3I2-OS-20-1

検索行動におけるプライバシ保護

Privacy preservation in information retrieval

荒井　ひろみ[*1]　清水　佳奈[*2]　浜田　道昭[*3]　津田　宏治[*2]　広川　貴次[*2]
Hiromi Arai　Kana Shimizu　Michiaki Hamada　Koji Tsuda　Takatsugu Hirokawa

佐久間　淳[*1]　浅井　潔[*2*3]
Jun Sakuma　Kiyoshi Asai

[*1]筑波大学 システム情報工学研究系
Department of Computer Science, University of Tsukuba

[*2]産業総合研究所 生命情報工学研究センター
Computational Biology Research Center, National Institue of Advanced Science and Technology

[*3]東京大学大学院新領域創成科学研究科
Graduate School of Frontier Science, The University of Tokyo

Privacy preserving data mining (PPDM), technical issues about analysis of private data, have emerged rapidly. Information retrieval among entities, such as different companies, in privacy preserving way is one of the major problem. In this paper, we introduce overview of current PPDM, espetially cryptographic techniques for secure distributed computation. As an application example, our system for searching chemical compound libraries using cryptographic techniques is presented.

1. はじめに

現在の情報化社会において，望まれない情報漏えいを防ぎつつ，情報共有の利益を得る仕組みの要請が高まっている．数多くの実社会情報の電子化と世界中に張り巡らされた通信ネットワークによってビジネスや個々人の生活は多くの利便性を享受している．例えば，検索サービス，通販，SNS，位置情報サービスなどおよびそれらの連携による個人情報の収集，蓄積はサービスのパーソナライゼーションや個人間のネットワーキングの活発化などの利益をもたらした．一方これらのサービスはサービスプロバイダや第三者による個人情報漏えいの可能性をはらんでいる．近年この危険性が広く認識されてきており，それに伴いサービスプロバイダによるユーザーのプライバシ保護の必要性が高まっている．

また従来から存在する個々の持つ情報の秘匿性が高い問題も再び注目を集めている．投票や企業間の交渉などの問題において，投票内容や企業の詳細な内情は秘密情報である．現在のデータの大規模化やビジネスの高速化，効率化の要請に伴いこのような問題についてもオンラインでの解決や電子的な高速処理への要請が出てきているといえる．

上記に取り上げたような問題は，分散した秘密の情報の利用として記述することができる．企業や個人などの情報を所有する主体を情報所有者，彼らの持つ漏えいが望まれない情報を秘密情報と呼ぶことにする．

分散した情報の秘匿性を考慮しなくてよい場合，これらは収集して情報検索やデータマイニングといった情報処理技術が用いて利活用することができる．例えばある事業者が小売店の販売履歴を収集して顧客の行動データを分析し，サービスの向上や新規顧客の開拓を行うことができる．この場合小売店の所有する情報は事業者に対して秘匿する必要がないため，それらを収集して利用することができる．

異なる情報所有者に分散されている情報を収集して利用したい場合，それらが秘密情報であり共有できない場合が想定される．例えば，通販業者の持つ顧客の購買情報と鉄道会社の持つ個人の移動履歴情報を統合したデータマイニングや複数の病院のカルテ情報を収集した治療効果分析などが想定される．

上記に示したように，秘密情報保護と情報共有の利益は大雑把にトレードオフの関係にある．これを解消するための，秘密情報の漏えいを防ぎつつそれらを利活用する技術提案が近年なされてきている．それらはプライバシ保護データマイニング（Privacy Preserving Data Mining, PPDM）と総称される．PPDM 技術を用いることによって情報処理の過程や公開される結果結果からの情報漏えいをコントロールできる．

連絡先: 荒井ひろみ（現　理化学研究所生命情報基盤研究部門基礎科学特別研究員），hiromi@base.riken.jp

2. プライバシ保護データマイニング技術

プライバシ保護データマイニングは大まかに以下の 3 つに分類できる．

- プライバシ保護データ出版
- 計算過程におけるプライバシ保護（狭義の PPDM）
- 出力プライバシの保護

プライバシ保護データ出版は収集した個人情報を公開する際に値の丸めや削除によって個人の秘密情報を守る技術である．国勢調査結果の公開などが応用例に挙げられる．代表的な研究には k 匿名化 [Sweeney 02]，l 多様性 [Machanavajjhala 07] などが挙げられる．

