プロセスマイニングにおけるプライバシー保護データ公開の実践的側面

2020.7.16

マジッド・ラフィエイ，ウィル・ファン・デル・アールスト

抄録

プロセス発見や適合性チェックなどのプロセスマイニング技術は，情報システムで広く利用されているイベントデータを分析することで，実際のプロセスへの見通しを提供する．これらのデータは非常に価値のあるものだが，多くの場合，機密情報が含まれているため，プロセスアナリストは機密性と実用性のバランスをとる必要がある．最近研究者の間で注目を集めている，プロセスマイニングにおけるプライバシー問題は，その解決策を統合し，実世界で利用できるようにするためのツールによって補完されるべきである．本論文では，プロセスマイニングにおける最先端のプライバシー保護技術を実装した Python（パイソン）ベースのインフラストラクチャを紹介する．

インフラストラクチャは，Web ベースのツールとして，統合された単一の技術から技術の集合体まで，使用方法の階層を提供する．私たちのインフラストラクチャは，標準的なものとそうでないものの両方を管理する．プライバシー保護技術の結果として得られる標準的なイベントデータ，また，機密データを保護するために適用された変更を追跡するために，明示的なプライバシーメタデータを保存する．

キーワード：

責任あるプロセスマイニング・プライバシー保護・プロセスマイニング・イベントデータ

1. 導入

プロセスマイニングでは，イベントログの形で保存されていることが多いイベントデータを使用して，実際のビジネスプロセスに対する事実に基づいた洞察を提供する．プロセスマイニングの基本的なタイプは，プロセス発見，適合性チェック，プロセス強化の 3 つである[1]．イベントログはイベントの集合体であり，各イベントはその属性によって記述される．プロセスマイニングに必要な主な属性は，ケース ID，アクティビティ，タイムスタンプ，リソースである．イベント属性の中には個人を指すものもある．例えば，ヘルスケアの環境では，ケース ID 属性はデータが記録されている患者を指し，リソース属性は患者のために活動を行う従業員，例えば看護師や外科医を指す．

プロセスマイニングにおけるプライバシーの問題は，個人のデータがイベントログに含まれている場合に浮き彫りになる．欧州一般データ保護規則（GDPR）[12]などの規制によると，組織はデータを分析する際に個人のプライバシーを考慮に入れることが義務付けられている．最近では，プライベートデータを責任を持って解析する必要性から，

プロセスマイニングにおけるプライバシー問題への関心が高まっている[10,6,4,5]．[2]では，[4]や[5]で紹介したプライバシー保護技術を実装した Web ベースのツール ELPaaS を紹介している．ELPaaS はユーザから必要なパラメータを取得し，その結果をユーザのメールアドレスに CSV 形式で提供する．

Fig. 1 はプロセスマイニングにおけるプライバシーの保護について，2 つの主要な活動を含む一般的なアプローチを示している．それらの活動はプライバシー保護データ公開(PPDP)とプライバシー保護プロセスマイニング(PPPM)である．PPDP は，イベントデータの中にあるレコード所有の身元や機密データを隠してプライバシーを保護することを目的としている．PPPM は，従来のプロセスマイニングアルゴリズムを拡張し，いくつかの PPDP 技術から得られる非標準データで動作するようにすることを目的としている．PPPM アルゴリズムは，対応する PPDP 技術と密接に結合されていることに注意しなければならない．

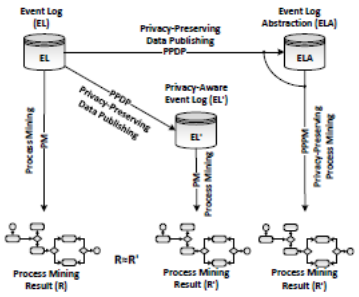

Fig. 1: The general approach of privacy in process mining.

