G. Ninaus, F. Reinfrank, M. Stettinger, and A. Felfernig. Content-based recommendation techniques for requirements engineering. *AIRE*, pp. 27–34, 2014.

[19] X. Yang, Y. Guo, and Y. Liu. Bayesian-inference based recommendation in online social networks. *KDE*, Vol. 24, No. 4, pp. 642–651, 2013.

[20] L. Nguyen. A new approach for collaborative filtering based on bayesian network inference. *IC3K*, 2015.

[21] 独立行政法人情報処理推進機構. アイデンティティ管理技術解説. 2013.

[22] 下江達二. アイデンティティ管理関連技術の進展と変遷. 人工知能学会誌, Vol. 24, No. 4, pp. 504–511, 2009.

[23] E. Chen, Y. Pei, S. Chen, Y. Tian, R. Kotcher, and P. Tague. Oauth demystified for mobile application developers. *CCS*, pp. 892–903, 2014.

[24] R. Yang, W. C. Lau, and T. Liu. Signing into one billion mobile app accounts effortlessly with oauth2.0. *blackhat Europe*, 2016.

[25] 一般社団法人全国銀行協会. オープン API のあり方に関する検討会報告書-オープン・イノベーションの活性化に向けて. オープン API のあり方に関する検討会, 2017.

[26] 中村啓佑. Oauth2.0 に対する脅威と対策：金融オープン API の一段の有効活用に向けて. 金融研究, Vol. 37, No. 3, 2018.

[27] A. C. Yao. Protocols for secure computations. *SFCS*, pp. 160–164, 1982.

[28] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game or a completeness theorem for protocols with honest majority. *STOC*, pp. 218–229, 1987.

[29] O. Goldreich. *Foundations of Cryptography*, Vol. 1 of *Basic Tools*. Cambridge University Press, 2001.

[30] M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. *STOC*, pp. 1–10, 1988.

[31] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools for privacy preserving distributed data mining. *SIGKDD Explorations*, Vol. 4, No. 2, pp. 28–34, 2002.

[32] B. Yang and H. Nakagawa. Computation of ratios of secure summations in multi-party privacy-preserving latent dirichlet allocation. *PAKDD 2010*, pp. 21–24, 2010.

[33] M. Aliasgari, M. Blanton, Y. Zhang, and A. Steele. Secure computation on floating point numbers. *NDSS*, pp. 24–27, 2013.

[34] R. Cramer, I. B. Damgård, and J. B. Nielsen. *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, 2015.

[35] S.-K. Hong, H. Kim, S. Lee, and Y.-S. Moon. Secure multiparty computation of chi-square test statistics and contingency coefficients. *BigDataSecurity*, pp. 53–57, 2017.

[36] O. Catrina. Round-efficient protocols for secure multiparty fixed-point arithmetic. *COMM*, pp. 431–436, 2018.

[37] D. Malkhi, N. Nisan, B. Pinkas, and Y. Sella. Fairplay – a secure two-party computation system. *SSYM*, pp. 3–20, 2004.

[38] C. Gentry. Fully homomorphic encryption using ideal lattices. *STOC*, pp. 169–178, 2009.

[39] B. Chen and N. Zhao. Fully homomorphic encryption application in cloud computing. *ICCWAMTIP*, pp. 471–474, 2014.

[40] M. Yasuda, T. Shimoyama, J. Kogure, K. Yokoyama, and T. Koshiba. Practical packing method in somewhat homomorphic encryption. *DPM*, 2013.

[41] J.-S. Coron, T. Lepoint, and M. Tibouchi. Scale-invariant fully homomorphic encryption over the integers. *PKC*, 2014.

[42] P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. *EUROCRYPT*, Vol. 1592, pp. 223–238, 1999.

[43] R. Cramer, I. Damgård, and J. B. Nielsen. Multiparty computation from threshold homomorphic encryption. *EUROCRYPT*, pp. 280–300, 2000.

[44] M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *TKDE*, Vol. 16, No. 9, pp. 1026–1037, 2004.

[45] R. Agrawal, A. Evfimievski, and R. Srikant. Information sharing across private databases. *SIGMOD*, pp. 86–97, 2003.

[46] L. Kissner and D. Song. Privacy-preserving set operations. *CRYPTO*, pp. 241–257, 2005.

[47] K. Hu and W. Zhang. mPSI: Many-to-one private set intersection. *CSCloud*, pp. 187–192, 2017.

[48] G. Ateniese, E. De Cristofaro, and G. Tsudik. (if) size matters: Size-hiding private set intersection. *PKC*, Vol. 6571, pp. 156–173, 2011.

[49] E. De Cristofaro, J. Kim, and G. Tsudik. Linear-complexity private set intersection protocols secure in malicious model. *ASIACRYPT*, pp. 213–231, 2010.