計算過程におけるプライバシ保護は情報所有者が秘密情報を公開することなく，それらの情報を集めてデータマイニングを行った場合と同等もしくは近い有用な結果を得る方法である．”個々人の位置情報は秘匿したままでの訪問地推薦”，”患

者の情報を秘匿した遺伝子と疾患の因果関係の分析 " などが想定される応用例として挙げられる．これは一見不可能に思われるが，暗号や摂動などのテクニックを用いた分散計算を行うことにより達成される．研究例として 2 者間のプライバシを保護する暗号を用いた決定木学習 [Lindell 02]，摂動を用いた複数人の情報を用いた決定木学習 [Agrawal 00] などが挙げられる．

出力プライバシの保護は，情報を分析した結果から，そこに含まれる秘密情報を推測されないようにデータを改変する技術である．" SNS での友人推薦において推薦情報から他人の秘密の友人関係の推測を防ぐ "，" 購買履歴からデータマイニングした頻出購買パターンから個人の購買履歴の推測を防ぐ " などが想定される応用例として挙げられる．何かしらの情報を出力した場合にすべての入力情報を守ることは不可能であるため，あらかじめ設定されたある基準をを満たすための方法を提示している．代表的な研究には出力にノイズを入れることにより入力の閾値以上の不確定性を保障する differential privacy[Dwork 06]，出力から推測できる入力の情報量を評価するプライバシ計量 [Agrawal 00] などがある．

本稿では計算過程におけるプライバシ保護を扱う．特に検索行動におけるプライバシを題材として現存の秘匿検索技術の紹介および課題，暗号を用いたプライバシ保護計算の要素技術の紹介および筆者らが提案する化合物データベースに対する秘匿検索プロトコルの概要を紹介する．

3. 情報検索におけるプライバシ

情報検索は，データベースを持つサーバおよび検索を行うクライアントのやりとりとして記述できる．このような検索問題において，サーバおよびクライアントどちらかもしくは両方の情報が秘密情報である場合が考えられる．例えば，クライアントが企業や研究者でありデータベースが特許や学術情報の場合，クライアントの検索質問は未発表の製品開発や研究トピックに関する情報を含むため秘密性が高いと考えられる．また，検索対象のデータベースの情報が知的財産などの理由から秘密である場合も想定される．しかし，検索を行うことの利得としてクライアントは望むデータを手に入れることができ，またサーバは検索サービスを提供することができる．このような背景からクライアント，サーバの秘密情報の保護を技術的に実現する問い合わせ方式の研究がおこなわれてきた．

情報検索において検索を行うクライアントのプライバシを守る技術としてこれまでに暗号分野で Private Information Retrieval (PIR) が研究されてきた．計算量からの漏えいを防ぐ方法が [Chor 95] をはじめとして提案されている．trivial な方法としてはデータベースをクライアントローカルにコピーする方法が考えられるが，それよりも通信量を抑えることを目的としている．

また，サーバ，クライアント双方のプライバシを守る方法として紛失通信 [Rabin 81] がある．紛失送信はクライアントの問い合わせ内容を守るとともにデータベースはクライアントへの回答以外は秘匿できる．この方式は計算量や通信量を要す．紛失送信を任意の関数計算に拡張した方式として秘匿関数計算 [Yao 86] がある．

本稿 5 章で示すプライバシ保護化合物検索法は準同型性公開鍵暗号を用いる点で上記と方式が異なっている．データベースがクライアントが望んだ情報をどれくらい持っているかを問い合わせに問題を限定しているが，代わりにこれらの既存研究よりも通信と計算量を低く抑えている．

