

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7584455号
(P7584455)

(45)発行日 令和6年11月15日(2024.11.15)

(24)登録日 令和6年11月7日(2024.11.7)

(51)国際特許分類		F I	
G 0 6 F	21/62 (2013.01)	G 0 6 F	21/62 3 4 5
G 1 6 H	10/00 (2018.01)	G 1 6 H	10/00
G 0 6 F	21/71 (2013.01)	G 0 6 F	21/71
H 0 4 L	9/10 (2006.01)	H 0 4 L	9/10 A
G 0 6 F	21/60 (2013.01)	G 0 6 F	21/60 3 2 0
請求項の数 26 (全68頁)			
(21)出願番号	特願2021-574925(P2021-574925)	(73)特許権者	519102141
(86)(22)出願日	令和2年6月22日(2020.6.22)		エヌフェレンス、インコーポレイテッド
(65)公表番号	特表2022-537300(P2022-537300 A)		N F E R E N C E , I N C
(43)公表日	令和4年8月25日(2022.8.25)		アメリカ合衆国 マサチューセッツ州 0
(86)国際出願番号	PCT/US2020/038987		2 1 4 2 , ケンブリッジ, スイート 4
(87)国際公開番号	WO2020/257783	(74)代理人	0 0 , ワン メイン ストリート
(87)国際公開日	令和2年12月24日(2020.12.24)		110001519
審査請求日	令和5年6月6日(2023.6.6)		弁理士法人太陽国際特許事務所
(31)優先権主張番号	63/012,738	(72)発明者	アラヴァムダン, ムラリ
(32)優先日	令和2年4月20日(2020.4.20)		アメリカ合衆国 マサチューセッツ州 0
(33)優先権主張国・地域又は機関	米国(US)		1 8 1 0 , アンダーバー, ダーラム ド
(31)優先権主張番号	62/984,989	(72)発明者	ライブ 4
(32)優先日	令和2年3月4日(2020.3.4)		マーウーガドス, カルティク
最終頁に続く			アメリカ合衆国 マサチューセッツ州 0
			2 1 4 2 , ケンブリッジ, アパートメン
			最終頁に続く

(54)【発明の名称】 個人のヘルスケアデータを用いて計算するためのシステムおよび方法

(57)【特許請求の範囲】
【請求項1】

セキュアエンクレーブを形成する分離されたメモリパーティションを構築するステップであって、前記セキュアエンクレーブは、1つまたは複数のプロセッサ上で実行されている1つまたは複数の不正なコンピューティングプロセスから分離して1つまたは複数のアプリケーションコンピューティングプロセスを実行するために、前記1つまたは複数のプロセッサに利用可能である、ステップと、

前記セキュアエンクレーブ内のソフトウェアを予めプロビジョニングするステップと、を含み、

該予めプロビジョニングされたソフトウェアは、前記1つまたは複数のプロセッサ上で前記1つまたは複数のアプリケーションコンピューティングプロセスの命令を実行させるように構成され、前記1つまたは複数のプロセッサが、

ーデータプロバイダから入力データを暗号化された形式で受信するステップと、

ー前記データプロバイダ以外のエンティティに対応する命令プロバイダから、前記入力データを解析するための1つまたは複数のプログラムを受信するステップと、

ー前記入力データを前記セキュアエンクレーブにロードするステップと、

ー前記1つまたは複数のプログラムを前記セキュアエンクレーブにロードするステップと、

ー前記セキュアエンクレーブを使用して前記入力データを解析するための前記1つまたは複数のプログラムを実行するステップと、

ー1つまたは複数の暗号鍵を使用して入力データを復号化するステップと、

－出力データを生成するために、復号化された入力データに基づいて前記1つまたは複数のアプリケーションコンピューティングプロセスを実行するステップと、

－前記1つまたは複数のアプリケーションコンピューティングプロセスの前記命令が前記受信した入力データ上で動作したことを証明する実行証明を生成するステップと、

－前記1つまたは複数の暗号鍵を使用して前記出力データを暗号化するステップと、

－前記暗号化された出力データおよび前記実行証明への外部アクセスを提供するステップと、を実行し、

前記入力データは一組の臨床データ記録を含み、

前記一組の臨床データ記録を前記セキュアエンクレーブにロードする前に前記入力データを非特定化するステップをさらに含み、該入力データを非特定化するステップは、前記入力データから1つまたは複数の個人またはエンティティを特定する情報を削除するステップを含む、

方法。

【請求項2】

前記入力データは、臨床データから個人識別情報（PII）を削除することによって非特定化される前記臨床データを含む、請求項1に記載の方法。

【請求項3】

前記セキュアエンクレーブ内の前記1つまたは複数のアプリケーションコンピューティングプロセスの前記入力データまたは前記命令に関連する復号鍵を、前記データプロバイダまたは前記命令プロバイダにそれぞれ提供するステップをさらに含む、請求項1に記載の方法。

【請求項4】

前記入力データを前記セキュアエンクレーブにロードする前に、入力データに関連する入力非特定化確率を前記入力データに関連付けるステップをさらに含み、前記セキュアエンクレーブは、前記入力非特定化確率を前記出力データ内に維持する、請求項1に記載の方法。

【請求項5】

前記入力データと前記入力データを解析するための前記1つまたは複数のプログラムとは、同じトランザクションで前記セキュアエンクレーブにロードされる、請求項1に記載の方法。

【請求項6】

前記セキュアエンクレーブを第2のセキュアエンクレーブ内で動作するウェブサーバに接続し、前記セキュアエンクレーブの前記出力データをウェブクライアントまたはウェブブラウザ上のウェブページとして表示させるステップをさらに含む、請求項1に記載の方法。

【請求項7】

セキュアエンクレーブ内の前記ウェブサーバによって処理されている前記入力データは、前記ウェブブラウザまたは前記ウェブクライアントからアクセスできない、請求項6に記載の方法。

【請求項8】

前記ウェブサーバを介して表示される前記出力データは、前記セキュアエンクレーブに関連する暗号化オブジェクトに関連付けられる、請求項6に記載の方法。

【請求項9】

前記入力データをキュレーションサービスから受信するステップをさらに含み、該キュレーションサービスは、前記入力データがプライバシーを保護していると判定する、請求項1に記載の方法。

【請求項10】

前記セキュアエンクレーブは、パイプラインを形成するために1つまたは複数の他のセキュアエンクレーブに通信可能に結合され、前記セキュアエンクレーブからの前記出力データは、前記パイプライン内の後続のセキュアエンクレーブへの入力データとして提供さ

10

20

30

40

50

れる、請求項 1 に記載の方法。

【請求項 1 1】

前記実行証明は暗号化オブジェクトを含む、請求項 1 0 に記載の方法。

【請求項 1 2】

前記暗号化オブジェクトは、前記セキュアエンクレーブの内容の表現である、請求項 1 1 に記載の方法。

【請求項 1 3】

前記セキュアエンクレーブの内容の前記表現は、前記入力データ上で不正なコンピュータプログラムが動作していないことを証明する、請求項 1 2 に記載の方法。

【請求項 1 4】

前記セキュアエンクレーブの内容の前記表現は、前記入力データに関連する入力非特定化確率が前記セキュアエンクレーブによって維持されたことを証明する、請求項 1 2 に記載の方法。

【請求項 1 5】

前記暗号化オブジェクトは、前記パイプライン内の前記 1 つまたは複数の他のセキュアエンクレーブの内容を表す 1 つまたは複数の他の暗号化オブジェクトとリンクされる、請求項 1 2 に記載の方法。

【請求項 1 6】

前記 1 つまたは複数のコンピューティングプロセスを前記セキュアエンクレーブ内で実行するステップは、前記入力データまたは命令の少なくとも 1 つに基づいて前記セキュアエンクレーブの固有のシグネチャを生成するステップをさらに含む、請求項 1 に記載の方法。

【請求項 1 7】

前記受信した入力データおよび前記 1 つまたは複数のアプリケーションコンピューティングプロセスの前記命令は、パイプライン内の前記セキュアエンクレーブにリンクされた他のいかなるセキュアエンクレーブからもアクセスできない、請求項 1 に記載の方法。

【請求項 1 8】

前記受信した入力データおよび前記 1 つまたは複数のアプリケーションコンピューティングプロセスの命令は、前記パイプラインのオペレータからアクセスできない、請求項 1 7 に記載の方法。

【請求項 1 9】

前記 1 つまたは複数の不正なコンピューティングプロセスは、少なくとも 1 つの特権ソフトウェア、特権ファームウェア、またはネットワークインターフェースプロセスを含む、請求項 1 に記載の方法。

【請求項 2 0】

非一時的メモリと、

前記非一時的メモリから命令を読み出すように構成された 1 つまたは複数のハードウェアプロセッサと、を含み、前記命令は、実行されると、前記 1 つまたは複数のハードウェアプロセッサに、

セキュアエンクレーブを形成する分離されたメモリパーティションを構築することであって、前記セキュアエンクレーブは、前記 1 つまたは複数のハードウェアプロセッサ上で実行されている 1 つまたは複数の不正なコンピューティングプロセスから分離して 1 つまたは複数のアプリケーションコンピューティングプロセスを実行するために、前記 1 つまたは複数のハードウェアプロセッサに利用可能である、構築することと、

前記セキュアエンクレーブ内のソフトウェアを予めプロビジョニングすることと、を含む動作を実行させ、

前記予めプロビジョニングされたソフトウェアは、前記 1 つまたは複数のプロセッサ上で前記 1 つまたは複数のアプリケーションコンピューティングプロセスの命令を実行させるように構成され、前記 1 つまたは複数のプロセッサに、

ーデータプロバイダから一組の臨床データ記録を暗号化された形式で受信することと、

10

20

30

40

50

－前記一組の臨床データ記録を解析するためのプログラムを、前記データプロバイダ以外のエンティティから受信することと、

－前記一組の臨床記録および前記一組の臨床データ記録を解析するための前記プログラムを前記セキュアエンクレーブにロードすることと、

－前記セキュアエンクレーブを使用して前記一組の臨床データ記録を解析するための前記プログラムを実行することと、

－1つまたは複数の暗号鍵を使用して前記一組の臨床データ記録を復号化することと、
－出力データを生成するために、前記復号化された前記一組の臨床データ記録に基づいて前記1つまたは複数のアプリケーションコンピューティングプロセスを実行することと、
－前記1つまたは複数のアプリケーションコンピューティングプロセスの前記1つまたは複数の命令が前記受信した入力データ上で動作したことを証明する実行証明を生成することと、

－前記1つまたは複数の暗号鍵を使用して前記出力データを暗号化することと、
－前記暗号化された出力データおよび前記実行証明への外部アクセスを提供することと、
－を実行させるものであり、

前記動作は、前記一組の臨床データ記録を前記セキュアエンクレーブにロードする前に前記一組の臨床データ記録を非特定化することをさらに含む、

システム。

【請求項21】

前記動作は、前記セキュアエンクレーブに関連する復号鍵を前記データプロバイダに提供することをさらに含む、請求項20に記載のシステム。

【請求項22】

前記動作は、前記セキュアエンクレーブを第2のセキュアエンクレーブ内で実行されるウェブブラウザに接続し、前記セキュアエンクレーブの前記出力データをウェブクライアントまたはウェブブラウザ上のウェブページとして表示させることをさらに含む、請求項20に記載のシステム。

【請求項23】

前記セキュアエンクレーブは、パイプラインを形成するために1つまたは複数の他のセキュアエンクレーブに通信可能に結合され、前記セキュアエンクレーブからの前記出力データは、前記パイプライン内の後続のセキュアエンクレーブへの入力データとして提供される、請求項20に記載のシステム。

【請求項24】

前記実行証明は暗号化オブジェクトを含む、請求項20に記載のシステム。

【請求項25】

前記1つまたは複数のコンピューティングプロセスを前記セキュアエンクレーブ内で実行するステップは、前記一組の臨床データ記録に基づいて前記セキュアエンクレーブの固有のシグネチャを生成するステップをさらに含む、請求項20に記載のシステム。

【請求項26】

1つまたは複数のハードウェアプロセッサによって実行されると、前記1つまたは複数のハードウェアプロセッサに動作を実行させる命令を格納する非一時的コンピュータ可読媒体であって、前記動作は、

セキュアエンクレーブを形成する分離されたメモリパーティションを構築することであって、前記セキュアエンクレーブは、前記1つまたは複数のハードウェアプロセッサ上で実行されている1つまたは複数の不正なコンピューティングプロセスから分離して1つまたは複数のアプリケーションコンピューティングプロセスを実行するために、前記1つまたは複数のハードウェアプロセッサに利用可能である、構築することと、

前記セキュアエンクレーブ内のソフトウェアを予めプロビジョニングすることと、
を含み、

前記予めプロビジョニングされたソフトウェアは、前記1つまたは複数のプロセッサ上で前記1つまたは複数のアプリケーションコンピューティングプロセスの命令を実行させ

10

20

30

40

50

るように構成され、前記1つまたは複数のプロセッサに、
ーデータプロバイダから一組の臨床データ記録を暗号化された形式で受信することと、
ー前記一組の臨床データ記録を解析するためのプログラムを、前記データプロバイダ以外のエンティティから受信することと、
ー前記一組の臨床記録および前記一組の臨床データ記録を解析するための前記プログラムを前記セキュアエンクレーブにロードすることと、
ー前記セキュアエンクレーブを使用して前記一組の臨床データ記録を解析するための前記プログラムを実行することと、
ー1つまたは複数の暗号鍵を使用して前記一組の臨床データ記録を復号化することと、
ー出力データを生成するために、前記復号化された前記一組の臨床データ記録に基づいて前記1つまたは複数のアプリケーションコンピューティングプロセスを実行することと、
ー前記1つまたは複数のアプリケーションコンピューティングプロセスの前記1つまたは複数の命令が前記受信した入力データ上で動作したことを証明する実行証明を生成することと、
ー前記1つまたは複数の暗号鍵を使用して前記出力データを暗号化することと、
ー前記暗号化された出力データおよび前記実行証明への外部アクセスを提供することと、
ーを実行させるものであり、
前記動作は、前記一組の臨床データ記録を前記セキュアエンクレーブにロードする前に前記一組の臨床データ記録を非特定化することをさらに含む、

非一時的コンピュータ可読媒体。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願の相互参照

本出願は、2019年6月21日に出願された、「Systems and Methods for Selective Information Masking in Text」と題する米国仮出願第62/865,030号、2020年1月16日に出願された、「Systems and Methods for Retrieving Information Responsive to a Query」と題する米国仮出願第62/962,146号、2020年3月4日に出願された、「Systems and Methods for Selective Information Masking in Text」と題する米国仮出願第62/984,989号、2020年3月4日に出願された「Pipelined Federated Architecture for Computing with Private Healthcare Data」と題する米国仮出願第62/985,003号、および2020年4月20日に出願された、「Systems and Methods for Augmented Curation and Temporal Discrimination of Health Records」と題する米国仮特許出願第63/012,738号の米国特許法第119条(e)に基づく優先権を主張し、これらの各々は、その全体が参照により本明細書に組み込まれる。

【0002】

本発明は、一般に、消費者または規制によって私的とみなされる情報を含むことができるデータを処理することに関する。

【背景技術】

【0003】

病院、ヘルスケアプロバイダおよび介護者は、患者から大量のデータを収集する。それは、ヘルスケアが一般の人々に提供されるプロセスの必要な部分である。典型的には、患者は、自身の病気の治療を受ける一部としてデータを介護者に提供する。このデータは、介護者によって記憶され、とりわけ研究目的で後に使用され得る。別の典型的なシナリオでは、1つまたは複数のデバイス、例えばパルスオキシメータ、グルコースモニタ、スマ

ートウォッチ、フィットネスブレスレットなどを介して消費者からデータを収集することができる。そのようなユースケースでは、収集されたデータは、連続的にまたは一定期間にわたって患者の健康を解析するために使用されることが多い。その結果、膨大な量の患者情報がサービスプロバイダによって蓄積され得る。

【発明の概要】

【発明が解決しようとする課題】

【0004】

介護者およびサービスプロバイダによって収集された患者データの多くの態様は、プライバシー規制の対象となり得る。患者から収集された処理データの有用性および利益は明らかであり、一般に認められている。しかしながら、特にデータを使用して患者を識別することができる場合、ユーザデータのプライバシーを維持する懸念が高まっている。このような懸念は、1996年に米国議会によって最初に通過したHIPAA (Health Insurance Portability and Accountability Act: 医療保険の携行性と責任に関する法律) 規制の根拠である。他の多くの国も同様の規則および法律を公布している。一般に、HIPAAおよび他の規制は、公衆の構成員の識別またはそれらの物理的属性もしくはバイオメトリックデータの詳細をもたらし得る個人情報情報の公開を制限する。

【課題を解決するための手段】

【0005】

したがって、個々の患者または消費者のプライバシーを維持するという制約の下で、生物医学的（および他の種類の）データが計算プロセスによって解析されることを可能にする必要がある。そのようなシステムおよび方法は、結果として、社会にとって商業的、社会的および科学的に大きな利益となる。

【図面の簡単な説明】

【0006】

【図1A】プライベートデータを含むデータセットを処理するための連手法法を示す図である。

【0007】

【図1B】いくつかの実施形態による、セキュアコンピューティング環境のアーキテクチャの簡略図である。

【0008】

【図2】いくつかの実施形態によるセキュアエンクレーブの一般的なアーキテクチャの簡略図である。

【0009】

【図3】いくつかの実施形態によるデータセットに適用可能な例示的なポリシーの簡略図である。

【0010】

【図4A】ポリシーがセキュアエンクレーブ内で適切にプログラムされることを保証するための例示的なオーケストレーションの簡略図である。

【0011】

【図4B】ポリシー制約に従ってデータが計算されることを保証するための例示的なオーケストレーションの簡略図である。

【0012】

【図5】信頼チェーンを拡張するための技術の使用、すなわち、計算の連合パイプラインに関連する認証の簡略図である。

【0013】

【図6】拡張された信頼チェーンの検証の簡略図である。

【0014】

【図7】いくつかの実施形態による非集中型信頼モデルの簡略図である。

【0015】

10

20

30

40

50

【図 8】いくつかの実施形態による、共同でプライベートデータを受信、格納、および処理する企業の連合グループのアーキテクチャの簡略図である。

【0016】

【図 9】いくつかの実施形態による、共同でプライベートデータを受信、格納、および処理する企業の連合グループのアーキテクチャの簡略図である。

【0017】

【図 10】いくつかの実施形態による、共同でプライベートデータを受信、格納、および処理する企業の連合グループのアーキテクチャの簡略図である。

【0018】

【図 11】いくつかの実施形態による情報マスキングのためのシステムの簡略図である。

10

【0019】

【図 12】いくつかの実施形態による、情報マスキングのための制御フローの簡略図である。

【0020】

【図 13】いくつかの実施形態によるエンティティのタグ付けプロセスおよび欠落したエンティティの処理の簡略図である。

【0021】

【図 14】いくつかの実施形態による、有効性をマスキングするためのテストングスイート診断フロー方法を示す簡略図である。

【0022】

20

【図 15】いくつかの実施形態による情報検索のためのシステムの簡略図である。

【0023】

【図 16】いくつかの実施形態による情報検索のためのデータフローの簡略図である。

【0024】

【図 17】いくつかの実施形態による、フラグメント検索のためのデータフローの簡略図である。

【0025】

【図 18】いくつかの実施形態による、クエリに応答して情報を検索するための方法の簡略図である。

【0026】

30

【図 19】いくつかの実施形態による、フラグメント検索を実行するための方法の簡略図である。

【0027】

【図 20 A】いくつかの実施形態による情報検索システムのグラフィカルインターフェースのスクリーンショットを示す図である。

【図 20 B】いくつかの実施形態による情報検索システムのグラフィカルインターフェースのスクリーンショットを示す図である。

【0028】

【図 21】いくつかの実施形態による情報検索のための方法の簡略図である。

【0029】

40

【図 22 A】いくつかの実施形態による情報検索システムのグラフィカルインターフェースのスクリーンショットを示す図である。

【図 22 B】いくつかの実施形態による情報検索システムのグラフィカルインターフェースのスクリーンショットを示す図である。

【0030】

【図 23】いくつかの実施形態による、健康記録の強化された管理のための方法の簡略図である。

【0031】

【図 24】いくつかの実施形態による、健康データの時間的判別のための方法の簡略図である。

50

【0032】

【図25-1】いくつかの実施形態による、時間的に整列された一組の健康記録に基づいて生成された視覚化の簡略図である。

【図25-2】いくつかの実施形態による、時間的に整列された一組の健康記録に基づいて生成された視覚化の簡略図である。

【0033】

開示された主題の様々な目的、特徴、および利点は、同様の符号が同様の要素を識別する以下の図面に関連して考慮される場合、開示された主題の以下の詳細な説明を参照してより完全に理解することができる。

【発明を実施するための形態】

【0034】

真に驚くべき量の情報が、健康状態、習慣、環境、周囲、および家庭に関する患者および消費者から収集されている。この情報は、機械学習および人工知能モデルを利用するコンピュータプログラムによってますます処理されている。そのようなコンピュータプログラムは、消費者の健康状態、疾患の発生および治療、ユーザの行動などの解析および予測において著しい進歩を示している。さらに、収集されたデータは、患者の生体認証および他の個人識別属性を含む可能性があるため、そのようなコンピュータプログラムは、患者および消費者の身元を知ることができるという懸念が高まっている。したがって、個人属性を含むヘルスケアデータの解析に関心のある企業は、個人のプライバシーを維持し、個人および個人データに関するHIPAA（医療保険の携行性と責任に関する法律1996）規制などの関連規制を遵守することに関心がある。

【0035】

HIPAAに加えて、欧州連合のGDPR（General Data Protection Regulations）、PSD2（Revised Payment Services Directive）、CCPA（California Consumer Privacy Act 2018）など、多くの他の規制が様々な管轄区域で制定されている。

【0036】

以下の説明では、「ユーザ情報」、「個人情報」、「個人の健康情報（「PHI」）」、「ヘルスケア情報のデータまたは記録」、「識別情報」、およびPII（個人を識別可能な情報）という用語は、互換的に使用され得る。同様に、「電子健康記録（「EHR」）」および「データ記録」という用語は互換的に使用され得る。

【0037】

プライベートデータを処理する1つの手法は、データセットのすべての記録を暗号化することである。暗号化されたテキストは、暗号文と呼ばれることもある。復号化されたテキストは平文とも呼ばれる。暗号化は、類推によって、データセットの記録をロックされたボックスに入れるものとして説明することができる。その後、ロックされたボックスの記録へのアクセスは、ロックされたボックスへの鍵によって制御される。その概念は、許可されたエンティティのみが（復号化）鍵へのアクセスを許可されるということである。

【0038】

いくつかの規制（例えば、HIPAA）は、ヘルスケアデータが暗号化形式で記憶されることを要求する。これは、「静止時暗号化」と呼ばれることもある。

【0039】

しかしながら、悪意のあるエンティティは、復号鍵にアクセスしたり、計算機構を使用して復号鍵を推測／憶測したりする可能性がある。後者の可能性は、暗号化／復号化技術が十分に強力ではない（例えば、鍵の長さ（鍵を構成するビット数）は、計算攻撃に耐えるのに十分な長さではない）場合、または鍵が失われたかもしくは安全に格納されていない場合に可能性が高くなる。

【0040】

暗号化および他のそのようなセキュリティ技術は、計算攻撃者が、基礎となるデータへ

10

20

30

40

50

のアクセスを得るために、特定の量のリソース（コンピュータ時間、メモリ、および計算能力）を費やす可能性が高いという予想に依存し得る。暗号鍵の長さは、暗号化を解除するために必要な計算リソースの量を増やすために使用される変数の1つである。

【0041】

強力な暗号化技術であっても、個人データの処理に関連するセキュリティ上の課題を解決することはできない。例えば、暗号化されたデータセットを処理している企業は、データセットをコンピュータにロードし、データセットを復号化し、データセットの記録を処理し、データセットを再暗号化することができる。この例では、データセットの1つまたは複数の記録は、処理中に（平文に）復号化される。悪意のあるエンティティは、平文記録が処理されている間にコンピュータにアクセスし、個人情報情報の漏洩につながる可能性がある。すなわち、処理を目的としてデータを復号化すると、「ランタイム」脆弱性が導入される。

10

【0042】

したがって、個人データを処理するための改善された技術を開発することが望ましい。

【0043】

完全準同型暗号化（FHE）は、暗号化されたデータを復号化せずに使用して計算するための手法を記述する。すなわち、暗号化されたデータ要素

【数1】

$$x_1^e, x_2^e, \dots$$

20

が与えられると、暗号化された結果

【数2】

$$(y_1^e, y_2^e, \dots)$$

をもたらす関数

【数3】

$$f(x_1^e, x_2^e, \dots)$$

30

を計算する。そのような計算の入力、出力、および処理段階は暗号化されたデータ要素のみを扱うので、漏洩の確率は最小限に抑えられる。暗号化技術の（数学的）根拠が十分に強い場合、例えば量子コンピュータなどの非常に強力なコンピュータが使用されたとしても、鍵の推論／推測は実行不可能な計算になる可能性がある。

【0044】

しかしながら、FHEデータセットを用いて計算するための従来の技術は、非実用的であるという点で非効率的であり得る。2009年に報告された計算では、FHEデータセット上で実行される計算は、暗号化されていないデータ計算よりも100兆倍遅い（Amesh Divatia, <https://www.darkreading.com/attacks-breaches/the-fact-and-fiction-of-homomorphic-encryption/a/d-id/1333691> および Priyadarshan Kolte, <https://baffle.io/blog/why-is-homomorphic-encryption-not-ready-for-prime-time/>を参照されたい）。

40

【0045】

さらに、基本的なFHE機能を提供するFHEライブラリを使用するために、既存のアプリケーションコードを書き換える必要がある場合がある。

【0046】

セキュアエンクレープは、機密データをコンピュータ内で実行されている他のプロセス

50

に公開することなく、機密データを復号化してメモリ内で処理することができるコンピューティング環境を記述する。データは、他のプロセスおよびネットワークから「隔離された」コンピューティング環境で復号化および処理される。このような環境の保護は、後述する方法で復号鍵を保護することによってさらに強化することができる。

【0047】

セキュアエンクレーブの技術は、FHE技術よりも効率的であり得る。

【0048】

場合によっては、セキュアエンクレーブを含むコンピュータは、セキュアコンピュータとも呼ばれ得る。セキュアコンピュータは、1つまたは複数のセキュアエンクレーブ、例えば、コンピュータで実行される各アプリケーションに対して1つのセキュアエンクレーブを含むことができる。

10

【0049】

一般に、エンクレーブを他のプロセスおよび他のエンクレーブから確実に隔離することは、セキュアエンクレーブ技術の目標である。

【0050】

セキュアエンクレーブは、ハードウェア（CPU、メモリ、レジスタ、キャッシュなど）および／またはソフトウェア（プログラムされた回路）で構成された隔離された環境である。セキュアエンクレーブは、コールゲートまたはファイアウォールと呼ばれることもある、特別に構成されたハードウェアおよびソフトウェア要素を介してアプリケーションプログラムによってアクセス可能である。セキュアエンクレーブへのアクセスは、暗号鍵を介して制御することができ、暗号鍵の一部は、製造時に構成されるハードウェア要素に存在することができる。悪意のあるエンティティは、セキュアエンクレーブのブートプロセス中に鍵を抽出しようと試みる可能性がある。鍵を抽出するためのリバースエンジニアリングまたは他のそのような攻撃は、繰り返される鍵要求を禁止すること、および／またはそのような要求間の時間を長くすることによって阻止することができる。場合によっては、一組の鍵は、ハードウェア要素の特定の組に関連付けられ得る。

