

一般口演

一般口演4

情報セキュリティとプライバシー・ネットワーク

2018年11月23日(金) 10:15 ～ 11:45 H会場 (福岡サンパレスHパレスホール)

[2-H-1-3] 多施設間の安全なデータ収集を目的とした情報基盤におけるプライバシーリスク評価の試み

○田中 勝弥¹, 山本 隆一² (1.東京大学大学院医学系研究科, 2.医療情報システム開発センター)

[はじめに] 現在多くの施設で導入されている SS-MIX2標準化ストレージは、ファイルシステムとしての制約から、患者横断的な診療データの検索に時間がかかる難点があり、多施設間横断的な診療データの収集に用いるためには、別途データベースを構築し双方を同期させて運用するなどの追加コストが発生する。また、抽出されたデータセットの安全な流通にあたっては、匿名加工基準に準拠しつつも、抽出ごとにデータセットを評価する運用が不可欠である。[方法] これまでに提案した、FUSE(File System in User space)をベースとしたリレーショナルデータベース上に実装可能な SS-MIX2標準化ストレージを用い、Bloom Filterによるストレージ間の重複レコード突合機能、抽出されたデータセット内の指定属性をキーにした重複度評価が可能な機能を試作した。東京大学医学部附属病院の SS-MIX2標準化ストレージ内の全メッセージを匿名加工したデータを複数の上記ストレージにランダムに分散させ、仮想的に多施設ストレージを構築し、ストレージ間のレコード群の突合を検証した。なお、ストレージ間のレコード群は10%ずつ重複させた。[結果] 試作したテストベッドにより、100万件程度のレコードをストレージ間の突合・データ抽出が1分程度で完遂可能なことを確認した。また、抽出されたデータセットに対して、対象となる属性を指定したうえで、上限値・下限値の指定、分割幅の指定、レコード間の重複度評価、出力時の重複度指定、が可能なユーザインターフェイスを作成し、試作機の動作確認が完了した。[おわりに] これまでに提案した FUSEベースの SS-MIX2標準化ストレージを応用することにより、ストレージ間の突合、抽出、データセットの評価という安全なデータ収集基盤としての一連の試作が完了した。より具体的な収集事例への適用について今後検討を進めたい。

多施設間の安全なデータ収集を目的とした情報基盤における プライバシーリスク評価の試み

田中 勝弥^{*1}、山本 隆一^{*2}

*1 東京大学大学院医学系研究科、*2 医療情報システム開発センター

Attempt to evaluate privacy risk using secure data collection infrastructure for data analysis among medical institutions

Katsuya Tanaka^{*1}, Ryuichi Yamamoto^{*2}

*1 Graduate School of Medicine, The University of Tokyo, *2 Medical Information System Development Center

SS-MIX2 standardized storage, which is currently installed in many medical institutions, has difficulty with using for data analysis across patients, due to constraints as a file system, and also the same with cross-institutional analysis. For secure data collection among medical institutions, it is essential to establish an infrastructure while conforming to anonymization processing standards. We have developed an SS-MIX2 standardized storage that can be implemented on the relational database based on the FUSE (File System in User space), which can detect the duplicate record collision between storages by Bloom Filter. Using this storage, we are developing prototypes that can evaluate the degree of privacy invasion risk evaluating the multiplicity of attributes in the extracted data set. The prototype test bed confirmed that about 1 million records processing can be accomplished in 1 minute or less between 3 data storages. In addition, for the evaluation of extracted data set, we have developed a user interface by which we can designate the target attribute, the upper and lower limit values, the division width at the time of data extraction. We will apply the developed application services to a more specific collection in the future.

Keywords: Information Security, Privacy Protection, SS-MIX2, HL7

1. はじめに

現在多くの施設で導入されている SS-MIX2 標準化ストレージは、ファイルシステムとしての制約から、患者横断的な診療データの検索に対しては、無数のファイル群を横断的にパースすることに起因して、膨大な時間がかかるため、多施設間横断的な診療データの収集に用いるためには、別途、目的に応じたインデックスデータベースを構築し双方を同期させて運用するなどの追加コストが発生する。標準規格で定められている拡張インデックスデータベースは、標準規格上はオプションの扱いであり、利用者側からの要求がない限り、標準装備されない機能である¹⁾。とくに、SS-MIX2 標準化ストレージ自体を遠隔地にバックアップするためには、差分同期を行う方法が最善であるが、ストレージ装置の機能を利用した差分同期を行う場合を除いては、差分の抽出そのものに時間を要し、拡張インデックスデータベースの構築なしに効率的なバックアップが実現できない。同様の状況が、次世代医療基盤法における匿名加工認定事業者の運用として想定されるスキームのような、大規模な SS-MIX2 標準化ストレージデータ収集時にも発生する。この場合には、SS-MIX2 標準化ストレージをバックアップする形で匿名加工認定事業者に集積し、中央集約された状態で、後段の匿名加工およびデータ解析を実施する流れとなるが、基本的には全データのバックアップが行われる形態であり、無数の HL7 ファイル群の収集が必要となる。トランザクションログを利用した収集を行う形態が試行されている。