[50] B. Goethals, S. Laur, H. Lipmaa, and T. Mielikäinen. On private scalar product computation for privacy-preserving data mining. *ICISC*, pp. 104–120, 2004.

[51] 佐久間淳, 小林重信. プライバシを保護した内積比較プロトコルの提案. 情報処理学会研究報告コンピュータセキュリティ, Vol. 2006, No. 81, pp. 257–264, 2006.

[52] 菊池浩明. ランダムプロジェクションを用いた秘匿内積プロトコル次元数削減. コンピュータセキュリティシンポジウム, pp. 891–898, 2014.

[53] 千田浩司, 五十嵐大, 濱田浩気, 高橋克巳. 匿名等結合プロトコルとその応用. 暗号と情報セキュリティシンポジウム, 2011.

[54] 桐淵直人, 五十嵐大, 諸橋玄武, 濱田浩気. 属性情報と履歴情報の秘匿統合分析に向けた秘密計算による高速な等結合アルゴリズムとその実装. コンピュータセキュリティシンポジウム, pp. 1072–1078, 2016.

[55] J. Vaidya, M. Kantarcioglu, and C. Clifton. Privacy-preserving naive bayes classification. *The VLDB Journal*, Vol. 17, No. 4, pp. 879–898, 2008.

[56] 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. コンピュータセキュリティシンポジウム, pp. 567–572, 2010.

[57] C. C. Aggarwal and P. S. Yu. *Privacy-Preserving Data Mining Models and Algorithms*, Vol. 34. Springer, 2008.

[58] I. Ioannidis, A. Grama, and M. Atallah. A secure protocol for computing dot-products in clustered and distributed environments. *ICPP*, pp. 379–384, 2002.

[59] M. J. Freedman, K. Nissim, and B. Pinkas. Efficient private matching and set intersection. *EUROCRYPT*, pp. 1–19, 2004.

[60] J. Canny. Collaborative filtering with privacy. *SP*, pp. 45–57, 2002.

[61] J. Canny. Collaborative filtering with privacy via factor analysis. *SIGIR*, pp. 238–245, 2002.

[62] 多田美奈子, 菊池浩明. アイテム間類似度に基づくプライバシ保護協調フィルタリン

グの提案. 情報処理学会論文誌, Vol. 51, No. 9, pp. 1554–1562, 2010.

[63] J. Zhan, C.-L. Hsieh, I-C. Wang, T.-S. Hsu, C.-J. Liau, and D.-W. Wang. Privacy-preserving collaborative recommender systems. *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 40, No. 4, pp. 472–476, 2010.

[64] A. Friedman, B. P. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky. Privacy aspects of recommender systems. In *Recommender Systems Handbook*, chapter 19, pp. 649–688. Springer, 2015.

[65] A. Jeckmans, Q. Tang, and P. Hartel. Privacy-preserving collaborative filtering based on horizontally partitioned dataset. *CTS*, pp. 439–446, 2012.

[66] F. McSherry and I. Mironov. Differentially private recommender systems, building privacy into the netflix prize contenders. *SIGKDD*, pp. 627–636, 2009.

[67] V. Nikolaenko, S. Ioannidis, U. Weinsberg, M. Joye, N. Taft, and D. Boneh. Privacy-preserving matrix factorization. *ACM CCS*, pp. 801–812, 2013.

[68] 佐久間淳. スペクトル差分プライバシーに基づくプライバシー保護推薦アルゴリズム. 人工知能学会 第 26 回全国大会, 2012.

[69] A. Basu, J. Vaidya, H. Kikuchi, and T. Dimitrakos. Privacy-preserving collaborative filtering for the cloud. *CloudCom*, pp. 223–230, 2011.

[70] E. Aïmeur, G. Brassard, J. M. Fernandez, and F. S. M. Onana. ALAMBIC: a privacy-preserving recommender system for electronic commerce. *IJIS*, Vol. 7, No. 5, pp. 307–334, 2008.

[71] R. Cissée and S. Albayrak. An agent-based approach for privacy-preserving recommender systems. *IFAAMAS*, pp. 319–326, 2007.

[72] L. Sweeney. k-anonymity: A model for protecting privacy. *IJUFKS*, Vol. 10, No. 5, pp. 557–570, 2002.

[73] N. Ammar, Z. Malik, B. Medjahed, and M. Alodib. K-anonymity based approach for privacy-preserving web service selection. *ICWS*, pp. 281–288, 2015.

[74] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *TKDD*, Vol. 1, No. 1, 2007.

[75] K. Oishi, Y. Tahara, Y. Sei, and A. Ohsuga. Proposal of l-diversity algorithm considering distance between sensitive attribute values. *SSCI*, 2017.