20

【0051】

セキュアエンクレーブに導入されるデータ（およびコンピュータプログラム）が暗号化されることを要求することによって、追加の保護を達成することができる。さらに、セキュアエンクレーブから出力されたデータも暗号化される。セキュアエンクレーブに一旦注入されると、暗号化されたデータは、セキュアエンクレーブ内で復号化され、処理され、結果は出力に含んで暗号化され得る。したがって、隔離されたセキュアエンクレーブは、上述した実行時の脆弱性の問題を解決する。

30

【0052】

セキュアエンクレーブ内のデータを保護する追加の手段は、復号鍵がセキュアエンクレーブの外部で知られないように保護することによって、セキュアエンクレーブ内のデータを復号化するプロセスをより安全にすることを要求することによって導入することができる。すなわち、セキュアエンクレーブインフラストラクチャの外部のエンティティは、復号鍵にアクセスすることを禁止される。

【0053】

このようにして、注入剤がセキュアエンクレーブのファイアウォールの制約を満たすときに、暗号化されたデータをセキュアエンクレーブに注入することができる。セキュアエンクレーブは、注入されたデータを復号化してそれを処理するために使用され得る復号鍵を含む。セキュアエンクレーブは、結果を出力する前に、セキュアエンクレーブ内で利用可能な暗号鍵を使用して処理アクティビティの結果を暗号化することができる。

40

【0054】

個人データを保護する問題に対処する別の技術は、データを非特定化または匿名化することである。この技術は、個人データをランダムデータで置き換えること、例えば社会保障番号をランダムな数字で置き換えることに依存している。そのような技術は、構造化データセットにおいて使用され得る。例えば、患者の名前、社会保障番号、および心拍数を

50

含む構造化データセットは、属性「名前」および「社会保障番号」の値を非特定化することによって匿名化することができる。

【0055】

構造化データセットにおける非特定化技術は、以下のように処理能力の損失をもたらす。

【0056】

構造化データセットは、多くの場合、最大の処理上の利点を得るために他の構造化データセットと組み合わせる必要がある。例として、2つの構造化データセット（氏名、SS #、心拍）および（名称、SS #、重み）を考える。2つのデータセットを組み合わせることにより、患者のより完全なデータ記録を得ることができる。すなわち、2つのデータセットに表された患者を関連付けることによって、2つのデータセットに固有の関係を利用することができる。2つのデータセットを非特定化するプロセスは、固有の関係を失う患者を匿名化することにつながる。

10

【0057】

上記の例を続けると、固有の関係を維持するために、非特定化を実行するエンティティは、2つのデータセット内の表された患者に同じランダムデータを割り当てることができる。すなわち、匿名化エンティティは、患者、例えばジョンが2つのデータセット内の特定のデータによって表されていることを知っている。これは、匿名化を行うエンティティの知識が脆弱になることを意味する。

【0058】

したがって、構造化データを非特定化すると、悪意のある計算エンティティによって悪用される可能性のある脆弱性が導入される可能性がある。

20

【0059】

従来の非特定化技術の別の欠点は、医療ノート、注釈、病歴、病理データなどの非構造化データセットには適用されないことである。大量のヘルスケアデータは、構造化されていないデータセットからなる。本開示の後の部分では、機械学習および人工知能技術を使用して非構造化データセットを非特定化する技術が開示される。

【0060】

構造化されていないデータセットを非特定化することの1つの結果は、結果として得られるデータセットがいくらかの残りのプライベートデータを含み得ることである。一実施形態では、構造化されていないデータセットの非特定化は、非特定化の有効性の尺度を導出する統計解析を受ける。すなわち、データセットが非特定化された確率の尺度を取得することができる。

30

【0061】

実施形態では、エンティティAは、確率尺度 p に対するデータセットを非特定化し、それをエンティティBに提供する。後者はまた、エンティティCから1つまたは複数のコンピュータプログラムを受信する。エンティティBは、エンティティCから受信したコンピュータプログラムを使用してエンティティAから受信したデータを処理し、処理の結果を別のエンティティDに提供する。（実施形態において、A、B、CおよびDは、原則として別個のエンティティであってもよく、実際には、エンティティA、B、CおよびDのうちの1つまたは複数は、相互の合意を通じて協力することができる。）

40

【0062】

本発明の実施形態は、エンティティBが、その処理がデータと関連付けられる確率 p を維持することをエンティティA（およびC、およびD）に保証することを可能にする。

【0063】

さらに、エンティティBを含まないプロセスでは、エンティティAは、そのデータセット上でエンティティCのコンピュータプログラムの使用を承認することができる。

【0064】

本発明の実施形態は、エンティティBが、問題のデータセットがエンティティCによって提供されるコンピュータプログラムによってのみ処理され、データセットが他のコンピュータプログラムによって処理されなかったことをエンティティC（およびA、およびD

50

）に保証することを可能にする。さらに、エンティティ B は、エンティティ C によって提供され、基礎となるデータセットを処理するために使用されたコンピュータプログラムが、いかなる方法でも変更、変更または修正されなかったこと、すなわち、処理中に使用されたコンピュータプログラムのバイナリ画像が、提供されたコンピュータプログラムのバイナリ画像と同一であったことを他のエンティティに保証することができる。すなわち、この有効化は、受信したコンピュータプログラムの起源を維持する。

【0065】

さらにまた、不可視性は、以下の条件を満たす性質に相当する。

1. エンティティ B は、エンティティ A によって提供されるデータセット、エンティティ C によって提供されるコンピュータプログラム、およびエンティティ D に提供される出力にアクセスできなかったことをエンティティ A、C、および D に保証することができる。

10

2. エンティティ B は、エンティティ A がデータセット A にのみアクセスし、C によって提供されるコンピュータプログラムまたはエンティティ D に提供される出力のいずれにもアクセスしなかったことをエンティティ C および D に保証することができる。

3. エンティティ B は、エンティティ C がそのコンピュータプログラムにのみアクセスし、A によって提供されるデータセットまたは D に提供される出力のいずれにもアクセスしなかったことをエンティティ A および D に保証することができる。

4. エンティティ B は、エンティティ A が提供したデータセット A にのみアクセスし、D に提供された出力または C によって提供されたコンピュータプログラムのいずれにもアクセスしなかったことをエンティティ C および D に保証することができる。

20

【0066】

さらに、上記の様々な保証は、暗号技術に基づいて、検証可能かつ偽造不可能なデータ機器、すなわち証明書の形態で提供される。

【0067】

図 1 A に示す本発明の実施形態は、第 1 のエンティティ 1 A 1 0 0 が、第 2 のエンティティ 1 A 1 0 1 から（所定の非特定化確率で）データセットを受信する点で発生し、第 3 のエンティティ 1 A 1 0 3 から受信されたコンピュータプログラムを使用して 1 つまたは複数のデータ処理段階 1 A 1 0 2 を通って延在し、処理の結果が第 4 のエンティティ 1 A 1 0 4 によって受信される点で終了する「計算の信頼チェーン」1 A 1 0 5 を構築することを可能にする。さらにまた、信頼チェーン 1 A 1 0 5 は秘匿性を満たす。したがって、信頼チェーンは、入力確率尺度、受信したコンピュータプログラムの起源、および不可視性特性を保存する概念を具体化する。

30

【0068】

一般性を失うことなく、説明を容易にするために、図 1 A の例示的な実施形態では、第 1 のエンティティ A は「オペレータ」とラベル付けされ、第 2 のエンティティは「データプロバイダ」とラベル付けされ、第 3 のエンティティは「プログラムプロバイダ」とラベル付けされ、第 4 のエンティティは「データサイエンティスト」とラベル付けされる。処理を実行する機器は、「連合パイプライン」とラベル付けされる。「連合」という用語は、パイプラインが複数のエンティティから入力を受信し、複数のエンティティに出力を提供することができることを示す。

40

【0069】

本開示は、とりわけ、データセットの入力非特定化確率、入力コンピュータプログラムの起源、ならびに計算に関与する様々なデータおよびコンピュータプログラムの不明性を維持する「連合パイプライン」（ソフトウェア技術および／またはハードウェア／ファームウェア構成要素を使用して実装される）について説明する。

【0070】

場合によっては、連合パイプラインから出力データセットまたは結果を取得したデータサイエンティスト（例えば、エンティティ 1 A 1 0 4、図 1 A を参照されたい）は、出力データセットを処理し、その結果を第三者と共有することを望むことがある。上記で説明したように、データサイエンティストは連合パイプラインから出力を受信するため、出力

50

は（一連の）認証、すなわち信頼チェーンに関連付けられることに留意されたい。データサイエンティストが受信した出力を処理し、それを第三者と共有することを望む場合、第三者は、信頼チェーンを新たに処理された結果に拡張することを求めることができる。

【0071】

すなわち、第三者は、連合パイプラインから受信した出力が実際に新しいコンピュータプログラムへの入力として使用され、第三者に提供された出力がプログラムによって出力されるという証明を取得することを望むことができる。すなわち、データサイエンティストは、連合パイプラインに関連する信頼チェーンを拡張するように第三者によって要求され得る。データサイエンティストが連合パイプラインに関連付けられていない場合、連合パイプラインシステムで使用される方法とは独立した信頼チェーンを拡張する方法が必要である。

10

【0072】

図5は、この課題を示す。データサイエンティストは、結果を共有する場合、データサイエンティストによって提供された可能性がある特定のコンピュータプログラム（P1）が実行されたこと、および認証を有する連合パイプラインによって提供された入力データセット（#1）のソースを受け入れ検証したことを結果の受取人に信頼してもらいたい。プログラムP1は、例えば、データセット#1の一部として提供されたシリアル番号を既知の外部データリポジトリと照合してもよい。プログラムP1の申し立てられた実行は、データセット#2をもたらす。

【0073】

20

さらに、データサイエンティストは、データセット#2が（データサイエンティストによって提供された可能性がある）コンピュータプログラムP2によって処理され、プログラムP2の申し立てられた実行が最終出力データセット（図5）をもたらしたと受取人が信頼することを望む可能性がある。

【0074】

D. Genkinらによる『Privacy in Decentralized Cryptocurrencies,』C. of the ACM, 2018は、参照によりその全体が本明細書に組み込まれ、プログラムP1およびP2の実行を検証するための例示的な技術を示している。ブルーバと呼ばれるソフトウェアモジュールは、プログラムP1およびP2を実行可能なコンピューティング環境を提供する。そのような実行時に、ブルーバは2つの出力、すなわち（1）プログラムP1およびP2の出力、ならびに（2）プログラムP1および／またはP2の実行の証明と呼ばれるデータオブジェクトを生成する。

30

【0075】

さらに、ブルーバは、任意の第三者に提供され得るベリファイア（図6を参照されたい）と呼ばれるソフトウェアモジュールも提供する。ベリファイアは、証明を入力とし、バイナリ「はい／いいえ」回答を出力する。回答「はい」は、問題のプログラムが実行され、入力された証明オブジェクトが生成されたことを示す。「いいえ」という応答は、申し立てられた実行の証拠を検証できなかったことを意味する。

【0076】

40

したがって、D. Genkinらは、コンピュータプログラムの申し立てられた実行が、申し立てられた実行の証明をベリファイアシステムに提出することによって検証され得るシステムおよび方法を示す。証明オブジェクトは暗号化オブジェクトであり、（申し立てられた実行が検証可能であるというメタステートメント以外の）基礎となるデータまたはプログラムに関する情報を漏洩しない。

【0077】

実施形態では、コンピュータプログラムPは、2つの企業E₁とE₂との間にポリシーを組み込むものとして合意されてもよい。旧企業E₁は、ここで、プログラムPを実行させ、上述のブルーバ技術を使用してその申し立てられた実行の証明 π を生成することができる。企業E₂は、 π を（上述の検証技術を使用して）検証し、プログラムPが実行され

50

たことを信頼することができ、それにより、合意されたポリシーが実装されたことを信頼する。

【0078】

図1Bは、アプリケーションの観点からのセキュアエンクレーブの論理アーキテクチャを示す。アプリケーション100は、それ自体のコード、データ、およびセキュアエンクレーブを含む。アプリケーション100は、(1)従来のコンピュータにおける典型的なアプリケーションとして動作する非セキュア部分と、(2)セキュアエンクレーブ内で動作するセキュア部分と、の2つの部分に論理的に分割される。アプリケーションの非セキュア部分のコードは、セキュアエンクレーブが作成され、特定のブート画像がセキュアエンクレーブにロードされて実行されることを要求することができる。その後、セキュアエンクレーブ内の実行終了時の制御は、呼び出しポイントに戻される。特権システム200(OS、BIOS、SMM、VMなどを含む)は、セキュアエンクレーブにアクセスすることが阻止される。

10

【0079】

いくつかの実施形態では、以下の方法を実行して、セキュアエンクレーブにコードおよびデータを取り込むことができる。

方法[セキュアエンクレーブの作成および投入]

- (1) アプリケーションのセキュア部分をコンパイルする、
- (2) セキュアエンクレーブを生成するためのコマンドを発行する(例えば、基礎となるハードウェア/OS命令セットを使用する)、
- (3) 予め指定されたライブラリから予めプロビジョニングされたコードをロードする、
- (4) ステップ1からのコンパイル済みコードをセキュアエンクレーブにロードする、
- (5) 適切な認証情報を生成する、
- (6) セキュアエンクレーブの画像および認証情報を保存する。

20

【0080】

いくつかの実施形態では、以下の方法を実行して、セキュアエンクレーブ内でコードを実行することができる。

方法[セキュアエンクレーブ内でコードを実行する]

- (1) アプリケーションの非セキュア部分(例えば、アプリケーション100)をセキュア画像と共にコンパイルする、
- (2) アプリケーションを実行する、
- (3) アプリケーションは、セキュアエンクレーブを作成し、画像をセキュアエンクレーブにロードする、
- (4) 様々な認証情報を検証する。

30

【0081】

セキュアエンクレーブのハードウェアおよびソフトウェアコンポーネントは、エンクレーブ内のコードおよびデータの完全性および機密性を保護することによってデータプライバシーを提供する。入口点および出口点は、アプリケーションコードをコンパイルするときに予め定義される。セキュアエンクレーブは、そのアプリケーションから暗号化されたデータを送信/受信することができ、暗号化されたデータをディスクに保存することができる。エンクレーブは、そのアプリケーションのメモリにアクセスすることができるが、その逆は真ではなく、すなわち、アプリケーションは、エンクレーブのメモリにアクセスすることができない。

40

【0082】

エンクレーブは、指定されたコンピュータ上で実行することができる自己完結型の実行可能ソフトウェアである。例えば、エンクレーブは、外部リソースまたは共有リソースを呼び出すのではなく、動作中に使用するリソース(例えば、コードライブラリ)を含んでもよい。場合によっては、エンクレーブに対してハードウェア(例えば、グラフィック処理ユニットまたは特定の量のメモリ)およびオペレーティングシステム(例えば、Linux(登録商標)バージョン2.7またはAlpineLinuxバージョン3.2)の

50

要件が指定されてもよい。

【0083】

図2は、いくつかの実施形態によるヘルスケアデータを処理するためのユースケースのシナリオを示す。データプロバイダは、ユーザの名前、住所、患者ID番号、郵便番号、および他のユーザ固有のデータなど、属性の一部が個人データであり得るデータ記録を含むデータベース200を提供する。データベース200は、場合によってはクラウドサーバ環境に存在する、集合的にパイプライン210と呼ばれる1つまたは複数のコンピュータに接続される。

【0084】

図2はまた、エンクレープ220の外部に存在する（プログラムプロバイダによって提供される）コンピュータプログラム270を示す。これは、前述したように、セキュアエンクレープに含まれていないため、非セキュアプログラムである。「セキュアエンクレープを作成および投入する」方法を使用して、プログラム270は、パイプライン210にセキュアエンクレープを作成し、そのセキュアアプリケーション部分であるアプリケーション230をそれに投入する。アプリ230はセキュアエンクレープ内にあるため、定義上、セキュアである。

【0085】

方法「セキュアエンクレープの作成および投入」で説明したように、予めプロビジョニングされたソフトウェアをセキュアエンクレープにロードすることもできる。SE220は、とりわけ、TLS（トランスポートレベルセキュリティ）接続のための1つのエンドポイントとして機能する予めプロビジョニングされたソフトウェア240-2を含む。TLS接続のための第2のエンドポイント240-1は、データベース200に存在する。（TLSの代わりに、任意の安全なネットワーク接続技術、例えばhttps、VPNなどを使用することができる）

【0086】

TLS接続は、アプリケーション230によってデータベース200からデータを取得するために使用され得る。アプリケーション230はまた、データ記録の受信を実行するためのプロキシ機構を含むことができる。

【0087】

さらに、SE220は、その機能が後述される予めプロビジョニングされたソフトウェアモジュールPA250（ポリシーエージェント）およびAC260（アクセスコントローラ）を含む。

【0088】

したがって、SE220のプログラムアプリケーション230は、TLSエンドポイント240-1および240-2を使用してデータベース200からデータを取得することができる。TLS技術は、転送されるデータが安全であることを保証する。データベース200は、暗号化されたデータ記録を含むことができる。したがって、アプリ230は、暗号化されたデータ記録を受信する。動作中、アプリケーション230は、受信したデータ記録を復号化し、そのプログラムされたロジックに従ってそれら进行处理する。（復号化を行う方法については後述する）

【0089】

上述した方法「セキュアエンクレープでコードを実行する」を使用して、アプリケーション230を呼び出すことができ、次いでデータベース200からデータを取り出して復号化することができる。処理の結果は、ポリシーエージェントPA250の制御下にあるエンティティラベル付きデータサイエンティスト280に向けられてもよい。一般に、PA250はポリシーマネージャ280と連携して動作する。PA250およびポリシーマネージャ280の機能および相互運用については、後により詳細に説明する。

【0090】

いくつかの実施形態では、ポリシーマネージャ280はそれ自体のセキュアエンクレープ290内に存在してもよい。

10

20

30

40

50

【0091】

図2は、2つのセキュアエンクレーブ220および290を含むパイプラインを示す。実施形態では、パイプラインは、1つまたは複数のセキュアエンクレーブを含むことができる。さらに、1つまたは複数のセキュアエンクレーブは、相互接続されてもよい（例えば、計算作業タスクを分散させるために）。例えば、1つまたは複数のセキュアエンクレーブは、計算タスクの同時実行を達成するためのマップ低減構成として知られるものを達成するために相互接続することができる。パイプラインは、1つまたは複数のコンピュータを使用して実装されてもよく、例えば、セキュアエンクレーブは、複数のコンピュータ、例えばクラウドサーバ環境に存在してもよい。図2は、エンクレーブに接続された単一のデータベース200を示す。実施形態では、1つまたは複数のデータベースを1つまたは複数のエンクレーブにそのように接続することができる。

10

【0092】

要約すると、計算タスクは、セキュア部分および非セキュア部分を有するアプリケーションプログラムとして符号化することによって達成され得る。呼び出されると、アプリケーションの非セキュア部分は、1つまたは複数のセキュアエンクレーブを作成し、そのセキュア部分をセキュアエンクレーブに注入し、その実行を呼び出す。アプリケーションのセキュア部分は、エンクレーブまたは他のエンクレーブに接続された（予めプロビジョニングされた）データベースからのデータにアクセスすることができる。そして、アプリケーションのセキュア部分は、受信したデータを復号化する。次いで、処理は、場合によっては相互接続されたエンクレーブの配置を利用して、アプリの論理に従って進行する。結果は、ポリシーエージェントを介してデータサイエンティストに提示される。

20

【0093】

データが決して解読されず、暗号化されたデータに対して処理が進行するFHEデータセット手法と比較して、図2に示す構成では、エンクレーブ内のデータは暗号化された形態であり、処理の前に復号化される。これは、結果が外部エンティティと共有される前に再暗号化されてもよい。したがって、図2の構成は、より効率的であり、FHEと比較して処理速度の改善を達成することができる。

【0094】

上述のパイプライン技術は、個人データおよび個人データを含み得るデータセットに対して計算を実行することを可能にする。パイプライン技術の一態様は、セキュアエンクレーブ内のデータ（およびプログラム）が監視不可能であること、すなわち、ポリシーマネージャ（またはそのコホート、ポリシーエージェント）によってポリシー制御を受けることである。さらに、プログラムの実行の結果として生成される出力はまた、ポリシーに従って方向付けられてもよい。

30

【0095】

例として、とりわけ、患者の体重、身長、生年月日、および住所を含むデータセットに記憶された個々の患者の肥満度指数（BMI）を計算するパイプラインで実行される計算を考える。次いで、計算は、様々な米国の郡にわたる平均BMIを計算するために進む。

【0096】

これらの計算は個人および個人の患者データを含むため、計算はプライバシー規制の対象となり得る。以下の例示的な例など、様々なタイプの出力が望ましい場合がある。（1）平均BMIが最も高い米国の5つの郡のデータセット、（2）「過体重」BMIを有する住所の5人の患者のデータセット、（3）マサチューセッツ州ノーフォーク郡からの郵便番号およびBMIを含む患者のデータセット、（4）マサチューセッツ州デダムからの25～45歳の「過体重」BMIを有する患者のデータセット、または（5）マサチューセッツ州デダムのアライドストリートからの体重、身長および年齢を含む患者のデータセット。いずれの場合も、計算への入力個人データおよび個人データを含み得るデータセットであり、出力は個人データおよび個人データも含み得るデータセットである。

40

【0097】

上記の第1の出力データセットは、郡人口のレベルに集約されたデータを列挙し、PI

50

データ属性を含まない。結果は、いかなる単一の個人のデータ記録からも独立している。結果は母集団に関係する。したがって、ポリシーは、そのようなデータセットが出力され得る、すなわち平文として出力され得ることを提供し得る。

【0098】

一方、上記の第2の出力データセットは、(1)個人を識別可能な情報、すなわち街路住所を含み、(2)データセット内の項目の数、すなわち出力セットのカーディナリティは小さい。悪意のあるエージェントは、そのようなデータセットから特定の個人を分離することができる可能性がある。この場合、そのような要求を許可しないようにポリシーを形成することができる。

【0099】

すなわち、出力されたデータセットのカーディナリティに制約を課す、プライバシーパラメータと呼ばれるパラメータKが提供されてもよい。したがって、そのP I E属性がK人未満の個人を識別する場合、出力されたデータセットは許可されない可能性がある。

【0100】

追加的または代替的に、出力データセットは、セキュアエンクレーブ内で、意図された受取人、例えば、データサイエンティストによって提出されたクエリに回答するコンピュータプログラムと共に、暗号化された形態で提供されてもよい。次いで、エンクレーブは、(非セキュア)アプリケーションプログラムを使用してエンクレーブ内の(セキュア)プログラムに問い合わせ、エンクレーブの応答を受信することができる。したがって、データサイエンティストは、患者データを見ることはできないが、彼のクエリに対する応答を受信することができる。さらに、安全なプログラムの応答は、出力データセットの選択および所定の「ビュー」のみを明らかにするように制約されてもよく、「ビュー」は、データベースシステムにおけるビューの一般的に受け入れられている概念に対応し得る。あるいは、出力データセットは、最初にF H Eを使用してデータセットを暗号化することによって、セキュアエンクレーブに封入することなくデータサイエンティストに提供されてもよい。

【0101】

上記の第3の出力要求では、データはある国の郵便番号にわたって集約されているため、そのような患者の数が十分に多い場合には、プライバシーの懸念を生じない可能性がある。そのような例では、出力データセットのサイズに制約を課すポリシーを形成することができ、例えば、出力データセットは、少なくとも20人の患者に関するデータを含まなければならない。同様のポリシーを第4および第5の出力要求にも使用することができる。

【0102】

いくつかの実施形態では、データセットのカーディナリティが課された制約された制限未満である場合、出力されたデータセットにランダムなデータ記録を追加することを提供できるポリシーを形成することができる。すなわち、最小サイズ、例えば20人の出力を達成するのに十分な記録が出力データセットに含まれるように制約が課される。

【0103】

出力要求(例えば、上記の第3、第4、および第5の出力要求)が一連の要求として発行され、出力が単一のエンティティ(例えば、データサイエンティスト)または結託して出力を共有する複数のエンティティによって収集される場合、さらなる課題が発生する可能性がある。出力要求は、より小さい母集団サイズに連続的に適用される計算データセットを計算するため、そのような「絞り込み」計算を使用して特定の個人に関する情報を取得する可能性がある。

【0104】

ますます増加する絞り込み(またはより正確な応答)のシーケンスが最終的に個々の情報を漏洩することが文献(Cynthia Dwork, Differential Privacy: A Survey of Results, International Conference on Theory and Applications of Models of Computation, 2008)に示されている。

10

20

30

40

50

【0105】

図3は、上述した様々なポリシーを示している。これらのポリシーは例示を目的としており、実装される実際のポリシーは異なってもよい。

【0106】

いくつかの実施形態では、ポリシーエージェントは、パイプラインの1つまたは複数のセキュアエンクレーブに予めプロビジョニングされたソフトウェアとして含まれるように構成され得る。ポリシーエージェントは、ポリシーマネージャ（後述）からそのポリシーを受信し、出力されたすべてのデータセットにそのポリシーを課す（そのいくつかの例は上記の説明で提供されている）。様々な（企業）当事者間の帯域外契約を使用して、当事者がポリシーエージェントに含まれる予めプロビジョニングされたポリシーを指定および表示できるようにすることができる。

10

【0107】

ポリシーエージェントソフトウェアはまた、エンクレーブ内で実行されるプログラムによって行われるすべてのアクセスおよび他のアクションを記録する、すなわちログする。

【0108】

ポリシーマネージャは、1つまたは複数のポリシーエージェントを管理するように構成され得る。ポリシーマネージャは後述する他の機能を遂行してもよい。簡単にするために、本開示は、マスタースレーブ構成においてパイプライン内のすべてのポリシーエージェントを管理するパイプライン用の単一のポリシーマネージャを示す。

【0109】

本開示はまた、例示目的のためにパイプラインのオペレータのドメインで実行されるポリシーマネージャを示し、様々な代替形態が可能である。いくつかの実施形態では、ポリシーマネージャは、オペレータ、データプロバイダ、プログラムプロバイダまたはデータサイエンティストのいずれかによって制御される任意のドメインに実装され得る。ポリシーマネージャが非集中型技術を用いて実装される場合、ポリシーマネージャの制御は上記ビジネスエンティティのうちの1つまたは複数にわたって非集中型とすることができる。本開示で使用する「非集中型」という用語は、ポリシーマネージャを制御するポリシーが単一の当事者ではなく複数の当事者によって提供され得ることを意味する。

20

【0110】

例えば、図7は、ポリシーマネージャの非集中型制御の1つの例示的な実施形態を示す。図7は、その行がグループを記述するポリシーマネージャに含まれる表を示す。グループは、協調エンティティおよびそれらの協調に関連する要素の集合である。協調エンティティのグループは、それらの個々のポリシーを介してポリシーマネージャの制御を行う。1行目は、アルゴリズムa1を提供するメンバーとしてA1という名前のエンティティを有するグループ1という名前のグループを示し、データd1を提供するD1という名前の別のメンバーを示す。二人のメンバーによって提供されるデータおよびアルゴリズムは処理されており、ロード画像はパイプラインにロードされるように準備されている。準備されたロード画像は安全なセキュアストレージに格納され、リンクL1を使用してアクセスすることができる。

