公募企画

公募企画シンポジウム6
安心・安全なビッグデータの流通プラットホームとセキュリティ基盤技術

2017年11月22日(水) 08:45 〜 10:45 C会場 (10F 会議室1001)

[3-C-1-PS6-3] プライバシに配慮したデータ流通のための基盤技術

清本 晋作 （株式会社KDDI総合研究所）

本発表では、 JST CREST「ビッグデータ統合利活用促進のためのセキュリティ基盤技術の体系化」の研究成果である、プライバシに配慮しつつデータの共有・流通を実現するための基盤技術について紹介する。

プライバシに配慮したデータ流通基盤技術

- 実用的な Personal Data Store (PDS) 実現に向けて-

清本晋作[*1]、中村徹[*1]、三本知明[*1]
*1 株式会社 KDDI 総合研究所

Privacy-Conscious Data Distribution Platform

- Towards Practical Personal Data Store -

Shinsaku Kiyomoto[*1], Toru Nakamura[*1], Tomoaki Mimoto[*1]
*1 KDDI Research Inc.

In this paper, we present a conceptual architecture for privacy-respecting data distribution and then design a platform based on the architecture. The platform consists of Personal Data Stores (PDSs) including Privacy Policy Manager (PPM) functionality, generation of datasets, privacy risk analysis, and anonymization algorithms for datasets. The architecture is a reference model for Information Bank.

Keywords: Privacy, Data, Personal Data Store, Security, Data Market, Privacy Policy Manager (PPM)

1. はじめに

情報通信技術の発展により、多種多様かつ大量のデータの収集・分析が可能となってきており、それにより新たなサービスが生み出されてきている。パーソナルデータの利活用を通して、社会変革を進める試みが検討されている。情報銀行コンソーシアム [1]などでは、プライバシセンシティブな個人の情報を集積し積極的に活用しようという方針のもと、様々な検討がなされている。そうした、プライバシセンシティブな情報の利活用においては、プライバシ・バイ・デザインのもと、プライバシに配慮した適切な運用が求められる。一方、個人のデータは個人のコントロールに委ねるという考えのもと、Personal Data Store (PDS)が検討されているが、個人で PDS を準備し運用することは容易ではない。本稿では、プライバシに配慮しつつ、データの利活用を推進するデータ流通基盤技術について述べる。

2. データ流通基盤のコンセプト

本節では、我々が検討を進めているデータ流通基盤のコンセプトを説明する。図1にアーキテクチャを示す。本流通基盤で想定しているのは、個人 PDS がクラウドサービス上で集積された集合型 Personal Data Store である。このコンセプトは、いわゆる「情報銀行」の具体的な実現モデルの 1 つとして位置づけられ、各個人が PDS を個別に構築運用するよりも実現性が高いアーキテクチャである。個人 PDS は、パーソナル情報の第三者提供に特化したデータ格納領域であり、クラウドサービス上で、セキュアに分割され構成される。個人 PDS には、様々なデータソースから、データ所有者(個人)が許可した情報のみが格納される。その際、データ所有者の希望を考慮して、個人特定性を低減するような加工を施して格納するケースも考えられる。また、データ所有者は、どのようなサービスであれば(さらなる集合匿名化を行った後に)第三者提供して良いか、どのような加工をすれば提供しても良いか、等の情報を、プライバシプリファレンスとして個人 PDS に登録しておく。集合型 PDS では、複数の個人 PDS から提供されたデータを結合させて集合データ(データセット)を生成し、データセットのプライバシリスクを評価し、リスクが基準を満たしていなければ適切に集合匿名化を施し、データを利用したい事業者に提供する。また、上記の個人 PDS から情報を提供する際に、それぞれ設定されたプライバシプリファレンスが考慮され、データの提供可否が判断される仕組みとなる。左記については、Privacy Policy Manager (PPM)[2]機能を具備することで実現される。次節では、データ流通基盤を構成する各機能について、説明を行う。