[76] N. Li, T. Li, and S. Venkatasubramanian. t-closeness; privacy beyond k-anonymity and l-diversity. *ICDE*, pp. 106–115, 2007.

[77] S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. *IEEE*

Trans. on Data Privacy, Vol. 7, No. 2, pp. 99–129, 2014.

[78] M. E. Nergiz, M. Atzori, and C. W. Clifton. Hiding the presence of individuals from shared databases. *SIGMOD*, pp. 665–676, 2007.

[79] P. Kooiman, L. Willenborg, and J. Gouweleeuw. PRAM: A method for disclosure limitation of microdata. *Research paper no. 9705, Statistics Netherlands*, 1997.

[80] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *TKDE*, Vol. 22, No. 11, 2010.

[81] H. Polat and W. Du. Privacy-preserving collaborative filtering using randomized perturbation techniques. *ICDM*, 2003.

[82] R. Parameswaran and D. M. Blough. Privacy preserving collaborative filtering using data obfuscation. *GRC*, pp. 380–386, 2007.

[83] S. Renckes, H. Polat, and Y. Oysal. A new hybrid recommendation algorithm with privacy. *Expert Systems*, Vol. 29, No. 1, pp. 39–55, 2012.

[84] R. Benedetti and L. Franconi. Statistical and technological solutions for controlled data dissemination. *NTTS*, Vol. 1, pp. 225–232, 1988.

[85] 瀧淳弘. 集計表におけるセル秘匿問題とその研究動向. 統計数理, Vol. 51, No. 2, pp. 337–350, 2003.

[86] C. Dwork. Differential privacy. *ICALP*, Vol. 4052, pp. 1–12, 2006.

[87] N. Phan, X. Wu, H. Hu, and D. Dou. Adaptive laplace mechanism: Differential privacy preservation in deep learning. *ICDM*, pp. 385–394, 2017.

[88] C. Skinner. Statistical disclosure control for survey data. *Handbook of Statistics*, Vol. 29, pp. 381–396, 2009.

[89] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. A John Wiley & Sons, 2012.

[90] D. Kifer and B. Lin. Towards an axiomatization of statistical privacy and utility. *PODS*, pp. 147–158, 2010.

[91] I. H. Witten and T. C. Bell. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. on Information Theory*, Vol. 37, No. 4, pp. 1085–1094, 1991.

[92] T. P. Minka. Estimating a dirichlet distribution. *available from <http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf>*, accessed 2014-11-28.

[93] 株式会社セブン・カードサービス. 電子マネー ナナコ nanaco. *available from <https://*

www.nanaco-net.jp>, accessed 2018-1-28.

[94] 株式会社セブン・カードサービス. セブン＆アイの電子マネー『nanaco』はおかげさまで 10 周年！ ニュースリリース, 2017.

[95] 株式会社 セブン&アイホールディングス. 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？ *available from <http://www.sej.co.jp/products/trivia/trivia_09.html*>, accessed 2014.

[96] 日本銀行. 決済システムリポート 2012-2013. *BOJ Reports & Research Papers*, 2013.

[97] 谷本啓, 本橋洋介. モデルのライフサイクルを考慮した大量予測モデル管理手法の検討. 人工知能学会 第 29 回全国大会, 2015.

[98] C. E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, Vol. 30, No. 1, pp. 50–64, 1951.

[99] G. J. Lidstone. Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *FA*, Vol. 8, No. 6, pp. 182–192, 1920.

[100] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrica*, Vol. 40, No. 3–4, pp. 237–264, 1953.

[101] A. Orlitsky and A. T. Suresh. Competitive distribution estimation: Why is good-turing good. *NIPS*, pp. 2143–2151, 2015.

[102] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modeling. *Computer, Speech, and Language*, Vol. 8, pp. 1–38, 1994.

[103] M. Song and C. D. Yoo. Multimodal representation: Kneser-ney smoothing/skip-gram based neural language model. *ICIP*, pp. 2281–2285, 2016.

[104] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. *SIGIR*, Vol. 51, No. 2, pp. 268–276, 2001.

[105] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS*, Vol. 22, No. 2, pp. 179–214, 2004.

[106] M. Smucker and J. Allan. An investigation of dirichlet prior smoothing's performance advantage. *CIIR Technical Report*, Vol. IR-548, , 2006.

[107] 正田備也, 高須淳宏, 安達淳. 混合ディリクレ分布を用いた文書分類の精度について. 情報処理学会論文誌, Vol. 48, pp. 14–26, 2007.

[108] Y. Han, J. Jiao, and T. Weissman. Does dirichlet prior smoothing solve the shannon entropy estimation problem? *ISIT*, pp. 1367–1371, 2015.

[109] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *TCC*, pp. 265–284, 2006.

[110] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, Vol. 9, No. 3–4, pp. 211–407, 2014.