30

【0111】

いくつかの実施形態では、ポリシーエージェントは、ポリシーマネージャを用いてその状態を記録してもよい。さらに、ポリシーマネージャは、レギュレータおよび／または第三者エンティティが個々のポリシーエージェントの記録された状態を検査することを可能にするように構築されてもよい。したがって、規制当局および第三者エンティティは、データセットが出力された制約を調べることができる。実施形態では、ポリシーマネージャの可能な実施態様はブロックチェーンシステムであり、その台帳は不変データ記録を含むことができる。

40

【0112】

上述のシナリオでは、ポリシーは、データサイエンティストがセキュアエンクレーブに囲まれた出力データセットを受信できることを指示することができる。これは、データセ

50

ット内のデータがデータサイエンティストにとって不透明であることを意味する。後者は、新たな要求をエンクレーブに注入することによって、エンクレーブ内の出力データセットに対して追加の出力要求を自由に実行することができる。これらの場合、出力されたデータセットがPIIデータを有していないか、またはプライバシーパラメータ制約に違反していない場合、データセットは制約されなくなり、データサイエンティストが利用できるようにすることができる。

【0113】

いくつかの実施形態では、データサイエンティストまたは他のリクエスタは、エンクレーブ内に含まれるデータセットの内容を閲覧することができる。エンクレーブのコンテンツは、エンクレーブをウェブブラウザに接続し、エンクレーブのコンテンツをウェブページとして表示させることによってリクエスタに利用可能にすることができる。これにより、要求者がブラウザの状態を保存またはコピーすることが防止される。しかしながら、場合によっては、リクエスタは、ブラウザページの視覚画像を取得することができる。

10

【0114】

いくつかの実施形態では、データサイエンティストは、データ要求を送信することができ、その後、キュレーションサービスを使用して管理される。キュレーションサービスがデータ要求をプライバシー保護であるとみなす場合、データ要求はエンクレーブ内のデータセットを使用して処理され得、出力されたデータセットは制約のないデータセットとしてデータサイエンティストに提供され得る。このようにして、キュレーションサービスは、提出されたデータ要求が良性であること、すなわち、データ要求がプライバシー規則に違反する出力を生成しないことを確認および保証する。

20

【0115】

上述したように、エンクレーブを使用してプライベートデータを処理することに関連するさらなる課題は、エンクレーブの内部のプロセスは不可解であるため、エンクレーブ内で実行される計算についてポリシーを提供できるかどうかである。例えば、図2に関して上記の一般的な説明に従って、セキュアエンクレーブ技術のユースケースを考える。暗号化されたデータを所有する第1の企業は、エンクレーブにデータを格納することができる。エンクレーブ内のデータは、パイプラインを提供する第2の企業によって使用するために処理および準備されてもよく、コンピュータプログラムは、第3の企業によって提供されるデータを処理する。ここで、データサイエンティストは、データ要求をエンクレーブに注入し、結果として出力されたデータセットを期待することができる。上記で説明したように、一例では、出力されたデータセットは、エンクレーブに囲まれたデータとしてデータサイエンティストに提供されてもよい。別の例では、出力されたデータセットは、データの暗号化されたストアとして提供されてもよい。後者の場合、データサイエンティストは、データへのアクセスを提供するために復号鍵を提供されてもよい。これらすべての行動は、第1、第2または第3の企業のいずれかによって事前に決定されたポリシーの影響を受ける。

30

【0116】

さらに、問題のポリシーは、データを処理し、データサイエンティストによって出力されたデータセットを受信するためのアクセスが認可されなければならないことを要求することができる。すなわち、データサイエンティストによるアクセスが認証されなければならない。それらの部分におけるデータサイエンティストは、データの完全性がデータ処理パラダイムにとって重要であるため、データ要求が指定されたデータプロバイダによって提供されたデータで動作することを保証されることを必要とする場合がある。特に、データサイエンティストが出力された結果を第三者と共有することを意図している場合、データサイエンティストは、入力データの完全性および結果が特定のデータ要求を実行することによって得られたという事実の前者を保証する必要がある。規制当局は、データを保存および処理するプロセス全体が透明でなければならず、調査および事後承認に利用可能にされなければならないことを要求することができる。

40

【0117】

50

上述した様々な懸念に対処するために、図4Aのワークフロー図に示すように、オーケストレーション方法を実行することができる。以下のエンティティがワークフローに関与する。(1) データプロバイダ、すなわち、データを所有するエンティティ、(2) オペレータ、すなわちパイプライン技術を提供するエンティティ、(3) データを処理するためのコンピュータプログラムを提供するプログラムプロバイダ、(4) データサイエンティスト、すなわち、出力された結果を取得したいエンティティ、および(5) ポリシーエージェントを制御するソフトウェアモジュールを含むことができるポリシーマネージャ。

【0118】

図4Aを参照すると、ステップ1、2、3、および4において、データプロバイダ、データサイエンティスト、プログラムプロバイダ、およびオペレータは、それぞれそれらのポリシーを指定する。ステップ5において、ポリシーマネージャはポリシーエージェントを開始する準備をする。ステップ6において、オペレータは、新しいパイプラインを作成し、ステップ7において、パイプラインの作成についてオーケストレーションの参加者に通知する。参加者は、データ、プログラム、およびポリシーをパイプラインに投入することができる。パイプラインは、予めプロビジョニングされたソフトウェアライブラリでも開始されることに留意されたい。

【0119】

図4Bを参照すると、データプロバイダと、パイプラインと、プログラムプロバイダによって提供されるセキュアアプリケーションプログラムと、ポリシーマネージャと、データサイエンティストと、ポリシーエージェントとの間のオーケストレーションが示されている。

⇒ ステップ1. ポリシーマネージャは、図4Aのステップ5で準備したポリシーエージェントを開始する。

⇒ ステップ2. セキュアアプリケーションが処理要求を開始する。

⇒ ステップ3. 開始要求を記録する。

⇒ ステップ4. ポリシーエージェントが、処理要求に関連する適切なポリシーおよびアクセス認証情報を選択する。

⇒ ステップ5. ポリシーエージェントが(ポリシーマネージャの助けを借りて)認証情報を検証する。認証情報が満たされない場合には、要求は終了する。

⇒ ステップ6. パイプラインが処理要求を実行し、データを格納する。

⇒ ステップ7. パイプラインが、要求された出力が利用可能であることをデータサイエンティストに通知する。

鍵管理

【0120】

公開鍵暗号は、通常、秘密鍵および公開鍵と呼ばれる一対の相補鍵に依存する。後者は、任意の関係者に分配されてもよい。前者、すなわち秘密鍵は常に秘密に保たれる。例えばAliceによって配布された公開鍵を使用して、別の当事者、例えばBobは、Aliceのみが自身の秘密鍵を使用してメッセージを復号化することができるという認識において、メッセージを暗号化してAliceに安全に送信することができる。Bobによって暗号化されたメッセージを復号化するために他の鍵を使用することはできない。前述したように、秘密鍵の所有権は主要な関心事であり、いくつかの技術がこのトピックに関連する文献で議論されている。

【0121】

セキュアエンクレーブ技術は、(公開鍵に対応する)秘密鍵が常にセキュアエンクレーブ内に存在することを保証することによって、秘密鍵の所有権の問題に対処するために使用され得る。これは、例えば、第1のセキュアエンクレーブを作成し、それを秘密鍵と公開鍵のペアを作成する公開鍵/秘密鍵暗号化ソフトウェアで予めプロビジョニングすることによって達成することができる。そのようなソフトウェアは、オープンソースのリポジトリを介して利用可能である。次いで、第2のセキュアエンクレーブに存在するコンピュータプログラムは、第1のエンクレーブに、必要な秘密鍵のコピーを(セキュアチャネル

を使用して) 提供するよう要求することができる。したがって、秘密鍵は、セキュアエンクレープインフラストラクチャの外部には決して存在せず、常にセキュアエンクレープ内に存在し、セキュアチャンネルを使用してそれらの間で送信される。

【0122】

いくつかの実施形態では、ポリシーマネージャに公開／秘密鍵ソフトウェアを予めプロビジョニングし、図2(280を参照)に示すようにポリシーマネージャをセキュアエンクレープに封入してもよい。

【0123】

次いで、セキュアエンクレープは、そのポリシーエージェントに秘密鍵を要求することができる。ポリシーエージェントは、上述したように、ポリシーマネージャと連携して動作し、ポリシーマネージャからポリシーマネージャを要求することができる。セキュアエンクレープで実行されるコンピュータプログラムは、データプロバイダから受信することができる暗号化データを復号化するために秘密鍵を必要とする場合がある。それは、そのポリシーエージェントに要求することができ、次いで、ポリシーエージェントは、復号化目的のために必要な秘密鍵を提供することができる。

【0124】

先に説明したように、ハッシュ関数またはハッシュアルゴリズムと呼ばれる暗号化技術が存在し、これは、メッセージと呼ばれることが多い平文の文字列を取り、それを16進数のシーケンス、すなわち、数字のシーケンス[0~9、A~F]として暗号化することができる。公開されているハッシュ関数の例は、MD5、SHA-256、SHA-512である。後者の2つの関数は、それぞれ長さ256および512の鍵を使用する。上述したように、鍵の長さは、悪意のある攻撃に耐える暗号技術の強度を保証する要因である。

【0125】

平文を16進数にマッピングするハッシュ関数の1つの特性は、異なる平文を同じ桁にマッピングしないことである。したがって、平文は、固有のシグネチャ、すなわち平文で動作するハッシュ関数の出力を入力として有することができる。

【0126】

プログラムおよびデータを含むセキュアエンクレープが平文を含むとみなすことができる場合、すべてのセキュアエンクレープが固有のシグネチャを有することになる。したがって、セキュアエンクレープの内容に適切なハッシュ関数を適用することによって、そのエンクレープのシグネチャが取得される。シグネチャは、他の異なるセキュアエンクレープがそのシグネチャを有しないという点で一意である。

【0127】

セキュアエンクレープに既知のコンピュータプログラムおよび既知のデータセットが読み込まれる場合には、そのセキュアエンクレープのシグネチャを使用して、セキュアエンクレープのシグネチャを以前に格納されたシグネチャと比較することによって、セキュアエンクレープが既知のデータセット上でプログラムを実行している(または実行された)ことをアサートすることができる。

【0128】

したがって、データプロバイダは、エンクレープのシグネチャが提供される場合、そのデータセットが破損していないかまたは変化しておらず、所定のプログラムによって操作されることが保証され得る。

【0129】

同様に、プログラムプロバイダは、そのプログラムが破損せず不変であることを保証され得る。データサイエンティストは、その出力が所定のデータに対する所定のプログラムによる処理の結果であることを保証され得る。

【0130】

ポリシーマネージャは、関連する復号鍵へのアクセスを拒否することによってオペレータがセキュアエンクレープのコンテンツにアクセスすることを許可しないようにプログラムされ得るので、パイプラインのオペレータはセキュアエンクレープのコンテンツを閲覧

10

20

30

40

50

または編集することができない。

【0131】

本開示では、セキュアエンクレーブは、シグネチャを作成するためにポリシーマネージャによって呼び出され得るハッシュ関数を計算するためのソフトウェアで予めプロビジョニングされ得る。次いで、ポリシーマネージャは、様々なエンティティ、例えばデータプロバイダまたはプログラムプロバイダに要求されると、これらのシグネチャを証明書として提供するようにプログラムされ得る。

【0132】

ここで図10を参照すると、初期データセット1001は、セキュアエンクレーブ1001に格納することができ、そこで処理してデータセット1010として出力することができる。データセット1010は、セキュアデータ層1009に存在する。データサイエンティストによって、エンクレーブ1002に1つまたは複数のアプリケーションを注入することができ、データセット1010をそのようなアプリケーションに提供することができる。処理時に、出力されたデータセットは、出力1008として格納され得る。後者は、企業1005が（独自の）アプリケーションを使用して結果を処理しデータセット1007として出力することができるエンクレーブ1004にさらに注入することができる。なお、出力データセット1007は暗号化されている。

【0133】

したがって、企業1005は、エンクレーブ1003に注入されたアプリケーションを実行するか、またはデータセット1008を異なるエンクレーブ1004に受け取り、その中で独自のアプリケーションを実行するかの選択を有する。

【0134】

すなわち、一連のエンクレーブ1001、1002、1003、および1004（図10）を組み立てることができ、各エンクレーブは、セキュアデータストア1009から暗号化データを受信し、次のラインエンクレーブのためのセキュア（暗号化された）データセットを順に生成する。したがって、元のデータ所有者1000は、処理のためにそのデータ1011を第三者、すなわち企業1005に提供することができ、個人データまたは個人データが漏洩しないことを保証することができる。

【0135】

企業1005は、適切なデータ要求プログラムが実行され、入力データの起源が確認された証明書と共に、データセット上で独自のデータ要求を実行し、処理の結果を顧客に提供する柔軟性を有する。企業1005は、データセット1007の所有権を引き受けることができるが、その後、そのプライバシーに対する法的責任を引き受ける。

【0136】

図10は、一連のエンクレーブを示し、各エンクレーブは、中間セキュアデータ層を介して別のエンクレーブに接続されている。しかしながら、実施形態では、図9に示すように、いくつかのエンクレーブ909および910も同時に実行することができる。さらに、すべてのコードがエンクレーブ内に存在するとは限らず、エンクレーブは、必要に応じて非セキュアコードを含むコンピューティング環境と混合されてもよい、902（図9）を参照。

【0137】

すべてのエンクレーブに利用可能なセキュアデータ層と共に、セキュアメッセージング904、アクセス制御およびポリシーエージェント通信905、ならびに暗号鍵906の交換のために追加の層を提供することができる。これらの追加の通信層は、エンクレーブが互いに漏洩することなく安全に様々な種類のデータを交換できるように設けられる。

【0138】

図8に示す例示的な実施形態を参照すると、第1の企業800は、データセット2Aを取得するために非特定化および匿名化することができるデータセット1を所有する。前述したように、非特定化手順は完全に有効ではない可能性があり、データセット2Aは依然

10

20

30

40

50

として何らかの個人データおよび個人データを含み得る。第1の企業は、第2の企業890による処理のために利用可能にされ得るように、セキュアデータ層810において、2Bとして示されるデータセット2Aのコピーを提供する。

【0139】

企業890は、データセット2Bを受信し、それをエンクレーブ802に格納させ、そこで処理し、さらなる処理の準備をさせることができ、その後、データセット850としてセキュアデータ層810に格納される。

【0140】

エンクレーブ802は、データセット850がエンクレーブ803から出力され、エンクレーブ802への入力として提供されることを意味するエンクレーブ803にパイプライン化される。ここで、エンクレーブ803内のアプリケーションは、データを処理し、出力データセット809として生成することができる。

【0141】

次に、エンクレーブ803は、企業899によって管理されるネットワーク内に存在するエンクレーブ804にパイプライン接続される。すなわち、エンクレーブ803は企業890によって管理され、エンクレーブ804は企業899によって管理される。後者の企業は、追加のデータ811をエンクレーブ804に注入し、また、データセット811を生成するために、入力データセット809と共にデータセット805を処理するためのアプリケーションを注入することができる。計算の結果は、ポリシーエージェント／マネージャの指示に従って、企業899のデータサイエンティストがアクセスできるようにすることができる。

【0142】

図8はまた、企業899が、エンクレーブ804に注入される（おそらく処理の結果として得られる）エンクレーブ803からのデータを提供することができる例示的な実施形態849を示す。これにより、得られた結果をさらなる処理のために再導入すること、すなわち、結果のさらなる処理のためのフィードバックループを可能にすることが可能になる。

【0143】

前述の説明では、様々な実施形態が、複数の当事者によるデータの協調的な記憶、処理、および解析のためのシステムおよび方法を示している。例えば、図8は、協働する3つの企業800、890および899を示す。企業800はデータを提供し、企業890はデータをエンクレーブに格納するインフラストラクチャを提供し、企業899は特定のデータ要求をエンクレーブに注入することによってデータを処理する。一実施形態では、第1の企業によって提供されたデータが協調合意の下で第2の企業に利用可能にされることを保証するために、企業のうちの1つ、例えばインフラストラクチャを提供する企業が信頼される中央信頼モデルが使用される。すなわち、信頼できる企業は、データアクセスおよびデータ処理が様々な所有権および処理上の懸念に従うことを保証する。データプロバイダは、自らのデータが承認された企業によってのみ処理されることを保証されることを望む。データ処理者は、データ要求を非公開にし、処理要求の詳細を競合企業と共有しないことを望む。そのような懸念の維持は、信頼できる企業に転置され得る。そのような実施形態は、集中型信頼モデルと呼ばれることがあり、すなわち、1つの企業またはエンティティに信頼が置かれる。

【0144】

別の実施形態では、複数の企業が信頼されている非集中型信頼モデルが提供されてもよい。そのような信頼モデルは、データプロバイダがデータを提供し、アナライザが提供されたデータを処理するデータ要求、すなわちコンピュータプログラムを提供するオープンな市場において特に適している可能性がある。単一の企業またはエンティティは、非集中型モデルで信頼されるべきではない。むしろ、データおよびアルゴリズムプロバイダを支配する制約が維持されていることを検証するために任意の第三者がアクセスすることができるオープンに利用可能な構造が提供される。

【0145】

図7は、非集中型信頼モデルの1つの例示的な実施形態を示す。図7は、行がグループを記述する表を示す。グループは、協調エンティティおよびそれらの協調に関連する要素の集合である。1行目は、プログラムa1を提供するメンバーとしてA1という名前のエンティティを有し、データd1を提供するD1という名前の別のメンバーを有するグループ1という名前のグループを示している。2つの部材によって提供されたデータおよびアルゴリズムは処理されており、ロード画像はエンクレーブにロードされるように準備されている。準備されたロード画像は安全なセキュアストレージに格納され、リンクL1を使用してアクセスすることができる。

【0146】

上述したように、画像をエンクレーブにロードするためには、データを暗号化するために特定の暗号化鍵が必要である（その対応する復号鍵は、データを復号化するためにエンクレーブによって使用される）。

【0147】

前述の実施形態は例示であり、多くの追加の代替的な実施形態が可能であることを理解されたい。いくつかの実施形態では、上記の連合パイプラインの少なくとも一部は、Intel Security Guard Extensions (SGX) などの保護メモリを提供するハードウェアおよびファームウェア上で実行されてもよく、その実施態様の詳細は、<https://www.intel.com/content/www/us/en/architecture-and-technology/software-guard-extensions.html>で説明されている。いくつかの実施形態では、連合パイプラインの少なくとも一部は、AMDセキュア暗号化仮想化 (SEV) などの分離された仮想マシンを作成する仮想化ソフトウェアを使用して実行することができ、その実施態様の詳細は、<https://developer.amd.com/sev/>で説明されている。いくつかの実施形態では、連合パイプラインは、<https://aws.amazon.com/kms/>でさらに詳細に説明されているAmazon AWS Key Management Service (KMS) などの鍵管理サービスを使用して暗号鍵を管理することができる。しかしながら、ハードウェア、ファームウェア、仮想化ソフトウェア、および鍵管理サービスのこれらの例は、異なる所有権、ポリシー、および認証を有する連合パイプラインを作成するために使用できる暗号プロトコルに基づく分離されたソフトウェアプロセスを独立して作成することはできない。したがって、いくつかの実施形態では、基礎となるハードウェア/ファームウェア、オペレーティングシステム、鍵管理、および暗号アルゴリズムを使用して、セキュアエンクレーブなどのセキュアかつプライベートな分離されたプロセスを達成することができるミドルウェア（例えば、ソフトウェアの層）を提供することができる。

【0148】

いくつかの実施形態では、セキュアエンクレーブを互いに連結してパイプラインを形成することができる。そのような実施形態と一致して、計算はサブタスクに分割することができ、サブタスクはパイプライン内で、パイプラインの配置に基づいて同時にもしくは順次に、またはその両方で処理される。

【0149】

いくつかの実施形態では、認証サービスはパイプラインに関連付けることができる。認証サービスは、パイプラインの開始からパイプラインの終了までの信頼チェーンを確立し、パイプラインの内部内容が外部エンティティに観察できない場合でも、外部エンティティの保証を提供する。いくつかの実施形態では、関連するパイプライン自体を拡張することなく、信頼チェーンをさらに拡張することができる。

【0150】

ヘルスケアデータを扱う1つの方法は、個人データ属性を匿名化またはマスキングすること、例えば、それが処理または解析される前に社会保障番号をマスキングすることである。本開示のいくつかの実施形態では、ヘルスケア記録から個人情報情報をマスキングおよび

10

20

30

40

50

非特定化するための方法を用いることができる。これらの方法を使用して、ヘルスケア記録を含むデータセットは、そのデータ属性の様々な部分をマスクまたは非特定化することができる。したがって、結果として得られるデータセットは、1人または複数の特定の個人を識別することができる個人情報または個人情報を含まなくてもよい。

【0151】

しかしながら、ヘルスケア記録の性質を考えると、ヘルスケアデータセットを完全に匿名化または非特定化することは不可能であり得る。例えば、データセットは、その後デジタル的に転写されるヘルスケアプロバイダによって取られた同時期の（手書きの）ノートを含むことができる。多くのヘルスケアデータセットは、アナログデータ、例えば病理データをデジタル化することによって得られる。したがって、データセットは、特に多数の記録を含む場合、匿名化または非特定化されていない私的情報および個人情報を含むことができる。

【0152】

特定のアプリケーションドメインについてのテキストにおける情報マスキングは、非常に高いレベルの性能、特にリコール（モデルによってタグ付けされた機密エンティティの数と機密エンティティの総数との間の比）を要求する。例えば、電子健康記録（EHR）内の名前、場所、電話、住所などの患者識別情報をマスクして研究目的で公開するには、EHR記録を研究用に公開するために非常に高いレベルのリコールが必要である。米国の病院および医療研究センターでは、EHRの採用が9%（2008年）から96%（2017年）に急増している。臨床試験データセットは、EHRがその一部である現実の証拠データのほんの一部を表しており、2003年以来、Big Pharmaごとの約3,490回の介入臨床試験において、2003年以来、Big Pharmaごとに約100万人の患者が治療されている。このレンズを通して、米国および世界の主要なヘルスケア組織のすべてにわたってヘルスケアプロバイダ独自のEHRデータセット（構造化されていないデータおよび構造化されたデータ）を合成することは、研究開発（R&D）データサイエンスの現代のカーネルとして浮上し、臨床試験データセットおよび純粋に構造化されたEHR／クレームデータベースへの現在の依存に取って代わる。適切に行われた場合、管理されたEHRデータは、R&D目的に使用される臨床データを大幅に強化／増強することができ、さらにはバイオ医薬品会社が、追加の臨床試験を実行する必要なく追加の承認およびライン拡張を得ることができるように、それらの薬物が現実世界で有効であることを規制当局（FDAなど）に実証するための代替方法を提供することができるので、バイオ医薬品会社にとって重要な資産となり得る。

【0153】

しかしながら、そのような堅牢な「臨床ゲノム」データセットの生成に関連するいくつかの課題がある。まず、データの所有権およびデータの非特定化は、これらの課題のうちの2つである。EHRデータから識別可能な患者情報は患者が所有しているが、非特定化された患者レベルのデータを合成し市販することができる。機関（病院等）は、非特定化されたEHRデータの現在の「所有者」または「カスタディアン」であるため、非特定化プロセスが堅牢な方法で行われる限り、そのデータに対する最大限の制御／活用を行うことができる。後者は、個人情報ユビキタスであり（臨床ノート、病理報告書などに記載されている）、さらにその個人情報が非常に多様な方法で利用可能であるため、「深いEHRデータ」を扱うときの複雑なタスクである。

【0154】

非特定化のためのプロセスは、患者の健康情報（「PHI」）としてフラグが立てられた単語、フレーズ、または数字が、患者のプライバシーまたは機密性を危険にさらすことのない特定不可能なブレースホルダに置き換えられることを必要とする。これらの要件の例は、米国HIPAAプライバシー規則第164.514節によって定義されたセーフハーバー法に概説されているものであり、参照によりその全体が本明細書に組み込まれ、マスクされなければならない18カテゴリのPHI識別子を指定している。セーフハーバー法はまた、健康情報の非特定化のためのプライバシー規則に含まれる要件をどのように実

施するのが最良であるかを事業者に知らせるための、参照によりその全体が本明細書に組み込まれる関連ガイドラインを含む。以下のいくつかの実施形態の説明は、典型的にはセーフハーバー法に従ってマスクされた患者データを対象としているが、本明細書に記載のシステムおよび方法は、これらの実施形態を超えてデータ記録にも同様に適用することができる。

【0155】

NLP（自然言語処理）の現在のディープラーニングモデルは、自発的には、本アプリケーションに必要とされるこれらの高い性能基準を満たさない。それらがこれらの規格を満たさない理由の1つは、これらのモデルが教師あり学習のために大量のラベル付きデータを必要とするという事実である。特定のエンティティタイプのための公的に利用可能なラベル付きデータは、場合によっては活用することができるが、これらのモデルが必要な高レベルのリコールに達するためには、実際には十分なドメイン固有のラベル付きデータがしばしば必要とされる。精度はマスキングの観点から主要な目的ではないが、マスキングされたデータが研究に何らかの材料を使用するためには、精度が高くなければならない。有効性の要件は、精度と想起の両方について高く、想起は通常より高い（例えば、最小許容精度は99.999%以上であってもよい）。

【0156】

機密情報を明らかにするエンティティの大規模マスキングは、EHRアプリケーションに限定されない。例えば、政府機関による機密情報を含む文書の公開は、小規模での人間のキュレーションによって実際には解決されることが多いが、対処するための同じ問題を有する。

【0157】

ディープラーニングモデルの前の辞書ベースの方法はさらに短くなり、人間のキュレーション／検証を必要とし、大規模マスキングアプリケーションはスケールのためにほとんど実用的ではなくなる。

【0158】

本開示は、以下に説明するように、これらの制限の1つまたは複数に対処することができる。

【0159】

図11は、いくつかの実施形態による情報マスキングのためのシステムの簡略図である。システムが情報を選択的にマスキングするように訓練されると、システム1102に供給されるテキスト入力1101は、単一の単語または複数の単語のいずれかである関心のあるエンティティのサブセットが選択的にマスクされる（一般的なプレースホルダのトークンで置き換えられる）出力1103をもたらす。図11に示す例では、人、場所、日付、および年齢の4つのエンティティタイプがマスクされている。テキスト入力1101は、テキストベースの情報の他のリポジトリの中でも、EpicまたはCernerなどの第三者記録データベースから引き出されたEHRのコーパスの部分を表すことができる。

【0160】

図12は、いくつかの実施形態による、情報マスキングのための制御フローの簡略図である。制御フローは、（対象のエンティティにわたって）未知の確率分布を有する大きなデータセット内の情報をマスキングする方法を説明しており、そのサイズを考慮すると完全にサンプリングするには実用的ではない。本開示に記載された方法は、小さなデータセットにわたって未知の分布を自動的に包含する。

【0161】

図12は、2つのデータセット1200aおよび1200bを示す。データセット1200aは、マスクされるべき1つまたは複数のエンティティタイプのためのラベル付きデータセットである。データセット1200aおよび1200bは、特定のエンティティタイプに関連する行および／または列を有する表形式を使用して編成された構造化データセットであってもよい。さらに、データセット1200aおよび1200bは、構造化データフォーマットの特定のセル内に非構造化データを含むことができ、非構造化データは、