一方で、目的が特化された多施設間での小規模な診療データの収集・解析には、次世代医療基盤法で示される全データの収集は本来不要であり、必要なデータ項目のみが患者横断的に収集可能であればよい。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」では、宮地らが多施設間のセキュアなクロス集計やデータ収集

を目的とした PSI (Private Set Intersection) ライブラリ²⁾、清本らがプライバシーリスク評価を目的としたライブラリ (以下、PIA ライブラリ) を開発してきており、このような用途についての実現が期待される。

2. 目的

本研究は、SS-MIX2 標準化ストレージに改良を加え、多施設間の安全なデータ収集を PSI ライブラリの適用により実施可能な SS-MIX2 標準化ストレージを適用した情報収集基盤を構築することを最終目標として、そのプロトタイプ開発を行うものである。本情報収集基盤は、クラウドでの利用を想定し、クラウド上でのデータの授受はすべて暗号化が施された環境下で動作することを目指す。本稿では、PSI サービスにより複数のストレージから抽出したデータセットに対し、プライバシーリスク評価を行う一連の操作を実現可能とする Web サービスのプロトタイプを構築したので報告する。

3. 方法

これまでに開発した仮想ファイルシステムの1つである FUSE (Filesystem in Userspace) を採用した SS-MIX2 標準化ストレージ³⁾を用い、PSI、PIA ライブラリを適用する。FUSE としては PostgreSQL に対応した pgfuse を採用した。最終的に安全なデータ収集を行うためのサービスをクラウドで運用することを想定し、PSI、PIA 各ライブラリによるデータ収集基盤のノードとなるサーバ・クライアントは、SOAP により Web サービスとして構成した。実験システムの構成を図1に示す。

検証用データは、東京大学医学部附属病院が保有する SS-MIX2 標準化ストレージのデータを匿名加工した HL7v2 形式のデータをとランダムに3つのストレージに分散させて、仮想的に多施設環境を構築した。各ストレージあたり 100 万

件規模の検体検査結果メッセージを格納し、PSI ライブラリによるデータセット作成を行うとともに、抽出されたデータセットをワンストップでリスク評価機能に適用可能なユーザインターフェイスを開発し、試験評価を行う。なお、SS-MIX2 標準化ストレージ間の患者は擬似的に人数カウントとして 10%重複させて評価を行った。

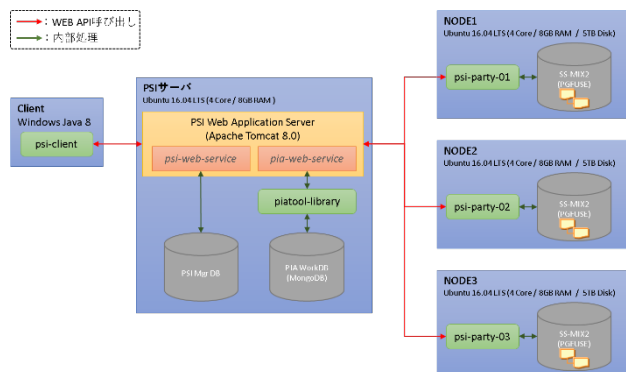


図1 実験システムの構成

プライバシーリスク評価機能は別の Web サービスとして構成し、PSI によるデータ抽出、取得後のデータ加工処理として操作が可能となる位置づけとした。リスク評価機能のシステム動作フローを図2に示す。抽出後データセットの各データ項目における最大値や最小値、データ個数などを画面で確認しながら、トップコーディング、ボトムコーディング、汎化処理（指定した分割精度による数値データの加工）を構築した。

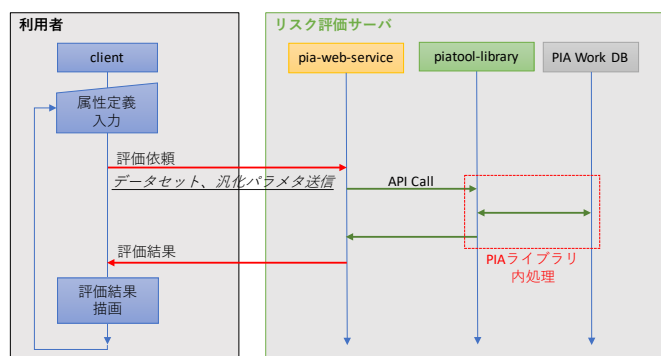


図2 プライバシリスク評価サービスの動作フロー

4. 結果

4.1 PSI サービスの性能評価

表1は、レコード数を変化させながら、3つのストレージで PSI サービスによるレコード間の突合を行った各処理時間の推移である。レコード数とともに処理時間は単調増加するが、100 万件規模の対象レコードに対して、検索・突合・抽出にかかる時間は約1分程度であった。