[111] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. *STOC*, pp. 75–84, 2007.

[112] X. Xiao and Y. Tao. Output perturbation with query relaxation. *VLDB*, Vol. 1, No. 1, pp. 857–869, 2008.

[113] N. Mohamme, R. Chen, B. C. M. Fung, and P. S. Yu. Differentially private data release for data mining. *KDD*, 2011.

[114] P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler. GUPT: Privacy preserving data analysis made easy. *SIGMOD*, pp. 349–360, 2012.

[115] 五十嵐大, 高橋克巳. 注目のプライバシー Differential Privacy. コンピュータソフトウェア, Vol. 29, No. 4, pp. 40–49, 2012.

[116] C. Dwork and A. Smith. Differential privacy for statistics: What we know and what we want to learn. *JPC*, Vol. 1, No. 2, pp. 135–154, 2009.

[117] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. *EUROCRYPT*, Vol. 4004, pp. 486–503, 2006.

[118] S. P. Kasiviswanathan and A. D. Smith. A note on differential privacy: Defining resistance to arbitrary side information. *IACR*, 2008.

[119] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. *ASIACCS*, pp. 32–33, 2012.

[120] F. Liu. Generalized gaussian mechanism for differential privacy. *TKDE*, 2018.

[121] C. Dwork. Differential privacy in new settings. *SODA*, pp. 174–183, 2010.

[122] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. *STOC*, pp. 715–724, 2010.

[123] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias. Differentially private event sequences over infinite streams. *VLDB Endowment*, pp. 1155–1166, 2014.

[124] G. Cormode, M. Procopiuc, D. Srivastava, and T. T. L. Tran. Differentially private publication of sparse data. *ICDT*, 2011.

[125] N. Shlomo, L. Antal, and M. Elliot. Measuring disclosure risk and data utility for flexible table generators. *JOS*, Vol. 31, No. 2, pp. 305–324, 2015.

[126] S. Zhou, K. Ligett, and L. Wasserman. Differential privacy with compression. *ISIT*, pp. 2718–2722, 2009.

[127] K. Chaudhuri, A. D. Sarwate, and K. Sinha. A near-optimal algorithm for differentially-private principal components. *JMLR*, 2013.

[128] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: Optimal bounds for privacy-preserving principal component analysis. *STOC*, pp. 11–20, 2014.

[129] GroupLens. MovieLens 1M Dataset. *available from <http://grouplens.org/datasets/movielens/1m>*, 2003.

[130] N. Craswell. Precision at n. *Encyclopedia of Database Systems*, pp. 2127–2128, 2009.

[131] F. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. *Communications of the ACM*, Vol. 53, No. 9, pp. 89–97, 2010.

[132] N. J. Nilsson. *Learning Machines*. McGraw-Hill, 1965.

[133] 石井健一郎, 前田英作, 上田修功, 村瀬洋. わかりやすいパターン認識. オーム社, 1998.

付録

関連論文の印刷公表の方法および時期

学術論文

1. **全著者名**: **山口 高康**, 寺田 雅之
   論文題目: 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
   印刷公表の方法および時期: 情報処理学会論文誌, Vol. 56, No. 9, pp. 1754-1769, Sep. 2015.

査読付き国際会議発表論文

1. **全著者名**: **Takayasu Yamaguchi**, Hiroshi Yoshiura
   論文題目: Inter-Organization Privacy-Preserving Recommender System Using Large-Scale Distributed Data
   印刷公表の方法および時期: Proceedings of 14th International Conference on Business and Information, pp. 188-204, Jul. 2017.
2. **全著者名**: **Takayasu Yamaguchi**, Hiroshi Yoshiura
   論文題目: A Strategy for Mitigating Information Loss in Anonymization by Using Nature of Personal Data
   印刷公表の方法および時期: Proceedings of 11th International Workshop on Informatics, pp. 255-260, Sep. 2017.

国内口頭発表

1. **全著者名**: **山口 高康**, 寺田 雅之
   論文題目: セキュアマッチングとナイーブベイズ識別器を用いたプライバシー保護リコメンド方式
   印刷公表の方法および時期: 情報処理学会 第 51 回コンピュータセキュリティ研究会報告, pp. 1-6, Dec. 2010.
2. **全著者名**: 千田 浩司, 寺田 雅之, **山口 高康**, 五十嵐 大, 濱田 浩気, 高橋 克巳
   論文題目: 統計的開示制御を考慮したセキュアマッチングプロトコル
   印刷公表の方法および時期: 情報処理学会 第 52 回コンピュータセキュリティ研究会報告, pp. 1-6, Mar. 2011.