自由形式のテキストまたはセンテンスであり得る。エンティティタイプは、他の識別特性の中でも、個人または組織の名前、場所、年齢、日時、電話番号、ポケットベル番号、臨床識別番号、電子メールおよびIPアドレス、ウェブURL、車両番号、物理アドレス、郵便番号、社会保障番号、および生年月日を含むことができる。エンティティタイプについての任意の辞書を、指定されたエンティティタイプについての1200aに追加することができる。サブワードを使用して単語を表すBidirectional Encoder Representations from Transformers (BERT)のようなサブワードベースのモデルは、それらのサブワードを共有するエンティティにタグ付けするために辞書内の単語を構成するサブワードを活用することができる。未知の分布データセット1200bは、特定のエンティティタイプを選択的にマスキングする必要があるデータである。マスクされる必要があるすべてのエンティティタイプが事前にラベル付けされたデータセット1200aを有する必要はない。データセット1200bのサンプルを使用して、ブートストラップするためのラベル付きデータセット1200aを作成することができる。場合によっては、データセット1200bのサンプルは、システム管理者または主題専門家によって手動で作成されてもよく、一方、他のサンプルは、パターンマッチング規則、正規表現、辞書、および公開データベース検索を使用してPHI要素を識別する機械学習またはルールベースのアルゴリズムを使用して作成されてもよい。例えば、ルールベースのアルゴリズムは、「年月日」（例えば、XX/XX/XXまたはXX/XX/XXXX）の形式で提示された日付または10桁の形式で提示された電話番号（例えば、(XXX)XXX-XXXX）などの標準的な形式の一連の情報のみに基づいてもよい。これらの標準フォーマットに基づいて、ルールベースのアルゴリズムは、パターンを識別し、潜在的に識別する情報を一般的なプレースホルダに置き換えて情報をマスキングすることができる。

【0162】

対象のエンティティは、マスクされる必要があるエンティティに限定される必要はない。モデルがマスクされる必要があるエンティティに対して高いリコールを有するだけでなく、研究に必要なエンティティ（例えば、薬物、疾患など）に対しても高い精度を有することを保証するために、より大きな一組のエンティティとすることができる。関心のあるエンティティのサブセットのみがマスクされる。残りのエンティティは、検出されている間、モデル性能評価にのみ使用される。

【0163】

各エンティティタイプのラベル付きデータセットは、そのエンティティタイプのモデル1202を微調整するために使用される。好ましい実施形態は、各エンティティのタグ付けのための少なくとも1つの別個のモデルを説明しているが、複数のエンティティにタグ付けするモデルを排除するものではない。

【0164】

本開示の好ましい実施形態では、各タグ付けモデルは、アテンションベースのモデル、例えば、参照によりその全体が本明細書に組み込まれるDevlin, et al., 'BERT: Pre-training of deep bidirectional transformers for language understanding,' arXiv preprint arXiv:1810.04805に記載されているBERTモデルである。しかしながら、シーケンスモデル（例えば、ロングショートタームメモリネットワーク(LSTM)、条件付きランダムフィールドを有するLSTM(LSTM-CRF)、またはリカレントニューラルネットワーク(RNN)などの他のモデルを、エンティティにタグ付けするために使用することができる。BERTを使用する場合、各エンティティタイプに対して、エンティティタイプに最適な事前訓練モデルが選択される。例えば、人物、場所などのエンティティにタグ付けする場合、Wikipediaのような一般的なコーパス上で教師なしで訓練されたモデルで十分であり得る。いくつかの実施形態では、事前訓練モデルは、とりわけ、健康科学ジャーナル、専門出版物、ピアレビュージャーナル出版物、またはオペレータコンパイル型データベースなどのモデル

訓練を增强するために、他の既存の公的に利用可能なデータベースに基づくことができる。疾患、遺伝子などのエンティティにタグ付けする場合、科学出版物／生物医学コーパスで訓練されたモデルは、ドメイン固有の言語的特徴をモデルに組み込むか、そうでなければ分野で広く普及していない可能性がある新興用語を識別するのを助けることによって、より良い結果をもたらすことができる。

【0165】

テキストデータの大きなコーパスからテキストのテンプレートまたはパターンを発見するための1つの方法は、統計的名称付きエンティティ認識（NER）モデルをコーパス内の個々の文に適用することによって生成された推論を含むことができる。テキストのパターンが頻繁に繰り返される大きなコーパスでは、統計的方法は、パターンのすべてのインスタンスを捕捉するとは限らない。例えば、`Electronically signed by: SMITH, JOHN C on 01/02/1980 at 12:12 PM CST` という文では、「SMITH, JOHN C」が人物のエンティティとして検出される可能性があるが、`Electronically signed by: DEWEY, JONES K on 01/02/1980 at 12:12 PM CST` という非常に類似した文では、「DEWEY, JONES K」が完全に人物として検出されない可能性がある。そのような状況では、パターンベースの方法がより良好に機能する。`Electronically signed by: [A-Za-z] + , [A-Za-z] + [A-Za-z] + on ¥d+ / ¥d+ / ¥d+ at ¥d+ : ¥d+ PM CST` のような正規表現構文は、すべてのそのようなケースを捕捉する。しかし、パターンの識別は、特に手動で行われる場合、コーパス内の十分に大きな文のサンプルの検査を含むため、不可能ではないにしても困難な作業になる。NER予測は、それらのタイプによって符号化されたエンティティ値を有する文テンプレートを生成し、それらの符号化を構文パターンにマッピングするために使用することができる。これは、パターンマッチングテンプレート候補の生成を自動化する方法を提供する。

【0166】

一実施形態によれば、NERモデルは、図13に関して以下でさらに説明する前処理の形態として、最初にコーパスを個々の文に分割することによって使用することができる。これは、自然言語ツールキット、spaCy、CoreNLP、TextBlob、Gensimなどの自然言語処理ツールを使用することによって完了することができる。いくつかの実施形態では、コーパスを文に分割することは、例えば、同じ文がコーパス内で複数回発生する場合に、処理されるテキストの全体量を減らすことができる。したがって、コーパスを文に分割し、得られた一組の文を重複排除することにより、計算効率を向上させることができる。いくつかの実施形態では、1つまたは複数の短い文（例えば、単語または用語の数がしきい値より少ない文）を組み合わせることができ、これは機械学習モデルに追加のコンテキストを提供することができ、NERタグ付け性能を改善することができる。一例として、`Performed by pet` という文において、「pet」という用語は、動物のペット、または（他の可能性の中でも）大文字使用を伴わない「Peter」という名前の人の頭字語のいずれかを意味することができる。しかしながら、前の文が「硬膜外が4時間後に摩滅した」であった場合、これは、「ペット」が硬膜外を投与した人の固有名詞を指す可能性が最も高いと推測するための機械学習モデルにとって適切なコンテキストを提供する。

【0167】

コーパスを文（または複数文の節）に分割した後に、各固有の文は、アルファベット、数字、および英数字と関連するトークンを含む構文テンプレートにマッピングされてもよい。例えば、文字の連続したシーケンスは、トークン「W」を使用して識別することができ、数字のみのシーケンスは、トークン「N」を使用して識別することができ、英数字のシーケンスは、トークン「A」を使用して識別することができる。例えば、`Electronically signed by: SMITH, JOHN C on 01/02/1980 at 12:12 PM CST` という文は、`WWW:W, W`

W W N/N/N W N:N W W' 'となる。そして、システムは、上記のトークンを使用して生成された各固有のテンプレートに整列するコーパス内の文のリストを作成し、統計的NERモデルを適用して、選択されたテンプレートに整列するコーパス内の文の数を判定することができる。しかしながら、このトークンベースの手法を使用すると、文内のすべての単語、数字、および英数字文字がトークン化され、そのうちのいくつかはマスクされる必要がない。システムはまた、テンプレート内で識別されたエンティティタイプに基づいてトークンを符号化する。トークン符号化技術は、内側、外側、開始（IOB）またはIOB2フォーマットを含むことができる。例えば、このステップに続いて、原文' 'Electronically signed by:SMITH, JOHN Con 01/02/1980 at 12:12 PM CST' 'は' 'Electronically signed by:B-PER, I-PER I-PER on B-DATE at B-TIME PM CST' 'になる。

【0168】

多数のサンプルが同じテンプレートにマッピングされる場合、システムはそのテンプレートをルールになる候補として識別し、異なるテンプレートに一致するルールの曖昧さを低減する。候補テンプレートをルールに遷移させる前に、システムは、トークン識別された文出力をIOBまたはIOB2符号化出力と相互参照して、ルール内でマスクされなければならないエンティティを識別することによって、テンプレート内のIOBまたはIOB2タグを対応する構文トークンに戻すことができる。IOB2フォーマットをトークンで置き換えることにより、トークンに基づく単純化された規則を可能にすることができる。例えば、フォーマットされたテンプレート' 'Electronically signed by:B-PER, I-PER I-PER on B-DATE at B-TIME PM CST' 'は、名前付きエンティティを識別するための規則またはパターンテンプレートである' 'Electronically signed by:W, W W on N/N/N at N:N PM CST' 'になる。パターンテンプレート内の構文トークンは、標準的な構文パターンマッチング方法に直接変換される。例えば、「W」は、正規表現「[A-Za-z]+」、「N」から「¥d+」にマッピングすることができる。したがって、一実施態様では、上記の例のパターンテンプレートは正規表現' 'Electronically signed by:[A-Za-z]+,[A-Za-z]+[A-Za-z]+on ¥d+ / ¥d+ / ¥d+at ¥d:¥d PM CST' 'になる可能性がある。

【0169】

代替的な実施形態では、上述したように、コーパスを個々の文に分離することに続いて、システムは、上記の説明と同様に、抽出されてIOBまたはIOB2タグで置き換えられたエンティティ用語を有する各文の出力を生成する統計的NERモデルを介して個々の文を最初に渡すことができる。次に、システムは、タグ付けされたエンティティタイプまたはタグ付けされたエンティティカウントの指定されたしきい値カウントをもたらないテンプレートを除外することができる。例えば、システムオペレータは、エンティティタイプに関係なく、少なくとも2つのエンティティタイプを有するテンプレートのみが規則に適格であるか、または潜在的に少なくとも4つのタグ付けされたエンティティを有するテンプレートであると指定することができる。しきい値カウントが満たされると、システムは、IOBまたはIOB2構文トークンを選択されたテンプレートにマッピングし、上述した同じアルファベット、数字、および英数字のトークン「W」、「N」、および「A」を使用してパターンテンプレートを作成することができる。パターンテンプレートを生成した後に、システムは、パターンテンプレートが固有の統計的NERクラスに由来するかどうかを評価し、ルールを作成するためにそれらの固有のパターンテンプレートを選択することができる。各固有のパターンテンプレートは、特定のパターン化シーケンスに続く文の数、または異なる数のエンティティまたはエンティティタイプを有する数に基づいてもよい。例えば、パターンテンプレートのすべてのインスタンス' 'Electron

ically signed by: W, W W on N / N / N at N : N PM CST ' ' が NER テンプレートのインスタンス ' ' Electronically signed by: B - PER, I - PER I - PER on B - DATE at B - TIME PM CST ' ' から生成された場合、それはルール作成のために選択される。そうすることにより、システムオペレータは、コーパス内のエンティティタイプを誤って識別することによって不一致データマスキングをもたらす可能性が低いものにルールベースのテンプレートを制限することができる。機密情報をマスキングするなどのいくつかの実施形態では、オペレータが複数のエンティティを同じものとして扱うことが有用であり得る。一致する異なるテンプレートは、それらが一致の最小数 / パーセンテージのいくつかのしきい値を潜在的に通過した後に、手動で検査することができる。例示的な IOB / IOB2 構文トークンを使用して説明されているが、この規則生成方法は、コーパス内のエンティティおよびエンティティタイプを示すために他の形式のシンタックスまたはトークンシステムを使用することができる。

【0170】

訓練およびルール開発はまた、エンティティタイプに基づいて個々のモデルを評価する前に、一連のカテゴリに基づいてコーパス内の文書または記録を手動で分類するオペレータからなることができる。データ記録は、記録内の特定のタイプの PHI が類似して識別可能である可能性がより高くなるように、セグメント記述、サービス記述、およびサービスイベントなどのカテゴリに分類することができる。例えば、特定の記録のセグメント記述は、とりわけ、経過ノート、相談、退院指示書、患者要約、または医師のノートなどのノートの性質に基づくことができる。サービス記述は、医療腫瘍学、整形外科手術、皮膚科学、家庭医学、または救急科、ならびに組織または企業の他のファセットを含む治療の部門または領域を識別することができる。サービスイベントは、臨床外来患者、入院患者、緊急事態、検査室、内分泌学相談、日中手術、または他の文書作成設定など、文書が作成された設定を識別することができる。

【0171】

訓練文書のコーパスが分類されると、個々の検証セットを個々の分類内で使用して、データ記録の反復位置に PHI を一般的に含む反復データフォーマットを識別することができる。例えば、分類は、外来設定で収集された皮膚科サービス記述が一般に、文書の第1の文内の個人名情報の形態の PHI を含むという識別をもたらす。分類はまた、記録タイプごとの患者ノートの数の分布を識別するために使用することができ、各ノートは特定の形態の PHI データを含むことができる。分類はまた、オペレータが分類メトリックのいずれか1つに基づいて記録当たりの PHI 要素の平均数を識別することを可能にし得る。分類に基づいて PHI の有病率を識別することにより、オペレータは、ノートタイプ (PHI が豊富なノートタイプとも呼ばれる) およびモデル訓練に優先順位を付けて、大量の PHI データを含む記録に焦点を合わせることができる。

【0172】

エンティティタイプに対して特定の事前訓練されたモデルが選択されると、モデルはブートストラップ訓練セット 1201 で微調整される。ブートストラップ訓練セット 1201 は、図 14 に関連して以下でさらに説明するように、訓練サンプルが訓練サンプルの初期セットに連続的に追加される反復プロセスを使用して更新され得る。いくつかの実施形態では、既存の訓練サンプルは、以前のモデルで識別されたエラーに基づいて選択されたコーパスから選択された個々の文によって補完されてよい。モデルが微調整された後に、未知の分布 1200b からサンプル評価 / テストセットが引き出され (1203)、リコール、精度、および F スコアレベルを評価することによって各モデルの有効性が評価される (1204)。このステップは、エンティティタイプごとに実行され、リコール / 精度スコアが計算される。

【0173】

平均リコールレベル、精度レベル、および F スコアレベルが、n 個のテスト (n はある数) にわたって必要なレベルと少なくとも同程度に良好である場合、そのエンティティ訓

10

20

30

40

50

練は完了したとみなされる。必要なレベルは、マスキング手順後に利用可能にされる情報を識別するためにオペレータが有するリスク許容度に応じて、システムオペレータによって決定されるしきい値成功率または比率であってもよい。すべてのエンティティが所望のしきい値1207に達すると、データセット1200bに対して完全なマスク実行1208が行われる。マスクされたセットに対してサンプリングが行われ、リコール／精度の最終的な測定値について検査される209。

【0174】

エンティティタイプの平均リコールスコアがそのエンティティタイプに対して必要なしきい値(1205)を上回っていない場合には、未知の分布のラベルなしデータセット1206からサンプルが引き出され、ステップ1204でモデルの有効性を評価するために同じ方法を使用して再び評価される。ステップ1206において、しきい値要件を満たさなかった失敗した文は、テンプレートサンプルを作成するために使用され、エンティティタイプ内で関連する辞書を使用して、失敗した文内の実際のエンティティインスタンスを辞書からの同等のエンティティインスタンスで置き換えるように、訓練セット1206に追加される。辞書を使用してエンティティタイプを置き換えることにより、これは、以前に故障したエンティティインスタンスが訓練の将来の反復において適切に説明されることを保証する。

【0175】

図13は、いくつかの実施形態によるエンティティのタグ付けプロセスの簡略図である。入力テキスト1301は、各エンティティタガーに供給される前に最小限前処理される。いくつかの実施形態では、入力テキスト1301は、同時に処理される患者ノートのバッチを含むことができる。前処理の一部として、システムは、ノートのバッチを分解し、個々の文に文識別子(「文ID」)を割り当てて、各ノートからの個々の文を個別に処理することができる。後述する後のコンパイルのために文IDの記録を保持する。さらに、前処理の一部として、システムは、複数の文が同じパターンまたは情報を含む場合、文IDを使用して追跡されたバッチ内の文を関連付けることができる。このようにして、システムは、識別情報を正確にマスキングするために関連する文の1つの代表文を処理するだけでよく、後に関連する文の各々に対して同じマスキングを実行するだけでよい。前処理は、エンティティ依存またはエンティティ非依存であり得る。例えば、人物タガーへの入力1つの経路は、ケーシングが保存されているテキストとすることができ、同じエンティティタイプの別のエンティティタガーへの入力のための別の経路は、ケーシング正規化を実行することができる(任意選択的に、すべての名詞形を上部ケーシングで開始し、続いて小文字のテキストで開始するようにPOSTタガーを使用する)。いくつかの実施形態では、段落の区切りおよび／または句読点を使用して、患者ノートのバッチのテキストを個々の文に分離することができる。さらに、非構造化テキスト文字列は、テキストファイル、データベース、または任意の他の一般的に使用されるデータストレージリポジトリシステムから分離することができる。

【0176】

いくつかの実施形態では、PHIを含まないと一般的に識別され、したがってマスキングを必要としない共通のまたは繰り返されるフレーズを識別するために、タガーモデル1302~1305に文を導入する前に、ホワイトリストタガー(図示せず)を使用することができる。例えば、ホワイトリストに登録されたエントリは、コンピュータ生成テキスト(例えば、「服薬に問題がある場合は医師に知らせてください」)、医師の筆記体の反復フレーズ(例えば、「疼痛症状：なし」)、またはセクションヘッダなどの臨床ノートの共有要素(例えば、「現在の病気の病歴」)を含むことができる。これらのホワイトリストセンテンスは、PHIまたは他の識別情報を公開するリスクがほぼゼロパーセントであるため、処理されるデータから削除され、後のコンパイルのためにアグリゲータ1306にルーティングされ得る。ホワイトリストタガーを作成するために、各固有の文の有病率カウントが、手動レビューステップ後の最高カウントに基づいて計算され、識別され得る。システムの一部としてホワイトリストタガーを使用すると、PHIを含まないことが

10

20

30

40

50

知られているデータのタグ付けおよびマスキングを必要としないことによって計算リソースが節約される。

【0177】

図13に示すように、複数のエンティティタガーモデルは各エンティティタイプにタグ付けし、タガーモデルの数およびタイプは、そのエンティティタイプに必要なリコール、精度、およびFスコアレベルなどに依存する。例えば、タガーモデル1302は、個人または組織の名前エンティティに合わせて調整され、タガーモデル1303は、日付エンティティに合わせて調整され、タガーモデル1304は、年齢エンティティに合わせて調整され、タガーモデル1305は、疾患エンティティに合わせて調整される。タガーモデルのいくつかは、ルールベースの手法、ディープラーニングモデル、およびパターンベースのモデルなどの異なる訓練手法を使用して訓練することができる。そうすることにより、エンティティタイプごとに多様なタグ付け手法が作成され、システムがエンティティ情報の識別を見逃す可能性がさらに低減される。

【0178】

好ましい実施形態では、各エンティティタイプは1つまたは複数のエンティティタガーモデル（エンティティタイプのタガーの数は一定ではない）によってタグ付けされるが、他の実施形態は複数のエンティティにタグ付けするモデルを有することができる。いくつかの実施形態では、特定のエンティティに関連する1つまたは複数のエンティティタガーモデルは、構造化されていないEHRデータ内のエンティティに関連する異なる相補的特徴を識別するように微調整され得る。例えば、名前エンティティに合わせて調整されたタガーモデルの場合、あるモデルはテキスト内の名前を識別することに焦点を合わせることができ、別のモデルはアドレスおよび位置に関連する名前に合わせて調整することができ、または別の場合、追加のタガーモデルは、前処理を必要とせずに名前のケース入りおよびケースなしの変形に焦点を合わせることができる。

【0179】

モデル1302、1303、1304、1305からのタグ付けされた出力は、すべての所望のエンティティがプレースホルダのタグで置き換えられた文を出力するアグリゲータ1306に供給される。アグリゲータは、誤って形成されたエンティティ配列の修正（例えば、I、IなどがB、I配列に変換される）、タグ付けの競合の解決（疾患と人の両方にタグ付けされた同じ用語—例えばパーキンソン病患者）、ミスアライメントされたエンティティタイプのスパンの処理、および最終的にタグ付けされたタームのプレースホルダのタグへの置き換えなどの複数の機能を実行する。アグリゲータは、エンティティタイプに必要なリコール、精度、Fスコアに基づいて、フィルタリングされるエンティティに優先順位を付けることができる。

【0180】

次いで、アグリゲータ1306からの出力は、ドレッグフィルタ1307、1308が代表的なドレッグフィルタのカスケードを通過する。ドレッグフィルタは、特定のエンティティタイプに対する最終的な処理フィルタとして機能して、前の処理ステップでミスした個々のエンティティが、PHIにアクセスするべきではないユーザへの出力として生成されないことを保証することができる。例えば、ドレッグフィルタカスケード1307は、人物エンティティタイプに合わせて調整され、ドレッグフィルタカスケード1308は、日付エンティティタイプに合わせて調整される。タガー1302～1305に関連するエンティティタイプに向けて追加のドレッグフィルタを含めることができる。ドレッグフィルタ1307、1308は、上記のステップでタグ付けされなかった用語／フレーズをフィルタリングする。ドレッグフィルタ1307、1308は、PHI強化ノートタイプに基づくルールベースのテンプレートを使用して、タガーモデル1302～1305によって識別されなかった追加のPHIを除外することができる。ルールベース型テンプレートは、PHIデータを最もよく識別するために、データ記録内の個々の文構造に合わせて調整されてもよい。各ドレッグフィルタカスケード1307、1308は、異なるエンティティタイプに向けられ、複数のルールベースのテンプレートを含むことができる。ドレ

10

20

30

40

50

ッグフィルタ1307、1308はまた、パターンマッチングフィルタまたは同様の手法を使用してPHIデータを識別することができる。最終出力1309は、元の文と、個々の文の各単語に対するタグの関連する配列（例えば、IOB2フォーマット）であり、所望のマスクされた出力をもたらす。最終出力1309の前に、システムはまた、モデル1302～1305にタグ付けすることによって実行されるタグ付けの前に記憶された文IDを使用してバッチデータセットから各ノートをコンパイルすることができる。

【0181】

フィルタドレグは入力文を取り込み、入力文は、IOBまたはIOB2フォーマットを使用して文のフレーズの始まりおよび終わりをマークするフレーズタガーを介して送信される。タグ付けされなかったそれらの用語／フレーズについては、存在する場合にはエンティティの辞書を使用して、各エンティティタイプについて辞書内の用語のベクトル表現とのタグ付けされていないフレーズのベクトル表現の類似性を見つける。これは、その後に、前のステップで見逃された可能性があるエンティティをさらにフィルタリングするために使用される。

【0182】

例えば、センテンスに出現した名前 ' ' j a y e n d r a n b a l a k r i s h n a n ' ' が、人（偽陰性）または任意の他のエンティティタイプ（偽陽性）としてタグ付けされていない場合、そのフレーズはフィルタカスケード段階1307／1308でピックアップされ、各エンティティタイプの辞書用語のベクトル表現と一致する。類似性測度が何らかのしきい値を超える場合、フレーズは文からトークンまたはトークン表現に置き換えられる。BERTモデルは、サブワードから単語を構築する能力のために、フレーズのベクトル表現を作成するために使用することができる。いくつかの実施形態では、訓練モデルおよび／またはドレグフィルタの両方は、関係の強さによって順序付けられた入力語間の関連付けを使用して、関連付けをグループ化し、PHIとして不適切にタグ付けされた可能性がある発展中の単語またはフレーズを識別することができる。例えば、入力「ECOG」は、公的に入手可能な文献に基づいて上位の疾患に関連付けることができるが、システムは、「ECOG」を人物の名前としてタグ付けしている可能性がある。「ECOG」と一般的に関連する疾患用語との間の関係を検討することにより、システムは「ECOG」を適切に識別し、それをPHIとして治療しないことができる。

【0183】

次いで、カスケードドレグフィルタ段階1308からの出力は、モデル精度／精度を評価するために使用される。これは、システムオペレータによる最終出力1309の手動サンプリング、または以下に説明する自動化プロセスを使用して完了することができる。

【0184】

いくつかの実施形態では、出力データで識別されたタグ付けされたエンティティは、様々な方法でマスクされてもよい。例えば、タグ付けされたエンティティは、削除されてもよく、空白文字で置き換えられてもよく、エンティティのタイプを識別するプレースホルダで置き換えられてもよい。いくつかの実施形態では、タグ付けされたエンティティの値は、同じタイプのランダム情報に変更され得る。例えば、タグ付けされたエンティティが電話番号である場合、エンティティはランダムな（または難読化された）電話番号に変更され得る。この手法は、エンティティタグ付けモデルが検出できないプライベート情報を難読化し得る。例えば、エンティティタグ付けモデルが、所与の一組の記録内の10個の電話番号のうちの9個を検出してランダム化するが、10番目を逃した場合、データのエンドユーザは、どの電話番号が実際の者であり、どれが難読化されているかを知らず、それによって、タガーが検出しなかった電話番号の保護形式を提供する。

【0185】

いくつかの実施形態では、非特定化が完了した後であっても、オペレータシステム内で患者データ記録を全体的に関連付けるために、特定の患者IDフィールドを保持することができる。例えば、マスキング後に情報が公開されているにもかかわらず、オペレータが関連する患者ファイルを遡及的に引き出したい場合には、マスクされた文書のコーパスに

10

20

30

40

50

患者摂取IDを保持することができる。患者IDはまた、患者の生年月日、住所、病室番号、電話番号、または電子メールアドレスなどの他の形態をとることもできる。そうすることにより、オペレータは、各個々の記録インスタンスが独立してマスクされた後であっても、匿名化されたデータの有用性を最大にするために、データ内に有用な医療および生物学的信号を保持することができる。患者および他の識別子は、そのような目標を達成するためにこれらの個々の記録間のリンクを提供することができる。

【0186】

患者IDがデータ記録インスタンスに保持される場合、いくつかの実施形態は、患者IDを有する患者記録の暗号化バージョンにアクセスすることができる、組織の個人または指定メンバーなどの対象エンティティに対する許可を含むことができる。例えば、カバーされるエンティティは、共通の患者識別子を有する患者データ記録のバージョンにアクセスするための暗号化鍵、パスワード、または多段階認証方法などの特定の認証情報を有することができる。他の実施形態では、患者IDはマスキング手順後にすべての文書に保持されてもよいが、患者IDをその患者の完全な一組の医療記録に関連させるデータベースは、暗号化データファイルに格納される（患者データの「リンク」とも呼ばれる）。暗号化方法は、AES-256暗号化を含み得る。データ記録はまた、ソルトを使用して暗号化されてもよく、またはデータをハッシュするために追加の一方向関数を使用したランダムなデータシーケンスが使用されてもよい。そうすることにより、暗号化ハッシュの繰り返し推定を実行するために効率的なコンピューティングシステムを使用して総当たり攻撃によってデータが危険にさらされることが防止される。いくつかの実施形態では、安全なハッシュは、システムがハッシュを患者IDと関連付ける一方向追跡であり、患者データ記録を再構築するために、入力IDと出力ハッシュIDを別々のデータリポジトリまたは記憶場所に別々に格納する必要がある。これは、とりわけ、SHA-512ハッシュを含み得る。これらの実施形態では、カバーされるエンティティは、生成された場合、暗号鍵、ソルト、および任意の出力マップへのアクセスを制御することができる。