表1 PSI サービスの処理性能測定結果

レコード数	20,000	40,000	80,000	160,000	320,000	640,000	960,000	1,280,000
検索	0.1	0.2	0.2	0.3	0.6	1.0	1.5	2.0
Bloom Filter処理	0.5	0.5	1.0	2.0	2.9	7.4	13.5	20.7
抽出	3.3	3.0	3.1	3.4	4.0	10.1	15.4	23.5
計	3.8	3.7	4.3	5.7	7.4	18.5	30.4	46.2

4.2 プライバシリスク評価サービス

今回開発したプライバシーリスク評価を行うためのユーザインターフェイスを図3に示す。PSI サービスにより抽出されたデータセットを読み込み、各データ項目の最大値や最小値が確認可能となっており、数値データに関しては、上限・下限・分割単位を指定した加工が可能である。また、画面上で指定した属性値群を用いた、対象データセットの加工後の重複度演算が可能となっており、プライバシーリスク評価のための指標が確認可能なインターフェイスを有している。

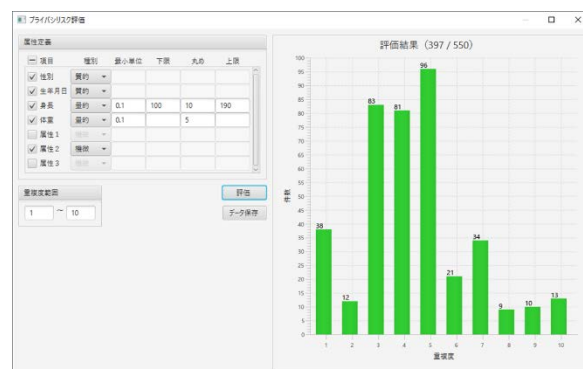


図3 プライバシリスク評価ツールの外観

5. 考察

本稿では、著者が所属する大学病院の診療データから仮想的に複数の SS-MIX2 ストレージを作成し、模擬的な多施設分散環境を作成し、PSI サービス、プライバシーリスク評価サービスの試験実装を行った。

PSI サービスに関しては、100 万件規模のレコード群に対して、分オーダの突合が行える性能が確認できた。今回は、SS-MIX2 標準化ストレージ内の診療データを対象にしたが、拡張ストレージにレセプト、DPC データを配置し、これらとの連結検索を可能とするストレージの開発を計画している。

プライバシーリスク評価サービスに関しては、10 万件規模の抽出後データセットに対して、数個の属性値を指定した場合、最大で数十秒程度のレスポンスで動作している。性能としては問題がないが、数値データの加工、重複度評価に関して、サービス側に機能を集中配置した構成であり、原データをサービス側へ暴露しなければ機能しない。データの秘匿の観点では課題を残した状態であり、機能配置を再考する必要がある。

なお、開発したサービス群は、SOAP を採用した Web サービスとして構成されており、今後 AWS などのクラウドへの実装を検討する。

6. おわりに

本稿では、SS-MIX2 標準化ストレージに蓄積されるデータを多施設間で安全に収集、検索するための基盤として PSI が適用可能な標準化ストレージを開発し、さらに抽出したデータセットのプライバシーリスク評価を行うプロトタイプの実装を行った。なお、本研究は、JST CREST グラント番号 JPMJCR1404 「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」により実施した。また、患者データの利用にあたっては、東京大学医学部倫理審査委員会の承認（審査番号：11787、課題名：患者横断検索が可能な SS-MIX2 標準化ストレージ向けプラットフォームの開発）をえて行った。

参考文献

- 1) 日本医療情報学会(JAMI):医療情報の標準化に関する情報・資料など Available from: <http://www.jami.jp/jamistd/smix2.php>.
- 2) Miyaji A, Nakasho K, Nishida S. Privacy-Preserving Integration of Medical Data. Journal of medical systems. 2017;41(3):37.
- 3) 田中 勝弥, 山本 隆一. 多施設間の安全なデータ収集を目的とした SS-MIX2 標準化ストレージ開発の試み. 医療情報学連合大会論文集. 2017;37 回:592-3.