本論文では，PPDP を中心に，プロセスを安全に発見するためのコネクタ法[9,10]，プライバシーを考慮したロールマイニングのための分解法[6]，プロセスマイニングのための TLKC-privacy モデル[8]など，最先端のプライバシー保護技術を提供するツールを紹介する．また，[7]で提案されているプライバシーメタデータも，提供されるプライバシー保全技術に組み込まれている．さらに，PM4Py-WS (PMTK) [3] を通じて，プロセスマイニングの環境でのプライバシー保護技術をウェブベースのインターフェイスを用いて紹介しており，特有の例として PPPM をサポートする既存のプロセスマイニングツールにプライバシー保護技術を追加できることを示している．

本論文の残りの部分は以下のように構成されている．第 2 節では，このツールの機能性と特徴を示す．第 3 節では，ツールの成熟度と可用性について概説し，第 4 節で本論文を締めくくる．

2. 機能と特徴

本節では，Django フレームワーク(https://www.djangoproject.com/)を用いて Python で書かれた他に比を見ないウェブベースツールである PPDP-PM の主な機能と特徴を示す．本ツールは，イベントデータ管理，プライバシーを考慮したロールマイニング，コネクタ方式，TLKC プライバシーの 4 つの主要なモジュールから構成されている．イベントデータ管理モジュールには，イベントログ抽象化 (ELA) [7]と呼ばれる，非標準的なイベントデータまたは，標準的な XES イベントログ(http://www.xes-standard.org/)で取り計らうイベントデータをアップロードして管理するための 2 つのタブがある．この

Fig. 2: The privacy-aware role mining page in PPDP-PM.

モジュールでは，プライバシー保護技術の入力としてイベントログを設定することができる．プライバシーを考慮したロールマイニングモジュール（Fig. 2）は，固定値，選択的，周波数ベースの 3 つの異なる技術をサポートする分解方法を実装している[6]．技術を適用した後，対応する「出力」セクションでは，XES 形式のプライバシー対応イベントログが提供される．生成されたイベントログは，誰が何を実行するかを公開することなく，リソースからロールマイニングするためのデータ有用性を保持する．

コネクタ・メソッドは，直帰グラフを発見するための暗号化ベースの方法を実装している [9,10]．これは，トレースをデータ構造に安全に格納された直帰関係のコレクションに分解する．この手法を適用した後，プライバシーを考慮したイベントデータは，対応する「出力」セクションに ELA 形式の XML ファイルとして提供される [7]．TLKC プライバシーモジュールは，セット，マルチセット，シーケンス，相対の 4 種類の背景知識を仮定して，グループベースのプライバシー保証を提供するプロセスマイニングのための TLKC-privacy モデル [8] を実装している．T はプライバシー対応イベントログのタイムスタンプの精度，L は背景知識の力，K は k-匿名性の定義[11]の k，C は同値クラスの敏感な属性値に関する信頼度の境界を表す．この方法を適用すると，XES 形式のプライバシーを考慮したイベントログが作成され，プロセス発見やパフォーマンス分析のためのデータの有用性が保存される．また，オープンソースのプロセスマイニングツールの環境でも，同じプライバシー保護技術を提供する．図 3 は，プロセスマイニングアルゴリズムをプライバシー対応イベントデータに直接適用できる PMTK のプライバシー統合のホームページのスニペットを示している．ツール内の各プライバシー保護技術は，イベントログ上で異なる技術の同時実行を可能にする Django アプリケーションとして実装されている．このアーキテクチャにより，プロジェクト全体の保守が容易になり，新しい技術は独立したアプリケーションとして簡単に統合することができる．プライバシー保護技術の出力は，各技術ごとに独立して提供され，イベントデータ置き場にダウンロードまたは保存することができる．PPDP-PM は，プライバシー保全技術のサイクルを提供するように設計されており，すなわち，イベントデータ置き場に追加されたプライバシー対応イベントデータは，標準的な XES イベントログの形式であれ

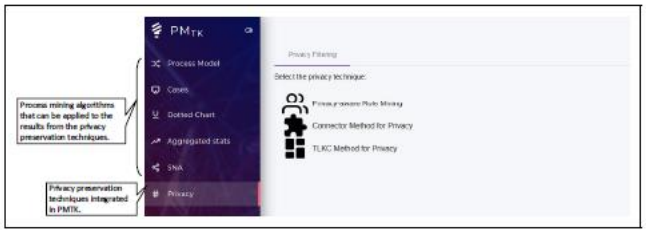

Fig. 3: The home page of the privacy integration in PM4Py-WS (PMTK).