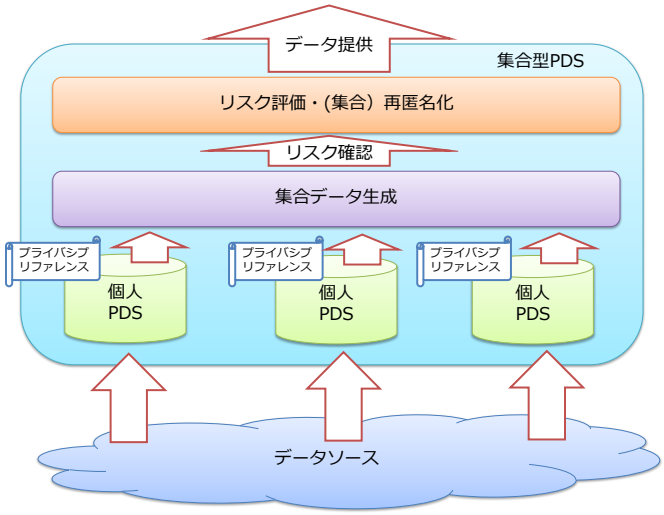

図1:データ流通基盤のコンセプト

3. データ流通基盤システム

データ流通基盤システムは、図 2 に示す通り、以下の機能が含まれる。各機能の詳細については、次節で説明する。

- 匿名化レベルによる匿名化機能:
  本機能は、PDS にデータを投入する前に加工を行う機能である。
- プライバシプリファレンス管理機能:
  本機能は、データ所有者のプライバシプリファレンスを管理し、プリファレンス情報を集合データ生成、匿名化処理等の際に提供する機能である。
- プライバシリスク評価・再匿名化機能:

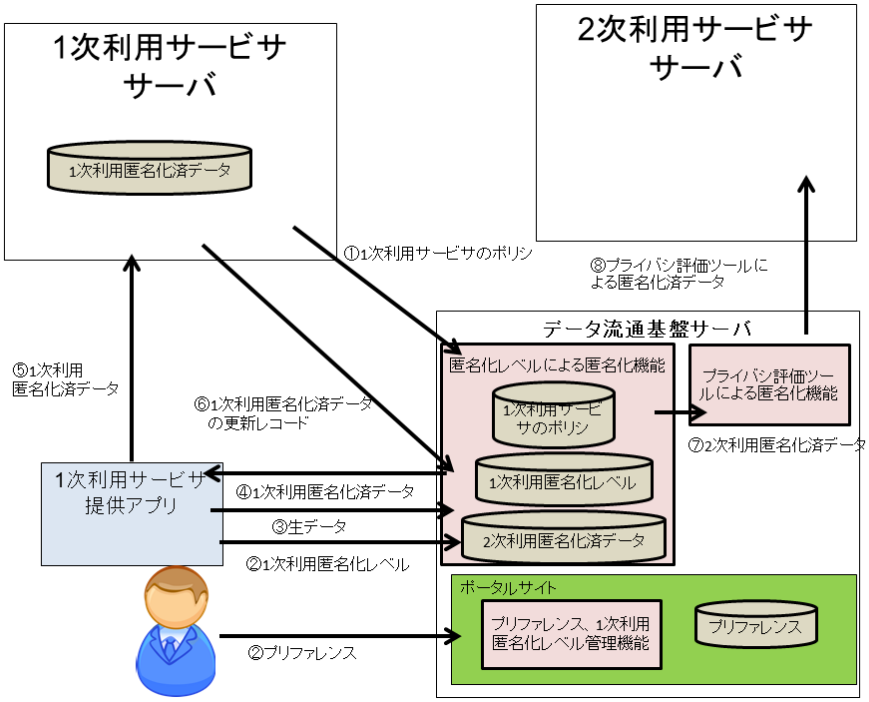

図 2:データ流通基盤システム

本機能は、集合データのプライバシリスクを評価し、また必要に応じて、再度匿名化処理を行う機能である。

2.1 匿名化レベルによる匿名化機能

各ユーザが設定した二次利用匿名化レベルや、一次利用サービサが設定した必要な情報における匿名化レベルに応じてデータを匿名化する機能である。機能としては、各ユーザが設定した二次利用匿名化レベルに応じて個人情報の匿名化を行い、DB に格納する機能と、一次利用サービサが要求する形に個人情報を匿名化してユーザに返す機能、を有する。二次利用匿名化レベルによる匿名化は、匿名化対象となる個人情報がどの匿名化の種類に該当するかを事前に定義し、設定したプリファレンスのレベルに応じて匿名化を実施する。匿名化レベルによる匿名化を行う際には各匿名化対象の情報が、先に記載したデータ種別のどれに該当するかを DB に定義する必要がある。匿名化を実施する際にはその DB にある情報やパラメータを参照することで、どのようにして匿名化を行うか決定する。以下に DB のイメージ例を記載する。匿名化対象の情報が増えるたびに DB に情報の追加を行い、どのように匿名化が行われるか定義する。