その他の研究業績

学術論文

1. **全著者名**: 寺田 雅之, 鈴木 亮平, **山口 高康**, 本郷 節之
   論文題目: 大規模集計データへの差分プライバシの適用
   印刷公表の方法および時期: 情報処理学会論文誌, Vol. 56, No. 9, pp. 1801-1816, Sep. 2015.
2. **全著者名**: 寺田 雅之, **山口 高康**, 本郷 節之
   論文題目: 匿名化個票開示への差分プライバシの適用
   印刷公表の方法および時期: 情報処理学会論文誌, Vol. 56, No. 9, pp. 1483-1500, Sep. 2017.
3. **全著者名**: Hiroaki Kikuchi, **Takayasu Yamaguchi**, Koki Hamada, Yuji Yamaoka, Hidenobu Oguri, Jun Sakuma
   論文題目: Study on Record Linkage of Anonymizied Data
   印刷公表の方法および時期: IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E101-A, No. 1, pp. 19-28, Jan. 2018.

査読付き国際会議発表論文

1. **全著者名**: Masayuki Terada, **Takayasu Yamaguchi**, Sadayuki Hongo
   論文題目: Providing Secure Integrity in Peer-to-Peer Oriented Mobile Environments
   印刷公表の方法および時期: Proceedings of 25th IEEE Workshops of International Conference on Advanced Information Networking and Applications, pp. 120-126, Mar. 2011.

査読無し国際会議発表論文

1. **全著者名**: **Takayasu Yamaguchi**, Shinji Sugawara, Tetsuya Miki
   論文題目: Relational Media System with Pointer for Network Lectures
   印刷公表の方法および時期: Proceedings of 7th Asia-Pacific Conference on Communications, pp. 169-172, Sep. 2001.

国内口頭発表

1. **全著者名**: **山口 高康**, 菅原 真司, 三木 哲也
   論文題目: リレーショナルメディアシステムのポインタと投影画像の位置合わせに関する研究
   印刷公表の方法および時期: 電子情報通信学会 ソサイエティ大会, p. 284, Sep. 2000.
2. **全著者名**: **山口 高康**, 菅原 真司, 三木 哲也
   論文題目: リレーショナルメディアシステムのポインタと投影画像の位置あわせに関する研究
   印刷公表の方法および時期: 電子情報通信学会 電子ディスプレイ研究会, Vol. 100, No. 605, pp. 7-12, Jan. 2001.
3. **全著者名**: **山口 高康**, 高畑 実, 本郷 節之
   論文題目: 位置情報を利用した情報ハンドリング技術に関する考察
   印刷公表の方法および時期: 情報処理学会 モバイルコンピューティングとユビキタス通信研究会, Vol. 49, No. 21, pp. 101-106, May 2002.
4. **全著者名**: **山口 高康**, 高畑 実, 本郷 節之
   論文題目: モバイルカメラを用いた視覚型情報検索技術に関する研究
   印刷公表の方法および時期: 画像電子学会 第 10 回 VMA 研究会, pp. 1-8, Jan. 2003.
5. **全著者名**: **山口 高康**, 高畑 実, 本郷 節之
   論文題目: モバイルカメラでの看板の撮影をトリガにした情報検索
   印刷公表の方法および時期: 電子情報通信学会 パターン認識・メディア理解研究会, Vol. 103, No. 515, pp. 19-24, Dec. 2003.
6. **全著者名**: 松岡 保静, **山口 高康**, 萩野 浩明, 金野 晃, 吉川 貴
   論文題目: ゼロ知識対話証明方式におけるメッセージ完全性保証機能の拡張
   印刷公表の方法および時期: 電子情報通信学会 ワイドバンドシステム研究会, Vol. 103, No. 715, pp. 161-165, Mar. 2004.
7. **全著者名**: **山口 高康**, 青野 博, 本郷 節之
   論文題目: モバイルカメラで撮影した看板画像の特徴量に関する考察
   印刷公表の方法および時期: 電子情報通信学会 パターン認識・メディア理解研究会, Vol. 104, No. 448, pp. 1-6, Dec. 2004.