【0187】

他の実施形態では、システムを使用して、PHIが画像自体にテキストとして埋め込まれているX線またはオプト相同画像などの画像に含まれるPHIをマスキングすることができる。そのような実施形態では、システムは、畳み込みニューラルネットワークを使用して、画像が作成されるオペレータの好ましいプログラムまたはアプリケーションに基づいて画像内の領域を自動的に検出することができる。いくつかの実施形態では、システムは、画像内の識別された領域からテキストを自動的に検出および抽出し、次いで、そのテキストを、文ごとにに関する上記の説明と同様のデータベースからの置換可能な匿名化テキストにマッピングすることができる。他の実施形態では、システムは、全スライド画像レベルのラベルを取り、パッチレベル（例えば、PHIを含有すると識別された画像の部分領域）で予測を行うように訓練された畳み込みニューラルネットワークからなるマルチインスタンス学習（MIL）を使用することができる。代表的なパッチサイズは、画像の2次元ピクセルサイズに基づくことができ、ピクセルレベルの注釈ではなくスライドごとのレベルでクラスを予測するために、パッチレベル予測に最大プーリングまたはリカレントニューラルネットワーク（RNN）を使用することができる。別の実施形態では、システムは、画像類似性方法を使用して、特定の画像内に見られるPHIを識別およびマスキングすることができる。そのようなモデルを訓練するために、システムオペレータは、訓練画像の特定の部分領域を選択し、解像度値を設定し、次いで、同じ解像度値でデータベースからすべての類似パッチを見つけるようにシステムに依頼することができる。例えば、トリプレット損失概念に基づく分類器ネットワークを最初に構築して、すべての解像度で各画像から（ベクトルとして）きめ細かい特徴を効率的に学習することができる。次いで、そのような画像由来特徴ベクトルのデータベースが作成される。画像の別個の倍率／解像度レベルを別個の画像自体として処理すること次に、システムは、最新の高性能高次元類似性検索方法（例えば、とりわけ、シャム型ニューラルネットワーク）を使用して、ユーザがクエリで使用したパッチに最も類似したパッチを見つけることができる。

10

20

30

40

50

【0188】

図14は、いくつかの実施形態による、有効性をマスキングするためのテストスイート診断フロー方法を示す簡略図である。テストスイート方法1400は、図12に記載されたものと同様に、オペレータがテスト環境で新しいモデルを訓練すること、またはライブシステムに以前に展開された個々のタギングモデルのパラメータを継続的にホーニングして、特定のエンティティモデルの精度、リコール、およびFスコアレベルを改善することの両方を可能にする。新しいモデルの場合、システムオペレータは、データセット内の特定のエンティティに関連するタグを含むラベル付きデータセット1401を訓練フェーズにおいてシステムに導入することができる。このラベル付きデータセットは、オペレータによって手動で識別されてもよいし、図13で説明したモデルタガー1302～1305によって生成された以前にラベル付きデータセットの積であってもよい。ステップ1402において、テストスイートの訓練段階において、ラベル付きデータセットは、初期タガーモデルV₀1403を訓練するために初期タガーモデルV₀1403に導入される。ステップ1404において、モデルV₀1403は、ラベルなしデータセット1409のサブセットを解析し、モデルV₀1403がラベルなしデータセット内で識別した一組のトークンを有するラベル付きデータセットの出力を生成する。次に、ステップ1405で出力が評価されて、タガーモデルタイプ（例えば、それがどのエンティティタイプ／タイプに合わせて調整されるか）を検証し、ラベル付きデータセットからタグ付けされたトークンのエラーを識別する。テストスイートは、モデルV₀1403出力の精度、リコール、およびFスコアレベルを決定し、システムオペレータによって決定されたしきい値に対して3つすべてのスコアを測定する。

【0189】

ステップ1405で精度、リコール、およびFスコアレベルが必要なしきい値を満たさない場合、テストスイートシステムは、ステップ1406で異なる、または更新されたラベル付きデータセットを選択し、ステップ1407で更新されたラベル付きデータセットをモデルV₀1403に再導入して、モデルが新しい文字を取り込み、モデルV_i1408をもたらす。更新されたラベル付きデータセットは、モデルが特定のエンティティタイプを認識するそれ自体の能力を反復的に改善するように、ステップ1404で知覚されたエラーに基づくことができる。再訓練に続いて、モデルV_i1408は、更新されたラベル付きデータセット内でモデルV_i1408が識別した新しい一組のトークンを有する更新されたラベル付きデータセットにタグ付けし、その出力は、ステップ1404および1405においてモデルV₀1403と同じ方法で評価される。いくつかの実施形態では、モデルV_i1408は、臨床ノートなどの報告されたデータのコーパスから取得されたラベルなしデータセット1409の一部に対して検証することができる。このプロセスは、再訓練されたモデルがステップ1405で精度、リコール、およびFスコアに必要なしきい値を満たすまで反復的に繰り返される。完了すると、検証済みモデルV_N1410を最終的に生成することができ、Nは、ステップ1405で適切なしきい値精度に達する前にモデルを完了するために必要な訓練フェーズ内の反復回数を表す。

【0190】

確認後に、モデルV_N1410は、ライブシステムに展開され、ステップ1411で、ラベルなしデータセット1409のより大きなコーパスから受信したデータにタグ付けするために使用され得る。ステップ1411におけるデータの処理は、図3に関して上述したものと同じであり、ステップ1412において最終的な匿名化データセットが得られる。

【0191】

前述の実施形態は例示であり、様々な代替形態が可能であることを理解されたい。いくつかの実施形態では、特定のエンティティタイプに対応する一組の単語（または複数単語用語）が与えられると、その組内の単語を含むテキスト文書またはコーパスから1つまたは複数の連続フレーズが抽出される。例えば、組織の種類に対応する一組の単語を考える。すなわち、集合は、例示的に、「肺」、「肝臓」、「脳」などの単語を含んでよい。セットはまた、「左」、「右」、および「後」などの用語を含むことができ、これらは組織

の文脈で追加の修飾語としてよく使用される。逆に、一組の単語は、P H Iまたは他の機密情報を表すことができる単語およびフレーズを省略することができる。特定のテキストフラグメント（例えば、「スミスさんは今日、左肺の手術のためにM G Hを訪問した」）が与えられると、エンティティ（例えば、「左肺」）に対応する一組の単語に見られる単語で構成されるフレーズが抽出される。特に、前述の例では、一組の単語がこのフレーズを直接含まない場合でも、「左肺」というフレーズが抽出される。むしろ、集合は構成語「左」および「肺」を含む。このようにして、患者識別可能コンテンツ（例えば、上記の例における名前「スミスさん」）が抽出されない間に、テキストから有用な情報が抽出され得る。このアプローチは、疾患、薬物などの多種多様な生物医学的エンティティに容易に拡張することができる。

10

【0192】

いくつかの実施形態では、名称付きエンティティ認識は、ラベル付き訓練文を使用せずに監視されない方法で実行されてもよい。例えば、名前付きエンティティ認識は、マスキングされた言語モデル目標を使用してコーパスで訓練されたBERTモデルを使用して実行され得る。この技術の例示的な例は、' 'Unsupervised NER using BERT, ' 'Towards Data Science, Feb. 28, 2020 (<https://towardsdatascience.com/unsupervised-ner-using-bert-2d7af5f90b8a>) の記事にさらに詳細に記載されており、その全体が参照により本明細書に組み込まれる。

【0193】

個人のヘルスケアデータなどの情報を安全に処理および受信することに関連する上述の課題に加えて、情報のリポジトリから検索クエリに応答する情報を検索することは困難であり得る。情報は、文書の集合、データベース（例えば、構造化データベース、半構造化データベース、または非構造化データベース）、知識グラフなど、様々な方法で格納することができる。いくつかの情報検索ツールは、クエリ用語に基づいてリポジトリから文書または他の記録を検索するように設計されている。例えば、様々な公的に利用可能な検索エンジン（例えば、G o o g l eまたはP u b M e d）は、ユーザ提供のクエリ用語に関連するウェブページ、雑誌記事、書籍などを識別するように構成される。

20

【0194】

しかしながら、いくつかのシナリオでは、リポジトリ内の関連する記録を識別することは、ユーザのクエリに対する適切な応答をもたらさない場合がある。例えば、ユーザが集約された情報（例えば、特定の薬物が特定の疾患に対して有効であることが示されているかどうかを、その薬物に対して行われた研究の集合に基づいて決定するために）に基づいて決定または推論を行おうとしているとき、関連文書のリストを含む検索結果は、ユーザのクエリに明確かつ直接応答しない場合がある。例示すると、特定の薬物が特定の疾患に対して有効であることが示されているかどうかの回答を求めるユーザの上記の例では、ユーザは、どの文書が実際にクエリに応答しているかを判定するために、各関連文書（例えば、薬物および疾患に言及する各文書）を閲覧するという面倒なタスクを課される可能性がある。その後、応答文書のその後の手動解析を実行して、ユーザによって提示された最初の質問に対する回答を得ることができる。検索結果を手動で調べて所望の情報をつなぎ合わせ、適切な決定または推論を行うこのプロセスは、面倒で時間がかかる可能性がある。

30

【0195】

したがって、クエリに応答する情報を検索するための改善された技術、特に異種フォーマットで記憶された情報に適用することができる技術を開発することが望ましい。

【0196】

図15は、いくつかの実施形態による情報検索のためのシステム1500の簡略図である。システム1500は、ネットワーク1510を介して通信可能に結合された複数のデバイス1501~1509を含む。デバイス1501~1509は、一般に、パーソナルコンピュータ、モバイルデバイス、サーバなどのコンピュータデバイスまたはシステムを

40

50

含む。ネットワーク1510は、1つまたは複数のローカルエリアネットワーク（LAN）、ワイドエリアネットワーク（WAN）、有線ネットワーク、無線ネットワーク、インターネットなどを含むことができる。例示的に、デバイス1501～1509は、TCP/IPプロトコルまたは他の適切なネットワークプロトコルを使用してネットワーク1510を介して通信することができる。

【0197】

デバイス1501～1509のうちの1つまたは複数は、デジタル文書1521～1529を格納し、および／またはネットワーク1510を介してデジタル文書1521～1529にアクセスすることができる。例えば、図15に示すように、デバイス1501、1502、および1509は、それぞれデジタル文書1521、1522、および1529を格納し、デバイス1503は、ネットワーク1510を介してデジタル文書1521～1529にアクセスする。デジタル文書1521～1529は、ウェブページ、デジタルファイル、デジタル画像（ビデオまたはアニメーションの1つまたは複数のフレームを含む）などを含むことができる。例示的に、デジタル文書1521～1529は、HTML/CSS文書、PDF文書、ワードプロセッシング文書（例えば、Word文書）、テキスト文書、スライドショープレゼンテーション（例えば、パワーポイントの提示）、画像ファイル（例えば、JPEG、PNG、またはTIFF画像）などとしてフォーマットすることができる。デジタル文書1521～1529は、異種（例えば、異なるフォーマットまたはファイルタイプの）または同種（例えば、同じフォーマットまたはファイルタイプの）であり得、構造化データまたは非構造化データを含むことができる。一般に、デジタル文書1521～1529は、英数字、記号、絵文字、テキストの画像表現などを含むことができるテキストデータを含む。ネットワーク1510を介した効率的な記憶および／または送信のために、文書1521～1529は、ネットワーク1510を介した送信前または送信中に圧縮されてもよい。暗号化、認証（多要素認証を含む）、SSL、HTTPS、および他のセキュリティ技術などのセキュリティ手段も適用することができる。

【0198】

いくつかの実施形態によれば、デバイス1503は、デバイス1501、1502、および1509からデジタル文書1521～1529をダウンロードすることによって、デジタル文書1521～1529のうちの1つまたは複数にアクセスすることができる。さらに、デバイス1501、1502、または1509のうちの1つまたは複数は、デジタル文書1521～1529をデバイス1503にアップロードすることができる。デジタル文書1521～1529は、様々な時点で更新されてもよい。したがって、デバイス1503は、最新のコピーを取得するために、様々な間隔（例えば、周期的に）で複数回デジタル文書1521～1529にアクセスすることができる。

【0199】

図1～図14と一致するいくつかの実施形態では、デバイス1501～1509のうちの1つまたは複数は、セキュアデータストア810、901、および1009などのセキュアデータストアに対応するか、それを含むことができる。例えば、デジタル文書1521～1529のうちの1つまたは複数は、個人のヘルスケアデータを含むことができ、したがって、例えばプライバシー規制に準拠するためにマスクされた情報を含むことができる。マスクングは、図1～図14と一致する技術を使用して、情報マスクングシステム1102などの情報マスクングシステムによって実行することができる。いくつかの実施形態では、情報マスクングシステムは、システム1500の一部として含まれてもよく、別個のシステムであってもよく、または複数のシステムにわたって分散されてもよい。マスクングは、デバイス1503によってアクセスされる前、アクセス中、またはアクセスされた後に実行されてもよい。

【0200】

図15に示すように、デバイス1503は、メモリ1540（例えば、1つまたは複数の非一時的メモリ）に結合されたプロセッサ1530（例えば、1つまたは複数のハードウェアプロセッサ）を含む。メモリ1540は、情報検索プログラム1550に対応する

命令および／またはデータを格納する。情報取得プログラム1550は、プロセッサ1530により実行されると、クエリに応じた情報の取得に関する動作をプロセッサ1530に実行させる。いくつかの実施形態では、クエリは、デバイス1504のユーザによって入力（例えば、クエリ文字列）として提供され、ネットワーク1510を介してデバイス1503に送信されてもよい。その後、情報検索プログラム1550を使用して決定されたクエリに対する応答は、ネットワーク1510を介してデバイス1504に配信され、ユーザインターフェースを介してユーザにレンダリングされ得る。情報検索プログラム1550によって実装されるデータフローの例示的な実施形態は、図16～図17を参照して以下でさらに詳細に説明される。

【0201】

情報取得プログラム1550の実行中、プロセッサ1530は、1つまたは複数のニューラルネットワークモデル1560を実行してもよい。ニューラルネットワークモデル1560は、入力データに基づいて予測（例えば推測）を行うように訓練される。ニューラルネットワークモデル1560は、ニューラルネットワークモデル1560の複数の層および層間の関係を定義する構成1562を含む。層の例示的な例には、入力層、出力層、畳み込み層、密結合層、マージ層などが含まれる。いくつかの実施形態では、ニューラルネットワークモデル1560は、入力層と出力層との間に少なくとも1つの隠れ層を有するディープニューラルネットワークとして構成されてもよい。層間の接続は、フィードフォワード接続または反復接続を含むことができる。

【0202】

ニューラルネットワークモデル1560の1つまたは複数の層は、訓練されたモデルパラメータ1564に関連付けられる。訓練されたモデルパラメータ1564は、機械学習プロセスに従って学習された一組のパラメータ（例えば、人工ニューロンの重みおよびバイアスパラメータ）を含む。機械学習プロセスの間、ラベル化された訓練データがニューラルネットワークモデル1560への入力として提供され、訓練されたモデルパラメータ1564の値は、ニューラルネットワーク1560によって生成された予測が対応するラベルと所望のレベルの精度で一致するまで反復的に調整される。

【0203】

性能を向上させるために、プロセッサ1530は、グラフィック処理ユニット、テンソル処理ユニット、特定用途向け集積回路などを使用してニューラルネットワークモデル1560を実行することができる。

【0204】

デバイス1503は、データベース1570またはデジタル情報の別の適切なリポジトリに通信可能に結合することができる。例えば、データベース1570は、スキーマまたは他の論理的関係（例えば、リレーショナルデータベース）に従って編成されたコンテンツを有する構造化データベースとして構成されてもよい。いくつかの実施形態では、データベース1570は、非リレーショナルデータベース、半構造化データベース、非構造化データベース、キーバリューストアなどとして構成されてもよい。データベース1570は、デバイス1503に直接結合されるものとして示されているが、様々な他の構成も可能であることを理解されたい。例えば、データベース1570は、メモリ1540に格納されてもよく、ネットワーク1510を介してアクセスされてもよい。

【0205】

図16は、いくつかの実施形態への情報検索のためのデータフロー1600の簡略図である。図15と一致するいくつかの実施形態では、データフロー1600は、以下でさらに説明するように、システム1500の様々な構成要素および／または特徴を使用して実施することができる。図1～図14と一致するいくつかの実施形態では、データフロー1600は、個人のヘルスケアデータを含むか、またはそれに基づく情報を検索するように構成され得る。したがって、データフロー1600は、図1～図14と一致する技術を使用して実施することができる。例えば、データフロー1600は、1つまたは複数のコンテナ内に実装されてもよく、入力情報は、図1～図14を参照して説明したようにマスク

10

20

30

40

50

されてもよい。

【0206】

図16に示すように、アプリケーション1602は、情報検索システム1604に通信可能に結合されている。図15と一致するいくつかの実施形態では、情報検索システム1604は、情報検索プログラム1550に対応することができる。アプリケーション1602は、一般に、情報検索システム1604にクエリを提供し、情報検索プログラム1604からの応答を処理するように構成されたプログラムに対応する。例えば、アプリケーション1602は、ユーザからクエリを受信し、情報検索システム1604（例えば、APIを介して）にクエリを送信し、対応する応答を受信してレンダリングするウェブアプリケーションまたはモバイルアプリケーションに対応することができる。図15と一致するいくつかの実施形態では、アプリケーション1602は、デバイス1504上で動作するフロントエンド構成要素、デバイス1503上で動作するバックエンド構成要素などを含むことができる。いくつかの実施形態では、情報検索システム1604は、情報検索システム1604が様々なタイプまたはバージョンのアプリケーションと通信することを可能にする標準化されたAPIまたは他のインターフェースを提供することができる。いくつかの実施形態では、情報検索システム1604は、ユーザがアプリケーション1602をバイパスして直接情報検索システム1604にクエリを提供することを可能にするユーザインターフェースを提供することができる。

10

【0207】

情報検索システム1604は、ユーザの要求を満たすために使用される複数のモジュール1610～1650を含む。いくつかの実施形態では、モジュール1610～1650はそれぞれ、統合プログラムの構成要素であってもよい。いくつかの実施形態では、モジュール1610～1650は、互いに独立して動作し、標準インターフェースを介して互いに通信する独立したプログラム（例えば、マイクロサービス）であってもよい。情報検索システム1604は、分散させることができる。性能および並列性を高めるために、情報検索システム1604は、モジュール1610～1650の複数のインスタンスを含むことができる。

20

【0208】

情報検索システム1604のコントローラモジュール1610は、アプリケーション1602からクエリ（および／または他のタイプの要求）を受信して処理する。コントローラモジュール1610は、情報検索プログラム1604の1つまたは複数の他のモジュール（例えば、モジュール1620～1650）に結合され、要求を満たすために他のモジュール間で調整する。いくつかの実施形態では、要求を満たすプロセスは、要求のタイプに応じて異なり得る。

30

【0209】

情報検索システム1604の知識ベースモジュール1620は、情報間の様々なタイプの関係を識別する知識ベースへのアクセスを提供する。例えば、知識ベースモジュール1620は、所与の関係を共有することが知られている用語の集合（例えば、これらの用語は互いに同義語であってもよい）を格納することができる。いくつかの実施形態では、情報およびそれらの関連する関係は、知識グラフまたは他の適切なデータストレージ技術を使用して格納および検索することができる。

40

【0210】

一般に、知識ベースモジュール1620によって記憶された知識ベースは、関心のある主題に関して包括的であることが望ましい。包括的な一組の関係は、様々な技術を使用して識別および集約され得る。いくつかの実施形態では、知識ベースは、生物医学ドメインの場合は統一医療言語システム（UMLS）などの既存の知識ベースから開始し、次いでドメイン固有情報の他のソースをそれに集約することによって構築することができる。例えば、データは、外部データベース（例えば、公的に利用可能なデータベースおよび専有または顧客固有のデータベース）から集約されてもよい。集約データ間の関係は、ニューラルネットワークモデル（例えば、ニューラルネットワークモデル1560）または集約

50

データから関係を検索するように構成された他の情報検索方法を使用して識別することができる。

【0211】

情報検索システム1604の検索モジュール1630は、クエリ用語に基づいてテキスト（例えば、文書の集合、データベース記録、および／またはこれらに類似するもの）のコーパスを検索することができる検索エンジンを提供する。コーパスは、例えば、公開文献コーパス、臨床データコーパス（例えば、医師のノートおよび他の種類の患者の健康記録）、またはその両方を含むことができる。コーパスは、構造化データまたは非構造化データを含むことができる。いくつかの実施形態では、構造化データは、投与される薬剤の名称などの患者健康記録の1つまたは複数の構造化データフィールドを含むことができる。そのような情報は、例えば、そのような情報が保険目的または様々な他の形式的または法的機能に使用され得るため、患者の健康記録における構造化されていないエントリよりも信頼性が高いと期待され得る。したがって、患者の健康記録内の構造化データフィールドを使用すると、信頼性を向上させることができ、コーパスの少なくともサブセットが機械学習モデルの訓練（またはテスト）データを提供するようにラベル付けされる実施形態では、手動でタグ付けされるデータの量を減らすことができる。これは、感情分類器などのコーパスに基づく機械学習モデルの迅速かつ正確な開発および訓練を容易にすることができる。

10

【0212】

臨床データコーパス、または患者データを含む他のコーパスの使用に関連する1つの課題は、前述したプライバシー上の懸念である。したがって、図1～図14で説明した技術のうちの1つまたは複数を使用して、安全な処理環境を提供し、機密患者データを非特定化またはマスキングすることができる。例えば、図13のエンティティのタグ付けプロセスは、機密情報をマスキングするためにコーパス（または機密情報を含むコーパスの一部）に適用され得る。いくつかの実施形態では、コーパスの検索は、患者の状態のみに基づいてこれらの患者を識別しやすくするまれな状態の患者を除外するように抑制され得る。例えば、患者が最小しきい値サイズ未満のコホート（例えば、類似または同一の状態を有する患者の数）に関連付けられている場合、その患者の記録はコーパスで抑制され、患者の記録は検索結果のリストに返されない。

20

【0213】

いくつかの実施形態では、検索モジュール1630は、クエリ用語に関連すると判定されたコーパスから完全なテキスト文書またはデータベース記録を識別および検索することができる。しかしながら、前述のように、この手法には様々な制限がある。例えば、ユーザが集約された情報（例えば、特定の薬物が特定の疾患に対して有効であることが示されているかどうかを、その薬物に対して行われた研究の集合に基づいて決定するため）に基づいて決定または推論を行おうとしているとき、単に関連文書のリストを含む検索結果は、ユーザのクエリに明確かつ直接応答しない場合がある。したがって、この手法は、所望の情報をつなぎ合わせて適切な決定または推論を行うために検索結果のそれぞれを検査するためにユーザの一部に面倒で時間のかかる努力を必要とする場合がある。

30

【0214】

これらの制限に対処するために、検索モジュール1630は、一致する文書または記録のリストではなく（またはそれに加えて）、クエリ用語に一致するテキストフラグメントのリストを返すように構成され得る。このフラグメントベースの検索手法は、関心のある情報、例えば、ユーザが行おうとしている判定または推論に直接関連する情報を含む文書の局所化された部分（例えば、いくつかの単語、文、または段落）をもたらす。このようにして、文書が一般にユーザのクエリに関連しないが応答フラグメントを含む場合、文書全体が一致する文書のリストに含まれるのに十分に関連しない場合でも、応答テキストフラグメントが検索結果に返される。逆に、文書がクエリ用語と強く一致するが、クエリに直接応答するいかなるフラグメントも含まない場合、文書は、一致するテキストフラグメントのリスト内のいかなるフラグメントも生成しない可能性がある。さらに、所与の文書

40

50

または記録が2つ以上の一致するテキストフラグメントを含む場合、単一の文書または記録は、返されたテキストフラグメントのリストに複数のエントリをもたらすことができる。結果として、フラグメントベースの検索手法は、検索結果の全体的な関連性および完全性を改善することができる。フラグメント検索モジュールの例示的な実施形態は、図17を参照して以下でさらに詳細に説明される。

【0215】

情報検索システム1604の統計解析モジュール1640は、検索モジュール1630によって提供される検索結果のリストなどの情報検索システム1604の他のモジュールからの情報を統計的に解析するためのツールを提供する。ユーザから受信した要求の種類などの要因に応じて、広範囲の統計解析を実行することができる。例えば、統計解析モジュール1640は、検索モジュール1630からの検索結果のリストに現れる様々なエンティティおよび用語の統計的有意性を計算することができる（例えば、検索結果における所与の用語の出現回数のカウント、用語の他の用語との共起回数のカウント、用語の他の用語に対する有意性を比較するためのスコアまたはランキングなどである）。統計解析を実行する際に、統計解析モジュール1640は、情報検索システム1604の他のモジュールと通信し、そこから情報を検索することができる。統計解析モジュール1640を使用して計算され得る統計的有意性メトリックの例は、図18を参照して以下でさらに詳細に説明される。

【0216】

情報検索システム1604の機械学習モジュール1650は、検索モジュール1630によって提供される検索結果のリストなどの情報検索システム1604の他のモジュールからの情報に機械学習モデル（例えば、ニューラルネットワークモデル1560）を適用するためのツールを提供する。いくつかの実施形態では、機械学習モジュール1650は、検索結果のテキストを解析するための自然言語処理（NLP）パイプラインを含むことができる。NLPパイプラインは、NLPプリミティブ（例えば、トークン化、埋め込み、名前付きエンティティ認識などである）を含むことができる。さらに、NLPパイプラインは、否定表現ファインダ、感情分類器、エンティティ抽出器などを含むがこれらに限定されない、事前訓練されたルールベースのモデルまたは機械学習モデルを含むことができる。結果間の関係および関連付けを識別するために、NLPパイプラインの出力に対してさらなる統計的解析を実行することができる。

【0217】

図17は、いくつかの実施形態による、フラグメント検索のためのデータフロー1700の簡略図である。図16と一致するいくつかの実施形態では、データフロー1700は、検索モジュール1630を使用して実施することができる。

【0218】

コーパス1710は、1つまたは複数のテキスト文書またはデータベース記録の集合などのテキストの集合に対応する。例えば、コーパス1710は、デバイス1501～1509から受信した文書1521～1529に対応してもよく、および／またはデバイス1503によってローカルに格納された文書を含んでもよい。いくつかの実施形態では、コーパス1710は、メモリ1540、データベース1570、オンチップメモリ（例えば、キャッシュ）などに格納することができる。コーパス1710内の文書は、ネイティブフォーマット（例えば、デバイス1501～1509から受信したフォーマット）で格納することができ、または受信した文書に対して様々な前処理動作を実行して、文書の内容またはフォーマットを修正することができる。例えば、非テキストデータ（例えば、画像データ）および／またはメタデータが文書から削除されてもよく、テキストデータが文書から抽出されてもよい（例えば、光学文字認識によって）などである。文書のフォーマットは、統一されたフォーマットに変換されてもよく、または文書からのデータを使用してデータベース（例えば、データベース1570）を投入してもよい。いくつかの実施形態では、コーパス1710は動的に更新されてもよい。

【0219】

コーパス1710の内容は、一般的な主題（例えば、様々なトピックをカバーするニュース記事またはWikipediaエントリのコレクション）またはドメイン固有の主題に関連することができる。例示的に、コーパス1710は、生物医学の主題に関連し得る。例えば、コーパス1710は、生物医学分野に関連する雑誌記事、参考教科書、特許出願、ウェブサイトなどからのテキストを含むことができる。コーパス1710は、分子データベース、科学文献、保険文書、製薬会社のウェブサイト、ニュースフィード、規制情報（臨床試験、SEC提出、知財）などの多種多様なソースから引き出すことができる。図1～図15と一致するいくつかの実施形態では、コーパス1710は、上述した技術に従って処理および格納されたプライベートデータ（例えば、ヘルスケア記録）を含むことができる。