情報名	匿名化種別	匿名化パラメータ	匿名化情報参照先(木構造、住所の場合の別途アクセス先)
氏名	文字列	2 (頭から 2 文字のみ表示 他は*)	NULL
年齢	数値(丸め処理)	1 (頭から 1 文字	NULL
		のみ表示。他は 0)	
ユーザグループ	数値(グループ化)	NULL	user_group_tbl (参照先の DB 名)
住所	住所	NULL	address_tbl (参照先の DB 名)
商品購入店舗名	木構造	NULL	store_address_tbl (参照先の DB 名)
購買日付	日付	NULL	NULL
購買時刻	時刻	NULL	NULL

2.2 プリファレンス管理機能

プリファレンス管理機能では、デフォルト認可設定、第三者提供に関する同意、二次利用匿名化レベルの設定や更新を行う機能を有する。また、ポータルサイトで入力したプリファレンスの保管をデータ流通基盤サーバの DB が行い、ユーザがプリファレンスを変更した際にその結果を反映させる機能を有する。さらに、データ所有者がプライバシプリファレンスを設定する際に、その手間を低減させるため、各データ所有者のプリファレンス設定情報から、少数のプリファレンスに関する設定をするだけで、残りの全プリファレンスに関する設定を推測してくれる機能も有する。デフォルト認可設定、二次利用匿名化レベルからなる、プリファレンスに関する設問を、ユーザが全て設定することは非常に労力がかかる。そこで、少数のプリファレンスに関する設定を入力することで、残りを推測する機能の実装を行う。本機能は推測を行うために必要な、推測モデルの生成を、データ流通基盤サーバが管理しているプリファレンスの情報から行う機能である。推測モデル作成のアルゴリズムとしては SVM を使用する。

推測モデル生成機能は、モデル学習データと設定ファイルを入力として、最適な質問の組み合わせと、推測モデルを出

力する。学習データは、m×n の行列であり、m が被験者、n が質問 ID を表し、行列の値は対応する被験者のプリファレンスとなる。以下のように推測モデルの生成を行う。

① 設定ファイルで定義されているモデル生成に使用する学習データ数、モデル生成 に使用する評価データ数にしたがって、学習データからサンプルを抽出する（以後前者をモ デル生成学習データ、後者をモデル生成評価データと呼ぶ）。
② 設定ファイルで定義されている質問数にしたがって、質問 ID を選択する。
③ モデル生成学習データのうち、選択された質問 ID に対応するデータを用いて、選択されなかった質問 ID に対応するデータを推測する SVM モデルを生成する。
④ 生成された SVM モデルと、モデル生成評価データのうち選択された質問 ID に該当するデータを用いて、選択されなかった各質問 ID に対応する推測値を算出する。推測値と実際のモデル生成評価データの値を比較し、精度を算出する。
⑤ 全ての質問 ID の組み合わせについて、②-④を繰り返す。
⑥ 最も精度がよかった場合の質問 ID の組み合わせを最適な質問の組み合わせとし、 そのときの SVM モデルを推測モデルとして出力する。

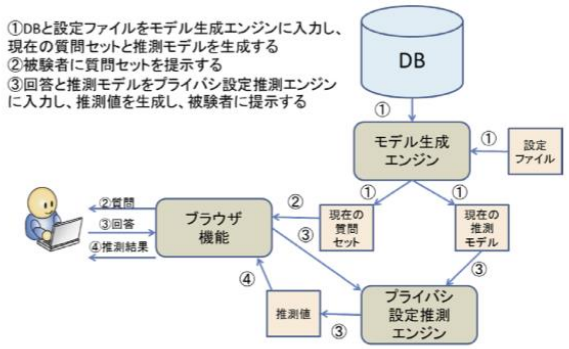

図 3:プライバシプリファレンスの推測決定機能

プリファレンス推測機能は、生成された最適な質問の組み合わせを入力として、入力に用いた質問 ID 以外のプリファレンスを推測する機能である(図3)。本機能では次のように推測値を出力する。なお、本機能の詳細については、別稿 [3]を参照されたい。