4. 秘匿分散計算

秘匿分散計算とは，秘密情報を持った複数の計算主体がうまくやり取りをすることで互いの入力を秘匿しながら出力を得る方法である．例として金持ちプロトコルを上げる．金持ち A と金持ち B が自分の資産総額 m_A，m_B を秘密にしたままどちらの資産が大きいか比較したいとする．この場合，比較関数を $f(\cdot)$ として A, B が協力して，m_A，m_B を開示することなく，それらを入力とする関数の値 $f(m_A, m_B)$ を A，B への出力として与える秘匿分散計算を行えばよい．

秘匿分散計算では暗号やかく乱を用いて，計算主体間でやりとりするメッセージから入力情報が得られないようにしつつ目的の計算を達成するプロトコルや計算方式を設計する．本章ではそのための要素技術を 2 つ紹介する．

4.1 加法準同型性公開鍵暗号

加法準同型性公開鍵暗号を用いると「鍵管理による安全な暗号文通信」「暗号化したままの演算」が可能である．

公開鍵暗号方式は鍵生成，暗号化，復号からなる．鍵生成において公開鍵 pk，対応する秘密鍵 sk が作られる．公開鍵 pk によって平文と呼ばれる整数値を暗号化できる．暗号文は対応する秘密鍵 sk を持つ者しか復号できない．暗号文もまた整数値である．また，暗号系は確率暗号とする．確率暗号を用いると，一つの整数値に対して対応する暗号文は大量にあり，ある暗号化においてそのうち一つが確率的に選択される．よって，同じ平文に同じ公開鍵を用いても暗号化の度に暗号文が異なるため，暗号文から平文を推測することはできない．以上の性質より，公開鍵暗号を用いた暗号文通信では，鍵管理によって暗号文を解読できる人間を制限できる．

平文を $m \in \mathbf{Z}_N$(N はセキュリティパラメータ) と書くことにする．セキュリティパラメータは鍵生成時に用いられ，平文のとる空間を指定する．$c = \mathrm{Enc}_{\mathsf{pk}}(m; l)$ は m の暗号文を，$m = \mathrm{Dec}_{\mathsf{sk}}(\mathrm{c})$ はその復号を表す．l は確率暗号に必要な乱数である．

加法準同型性公開鍵暗号では下式が成立する．

$$\mathrm{Enc}_{\mathsf{pk}}(m_1 + m_2; l) = \mathrm{Enc}_{\mathsf{pk}}(m_1; l_1) \cdot \mathrm{Enc}_{\mathsf{pk}}(m_2; l_2)$$

すなわち暗号文同士の積が平文の和の暗号文に対応し，それを応用してある整数値との積も計算できる．以降乱数 l は略記する．

単純な応用例を示す．n 人の集団 $A_n = \{a_1, a_2, \ldots, a_n\}$ において，それぞれの財布の中身 m_i の総計を出すことを考える．情報漏えいや盗聴を行わない鍵の管理者 a_0 を用いる．A_n の各人が計算主体となり，以下の手続きで計算を行う．

1. 鍵の管理者 a_0 が公開鍵 pk と秘密鍵 sk を生成し pk を公開する．
2. 各 a_i はそれぞれの持つ金額 m_i を pk を用いて暗号文 $c_i = \mathrm{Enc}_{\mathsf{pk}}(m_i)$ を得る．
3. a_1 が a_2 に暗号文 $c(1) = c_1$ を送る．
4. $i = 2$ から $i = n-1$ の順に，a_i は a_{i-1} から暗号文 $c(i-1)$ を受け取り，a_{i+1} に $c(i) = c(i-1) \times c_i$ を送る．
5. a_n は a_{n-1} から暗号文 $c(n-1)$ を受け取り，a_0 に $c(n) = c_0 \times, \ldots, \times c_n = \mathrm{Enc}_{\mathsf{pk}}(m_1 + \ldots, +m_n)$ を送る．
6. a_0 は $c(n)$ を秘密鍵 sk を用いて復号し，A_n に配る．