10

【0220】

図17に示すように、コーパス1710は複数のサブセットに分割される。各サブセットは、シャード1721～1729のうちのそれぞれのシャードに提供され得る。いくつかの実施形態では、シャード1721～1729間でコーパス1710を分割することは、分散コンピューティングリソース（例えば、分散プロセッサおよび／またはストレージシステムを使用する）を使用してコーパス1710の処理を容易にすることができる。例えば、シャード1721～1729のうちの1つまたは複数は、データセンタ内の異なるマシンおよび／または異なるデータセンタに配置されてもよい。いくつかの実施形態では、コーパス1710のサブセットの各々は、サイズがほぼ等しくてもよく、例えば、それらは同様の総ディスクスペースを占有してもよく、またはそれらは同様の数の文書を含んでもよい。

20

【0221】

シャード1721～1729の各々は、対応する反転リスト1731～1739を含む。反転リスト1731～1739のそれぞれは、コーパス1710の対応するサブセット内のトークン（例えば、単語）ごとに、コーパス1710のサブセット内のトークンの出現のリストを識別する。例えば、反転リスト1731～1739は、コーパス1710のサブセット内のトークンの各出現の位置（例えば、コーパス1710のサブセット内の各文書の連結に対応するテキストの連続した配列内の位置）を識別することができる。いくつかの実施形態では、反転リスト1731～1739は、トークンが発生する文書に対応する文書識別子、トークンの発生に対する文書内のオフセットなどを識別することができる。いくつかの実施形態では、反転リスト1731～1739内の各エントリは、各トークンの各出現に対する複数の位置識別子を含み得る。複数の識別子は、（1）連結された文書の連続した配列内のトークンの出現の配列インデックス、（2）出現の文書識別子、および（3）識別された文書内の出現までのオフセットを識別するトリプレットなどの適切なデータ構造に格納されてもよい。

30

【0222】

いくつかの実施形態では、反転リスト1731～1739は、トークンの効率的な検索を容易にするために順序付けられてもよい。例えば、反転リスト1731～1739は、コーパス1710のサブセットに対応するテキストの配列内の各トークンの位置の昇順に基づいて順序付けられてもよい。反転リスト1731～1739は、トークンに対応する整数が与えられると、反転リスト1731～1739を含むデータ構造がトークンの出現の対応するリストを効率的に返すように、各トークンに関連する整数値を使用してインデックス付けされ得る。

40

【0223】

シャード1721～1729の各々は、対応するフラグメント検索モジュール1741～1749をさらに含む。フラグメント検索モジュール1741～1749の各々は、フラグメントクエリ1752を受信し、反転リスト1731～1739からのデータにアクセスすることによってフラグメントクエリに対する応答を生成するように構成される。フラグメントクエリ1752は、検索アグリゲータ1754を用いて、フラグメント検索モジュール1741～1749へ配信され得る。次いで、検索アグリゲータ1754は、フ

50

ラグメント検索モジュール1741～1749によって生成された検索結果を受信し、集約することができる。その後、検索結果は、その後の解析1756に使用することができる。例えば、図16と一致するいくつかの実施形態では、解析1756は、知識ベース1620、統計解析モジュール1640、または機械学習モジュール1650のうちの1つまたは複数を使用して実行することができる。

【0224】

いくつかの実施形態では、フラグメントクエリ1752は、フラグメント検索の所望の検索基準を示す1つまたは複数のクエリパラメータを含む。例えば、フラグメントクエリ1752は、クエリパラメータ（例えば、AND、OR、およびNOTなどのブール演算子によって任意選択的に結合された、検索される1つまたは複数のトークン、単語、または複数単語句の組み合わせである）を含み得る。フラグメントクエリ1752はまた、フラグメント検索モジュール1741～1749によって返されたテキストフラグメントの所望のサイズを示すサイズパラメータを含んでもよい。フラグメントクエリ1752は、文書内のフラグメントが検索結果に含まれるための前提条件として文書が満たすべき1つまたは複数の基準を指定する文書パラメータをさらに含むことができる。例えば、文書パラメータは、適格な文書が指定された単一もしくは複数の語句（またはそれらの論理的組み合わせ）を含む基準、または適格な文書が文書メタデータと関連付けられる基準（例えば、著者名、出版年、文書ソース、文書タイプなど）を含むことができる。そのような実施形態と一致して、フラグメントクエリ352は、`<Frag Query, Fragment Size, Document Specifier>`として表されるデータ構造など、様々な検索パラメータを送信および処理するための適切なデータ構造を使用して表されてもよく、`Frag Query`はクエリパラメータを表し、`Fragment Size`はサイズパラメータを表し、`Document Specifier`は検索結果に含まれる文書の適格条件を表す。

【0225】

図18は、いくつかの実施形態による、クエリに応答して情報を検索するための方法1800の簡略図である。図15～図17と一致するいくつかの実施形態によれば、方法1800は、情報検索プログラム1550の実行中にプロセッサ1530によって実行されてもよい。例えば、方法1800は、コントローラモジュール1610を使用して実行することができる。

【0226】

プロセス1810において、クエリがコントローラ（例えば、コントローラモジュール1610）によって受信される。クエリは、1つまたは複数のトークン（例えば、単語、文など）、ブール演算子、制約、フィルタ、および様々な他のパラメータを含むことができる。いくつかの実施形態では、クエリは、アプリケーション1602などのアプリケーションによって送信された要求に含まれてもよい。そのような実施形態と一致して、要求はAPIを介して受信されてもよい。要求の受信に応答して、コントローラは、以下のプロセス1820～1850のうちの1つまたは複数に従ってクエリに応答する情報を収集することによって要求を処理および応答することができる。分散コンピューティング環境では、情報は、異なるネットワークノードに配置された様々なモジュールとの間で送信される通信を介して収集することができる。

【0227】

プロセス1820において、クエリは、1つまたは複数の関連用語を含むように拡張される。関連用語は、知識ベースモジュール1620の知識ベースなどの知識ベースから検索することができる。いくつかの実施形態では、関連用語は、元のクエリ内の用語と所定の関係を共有することができる（例えば、これらの用語は互いに同義語であってもよい）。このようにして、関連用語を検索することは、ユーザによって提供された用語と同様の意味を有する用語を含むようにクエリを広げることができる。

【0228】

プロセス1830において、検索結果のリストが、拡張されたクエリに基づいて検索さ

10

20

30

40

50

れる。検索結果のリストは、クエリに一致する文書またはパッセージ（またはそのような文書またはパッセージを識別する情報）を含む。検索結果のリストは、検索モジュール1630などの検索エンジンまたは検索モジュールを介して検索することができる。いくつかの実施形態では、検索結果は、コントローラによって提供されたクエリ用語に基づいて、テキストのコーパス（例えば、文書の集合、データベース記録、および／またはこれらに類似するもの）から検索され得る。例えば、検索結果は、フラグメントベースの検索の結果に対応し得る。この手法では、検索結果は、関心のある情報を含むフラグメント（例えば、いくつかの単語、文章、段落、または文書の他の局所化された部分）を含む。フラグメントベースの検索の例示的な実施形態は、図19を参照して以下でさらに詳細に説明される。

10

【0229】

いくつかの実施形態では、検索結果のリストは、クエリの「コンテキスト」と呼ばれてもよく、適切なデータ構造を使用して格納またはインデックス付けされてもよい。コンテキストは、コーパス内の一致するクエリ用語の近くに現れる用語（例えば、コーパス内の一致するクエリ用語のn個のトークンのウィンドウサイズ内に現れる用語）と共に、クエリ用語を含むコーパスのウィンドウを含む。コンテキストは、バイナリであっても非バイナリであってもよい。バイナリコンテキストでは、コーパス内の用語は、コンテキスト（例えば、それらがクエリ用語の出現のn個のトークン内にある場合）に含まれるか、またはコンテキストから省略される。非バイナリまたは「滑らかな」コンテキストでは、コーパス内の用語は、クエリ用語からの距離などの要因に基づいて重み付けされ得る（例えば、0と1との間の値が割り当てられる）。例えば、非バイナリコンテキスト内の用語に割り当てられた重みは、クエリ用語からの用語の距離に基づいて指数関数的に減衰することができる。

20

【0230】

プロセス1840において、検索結果のリスト内で、1つまたは複数のエンティティが任意選択的に識別される。例えば、生物医学的用途の文脈において、エンティティの例示的な例は、薬物、疾患、遺伝子、製薬会社、研究機関などの名称を含み得る。いくつかの実施形態では、1つまたは複数のエンティティは、知識ベースモジュール1620の知識ベースなどの知識ベースを参照することによって識別することができる。例えば、知識ベースは、検索結果のリスト内のエンティティを識別するために検索結果のリストをエンティティの集合と比較することができるように、エンティティの集合を記憶することができる。いくつかの実施形態では、名称付きエンティティ認識などの自然言語処理技術を使用して、検索結果のリスト内のエンティティを正確に識別することができる。

30

【0231】

プロセス1850において、検索結果のリスト内の1つまたは複数の用語またはエンティティについての統計解析が提供される。統計解析は、統計解析モジュール1640などの統計解析モジュール、機械学習モジュール1650などの機械学習モジュール、またはモジュールの組み合わせによって実行されてもよい。いくつかの実施形態では、統計解析は、関心のある各用語またはエンティティについて、関係スコア、有意性スコア、またはその両方を計算することを含んでもよい。そのような統計的解析は、検索結果における所与の用語の出現数のカウント、用語と他の用語との共起数のカウント、用語の他の用語に対する有意性を比較するスコアまたはランキングなどに基づくことができる。

40

【0232】

いくつかの実施形態では、有意性スコアは、所与の用語（例えば、トークン、m個の単語のフレーズ、エンティティなど）のユーザクエリとの関連性を捕捉する統計的有意性メトリックを計算することによって測定され得る。実例として、統計的有意性メトリックは、mが1である（すなわち、フレーズが単一の単語である）場合、以下の式を使用して計算することができる。

【0233】

【数4】

50

$$k \cdot \left(\ln \left[\frac{\frac{k}{n}}{p} \right] - 1 \right)$$

【0234】

ここで、kは、検索結果のリスト（例えば、クエリに応答して検索モジュール1630によって返されるテキストフラグメントのリスト）内のフレーズの出現数であり、nは検索結果のリスト内の単語の総数であり、pは比率 $N(w)/N$ であり、ここで、 $N(w)$ はコーパス内のフレーズの出現数であり、Nはコーパス内の単語の総数である。

10

【0235】

mが1より大きい（すなわち、フレーズが複数の単語のフレーズである）場合、以下の式を使用することができる。

【数5】

$$k \cdot \left(\ln \left[\frac{\frac{k}{n - (m-1)k}}{p} \right] - 1 \right) + n \cdot \ln[1 - p + p \cdot (1 - p)^{m-1}]$$

【0236】

検索結果のリストが非バイナリコンテキストを含むいくつかの実施形態では、統計的有意性メトリックは、検索結果のリスト内のエンティティの各外観に割り当てられた重みに基づいて調整され得る。例えば、上記の式におけるkの値は、検索結果におけるエンティティの出現数の重み付けされたカウントに対応し得る。

20

【0237】

プロセス1850で提供される上記の統計解析は例示的なものであり、様々な代替形態が可能であることを理解されたい。例えば、統計解析モジュール1640は、関連性スコアEMS(Q, t)および統計的有意性メトリックpVal(Q, t)を決定してもよい。これらの値は、m個の単語のフレーズtのユーザクエリQとの関連性を捕捉することができる。

30

【0238】

プロセス1860において、1つまたは複数の濃縮セットが、任意選択的に、1つまたは複数のエンティティの統計解析に基づいて生成される。いくつかの実施形態では、濃縮セットは、プロセス1850で識別された最上位エンティティ（例えば、最も高い有意性または関係スコアを有するエンティティ）を含んでもよい。例えば、生物医学の文脈では、1つまたは複数の濃縮セットは、一組の薬物、一組の疾患、および一組の遺伝子を含んでもよい。例えば、濃縮された一組の薬物は、最も重要であるか、またはユーザのクエリに関連すると判定されたn個の薬物を識別することができる。

【0239】

プロセス1870において、検索結果のリストまたは1つもしくは複数の濃縮セットのうちの少なくとも一方を含むクエリに対する応答が提供される。いくつかの実施形態では、応答は、アプリケーション1602などのアプリケーションに送信され、ユーザに表示されてもよい。応答は、ユーザが検索結果または濃縮セットと対話することを可能にするために、対話型ユーザインターフェース要素をユーザに提供することができる。例えば、濃縮セット内の項目について、ユーザは、項目上にホバーして、項目に関連する統計解析（例えば、有意性スコア、関係性スコアなど）を見ることができる。

40

【0240】

図19は、いくつかの実施形態による、フラグメント検索を実行するための方法1900の簡略図である。図15～図18と一致するいくつかの実施形態では、方法1900は、検索モジュール1630などの検索モジュールによって実行されてもよい。次いで、方

50

法 1 9 0 0 を使用して生成されたフラグメント検索結果は、方法 1 8 0 0 のプロセス 1 8 3 0 においてコントローラによって検索され得る。

【 0 2 4 1 】

プロセス 1 9 1 0 において、コーパス 1 7 1 0 などのコーパスは、複数のサブセットに分割される。コーパスは、複数のテキスト文書またはデータベース記録を含む。いくつかの実施形態では、コーパスのサブセットの各々は、サイズがほぼ等しくてもよく、例えば、それらは同様の総ディスクスペースを占有してもよく、またはそれらは同様の数の文書を含んでもよい。

【 0 2 4 2 】

プロセス 1 9 2 0 において、複数のサブセットの各々は、シャード 1 7 2 1 ~ 1 7 2 9 などの対応する複数のフラグメントに分散される。各シャードは、分散方式でコーパスの一部を処理する役割を担う。いくつかの実施形態では、シャード間でコーパスを分割することは、分散コンピューティングリソースを使用して（例えば、分散プロセッサおよび／またはストレージシステムを使用して）コーパスの処理を容易にすることができる。例えば、シャードの 1 つまたは複数のは、データセンタ内の異なるマシンおよび／または異なるデータセンタに配置されてもよい。

10

【 0 2 4 3 】

プロセス 1 9 3 0 において、各シャードについて、コーパスのそれぞれのサブセット内の 1 つまたは複数の文書が連結されてテキストシーケンスを形成する。例えば、テキストシーケンスは、1 つまたは複数の文書に関して連続していてもよい。

20

【 0 2 4 4 】

プロセス 1 9 4 0 において、シャードごとに、反転リストが生成される。反転リストは、コーパス内のトークン（例えば、語彙単語）ごとのエントリを含む。各エントリは、コーパス内のトークンの出現のリストを含む。例えば、出現のリストは、プロセス 1 9 3 0 において形成された配列内のトークンの各出現の位置を識別することができる。いくつかの実施形態では、反転リストは、トークンが発生する文書に対応する文書識別子、トークンの発生に対する文書内のオフセットなどを含むことができる。いくつかの実施形態では、反転リスト内の各エントリは、各トークンの各出現に対する複数の位置識別子を含んでもよい。複数の識別子は、（１）連結された文書の連続した配列内のトークンの出現の配列インデックス、（２）出現の文書識別子、および（３）識別された文書内の出現までのオフセットを識別するトリプレットなどの適切なデータ構造に格納されてもよい。

30

【 0 2 4 5 】

反転リストは、トークンの効率的な検索を容易にするために順序付けられてもよい。例えば、反転リストは、テキストの配列内の各トークンの位置の昇順に基づいて順序付けられてもよい。反転リストは、トークンに対応する整数が与えられると、反転リストを含むデータ構造がトークンの出現の対応するリストを効率的に返すように、各トークンに関連する整数値を使用してインデックス付けされ得る。

【 0 2 4 6 】

プロセス 1 9 5 0 において、フラグメントクエリ 1 7 5 2 のようなフラグメントクエリが、検索アグリゲータ 1 7 5 4 のような検索アグリゲータによって受信される。フラグメントクエリは、フラグメント検索のための所望の検索基準を示す 1 つまたは複数のクエリパラメータを含む。例えば、フラグメントクエリは、クエリパラメータ（例えば、AND、OR、および NOT などのブール演算子によって任意選択的に結合された、検索される 1 つまたは複数のトークン、単語、または複数単語のフレーズの組み合わせ）を含んでもよい。フラグメントクエリはまた、検索結果におけるテキストフラグメントの所望のサイズを示すサイズパラメータを含んでもよい。フラグメントクエリは、文書内のフラグメントが検索結果に含まれるための前提条件として、文書が満たすべき 1 つまたは複数の基準を指定する文書パラメータをさらに含むことができる。例えば、文書パラメータは、適格な文書が指定された単一もしくは複数の語句（またはそれらの論理的組み合わせ）を含む基準、または適格な文書が文書メタデータと関連付けられる基準（例えば、著者名、出版

40

50

年、文書ソース、文書タイプなど）を含むことができる。

【0247】

プロセス1960において、フラグメントクエリが複数のシャードに分散される。フラグメントクエリを受信すると、複数のシャードの各々は、それぞれの反転リストに対してルックアップを実行して一致するフラグメントを識別し、一組のフラグメント検索結果を生成する。いくつかの実施形態では、フラグメントクエリに含まれる検索基準（例えば、検索に含める適格な文書に対する制限）は、ルックアップ中に適用され得る。

【0248】

プロセス1970において、複数のシャードからの一組のフラグメント検索結果は、検索アグリゲータによって集約される。例えば、検索結果は、コンパイル、連結、ソート、ランク付けなどを行うことができる。検索結果を集約すると、検索結果は、さらなる解析のためにコントローラまたは別のモジュールに提供されてもよく、またはユーザに返されてもよい。

【0249】

図20Aおよび図20Bは、いくつかの実施形態による、情報検索システム1604などの情報検索システムのグラフィカルインターフェース2000のスクリーンショットである。図15～図19と一致するいくつかの実施形態では、グラフィカルインターフェース2000は、アプリケーション1602に関連付けられてもよい。例示目的のために、アプリケーションは、ユーザが生物医学領域で情報を検索するためのツールを提供するものとして示されている。

【0250】

図20Aに示すように、グラフィカルインターフェース2000は、ユーザが検索語を入力または入力することができる検索入力フィールド2010を含む。図20Aおよび図20Bの図において、検索語は「イマチニブ」と名付けられた薬物である。

【0251】

ユーザが検索語を入力したことに応答して、情報検索システムは、検索文字列に関連する情報を検索し、グラフィカルユーザインターフェース2000を介して情報を表示する。図15～図19と一致する実施形態では、情報は、コントローラ1610、知識ベースモジュール1620、検索モジュール1630、統計解析モジュール1640、および機械学習モジュール1650などの情報検索システムの様々なモジュールを使用して検索することができる。例えば、グラフィカルインターフェース2000は、検索語に一致する一致する記事2020（またはその一部）を表示することができる。この場合、一致する記事2020は、「イマチニブ」についてのWikipediaの記事に対応する。

【0252】

グラフィカルインターフェース2000は、検索語に関連するエンティティまたは概念の濃縮セット2031～2039をさらに表示することができる。いくつかの実施形態では、濃縮セット2031～2039は、検索語に対して最も有意であると判定されたエンティティ（例えば、統計的有意性スコアが最も高いエンティティ）を含み得る。図20Aに示すように、濃縮セット2031～2039は、イマチニブ2031に関連する疾患の集合、イマチニブ2032に関連する薬物の集合および薬物のクラス、ならびにイマチニブ2039に関連する細胞および組織の集合を含む。

【0253】

いくつかの実施形態では、グラフィカルインターフェース2000は、グラフィカルインターフェース2000に表示される1つまたは複数の用語または概念に関連する統計解析および／または機械学習解析を表示することができる。解析は、デフォルトで表示されてもよいし、ユーザが特定の用語または概念をホバーしたときに動的に表示されてもよい。いくつかの実施形態では、統計解析モジュール1640、機械学習モジュール1650、またはそれらの組み合わせを使用して、統計解析または機械学習解析を実行することができる。解析で使用される基礎となるデータは、知識ベース1620および／または検索モジュール1630から検索することができる。例えば、図20に示すように、ユーザが

様々なスコアリング指標および発生／共起指標（例えば、コーパスにおける用語の出現回数および検索用語との用語の共起回数）を含む用語「メシル酸イマチニブ」をホバーすると、一組の解析 2040 が表示される。

【0254】

いくつかの実施形態では、グラフィカルインターフェース 2000 は、ユーザ選択に応答して検索語に関する追加情報を表示するウィジェット 2050 を提供することができる。図 20B に示すように、ウィジェット 2050 は、検索語に関連する文献証拠（例えば、イマチニブの有効性に関する発表された研究の編集）を表示する。例えば、文書 2061～2069 は、検索語または関連用語を含む文書 2061～2069 内のテキストフラグメントと共に示す。このようにして、本アプリケーションの使用者は、イマチニブが手元の目的のために関連性があるかまたは臨床的に有効である可能性が高いかどうかを効率的に決定することができる。

【0255】

図 21 は、いくつかの実施形態による情報検索のための方法 2100 の簡略図である。図 15～図 19 と一致するいくつかの実施形態によれば、方法 2100 は、情報検索システム 1500 などの情報検索システムを使用して、生物医学領域内の関係を識別および解析する方法の例示的な例を提供することができる。方法 2100 におけるユーザの例示的な目的は、科学文献が第 1 のエンティティ（例えば、薬物、化学物質など）と第 2 のエンティティ（例えば、疾患、生体分子、遺伝子など）との間の特定の関係の証拠を示すかどうかを決定することである。特定の関係の例には、有害事象関係、ターゲット関係（例えば、薬物または化学物質が生体分子、遺伝子などをターゲットとする関係）、またはモダリティ関係（例えば、薬物または化学物質が特定の治療モダリティを有する関係）が含まれるが、これらに限定されない。実例として、以下に記載される方法 2100 の実施形態では、第 1 のエンティティは薬物に対応し、第 2 のエンティティは疾患に対応する。しかしながら、方法 2100 の様々な実施形態を使用して、様々な他の種類のエンティティ間の関係を解析することができることを理解されたい。

【0256】

プロセス 2110 において、薬物、疾患）、および関係のそれぞれを指定するクエリが受信される。いくつかの実施形態では、薬物、疾患、および関係は、グラフィカルインターフェースの入力形態を使用してユーザから収集することができる。

【0257】

プロセス 2120 において、クエリに応答して、クエリに一致するテキストフラグメントのリスト（例えば、薬物と疾患の両方に言及することによって）が取得される。いくつかの実施形態では、テキストフラグメントのリストは、検索モジュール 1630 などの検索システムを使用してコーパス（例えば、コーパス 1710）から検索され得る。いくつかの実施形態では、例えば、知識ベースモジュール 1620 を使用して、クエリ用語に関連する同義語および他の用語を識別することによって、検索の範囲を拡張することができる。次いで、同義語と一致するテキストフラグメントを検索結果に含めることができる。いくつかの実施形態では、検索の範囲は、検索される文書のタイプに対する制約（例えば、検索は、特定の時間枠の間に発行された特定のジャーナルからの文書に限定されてもよい）など、検索の範囲に対する制約を課すことによって制限され得る。

【0258】

プロセス 2130 において、各テキストフラグメントは、1 つまたは複数の文に分割される。いくつかの実施形態では、テキストフラグメントは、機械学習モジュール 1650 などの第 1 の自然言語処理または機械学習システムを使用して分割されてもよい。

【0259】

プロセス 2140 において、各テキストフラグメントについて、関係が成立するかどうか（例えば、薬物と疾患が有害事象関係を有するかどうか）が推測される。いくつかの実施形態では、推測は、機械学習モジュール 1650 などの第 2 の機械学習システムを使用して行われてもよい。例えば、第 2 の機械学習システムは、文のリストとしてフォーマッ

10

20

30

40

50

トされたテキストフラグメントを含む入力に基づいて有害事象関係が成り立つかどうかを識別するように訓練された機械学習モデルをホストすることができる。

【0260】

プロセス2150において、各テキストフラグメントについて、関係が成立することを推測が示すかどうかに基づいてメトリックが決定される。例えば、メトリックは、有害事象関係（または他の特定の関係）を示すと推測されるテキストフラグメントの割合を含んでもよい。

【0261】

プロセス2160において、メトリックおよびテキストフラグメントのリストのうちの1つまたは複数のテキストフラグメントを含む解析の結果が表示される。プロセス2140において指定された関係を示すと推論されたか否かに従って、1つまたは複数のテキストフラグメントをラベル付けすることができる。解析の結果を表示する例示的な実施形態は、図22Aおよび図22Bを参照して以下でさらに詳細に説明される。

【0262】

図21は単なる例示であり、様々な他の実施形態が可能であることを理解されたい。例えば、方法2100は、他の種類のエンティティ（例えば、薬物および疾患以外のエンティティ）間の関係を識別するために使用され得る。

【0263】

図22Aおよび図22Bは、いくつかの実施形態による情報検索システムのグラフィカルインターフェース2200のスクリーンショットである。図15～図21と一致するいくつかの実施形態によれば、グラフィカルインターフェース2200は、方法2100の間に生成された結果を表示することができる。図22Aにおいて、ユーザは、薬物ダサチニブが肺動脈高血圧症を有害事象として引き起こすかどうかを決定しようとする。表示された結果は、検索結果のテキストフラグメント316個のうち314個が有害作用を示すとラベル付けされたので、この有害事象関係が真であるという圧倒的な証拠があることを示している。一方、図22Bにおいて、使用者は、薬物イマチニブが有害事象として白血病に関連し得るかどうかを決定しようとする。ここで、圧倒的な証拠は、805個の一致するテキストフラグメントのうち31個のみが有害作用を示すとラベル付けされているので、そのような有害関係が存在しないことを示唆している。

【0264】

上記の開示が示唆するように、電子健康記録（EHR）および患者チャートなどの健康記録は、患者の状態に関する広範囲の情報を取り込む。健康記録は、医師によって書かれたメモのテキストなど、非構造化または半構造化フォーマットで表されることが多い。人間は、テキスト（例えば、英数字、記号、絵文字などのシーケンスとして表される単語、フレーズ、および他の用語である）の形式で提示される情報の意味を理解することができるが、コンピュータ技術は、一般に、人間が読める形式のテキストの意味論的意味を理解することができない。さらに、患者の健康記録に反映される患者の状態は、経時的に変化し得る。したがって、健康記録の拡張キュレーションおよび時間的識別のための技術を開発することが望ましい。

【0265】

本開示のさらなる背景は、*Augmented Curation of Unstructured Clinical Notes from a Massive EHR System Reveals Specific Phenotypic Signature of Impending COVID-19 Diagnosis*と題する論文（以下、「増強キュレーション論文」という）に提供されており、これは米国仮特許出願第63/012,738号に添付されており、その全体が参照により本明細書に組み込まれる。