8. **全著者名**: **山口 高康**, 青野 博, 本郷 節之
   論文題目: モバイルカメラで撮影した看板画像の学習・判別に関する考察
   印刷公表の方法および時期: 電子情報通信学会 パターン認識・メディア理解研究会, Vol. 104, No. 448, pp. 7-12, Dec. 2004.
9. **全著者名**: **山口 高康**, 青野 博, 本郷 節之, 松浦 幹太
   論文題目: 分類された情報セキュリティ対策に依存する脅威発生率を導入したリスクアセスメントモデル
   印刷公表の方法および時期: 情報処理学会 コンピュータセキュリティ研究会, Vol. 43, No. 33, pp. 7-12, May 2006.
10. **全著者名**: 千田 浩司, 寺田 雅之, **山口 高康**, 五十嵐 大, 濱田 浩気
    論文題目: セキュアマッチングを用いた組織間クロス分析
    印刷公表の方法および時期: コンピュータセキュリティシンポジウム, pp. 567-572, Oct. 2010.
11. **全著者名**: 齋藤 祐也, 森岡 康史, 佐野 洋介, 小泉 大輔, **山口 高康**, 寺田 雅之, 萩原 淳一郎
    論文題目: 重み付き k-Nearest Neighbor 法を用いたセクタ勢力範囲推定法の検討
    印刷公表の方法および時期: 電子情報通信学会 総合大会, pp. 50, Mar. 2012.
12. **全著者名**: 寺田 雅之, 鈴木 亮平, **山口 高康**, 本郷 節之
    論文題目: 大規模集計データへの差分プライバシーの適用
    印刷公表の方法および時期: コンピュータセキュリティシンポジウム, pp. 899-908, Oct. 2014.
13. **全著者名**: 菊池 浩明, **山口 高康**, 濱田 浩気, 山岡 裕司, 小栗 秀暢, 佐久間 淳
    論文題目: 匿名加工・再識別コンテスト Ice & Fire の設計
    印刷公表の方法および時期: コンピュータセキュリティシンポジウム, pp. 363-370, Oct. 2015.
14. **全著者名**: 森尾 淳, 牧村 和彦, **山口 高康**, 池田 大造, 西野 仁, 藤岡 啓太郎, 今井 龍一
    論文題目: 東京都市圏におけるモバイル空間統計とパーソントリップ調査の比較分析 –都市交通分野への適用に向けて–
    印刷公表の方法および時期: 土木計画学研究発表会, Vol. 52, Nov. 2015.
15. **全著者名**: 吉田 純土, 森尾 淳, 中野 敦, **山口 高康**, 池田 大造, 藤岡 啓太郎, 今井 龍一
    論文題目: 都市交通分野における携帯電話基地局データとパーソントリップ調査の組合せ分析に関する研究
    印刷公表の方法および時期: 土木計画学研究発表会, Vol. 53, May 2016.

16. **全著者名**: 寺田 雅之, **山口 高康**, 本郷 節之
    論文題目: 匿名化個票への差分プライバシー基準の適用に関する一考察
    印刷公表の方法および時期: 情報処理学会 コンピュータセキュリティ研究会, Vol. 73, No. 26, pp. 1-8, May 2016.
17. **全著者名**: **山口 高康**, 寺田 雅之, 吉浦 裕
    論文題目: 差分プライバシに基づく一括開示と対話開示のデータ有用性の評価
    印刷公表の方法および時期: 情報処理学会 コンピュータセキュリティ研究会, Vol. 74, No. 32, pp. 1-8, Jul. 2016.
18. **全著者名**: **山口 高康**, 寺田 雅之, 吉浦 裕
    論文題目: 差分プライバシに基づく一括開示と対話開示のデータ有用性の評価 – 多属性に関する考察–
    印刷公表の方法および時期: コンピュータセキュリティシンポジウム, pp. 1191-1198, Oct. 2016.
19. **全著者名**: 菊池 浩明, 小栗 秀暢, 野島 良, 濱田 浩気, 村上 隆夫, 山岡 裕司, **山口 高康**, 渡辺 知恵美
    論文題目: PWSCUP: 履歴データを安全に匿名加工せよ
    印刷公表の方法および時期: コンピュータセキュリティシンポジウム, pp. 271-278, Oct. 2016.
20. **全著者名**: 寺田 雅之, **山口 高康**, 本郷 節之
    論文題目: 高次元大規模データへの差分プライバシー適用のための最適精緻化法
    印刷公表の方法および時期: 暗号と情報セキュリティシンポジウム, pp. 1-8, Jan. 2017.

解説等

1. **全著者名**: **山口 高康**, 青野 博, 本郷 節之
   論文題目: モバイルカメラで情報を検索する対象判別技術
   印刷公表の方法および時期: NTT DOCOMO テクニカル・ジャーナル，Vol. 13, No. 3，pp. 6-10，2005.
2. **全著者名**: 大薮 勇輝，寺田 雅之，**山口 高康**，岩澤 俊弥，萩原 淳一郎，小泉 大輔
   論文題目: モバイル空間統計の信頼性評価
   印刷公表の方法および時期: NTT DOCOMO テクニカル・ジャーナル，Vol. 20, No.3，pp. 17-23，2012.