① 入力された回答と、各質問 ID に対応する SVM モデルを用いて、各質問 ID に対応する推測値を取得する。
② 取得した推測値を出力する。
③ 出た所有者が出力された推測値なるプライバシプリファレンスの確認を行い、問題が無ければ個人 PDS に登録する。

2.3 プライバシリスク評価・再匿名化機能

プライバシリスク評価・再匿名化機能は、複数のユーザの匿名化済みのデータの集合(データセット)において、個人が再識別されるリスクを自動で評価するとともに、リスクが基準を満たしていなければデータセットの再匿名化を行う機能である。例えば、入力したデータセットは K-匿名性が保たれているかどうかを評価するとともに、データセットの各個人情報を匿名化させることで、より匿名化の強度を高めることが出来る。本機能は、プライバシ評価機能と、再匿名化機能というサブコンポーネントから構成され、大規模な集合データに対しても適応可能となっている。また、匿名化処理については、複数の匿名化手法を組み合わせて適用できるようになっており、さらに複数の匿名化手法を組み合わせた場合でも、個人が再識別されるリスクを適切に評価できるようになっている。以下、リスク評価手法について述べる。

従来手法では、匿名化後のデータの一般公開を想定しているため、攻撃者の背景知識を想定することが困難であると仮定しており、非常に強い攻撃者を想定する必要があった。このような攻撃者の仮定のもと十分な安全性を維持しようとした場合、データの有用性を同時に維持することは困難であり、匿名化データの有効活用が思うように進まないことが考えられる。しかし匿名化したデータセットは、流通経路を明確にする、匿名化データへのアクセスを権限のあるユーザのみにするなど十分安全に運用することにより、攻撃者の背景知識を抑えることができることが知られている。我々は、このような現実的な仮定のもとに攻撃者が知り得る情報が限定されている場合についてのリスクを評価するためシミュレーションベースのリスク評価手法を導入する。リスク評価を行うにあたり、匿名化シミュレータを構築する必要がある。シミュレータは内部関数として匿名化機能(匿名化処理をシミュレートする機能)を呼び出す構成となるが、実際にどのような匿名化を実施したか、また想定される攻撃者がその加工手段を知り得るかどうかによって大きく異なる。我々の評価手法では複数の匿名化手法に対応するため、複数のシミュレータを構築し、それぞれに対してシミュレーションを実施して評価できるようにした。シミュレータの構築方法は多種多様であるが、実際の加工手段とあまりにも乖離しているシミュレータを用いた評価は実施する意味を持たないため、あくまで妥当なシミュレータを構築している. 本手法の詳細については、別稿 [4]を参照されたい。

4. おわりに

本稿では、プライバシに配慮しながら安全なデータ流通を実現するデータ流通基盤技術について述べた。今後は、本基盤を実装し、実証実験を行っていく予定である。

謝辞: 本研究は、JST CREST の支援を受けたものである。

参考文献

1) 情報銀行コンソーシアム, http://www.information-bank.net/
2) S. Kiyomoto, T. Nakamura, H. Takasaki, R. Watanabe, Y. Miyake, PPM: Privacy Policy Manager for Personalized Services, CD-ARES 2013: Security Engineering and Intelligence Informatics pp 377-392, 2013.
3) T. Nakamura, S. Kiyomoto, W. B. Tesfay, and J. Serna, Personalised Privacy by Default Preferences – Experiment and Analysis. In proc. of ICISSP 2016, pp. 53-62, 2016.
4) T. Mimoto, S. Kiyomoto, K. Tanaka, A. Miyaji, (p, N) – identifiability: Anonymity Under Practical Adversaries, Proc. of IEEE TrustCom 2017, to appear.