なお，秘密鍵を複数人で分散管理する閾値暗号系を用いることによって，鍵管理者による秘密情報への攻撃の危険を軽減することができる．

準同型性公開鍵暗号の利用には，鍵生成および解読，暗号同士の計算の時間を計算時間として要し，また暗号文のやり取りのために通信が生じる．プロトコルの設計によって，次の節で紹介する関数秘匿計算よりも計算量通信量ともに抑えられることが経験的に知られている．例えば本稿 5 章や [Arai 11] などでの評価を参照されたい．

4.2 関数秘匿計算

秘匿関数計算 (SFE) [Yao 86] は論理式で表現できる任意の関数計算を秘密に実行する暗号学的手法である．SFE では計算対象の関数から計算回路を作成する．その計算回路上で，他の計算主体や第三者には無情報であるが，積み重ねによって意図した情報のみ受け渡しができるようなデータのやり取りを行う．結果，入力データを秘匿したままの計算が可能となる．しかし，計算する関数や入力の数値によっては計算回路が非常に大きくなり，回路の作成および計算の実行が困難になる場合があることが経験的に知られている．

5. プライバシ保護化合物検索

5.1 背景

創薬研究の分野で，創薬ターゲット（薬のもととなる化合物）の発見を試みる場合，既知の化合物データベースから類似化合物を検索する必要がある．創薬ターゲットとして優良な化合物を収集したデータベースは Focused Library[Miller 06] と呼ばれ，販売されるケースも多い．化合物データベースに対するクライアントの具体的な検索行動は創薬ターゲットに関する情報が推測されかねないもので秘匿性が非常に高い．一方で商用の化合物データベースは知的財産としての価値があるため，データベースの所有者はその具体的な内容を購入者以外に漏えいさせたくないという要望を持つ．すなわち化合物データベース，クライアントの検索質問双方が秘匿性が高い問題である．

さらに創薬ターゲットや化合物データベースの内容は非常に厳重に管理されており，予期せぬ漏えいを防ぐためにデータベースをネットワークにつなぐことを禁じている場合が多い．そのため最低限の通信で検索を行うことが望まれる．

5.2 化合物の類似検索

化合物データベース検索の目的は，類似化合物の情報を収集することである．化合物間の類似度を扱う方法として本稿では化合物のフィンガープリントを用いた検索を用いる．化合物の情報は化合物の持つ物理的または化学的特徴 f_i の集合 $F = \{f_i\}$ を用い，それらの有無を示すフィンガープリントと呼ばれる固定長のビット列で表現できる．フィンガープリントの各桁はそれぞれ対応する特徴の有無を示しており，桁 i の値が 1 の場合は桁 i に対応する特徴 f_i を持ち，0 の場合はないとする．よって 2 つの化合物に対応するフィンガープリントが類似していれば，フィンガープリントに用いた部分構造の空間において化合物同士が似ていることを示す．

フィンガープリントの類似性を評価する指標として Tversky 係数がよく用いられる．化合物 a のビット 1 が立っている桁を要素に持つ集合をフィンガープリント集合 F_a とおく．$\overline{F}_a = F - F_a$ である．化合物 a と b のフィンガープリントについての Tversky 係数 $T_{a,b}$ は

$$T_{a,b} = \frac{|F_a \cap F_b|}{|F_a \cap F_b| + \alpha|F_a \cap \overline{F}_b| + \beta|\overline{F}_a \cap F_b|}. \qquad (1)$$

ここで，α，β はクライアントが設定するパラメータである．通常の化合物データベース検索では，Tversky 係数が高い化合物を検索結果としてクライアントが得る．

5.3 プライバシ保護化合物検索

5.1 節で示したようにデータベースを所有するサーバ，クライアント双方の情報が秘密情報であると想定した化合物データベース検索を扱う．

ここで，通常の検索のように検索結果のフィンガープリントをクライアントに提示してしまうと，クライアントに有用な情報がデータベース販売前に渡されてしまう．秘密検索によってデータベース販売の機会を設け，サーバ，クライアント双方に利得があるように以下のような ”お見合い” 問題を考える．