特許

1. 特願 2002-052044
   無線通信システム，無線通信方法，情報管理装置，情報管理方法，無線通信端末，無線通信端末制御方法，及びプログラム
   山口 高康，高畑 実，本郷 節之
   2002/2/27
2. 特願 2002-142030
   画像判別装置，及び画像判別方法
   山口 高康，高畑 実，本郷 節之
   2002/5/16
3. 特願 2002-142040
   サーバ装置，携帯端末，情報提供システム，情報提供方法，及び情報取得方法
   山口 高康，高畑 実，本郷 節之
   2002/5/16
4. 特願 2003-002517
   画像学習装置及びその学習方法
   山口 高康，高畑 実，本郷 節之
   2003/1/8
5. 特願 2003-092674
   電子メール管理装置，電子メール管理方法及び電子メール管理用プログラム
   山口 高康，藤田 将成，本郷 節之
   2003/3/28
6. 特願 2003-096089
   画像管理装置，画像管理方法及び画像管理用プログラム
   山口 高康，本郷 節之
   2003/3/31
7. 特願 2003-208486
   被検証装置，検証装置，被検証方法及び検証方法
   山口 高康，吉川 貴，金野 晃，松岡 保静，萩野 浩明
   2003/8/22
8. 特願 2003-403530

画像処理装置及び画像処理方法
山口 高康，本郷 節之
2003/12/2 (特許 4769416, 2011/6/24 登録)

9. 特願 2004-050489
画像処理装置及び画像処理方法
山口 高康，本郷 節之
2004/2/25 (特許 4741804, 2011/5/13 登録)

10. 特願 2004-056706
宛先推定装置及び宛先推定方法
山口 高康，本郷 節之，稲村 雄
2004/3/1

11. 特願 2004-307286
関連情報提供装置および関連情報提供方法
山口 高康，服部 篤人，青野 博
2004/10/21 (特許 4551179, 2010/7/16 登録)

12. 特願 2004-326994
画像処理装置及び画像処理方法
山口 高康，青野 博，本郷 節之
2004/11/10 (特許 4664047, 2011/1/14 登録)

13. 特願 2004-347153
共同購入実現装置及び共同購入実現方法
塚田 千佳子，**山口 高康**
2004/11/30 (特許 4391399, 2009/10/16 登録)

14. 特願 2006-133147
プログラム
山口 高康，青野 博，本郷 節之
2006/5/11

15. 特願 2006-281986
コンテンツ再生装置及びコンテンツ再生プログラム
山口 高康，寺田 雅之，本郷 節之
2006/10/16

16. 特願 2007-019970
コンテンツ利用装置及びコンテンツ利用方法

寺田 雅之，**山口 高康**，石原 武，本郷 節之
2007/1/30

17. 特願 2007-037106
耐タンパーデバイス，時刻情報提供装置及び情報状態制御システム並びに情報状態制御方法
杉尾 信行，野秋 浩三，**山口 高康**，大崎 憲嗣，青野 博，本郷 節之
2007/2/16 (特許 5101126, 2012/10/5 登録)

18. 特願 2007-124791
通信端末，送信制御システム，送信制御プログラム，及び送信制御方法
杉尾 信行，**山口 高康**，西 康裕，山田 恵，大崎 憲嗣，本郷 節之
2007/5/9 (特許 4996968, 2012/5/18 登録)

19. 特願 2007-128400
リコメンド装置及びリコメンド方法
山口 高康，杉尾 信行，丸山 ちひろ，大崎 憲嗣，本郷 節之
2007/5/14 (特許 4522430, 2010/6/4 登録)

20. 特願 2007-133222
判別装置及び判別方法
山口 高康，寺田 雅之，大崎 憲嗣，本郷 節之
2007/5/18

21. 特願 2007-230581
コンテンツ提供システム，コンテンツ提供方法及び通信端末
石原 武，**山口 高康**，伊東 秀昭，寺田 雅之，本郷 節之
2007/9/5

22. 特願 2007-286177
情報提供システム及び情報提供方法
山口 高康，江夏 俊輔，久保川 祐加，萩野 浩明
2007/11/2 (特許 5319909, 2013/7/19 登録)

23. 特願 2007-313614
情報提供サーバ，情報提供システム及び情報提供方法
山口 高康，江夏 俊輔，萩野 浩明，久保川 祐加
2007/12/4

24. 特願 2007-313615
情報提供サーバ，情報提供システム及び情報提供方法

山口 高康，江夏 俊輔，萩野 浩明，久保川 祐加
2007/12/4

25. 特願 2008-135882
通信端末，通信制御装置，通信ネットワーク，及び通信方法
石原 武，**山口 高康**，寺田 雅之
2008/5/23

26. 特願 2008-139870
情報端末，情報提供方法及び情報提供プログラム
吉村 健，秋永 和計，**山口 高康**
2008/5/28

27. 特願 2008-139880
情報端末，情報提供方法及び情報提供プログラム
吉村 健，**山口 高康**，鳥居 大祐
2008/5/28

28. 特願 2010-103734
機械学習方法および機械学習システム
鳥居 大祐，**山口 高康**，栄藤 稔
2010/4/28 (特許 5143182, 2012/11/30 登録)