【0266】

図23は、一部の実施形態による、健康記録の強化管理のための方法2300の簡略図である。健康記録の拡張キュレーションは、電子健康記録（EHR）、患者チャートなど

10

20

30

40

50

の生の健康記録を患者表現型（例えば、患者の症状、診断、治療などのスナップショット）の構造化表現に変換する。次いで、構造化表現を視覚化し、統計的または機械学習解析などのための入力として使用することができる。いくつかの実施形態では、生の健康記録は、例えばプライバシー規制（例えば、HIPAA）に準拠するために、患者の個人情報または使用できない可能性がある他の情報をマスキングするために非特定化されてもよい。そのような選択的情報マスキング技術の例は、上記で参照した米国仮特許出願第62／865,030号明細書および米国仮特許出願第62／985,003号明細書、ならびに本開示の図1～図14にさらに詳細に記載されている。健康記録の非特定化は、方法2300の前、間、または後に行うことができる。図1～図22Bと一致するいくつかの実施形態では、方法2300の少なくとも一部は、システム1500を使用して実行することができる。

10

【0267】

プロセス2310において、管理された一組の健康記録が受信される。プロセス2310の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。例えば、管理された一組の健康記録は、手動で管理された健康記録に対応することができる。生の（または非特定化された）健康記録を管理することは、特定の症状、疾患、投薬などの同義語などの関連する単語およびフレーズを識別およびグループ化することを含むことができ、その例は上記の増補された管理論文に記載されている。例えば、とりわけ、「SOB」、「息切れ」、および「呼吸困難」などの用語は、それぞれ、キュレーション中の症状エンティティ「息切れ」に対応するものとして識別され得る。健康記録におけるエンティティの所与の発生は、感情、例えば、症状の有無、医薬品の有効性または無効性などを識別する肯定的／否定的感情とさらに関連付けられ得る。このようにして、管理された健康記録は、患者の表現型の構造化された表現を提供することができる。キュレーションは、医師または生の健康データを解釈する際の専門知識または訓練を有する他の個人によって実行されてもよい。管理された一組の健康記録のサイズは、ニューラルネットワークモデルを訓練するために通常使用されるものと比較して小さくてもよく、したがって、あまり労働集約的でなく時間のかかるキュレーションで所望の精度を達成することができる。例えば、管理された一組の健康記録は、100人の患者のチャートからのデータ（または所与の症状、疾患、または他のエンティティを伝えるために使用される様々な単語およびフレーズを捕捉するのに十分な別の数の記録）に対応することができるが、典型的なニューラルネットワークモデルは、何千もの管理された記録を使用して訓練することができる。

20

30

【0268】

プロセス2320において、管理された一組の健康記録を使用してニューラルネットワークが訓練される。プロセス2320の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。例えば、ニューラルネットワークは、BERTベースのニューラルネットワークに対応することができる。いくつかの実施形態では、ニューラルネットワークは、プロセス2310で受信された管理された一組の健康記録よりも大きくなり得る健康データのコーパスを使用して以前に訓練されていてもよい。ニューラルネットワークは、所与のトークンに含まれるエンティティ（例えば、症状、疾患、投薬など）を抽出し、それらを肯定的／否定的感情に従って分類するニューラルネットワーク分類器に対応することができる。感情は、例えば自然言語処理（NLP）技術を使用して周囲の語法に基づいて判定され得る。例えば、ニューラルネットワークは、語句「患者は息切れを呈した」および「呼吸困難について陰性」がそれぞれ症状エンティティ「息切れ」を含むが、異なる感情（第1のフレーズは症状の存在を示し、後者は症状の非存在を示す）を有することを検出することができる。いくつかの実施形態では、管理された一組の健康記録は、ニューラルネットワークを訓練するために使用されたコーパス内のエンティティに加えてエンティティを含んでもよい。したがって、管理された一組の健康データは、予め訓練されたニューラルネットワークによって提供される感情解析を実行する際の精度を活用しながら、ニューラルネットワークが認識することが

40

50

できる一組のエンティティを拡張することができる。

【0269】

いくつかの実施形態では、ニューラルネットワークは、プロセス2320中の訓練の結果として特定の性能メトリックを達成することができる。例えば、性能メトリックは、ニューラルネットワークの精度に対応することができ、検査目的のために確保され、訓練に使用されない管理された一組の健康記録の一部などの検査データを使用して測定することができる。ニューラルネットワークの典型的な訓練セットと比較して管理された一組の健康記録のサイズが限られているため、ニューラルネットワークが達成する性能メトリックはそれに応じて制限され得る。

【0270】

プロセス2330において、未管理の一組の健康記録が受信される。プロセス2330の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。例えば、未管理の一組の健康記録は、生の電子健康記録または患者チャートを含むことができ、これは上述のように非特定化することができる。未管理の一組の健康記録は、トークン化されてもよく、例えば、文、単語、または他のテキストフラグメントに分割されてもよい。いくつかの実施形態では、未管理の一組の健康記録のサイズは、管理された一組の健康記録よりも大きくてもよい（例えば、より多くの記録、より多くの患者データの全体量、またはその両方を含むことができる）。

【0271】

プロセス2340において、未管理の一組の健康記録は、プロセス2320の訓練されたニューラルネットワークを使用して管理される。プロセス2340の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。プロセス2340は、プロセス2310および2340からの管理された一組の健康記録を含む拡張管理された一組の健康記録を生成することができる。未管理の一組の健康記録を管理することは、各記録の各トークンに対してエンティティ認識および感情解析を実行し、各記録でキャプチャされた患者表現型の構造化表現を生成することを含み得る。いくつかの実施形態では、プロセス140で生成された管理された健康記録は、検証、例えば、受け入れまたは拒否され、再分類され得る。検証は、管理された健康記録のエラーを識別するのに十分な専門知識を有する医師または他の熟練したまたは訓練された個人に管理された健康記録を提供することを含むことができる。個人は次に、管理された健康記録の精度を確認するか、管理された健康記録の修正バージョン（例えば、誤ったエントリが訓練された個人によって再分類されたバージョン）を提供することができる。一般に、時折発生する不正確さを検出して対処するために自動的に管理された健康記録を検証することは、基礎となる生の健康記録を手動で管理するよりも効率的に実行され得る。検証された管理された健康記録は、拡張キュレーションプロセスが進行するにつれて、ニューラルネットワークモデルを反復的に再訓練するために、または新しいニューラルネットワークモデルを訓練するために使用することができ、キュレーション効率およびモデル精度の向上につながる。プロセス2340は、管理された結果のフィルタリングを連続的に少なくすることを含む1つまたは複数の段階で実行することができ、これは、ニューラルネットワークが再訓練されるにつれて増加するキュレーションの精度を反映する。

【0272】

したがって、方法2300は、生の（および／または非特定化された）健康記録に基づいて患者表現型の構造化表現を生成する。これらの構造化表現は、生の健康記録よりも計算解析に適している場合がある。例えば、構造化表現は、健康記録の構造化データベースを生成するために使用されてもよい。データは、視覚化されてもよく、統計的または機械学習解析のための入力として使用されてもよい。構造化表現を使用して実行され得る計算解析の例は、上記で参照した米国仮特許出願第62／962,146号明細書および米国仮特許出願第62／985,003号明細書、増強キュレーション論文、ならびに図1～図22にさらに詳細に記載されている。

【0273】

いくつかの実施形態では、方法 2300 に基づいて取り込まれた健康記録の構造化データベースを使用して、濃縮解析を実行することができる。例えば、濃縮解析は、患者が疾患を有する可能性を示す一組の症状または他の患者特性など、疾患に関連する早期予測因子を同定することを含み得る。濃縮解析は、特定の患者コホートと密接に関連する属性を同定することなどによって、患者コホートを特徴付けることを含み得る。構造化データベースはまた、例えばクラスタリングアルゴリズムを使用して患者サブグループを識別するなど、患者情報の遡及的解析に使用することもできる。患者を区別する能力または患者を層別化する能力は、肺高血圧症などの広範囲の重症度または転帰に関連する（そして早期段階で区別することが困難であり得る）状態に特に有用であり得る。いくつかの実施形態では、異なる患者サブグループに対して異なる治療計画が選択されてもよい。例えば、より低いリスクに関連すると特定された患者サブグループよりも悪い転帰に関連する患者サブグループに対して、より積極的な治療計画が選択され得る。そのような調整された治療計画は、リソースのより効率的な使用およびより良好な全体的なケアの質をもたらし得る。

【0274】

さらに、方法 2300 は、新しい健康データが利用可能になるとリアルタイムで適用することができる。さらに、方法 2300 は、比較的少量の管理されたデータがプロセス 2310 で使用されて新しいエンティティを解析に追加するため、新しい疾患、症状、および治療（例えば、古い文献に頻繁に現れないもの）などの関心のある新しい概念に適用することができる。この手法の例示的な利点は、上記で参照した米国仮特許出願第 62/962,433 号、第 62/962,146 号、および第 62/985,003 号、ならび

【0275】

図 24 は、いくつかの実施形態による、健康データの時間的識別のための方法 2400 の簡略図である。図 1～図 22B と一致するいくつかの実施形態では、方法 2400 の少なくとも一部は、システム 1500 を使用して実行することができる。

【0276】

プロセス 2410 において、管理された一組の健康記録が時間的に整列される。プロセス 2410 の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。例えば、管理された一組の健康記録は、方法 2300 を使用して生成された管理された一組の健康記録に対応することができる。健康記録を時間的に整列させることは、特定の診断検査が行われた日付、特定の症状が最初に観察された日付など、所与の患者の健康履歴における所定の事象の発生を識別することを含むことができる。次いで、患者の健康記録は、所定の事象の発生日に対して整列される。例えば、所定の事象の発生日が 0 日目に指定されている場合、事象の 3 日前に記録された患者の表現型を -3 日目に指定することができ、事象の 4 日後に記録された表現型を 4 日目に指定することができる。患者の表現型の日付は、患者記録に関連するメタデータ（例えば、タイムスタンプ）、所与の記録内に含まれる情報（例えば、日付を示す医師のメモ）などに基づいて決定することができる。いくつかの実施形態では、日付は、自然言語処理技術を使用して決定されてもよい。例えば、記録が「患者は 3 日前に咳を始めた」というフレーズを含む場合、自然言語処理を使用して、咳の症状の開始日が記録の日の 3 日前であったと判定することができる。

【0277】

プロセス 2420 において、時間窓の選択が受信される。プロセス 2420 の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。例えば、時間窓は、精漿検査（COVID-19 の検査など）の前後の小さな時間窓（例えば、10～30 日）に対応することができる。時間窓は、研究されている条件に基づくことができる。例えば、肺高血圧症などのよりゆっくりと進行する状態を研究するために、診断前後のより大きな時間窓（例えば、1 年以上）を使用することができる。

【0278】

10

20

30

40

50

プロセス 2430 において、時間的に整列した一組の健康記録が視覚化される。例えば、所定の事象の N 日以内に特定の表現型（例えば、特定のシステムを提示する）を有する患者の数または割合を、時間窓全体にわたって時間の関数としてプロットすることができる。そのような視覚化は、医師または科学者による下流の意思決定に有用であり得る。例えば、図 25 は、いくつかの実施形態による、時間的に整列した一組の健康記録に基づいて生成された視覚化 2500 の簡略図である。

【0279】

プロセス 2440 において、時間的に整列した一組の健康記録は、健康記録の 1 つまたは複数の対照セットと比較される。プロセス 2440 の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。いくつかの実施形態では、時間的に整列した健康記録を 1 つまたは複数の対照セットと比較することを使用して、所与の状態（例えば、陽性 COVID 検査）を予測する表現型（例えば、顕著な症状）を識別することができる。制御セットは自動的に決定されてもよく、例示的に、（1）「慢性的な病気の患者」のランダムなセットの健康記録、（2）慢性疾患以外の理由で診療所を訪れる集団のランダムな一組の健康記録、および（3）検査されている状態に関連する患者の健康記録に対応してもよい。一例として、COVID-19 を研究する場合、第 3 の対照セットは、症状を COVID-19 患者と区別するために「インフルエンザ」患者を含んでもよい。

【0280】

プロセス 2450 において、時間的に整合された一組の健康記録に基づいて時間的濃縮が識別される。プロセス 2450 の例示的な実施形態は、例えば、上記で参照された増強キュレーション論文の「方法」セクションに記載されている。例えば、時間窓内の所与の日に陽性検査結果を示す患者の数と陰性検査結果を示す患者の数との比を使用して、時間的濃縮を定量化することができる。濃縮は、上記で参照した米国仮出願第 62/962,146 号および第 62/985,003 号ならびに図 1~22 にさらに詳細に記載されている。例えば、COVID-19 患者は、対照患者（例えば、インフルエンザ患者）と比較して、陽性検査に至る日に特定の症状（例えば、咳、下痢、味覚の変化など）を示す可能性がより高くなり得る。そのような知見は、（例えば、検査のために特定の症状を示す患者を優先することによって）臨床的意思決定またはリソース最適化に役立つ可能性が高い。時間的濃縮は、特定の結果のリスクに関連する因子の組み合わせを同定する共濃縮をさらに含み得る。そのような組み合わせは、様々な既知の統計解析方法を使用して特定することができる。

【0281】

本明細書に記載の主題は、本明細書に開示した構造的手段およびその構造的均等物を含むデジタル電子回路、またはコンピュータソフトウェア、ファームウェア、またはハードウェア、またはそれらの組み合わせで実装することができる。本明細書に記載の主題は、データ処理装置（例えば、プログラム可能なプロセッサ、コンピュータ、または複数のコンピュータ）によって実行される、またはその動作を制御するために、情報キャリア（例えば、機械可読記憶装置において）に有形に具現化された、または伝播信号に具現化された 1 つまたは複数のコンピュータプログラムなどの、1 つまたは複数のコンピュータプログラム製品として実装することができる。コンピュータプログラム（プログラム、ソフトウェア、ソフトウェアアプリケーション、またはコードとしても知られている）は、コンパイル言語またはインタプリタ言語を含む任意の形式のプログラミング言語で記述することができ、スタンドアロンプログラムとして、またはモジュール、コンポーネント、サブルーチン、もしくはコンピューティング環境での使用に適した他のユニットとして含む任意の形式で展開することができる。コンピュータプログラムは、必ずしもファイルに対応しているとは限らない。プログラムは、他のプログラムまたはデータを保持するファイルの一部、問題のプログラム専用の単一のファイル、または複数の協調ファイル（例えば、1 つまたは複数のモジュール、サブプログラム、またはコードの一部を格納するファイル）に格納することができる。コンピュータプログラムは、1 つのコンピュータ上で、また

10

20

30

40

50

は1つのサイトの複数のコンピュータ上で実行されるように、または複数のサイトにわたって分散され、通信ネットワークによって相互接続されるように展開することができる。

【0282】

本明細書に記載の主題の方法ステップを含む、本明細書に記載のプロセスおよび論理フローは、入力データを操作し、出力を生成することによって、本明細書に記載の主題の機能を実行する1つまたは複数のコンピュータプログラムを実行する1つまたは複数のプログラマブルプロセッサによって実行することができる。プロセスおよび論理フローはまた、例えばFPGA（フィールドプログラマブルゲートアレイ）またはASIC（特定用途向け集積回路）などの専用論理回路によって実行することができ、本明細書に記載の主題の装置は、専用論理回路として実装することができる。

10

【0283】

コンピュータプログラムの実行に適したプロセッサは、例として、汎用および専用マイクロプロセッサの両方、ならびに任意の種類のデジタルコンピュータの任意の1つまたは複数のプロセッサを含む。一般に、プロセッサは、読み出し専用メモリまたはランダムアクセスメモリまたはその両方から命令およびデータを受信する。コンピュータの必須要素は、命令を実行するためのプロセッサ、ならびに命令およびデータを格納するための1つまたは複数のメモリデバイスである。一般に、コンピュータはまた、データを記憶するための1つまたは複数の大容量記憶装置、例えば、磁気、光磁気ディスク、または光ディスクを含むか、それらからデータを受信するか、それらにデータを転送するか、それらの両方を行うように動作可能に結合される。コンピュータプログラム命令およびデータを具現化するのに適した情報キャリアは、例として半導体メモリデバイス（例えば、EPROM、EEPROM、およびフラッシュメモリデバイス）、磁気ディスク（例えば、内蔵ハードディスクまたはリムーバブルディスク）、光磁気ディスク、光ディスク（例えば、CDおよびDVDディスク）を含む、あらゆる形態の不揮発性メモリを含む。プロセッサおよびメモリは、専用論理回路によって補完され得るか、または専用論理回路に組み込まれ得る。

20

【0284】

ユーザとの対話を提供するために、本明細書に記載の主題は、情報をユーザに表示するためのディスプレイ装置、例えばCRT（陰極線管）またはLCD（液晶ディスプレイ）モニタと、ユーザがコンピュータに入力を提供することができるキーボードおよびポインティングデバイス（例えば、マウスまたはトラックボール）とを有するコンピュータ上に実装することができる。他の種類のデバイスを使用して、ユーザとの対話を提供することもできる。例えば、ユーザに提供されるフィードバックは、任意の形態の感覚フィードバック（例えば、視覚フィードバック、聴覚フィードバック、または触覚フィードバック）とすることができ、ユーザからの入力は、音響、音声、または触覚入力を含む任意の形態で受信することができる。

30

【0285】

本明細書に記載の主題は、バックエンド構成要素（例えば、データサーバ）、ミドルウェア構成要素（例えば、アプリケーションサーバ）、もしくはフロントエンド構成要素（例えば、ユーザが本明細書に記載の主題の実施態様と対話することができるグラフィカルユーザインターフェースまたはウェブブラウザを有するクライアントコンピュータ）、またはそのようなバックエンド、ミドルウェア、およびフロントエンド構成要素の任意の組み合わせを含むコンピューティングシステムに実装することができる。システムの構成要素は、任意の形態または媒体のデジタルデータ通信、例えば通信ネットワークによって相互接続することができる。通信ネットワークの例には、ローカルエリアネットワーク（「LAN」）およびワイドエリアネットワーク（「WAN」）、例えばインターネットが含まれる。

40

【0286】

個人のヘルスケアデータを用いて計算するための技術が開示されている。特定の一実施形態では、本技術は、セキュアエンクレープを形成する分離メモリパーティションを構築

50

するステップと、セキュアエンクレーブ内にソフトウェアを予めプロビジョニングするステップと、を含む方法として実現され得る。セキュアエンクレーブは、1つまたは複数のプロセッサ上で実行されている1つまたは複数の不正なコンピューティングプロセスから分離して1つまたは複数のアプリケーションコンピューティングプロセスを実行するために1つまたは複数のプロセッサに利用可能である。予めプロビジョニングされたソフトウェアは、暗号化された形式で入力データまたは1つまたは複数のアプリケーションコンピューティングプロセスのための命令のうちの少なくとも1つを受信することと、1つまたは複数の暗号鍵を使用して入力データまたは命令のうちの少なくとも1つを復号化することと、出力データを生成するために、復号化された入力データまたは命令のうちの少なくとも1つに基づいて1つまたは複数のアプリケーションコンピューティングプロセスを実行することと、1つまたは複数のアプリケーションコンピューティングプロセスが受信した入力データ上で動作したことを示す実行証明を生成することと、1つまたは複数の暗号鍵を使用して出力データを暗号化することと、暗号化された出力データおよび実行証明への外部アクセスを提供することと、によって、1つまたは複数のプロセッサ上で1つまたは複数のアプリケーションコンピューティングプロセスの命令を実行するように構成される。

10

【0287】

この特定の実施形態の他の態様によれば、本方法は、データプロバイダから入力データを受信するステップと、データプロバイダ以外のエンティティに対応する命令プロバイダから、入力データを解析するための1つまたは複数のプログラムを受信するステップと、入力データをセキュアエンクレーブにロードするステップと、入力データを解析するための1つまたは複数のプログラムをセキュアエンクレーブにロードするステップと、セキュアエンクレーブを使用して入力データを解析するための1つまたは複数のプログラムを実行するステップと、をさらに含む。

20

【0288】

この特定の実施形態の他の態様によれば、本方法は、一組の臨床データ記録をセキュアエンクレーブにロードする前に入力データを非特定化するステップをさらに含み、入力データを非特定化するステップは、入力データから1つまたは複数の個人またはエンティティを識別する情報を削除するステップと、セキュアエンクレーブ内の1つまたは複数のアプリケーションコンピューティングプロセスのデータまたは命令に関連する復号鍵を、データプロバイダまたは命令プロバイダにそれぞれ提供するステップと、入力データをセキュアエンクレーブにロードする前に、入力非特定化確率を入力データに関連付けるステップと、を含み、セキュアエンクレーブは、入力非特定化確率を出力データ内に維持する。

30

【0289】

この特定の実施形態の他の態様によれば、入力データは、臨床データから個人識別情報（PII）を削除することによって非特定化される臨床データを含み、入力データおよび入力データを解析するための1つまたは複数のプログラムは、同じトランザクションでセキュアエンクレーブにロードされる。

【0290】

この特定の実施形態の他の態様によれば、本方法は、第2のセキュアエンクレーブ内で実行されるウェブサーバにセキュアエンクレーブを接続するステップと、セキュアエンクレーブの出力データをウェブクライアントまたはウェブブラウザ上のウェブページとして表示させるステップと、をさらに含む。

40

【0291】

この特定の実施形態の他の態様によれば、セキュアエンクレーブ内のウェブサーバによって処理されている入力データは、ウェブブラウザまたはウェブクライアントからアクセスできず、ウェブサーバを介して表示される出力データは、セキュアエンクレーブに関連する暗号化オブジェクトに関連付けられる。

【0292】

この特定の実施形態の他の態様によれば、本方法は、管理サービスから入力データまた

50

は命令を受信するステップをさらに含み、管理サービスは、入力データまたは命令がプライバシーを保護していると判定する。

【0293】

この特定の実施形態の他の態様によれば、セキュアエンクレーブは、パイプラインを形成するために1つまたは複数の他のセキュアエンクレーブに通信可能に結合され、セキュアエンクレーブからの出力データは、パイプライン内の後続のセキュアエンクレーブへの入力データとして提供され、実行証明は暗号化オブジェクトを含み、暗号化オブジェクトはセキュアエンクレーブの内容の表現であり、セキュアエンクレーブの内容の表現は、入力データ上で不正なコンピュータプログラムが動作していないことを証明し、セキュアエンクレーブの内容の表現は、入力データに関連する入力非特定化確率がセキュアエンクレーブによって維持されたことを証明し、暗号化オブジェクトは、パイプライン内の1つまたは複数の他のセキュアエンクレーブの内容を表す1つまたは複数の他の暗号化オブジェクトとリンクされる。

10

【0294】

この特定の実施形態の他の態様によれば、1つまたは複数のコンピューティングプロセスをセキュアエンクレーブ内で実行するステップは、入力データまたは命令の少なくとも1つに基づいてセキュアエンクレーブの固有のシグネチャを生成するステップをさらに含み、受信した入力データおよび1つまたは複数のアプリケーションコンピューティングプロセスの命令は、パイプライン内のセキュアエンクレーブにリンクされた他のいかなるセキュアエンクレーブからもアクセスできず、受信した入力データおよび1つまたは複数のアプリケーションコンピューティングプロセスの命令は、パイプラインのオペレータからアクセスできず、1つまたは複数の不正なコンピューティングプロセスは、少なくとも1つの特権ソフトウェア、特権ファームウェア、またはネットワークインターフェースプロセスを含み、入力データは一組の臨床データ記録を含む。

20

【0295】

特定の一実施形態では、本技術は、テキストシーケンスを受信するステップと、テキストシーケンスを複数のエンティティタグ付けモデルに提供するステップと、複数のエンティティタグ付けモデルによって識別されたテキストシーケンスからタグ付けされたエンティティを集約するステップと、集約されたタグ付けされたエンティティに基づいてテキストシーケンス内の情報をマスキングするステップと、を含む情報マスキング方法として実現され得る。複数のエンティティタグ付けモデルの各々は、対応するエンティティタイプを有するテキストシーケンスの1つまたは複数の部分にタグ付けし、所定のしきい値を超える性能メトリックを達成するように訓練される。

30

【0296】

この特定の実施形態の他の態様によれば、テキストシーケンスは、電子健康記録の少なくとも一部を含み、複数のエンティティタグ付けモデルのうちの少なくとも1つは、エンティティタイプのエンティティにタグ付けするように訓練され、エンティティタイプは、個人名、組織名、年齢、日付、時間、電話番号、ポケットベル番号、臨床識別番号、電子メールアドレス、IPアドレス、ウェブURL、車両番号、物理アドレス、郵便番号、社会保障番号、または生年月日のうちの少なくとも1つを含み、複数のエンティティタグ付けモデルのうちの少なくとも1つは、ルールベースのアルゴリズムに基づいてエンティティをタグ付けし、複数のエンティティタグ付けモデルのうちの少なくとも1つは、テキストのシーケンスからの学習に基づく機械学習モデルを含み、テキストシーケンス内の情報をマスキングすることは、1つまたは複数のタグ付けされたエンティティを、1つまたは複数のタグ付けされたエンティティのタイプをマーキングするブレースホルダと置き換えることを含み、テキストシーケンス内の情報をマスキングするステップは、1つまたは複数のタグ付けされたエンティティの値をランダム化された値に変更するステップを含む。

40

【0297】

この特定の実施形態の他の態様によれば、情報マスキング方法は、テキストシーケンスの1つまたは複数の部分をホワイトリスト化するステップをさらに含み、1つまたは複数

50

のホワイトリスト化された部分は、集約されたタグ付けされたエンティティを通過させる複数のエンティティタグ付けモデルに1つまたは複数のドレグフィルタを介して提供されず、1つまたは複数のドレグフィルタの各々は、ルールベースのテンプレートまたはパターンマッチングフィルタの少なくとも1つに基づいて対応するエンティティタイプをフィルタリングするように構成される。ルールベースのテンプレートは、テキストシーケンスの1つまたは複数の部分の各々に対応する構文テンプレートにマッピングするステップと、テキストシーケンスの1つまたは複数の部分に基づいて1つまたは複数の候補構文テンプレートを推測する機械学習モデルに基づいて候補構文テンプレートを識別するステップと、候補テンプレートに対応するテキストシーケンスの部分内の1つまたは複数のタグ付けされたエンティティの各々に対応する構文トークンで置き換えることによって、候補構文テンプレートからルールベースのテンプレートを作成するステップと、によって作成され得る。

10

【0298】

別の特定の実施形態では、本技術は、薬物、疾患、および関係を指定するクエリを受信するステップと、クエリに回答して、薬物および疾患の各々に言及するテキストフラグメントのリストを取得するステップであって、テキストフラグメントのリスト内の各テキストフラグメントは、薬物および疾患に対応するトークンと、薬物と疾患との間および周囲の複数の追加のトークンと、を含む、ステップと、訓練された少なくとも1つのニューラルネットワークモデルを使用して、各テキストフラグメントについて、関係が成立するかどうかを推測するステップと、関係がテキストフラグメントのリスト間で成立することを推測が示す頻度に基づいてメトリックを決定するステップと、メトリックと、テキストフラグメントのリストの中の1つまたは複数のテキストフラグメントと、を含む方法として実現され得る。

20

【0299】

この特定の実施形態の他の態様によれば、テキストフラグメントのリストを取得するステップは、クエリに基づく第2のクエリを検索アグリゲータに送信して、検索アグリゲータに、第2のクエリを受信させ、第2のクエリを複数のシャードに分散させて、複数のシャードの各々にそれぞれの反転リストに対するルックアップを実行させ、複数のシャードの各々は、コーパスの対応する部分を処理して反転リストを形成し、反転リストは、コーパス内の各トークンの出現のリストを含み、一組の検索結果を生成し、複数のシャードからの検索結果を集約して、テキストフラグメントのリストを形成させるステップを含む。