ば，再度技術の入力として設定することができる．プロセスアナリストがプライバシーを意識したイベントログに適用された変更を認識し続けるために，プライバシーメタデータ[7]は，適用されたプライバシー保護技術の順序を指定する．さらに，本ツールでは，技術名，作成時刻，イベントログ名に基づいて，プライバシー対応イベントデータを一意に識別するためのネーミングアプローチを採用している．

3. 可用性と成熟度

前述の通り，PPDP-PM は Python で書かれたウェブベースのアプリケーションである．ソースコード，スクリーンキャスト，その他の情報は GitHub リポジトリ(https://github.com/m4jidRafiei/PPDP-PM)で入手可能．第 2 節で説明したプライバシー保護技術と，PMTK への統合は別々の Git リポジトリ(https://github.com/m4jidRa_ei/)としても利用可能．プライバシー保護技術の利用と統合を容易にするために，これらの技術は Python の標準パッケージとしても公開されている(https://pypi.org/)：pp-role-mining, p-connector-deg, ptlkc-privacy, p-privacy-metadata．我々のインフラストラクチャは，利用者がそれぞれの技術を独立して利用できるように，また，プライバシー保護技術を統合した PPDP-PM を無類のウェブベースのアプリケーションとして利用できるように，また，プライバシー保護技術を統合したプロセスマイニングツールで提供される技術を利用できるように，利用方法の階層を提供する．このツールの拡張性は，プライバシー保護技術や入力イベントログのサイズに応じて変化する．我々の試みに基づいて，我々のツールは実世界のイベントログ，例えば BPI チャレンジデータセット(https://data.4tu.nl/repository/collection:event logs real)を扱うことができる．しかし，産業規模での利用にはまだまだ改良の余地がある．PPDP-PM と PMTK への統合は Docker コンテナとしても提供されており，ユーザが簡単にホストすることができる: https://hub.docker.com/u/m4jid.

4. 結論

イベントデータには，多くの場合，プロセスアナリストが規制とは関係なく考慮する必要がある機密性の高い情報が含まれている．本論文では，プロセスマイニングにおけるプライバシー問題を扱うための Python ベースのインフラストラクチャを紹介する．プロセスマイニングにおけるプライバシー保護のためのデータ公開技術を実装したウェブベースのアプリケーションを紹介した．また，オープンソースのウェブベースのプロセスマイニングツールである PMTK にプライバシーを統合することを示した．また，他のプライバシー保護技術を統合できるように設計されている．本研究では，プロセスマイニングにおけるプライバシーと機密性の問題について，さまざまな観点から研究を行う予定であり，導入されたフレームワークに新しい技術が統合されることを想定している．また，他の研究者の方にも，それぞれの解決策を独立したアプリケーションとして，このフレームワークに統合していただくことをお願いしている．

参考文献

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016). https://doi.org/10.1007/978-3-662-49851-4

2. Bauer, M., Fahrenkrog-Petersen, S.A., Koschmider, A., Mannhardt, F., van der Aa, H., Weidlich, M.: Elpaas: Event log privacy as a service. In: Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019 (2019)

3. Berti, A., van Zelst, S.J., van der Aalst, W.M.P.: Pm4py web services: Easy development, integration and deployment of process mining features in any application stack. In: Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019 (2019)

4. Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRETSA: event log sanitization for privacy-aware process discovery. In: International Conference on Process Mining, ICPM 2019, Aachen, Germany (2019)

5. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-preserving process mining - di_erential privacy for event logs. Business & Information Systems Engineering 61(5), 595{614 (2019)

6. Ra_ei, M., van der Aalst, W.M.P.: Mining roles from event logs while preserving privacy. In: Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria. pp. 676{689 (2019)

7. Ra_ei, M., van der Aalst, W.M.P.: Privacy-preserving data publishing in process mining. In: Business Process Management Forum - BPM Forum 2020, Sevilla, Spain, September 13-18, 2020, Proceedings (2020)

8. Ra_ei, M., Wagner, M., van der Aalst, W.M.P.: TLKC-privacy model for process

mining. In: 14th International Conference on Research Challenges in Information Science, RCIS 2020 (2020)
9. Ra_ei, M., von Waldthausen, L., van der Aalst, W.M.P.: Ensuring con_dentiality in process mining. In: Proceedings of the 8th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2018), Seville, Spain (2018)
10. Ra_ei, M., von Waldthausen, L., van der Aalst, W.M.P.: Supporting condentiality in process mining using abstraction and encryption. In: Data-Driven Process Discovery and Analysis - 8th IFIP WG 2.6 International Symposium, SIMPDA 2018, and 9th International Symposium, SIMPDA 2019, Revised Selected Papers (2019)
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557{570 (2002)
12. Voss, W.G.: European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting. Business Lawyer 72(1) (2016)