29. 特願 2010-149340
需要予測装置及び需要予測方法
山口 高康，金野 晃，秋永 和計，中山 雄大，栄藤 稔，村瀬 淳
2010/6/30 (特許 5603678, 2014/8/29 登録)

30. 特願 2010-166071
ネットワーク評価支援装置およびネットワーク評価支援方法
中山 雄大，金野 晃，秋永 和計，**山口 高康**
2010/7/23 (特許 5411815, 2013/11/15 登録)

31. 特願 2010-269600
リコメンドシステム及びリコメンド方法
山口 高康，寺田 雅之
2010/12/2

32. 特願 JP2012/053194
端末数推計装置および端末数推計方法
寺田 雅之，**山口 高康**，岡島 一郎

2012/2/10

33. 特願 JP2012/053195
エリア範囲推定装置およびエリア範囲推定方法
山口 高康，寺田 雅之，萩原 淳一郎，岡島 一郎
2012/2/10

34. 特願 2013-028215
交通量算出装置，交通量算出方法
山口 高康，寺田 雅之
2013/2/15 (特許 5634544, 2014/10/24 登録)

35. 特願 2012-556947
エリア範囲推定装置およびエリア範囲推定方法
山口 高康，寺田 雅之，萩原 淳一郎，岡島 一郎
2013/2/25 (特許 5425319, 2013/12/6 登録)

36. 特願 12744253.1
端末数推計装置および端末数推計方法
寺田 雅之，**山口 高康**，岡島 一郎
2013/3/6

37. 特願 12744432.1
エリア範囲推定装置およびエリア範囲推定方法
山口 高康，寺田 雅之，萩原 淳一郎，岡島 一郎
2013/3/7

38. 特願 13/824562
エリア範囲推定装置およびエリア範囲推定方法
山口 高康，寺田 雅之，萩原 淳一郎，岡島 一郎
2013/3/18

39. 特願 13/825916
端末数推計装置および端末数推計方法
寺田 雅之，**山口 高康**，岡島 一郎
2013/3/25 (特許 8849307, 2014/9/30 登録)

40. 特願 2012-081872
リコメンド支援方法，リコメンド支援装置及びプログラム
山口 高康，寺田 雅之
2012/3/30

41. 特願 201280004227.2
エリア範囲推定装置およびエリア範囲推定方法
山口 高康，寺田 雅之，萩原 淳一郎，岡島 一郎
2013/6/19

42. 特願 2012-556946
端末数推計装置および端末数推計方法
寺田 雅之，**山口 高康**，岡島 一郎
2013/7/17 (特許 5613267, 2014/9/12 登録)

43. 特願 201280008348.4
端末数推計装置および端末数推計方法
寺田 雅之，**山口 高康**，岡島 一郎
2013/8/9

44. 特願 2013-218400
滞留目的推定装置および滞留目的推定方法
大薮 勇輝，**山口 高康**
2013/10/21 (特許 6175346, 2017/7/14 登録)

45. 特願 2013-220984
滞留位置推定装置および滞留位置推定方法
大薮 勇輝，**山口 高康**
2013/10/24 (特許 6169471, 2017/7/7 登録)

46. 特願 2015-093446
情報処理装置
青柳 禎矩，**山口 高康**
2015/4/30

47. 特願 2016-096209
プライバシー保護装置
寺田 雅之，**山口 高康**
2016/5/12

48. 特願 2016-217850
拠点推定装置
青柳 禎矩，**山口 高康**
2016/11/8

表彰

1. 情報処理学会 MBL 研究会 優秀論文賞
   “位置情報を利用した情報ハンドリング技術に関する考察”
   山口 高康, 高畑 実, 本郷 節之
   2003 年 7 月
2. NTT ドコモ社長表彰
   “モバイルカメラを使った対象判別技術の研究開発”
   山口 高康, 他 8 名
   2005 年 10 月
3. NTT ドコモ社長表彰
   “i モード検索アルゴリズムの開発と商用化実現”
   山口 高康, 他 5 名
   2010 年 10 月
4. NTT ドコモ社長表彰
   “モバイル空間統計による社会貢献アピール及び CEATEC AWARD グランプリ受賞に関する功績”
   山口 高康, 他 9 名
   2012 年 3 月
5. 情報処理学会 喜安記念業績賞
   “社会・産業の発展に寄与する新たな人口統計情報「モバイル空間統計」の実用化”
   山口 高康, 小林 基成, 鈴木 俊博
   2015 年 6 月
6. 情報処理学会 論文賞
   “大規模集計データへの差分プライバシの適用”
   寺田 雅之, 鈴木 亮平, **山口 高康**, 本郷 節之
   2016 年 6 月