定義 1 化合物データベースマッチング：化合物データベース D を持つサーバ S に対する，フィンガープリントの定義 F，クエリ化合物のフィンガープリント F_q，*Tversiky* 係数のパラメタ α，β，閾値 θ を用いたクライアント C の検索から，クライアント C はサーバ S が持つ，F_q との *Tversky* 係数が閾値 θ 以上の化合物の数を得る．

サーバ，クライアント双方が互いの秘密情報を開示せずに上記のマッチングを行う問い合わせ問題を以下に定義する．

Statement 1 プライバシ保護化合物検索：化合物データベース D を持つサーバ S，検索クエリ化合物 F_q，パラメタ α，β，θ を持つクライアント C によるプライバシ保護化合物によって，クライアント C はサーバ S に対する化合物データベースマッチングの結果のみを知ることができ，サーバ S はパラメタ α，β，θ 以外何の情報も得ない．

プライバシ保護化合物検索によって，クライアントは自らの検索条件を秘匿したまま，データベースが持つ検索条件に適した化合物の数を知ることができる．サーバ側は何の情報も得ないが，データベース販売の機会を得ることができる．またサーバはクライアントに渡す検索結果以外には秘匿するため，データベースの化合物情報はおおよそ守られる．

5.4 秘匿検索プロトコル

プライバシ保護化合物検索を実現する秘匿検索プロトコルを PPCS(Privacy Preserving Chemical compound Search) とし，その概要を示す．クライアントおよびサーバが計算主体となる加法準同型公開鍵暗号系を用いたアプローチをとる．

鍵管理は以下のように行う．クライアント C が鍵を生成し，公開鍵 pk のみをサーバ S に渡すとする．S は公開鍵 pk を用いてメッセージの暗号化が可能であるが，秘密鍵 sk を所有していないため，暗号文の復号はできない．そのため暗号文の通信において S に暗号文の中身は漏えいしない．C は pk，sk 両方を持つため暗号化復号化両方が可能である．

プロトコルの概要は以下の手続である．

1. C が鍵を生成，公開鍵 pk を S に送信する．
2. C はクエリ化合物 q のフィンガープリント F_q の各ビットを暗号化し，暗号文を並べたベクトル $\mathbf{x}$ に変換する．C は $\mathbf{x}$ およびパラメタ α，β，θ を S に送る．
3. S は $\mathbf{x}$ を受け取り，各 $F_b(b \in D)$ について，F_a の間の Tversky 係数 $T_{a,b}$ の暗号文を計算する．さらに $T_{a,b}$ と閾値 θ との大小を示す暗号文 c_b を計算する．
4. S は $\{c_b|b \in D\}$ をシャッフルして C に送信する．

5. C は暗号文を復号し，Tversky 係数が θ 以上の化合物の数を得る．

上記のプロトコルではサーバとクライアントのやり取りはすべて暗号文で行われるため，安全性が保障される．

詳細は講演において紹介する．

5.5 提案プロトコルの計算量

提案プロトコルの計算量をモデルデータを用いて検討した．PPCS を実装する暗号方式に Paillier 暗号 [Dåmgard 01] を用いた．また，プライバシ保護化合物検索は SFE を用いても実現できる．これを SFE-IR とし，Fairplay[Malkhi 04] を用いて実装し PPCS と計算量を比較した．

実験データセットとして，フィンガープリントとして 10 から 1000 のビットベクトル，フィンガープリント数が 1000 件のモデルデータを用いた．実験には 3.46GHz(CPU)，16GB(RAM) の Linux マシンを用いた．

実験結果を図 1 に示す．これより，PPCS は SFE-IR に比べ計算時間が数オーダー低いことがわかる．なお，SFE-IR はサーバ，クライアント双方に同じ計算時間を要する．また，SFE はフィンガープリントの長さにともない計算回路が大きくなり，200 以上の長さでは使用した実験環境で計算回路を構築することはできなかった．また，提案プロトコルは一往復の通信しか行わないのに対し，SFE-IR はその方式から複数回の通信を必要とする．これらの結果より，プライバシ保護化合物検索において PPCS は SFE-IR に比べスケーラブルであり，通信回数を極力抑えたいというユーザーの要請にも応えるものであるといえる．

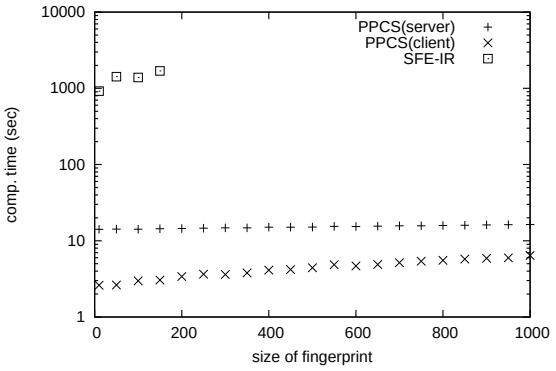

図 1: PPCS と SFE-IR の計算時間の比較

6. 終わりに

PPDM は秘密情報利活用の道を開く技術であり，本稿で紹介した事例をはじめ近年実社会問題への応用が始まりつつある．

学術研究におけるプライバシ保護データマイニング技術は強い安全性を保障するが，いざ実用化しようとすると計算量や通信量がネックになる場合がある．また，その方式がユーザーの要望や情報管理ポリシー，心象にそぐわないといった問題が生じる可能性も考えられる．そのため実用上では問題ごとに適した PPDM 方法を選択，設計する必要があると考えられる．

なお，秘密情報の漏えいを議論するためには計算過程の安全性に加えて出力プライバシの問題を検証する必要がある．本稿で紹介した化合物検索技術においてもデータベース側の出力における情報漏えいの議論が課題として残っている．

今後，プライバシ保護データマイニング技術の利用によって個人や企業などの所有する情報の安全な利活用がより活発になっていくと期待される．

参考文献

[Agrawal 00] Agrawal, R. and Srikant, R.: Privacy-preserving data mining, in *ACM Sigmod Record*, Vol. 29, pp. 439–450ACM (2000)

[Arai 11] Arai, H. and Sakuma, J.: Privacy preserving semi-supervised learning for labeled graphs, *Machine Learning and Knowledge Discovery in Databases*, pp. 124–139 (2011)

[Chor 95] Chor, B., Goldreich, O., Kushilevitz, E., and Sudan, M.: Private information retrieval, in *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pp. 41–50IEEE (1995)

[Dåmgard 01] Dåmgard, I. and Jurik, M.: A Generalisation, a Simplification and Some Applications of Paillier's Probabilistic Public-Key System, in *Public Key Crypt.*Springer (2001)

[Dwork 06] Dwork, C., McSherry, F., Nissim, K., and Smith, A.: Calibrating noise to sensitivity in private data analysis, *Theory of Cryptography*, pp. 265–284 (2006)

[Lindell 02] Lindell, Y. and Pinkas, B.: Privacy preserving data mining, *Journal of cryptology*, Vol. 15, No. 3, pp. 177–206 (2002)

[Machanavajjhala 07] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 1, No. 1, p. 3 (2007)

[Malkhi 04] Malkhi, D., Nisan, N., Pinkas, B., and Sella, Y.: Fairplay: secure two-party computation system, in *Proc. of the 13th USENIX Security Symposium*, pp. 287–302 (2004)

[Miller 06] Miller, J.: Recent developments in focused library design: targeting gene-families, *Current topics in medicinal chemistry*, Vol. 6, No. 1, pp. 19–29 (2006)

[Rabin 81] Rabin, M.: How to exchange secrets by oblivious transfer, Technical report, Technical Report TR-81, Harvard Aiken Computation Laboratory (1981)

[Sweeney 02] Sweeney, L., et al.: k-anonymity: A model for protecting privacy, *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, Vol. 10, No. 5, pp. 557–570 (2002)

[Yao 86] Yao, A.: How to generate and exchange secrets, in *Proc. of the 27th IEEE Annual Symposium on Foundations of Computer Science*, pp. 162–167 (1986)