30

【0300】

この特定の実施形態の他の態様によれば、反転リストは、コーパスを複数のサブセットに分割するステップと、複数のサブセットの各々に対応する複数のシャードに分散させるステップと、複数のシャードのうちの各シャードについて、テキスト配列を形成するためにそれぞれのサブセット内の1つまたは複数の文書を連結するステップと、それぞれのサブセット内の各トークンの出現のリストを決定するステップと、によって生成され、出現のリストは反転リストとして格納される。

【0301】

この特定の実施形態の他の態様によれば、本方法は、クエリを、薬物、疾患、または関係のうちの少なくとも1つに関連する1つまたは複数の関連用語を含むように拡張するステップをさらに含み、テキストフラグメントのリストは、拡張されたクエリに基づいて取得され、1つまたは複数の関連用語は、複数の用語間の関係を識別する知識ベースから検索される。

40

【0302】

この特定の実施形態の他の態様によれば、本方法は、テキストフラグメントのリスト内の1つまたは複数のエンティティを識別するステップと、1つまたは複数のエンティティの統計解析に基づいて1つまたは複数の濃縮セットを生成するステップであって、クエリに対する応答は1つまたは複数の濃縮セットを含み、1つまたは複数の濃縮セットは、1つまたは複数のエンティティの中で最も高い統計的有意性スコアを有する1つまたは複数

50

のエンティティのサブセットを含む、ステップと、をさらに含む。

【0303】

この特定の実施形態の他の態様によれば、テキストフラグメントのリストは、公開文献コーパスまたは臨床データコーパスのうちの少なくとも一方を含むコーパスから取得され、コーパスは、マスクされた情報を含む少なくとも1つのテキストシーケンスを含み、コーパスは、少なくとも1つの構造化データフィールドを含む構造化テキストデータを含み、コーパスは、最小コホートサイズ未満のコホートに関連付けられている少なくとも1つの患者記録に基づいて、応答から省略される少なくとも1つの患者記録を含む。

【0304】

この特定の実施形態の他の態様によれば、マスクされた情報は、少なくとも1つのテキストシーケンスを受信することと、テキストシーケンスを複数のエンティティタグ付けモデルに提供することと、複数のエンティティタグ付けモデルの各々は、所定のしきい値を超える性能メトリックを達成するように訓練される、提供することと、複数のエンティティタグ付けモデルによって識別されたテキストシーケンスからタグ付けされたエンティティを集約することと、集約されたタグ付けされたエンティティに基づいてテキストシーケンス内の情報をマスキングすることと、によってマスクされる。

【0305】

別の特定の実施形態では、本技術は、第1の管理された一組の健康記録を受信するステップであって、管理された一組の健康記録内の各健康記録は患者情報の構造化表現を含む、ステップと、第1の管理された一組の健康記録を使用して第1のニューラルネットワークモデルを訓練するステップと、第1の未管理の一組の健康記録を受信するステップと、訓練された第1のニューラルネットワークモデルを使用して第1の未管理の一組の健康記録を管理し、第2の管理された一組の健康記録を生成するステップと、第2の管理された一組の健康記録を使用して第2のニューラルネットワークモデルを訓練するステップであって、第2のニューラルネットワークモデルは、訓練に基づいて第2の性能メトリックを達成し、第2の性能メトリックは、第1の性能メトリックと比較して改善された性能を示す、ステップと、を含む方法として実現され得る。

【0306】

この特定の実施形態の他の態様によれば、本方法は、現在の反復のための未管理の一組の健康記録を受信するステップと、前の反復からの訓練されたニューラルネットワークモデルを使用して、現在の反復のための未管理の一組の健康記録を管理し、現在の反復のための管理された一組の健康記録を生成するステップと、現在の反復のための管理された一組の健康記録を使用して、現在の反復のためのニューラルネットワークモデルを訓練するステップと、を反復的に実行することをさらに含む。

【0307】

この特定の実施形態の他の態様によれば、本方法は、第2の未管理の一組の健康記録を受信し、訓練された第2のニューラルネットワークモデルを使用して第2の未管理の一組の健康記録を管理し、第3の管理された一組の健康記録を生成するステップと、第4の管理された一組の健康記録に基づいて健康記録の構造化データベースを取り込むステップであって、第4の管理された一組の健康記録は、第1の管理された一組の健康記録、第2の管理された一組の健康記録、第3の管理された一組の健康記録、または1つもしくは複数のニューラルネットワークモデルを反復的に訓練することによって生成された管理された一組の健康記録のうちの少なくとも1つを含み、1つまたは複数のニューラルネットワークモデルの各々は、前の反復中に生成された管理された一組の健康記録を使用して訓練される、ステップと、健康記録の構造化データベースに基づいて濃縮解析を実行するステップと、クラスタリング解析に基づいて複数の患者サブグループを判別し、複数の患者サブグループの各々に対して異なる治療計画を選択するステップと、第3の管理された一組の健康記録を時間的に位置合わせし、時間的に位置合わせされた一組の健康記録に基づいて1つまたは複数の時間的濃縮を識別するステップと、をさらに含む。

【0308】

この特定の実施形態の他の態様によれば、訓練された第2のニューラルネットワークモデルは、関心のある緊急の概念に基づいて、第3の管理された一組の健康記録を管理し、濃縮解析は、疾患に関連する疾患予測子を識別すること、または患者コホートを特徴付けることの少なくとも一方を含み、第3の管理された一組の健康記録は、陽性検査日、医療処置日、または薬剤の使用に関連する日のうちの少なくとも1つに基づいて時間的に整列され、1つまたは複数の時間的濃縮は、複数の時間窓内で識別され、1つまたは複数の時間的濃縮は、複数の因子に基づく少なくとも1つの共濃縮を含む。

【0309】

別の特定の実施形態では、本技術は、システムに通信可能に結合され、システム内で動作するように構成された少なくとも1つのコンピュータプロセッサを含むシステムとして実現され得、少なくとも1つのコンピュータプロセッサは、上述の方法のうちの1つまたは複数のステップを実行するようにさらに構成され、第2のニューラルネットワークモデルは、第1の管理された一組の健康記録と第2の管理された一組の健康記録とを組み合わせた一組の健康記録を使用して訓練される。

10

【0310】

別の特定の実施形態では、本技術は、非一時的なプロセッサ可読媒体と、媒体に格納された命令と、を含む製品として実現され得、命令は、情報マスキングシステムに通信可能に結合され、情報マスキングシステム内で動作するように構成された少なくとも1つのコンピュータプロセッサによって媒体から読み取り可能に構成され、それによって少なくとも1つのコンピュータプロセッサを、上述の方法のうちの1つまたは複数のステップを実行するように動作させる。

20

【0311】

開示された主題は、その適用において、上記の説明に記載されたまたは図面に示す構成の詳細および構成要素の配置に限定されないことを理解されたい。開示された主題は、他の実施形態が可能であり、様々な方法で実施および実行することができる。また、本明細書で使用される表現および用語は、説明のためのものであり、限定するものとみなされるべきではないことを理解されたい。

【0312】

したがって、当業者は、本開示の基礎となる概念が、開示された主題のいくつかの目的を実行するための他の構造、方法、およびシステムの設計の基礎として容易に利用され得ることを理解するであろう。したがって、特許請求の範囲は、開示された主題の趣旨および範囲から逸脱しない限り、そのような同等の構成を含むとみなされることが重要である。

30

【0313】

開示された主題を前述の例示的な実施形態で説明および図示してきたが、本開示は例としてのみなされており、開示された主題の趣旨および範囲から逸脱することなく、開示された主題の実施態様の詳細における多くの変更がなされ得ることが理解され、これは以下の特許請求の範囲によってのみ限定される。

40

【図面】
【図 1 A】

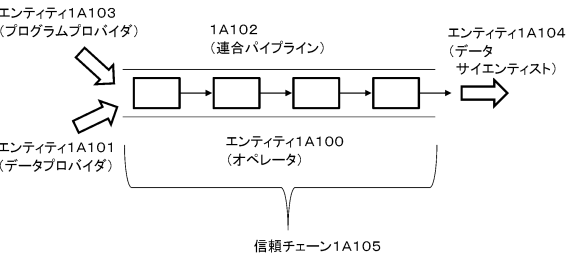


Fig. 1A

【図 1 B】

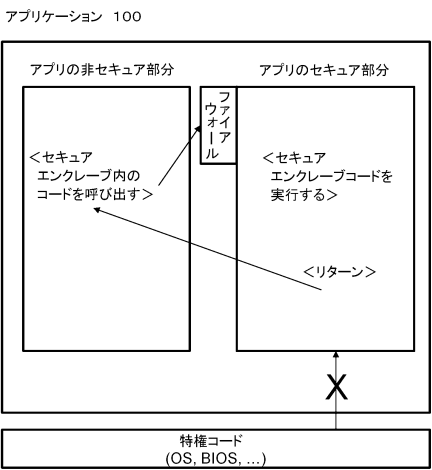


Fig. 1B

【図 2】

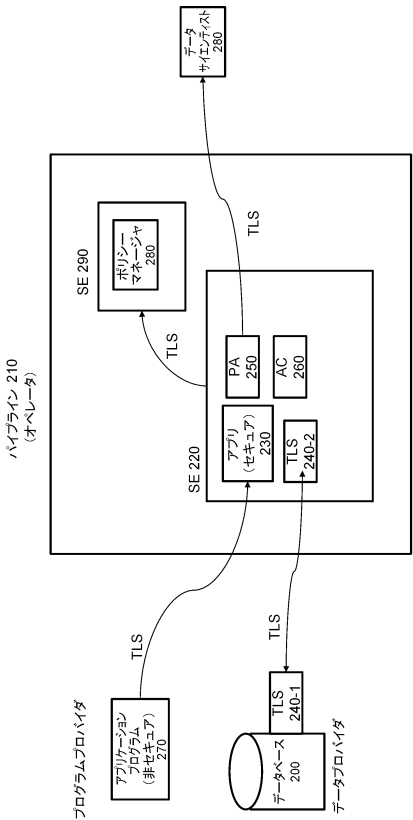


Fig. 2

【図 3】

出力データセット	ポリシー
PII属性を含まない	制約のないデータセット
PII属性を含む (K個の記録より多い)	セキュアエンクレープ (SE) 内の制約のないデータセット
PII属性を含む (K個の記録より少ない)	出力データセットをSEに封入する
	FHEを使用して出力データセットを暗号化する
	出力要求を禁止する
PIIデータを含み、出力要求は非プライバシー保護である	出力データセットをSEに封入する
	要求を禁止する

Fig. 3

10

20

30

40

50

【図 4 A】

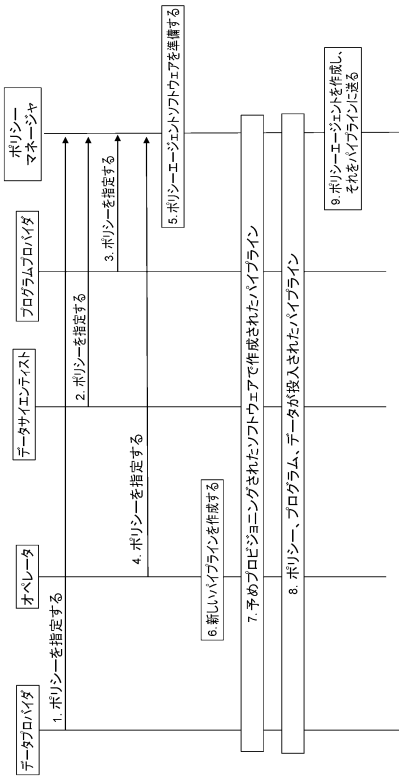


Fig. 4A

【図 4 B】

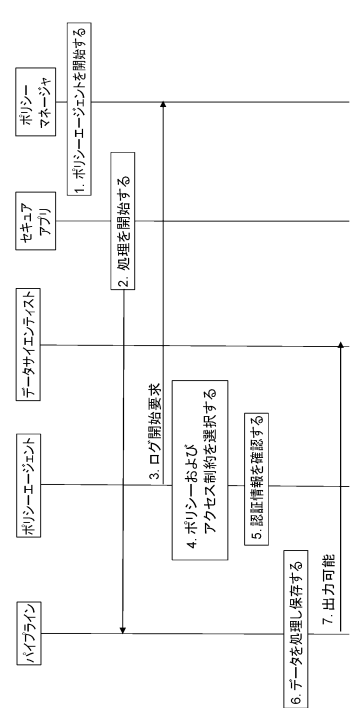


Fig. 4B

【図 5】

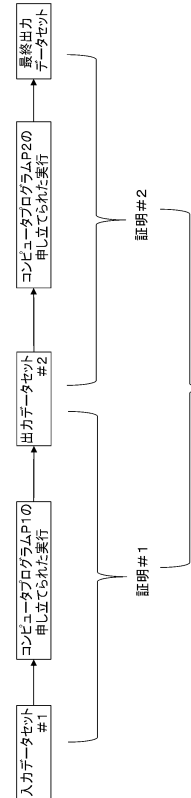


Fig. 5

【図 6】

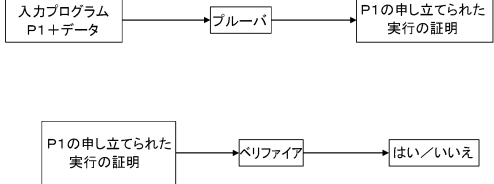


Fig. 6

【図 7】

	アルゴリズム プロバイダ	データプロバイダ	継プログラム	ロード画像へのリンク
グループ1	A1:a1	D1: d1	A1(k11), D1(k12)	L1
グループ2	A2:a2	D2: d2	A2(k21), D2(k22)	L2
...	...	...	...	...

Fig. 7

【図 8】

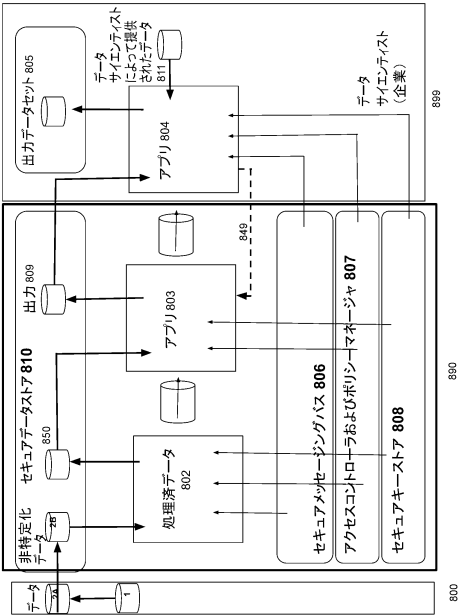


Fig. 8

【図 9】

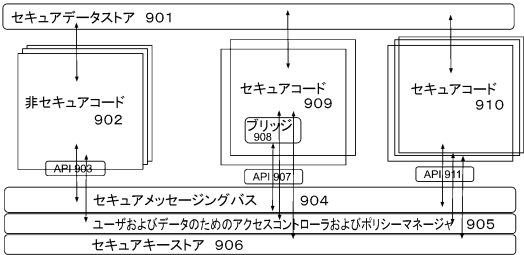


Fig. 9

【図 10】

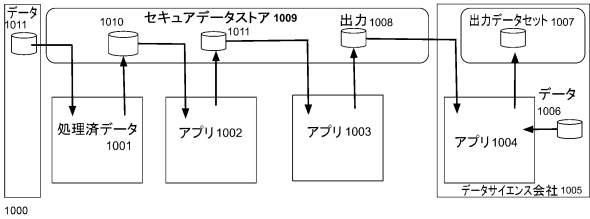


Fig. 10

【図 1 1】

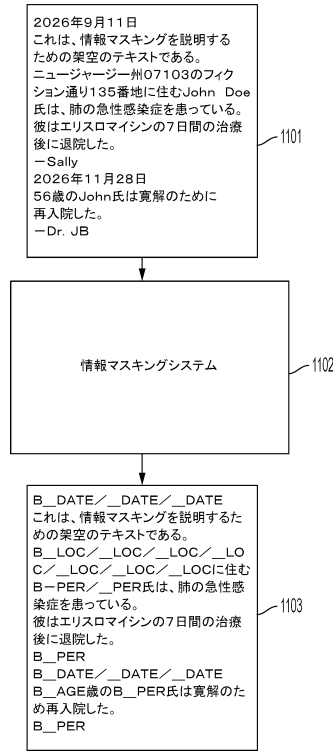


FIG. 11

【図 1 2】

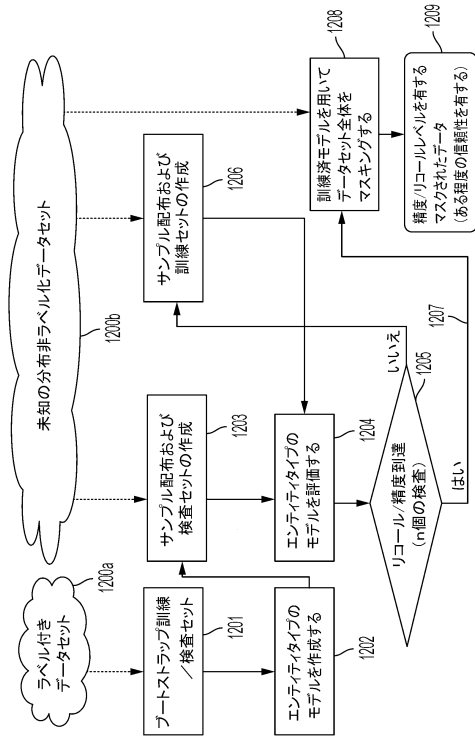


FIG. 12

【図 1 3】

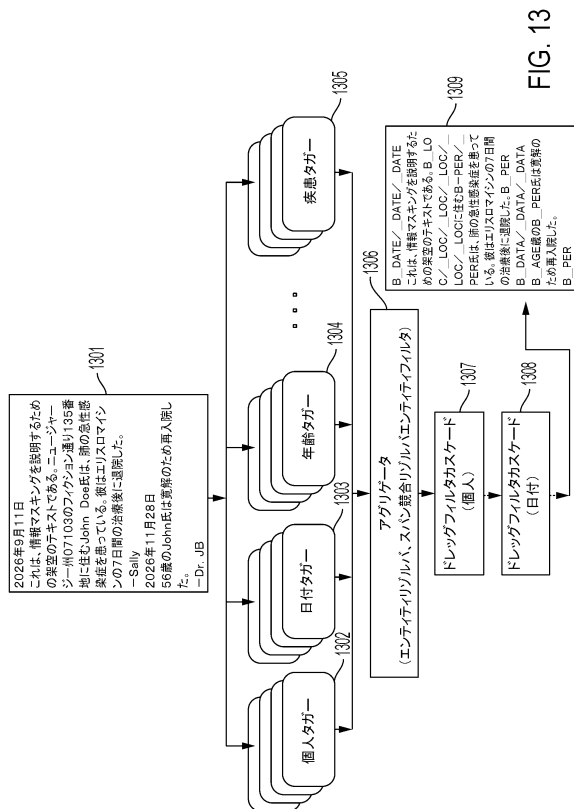


FIG. 13

【図 1 4】

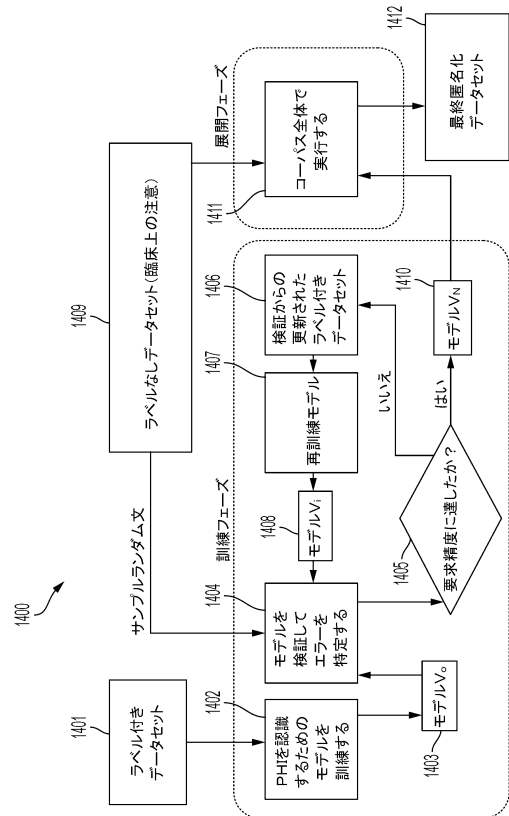


FIG. 14

【図 15】

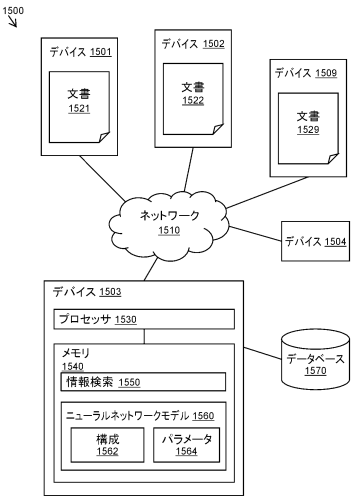


FIG. 15

【図 16】

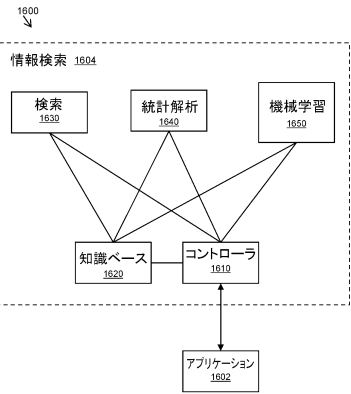


FIG. 16

【図 17】

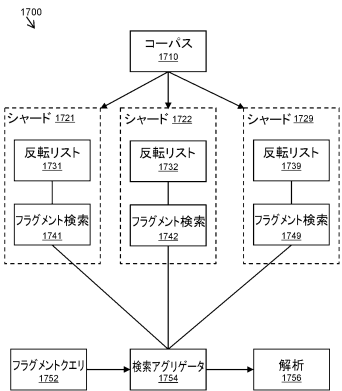


FIG. 17

【図 18】

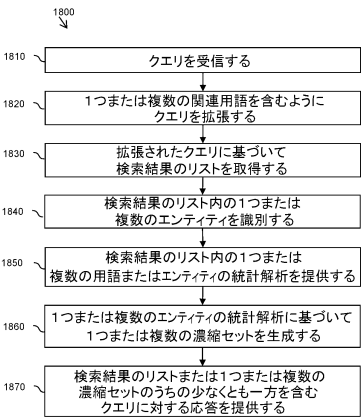


FIG. 18

10

20

30

40

50

【図 1 9】

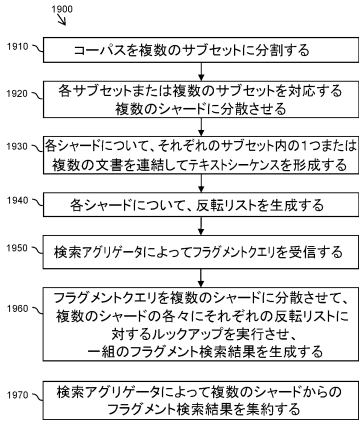


FIG. 19

【図 2 0 A】

FIG. 20A

【図 2 0 B】

FIG. 20B

【図 2 1】

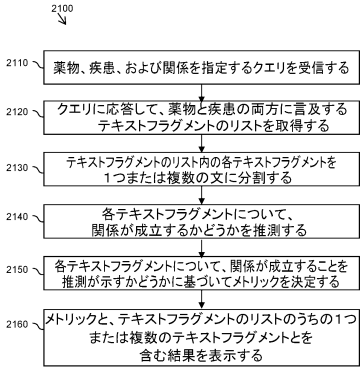
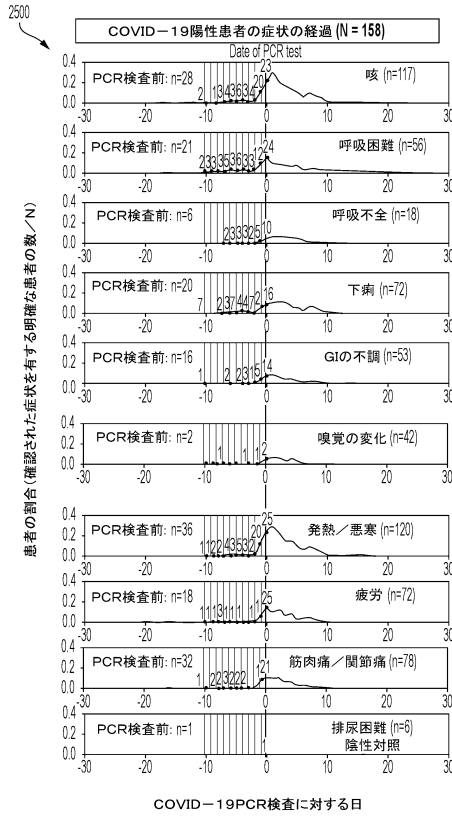
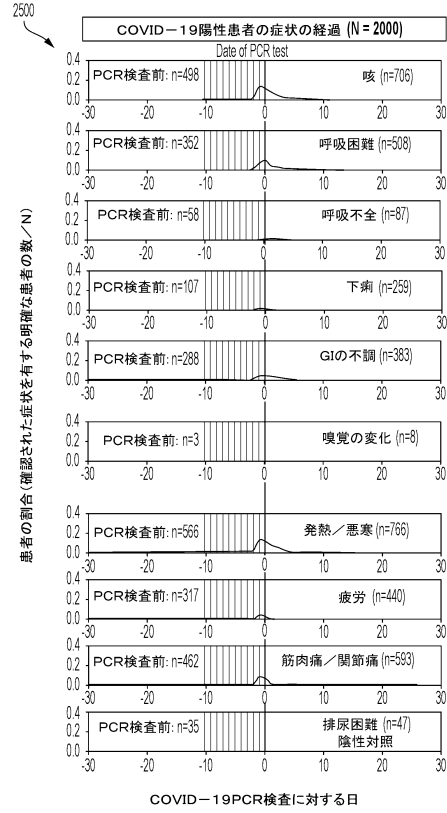


FIG. 21

【図 25-1】



【図 25-2】



10

20

30

40

50

フロントページの続き

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 62/962,146

(32)優先日 令和2年1月16日(2020.1.16)

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 62/865,030

(32)優先日 令和1年6月21日(2019.6.21)

(33)優先権主張国・地域又は機関

米国(US)

(31)優先権主張番号 62/985,003

(32)優先日 令和2年3月4日(2020.3.4)

(33)優先権主張国・地域又は機関

米国(US)

ト 604, ロジャーズ ストリート 33

(72)発明者 アルドハナリ, サンカー

アメリカ合衆国 マサチューセッツ州 01810, アンドーバー, スターリング ストリート 29

(72)発明者 アナンド, アカシュ

インド バーガルプル 813209, ポスト:ピルペインティ, エスビーアイ シェルマリ, シー
／オー マヘシュ クマール バランワル

(72)発明者 バーブ, ラケシュ

インド バンガロール 560001, パパンナ ストリート 2, ラハシエンダ 204

(72)発明者 サウンダララジャン, ヴェンカタラマナン

アメリカ合衆国 マサチューセッツ州 01810, アンドーバー, エイコーン ドライブ 18

(72)発明者 アワスティ, サミール

アメリカ合衆国 マサチューセッツ州 02130, ジャマイカ プレイン, アpartment 2,
ローズウェイ ストリート 39

(72)発明者 ラジャセクハラン, アジット

アメリカ合衆国 ニュージャージー州 08550, ウェスト ウィンザー, ル パーク コート 5

(72)発明者 ワグナー, タイラー

アメリカ合衆国 マサチューセッツ州 02127, ボストン, ユナイト 108, ウェスト ブロ
ードウェイ 45

(72)発明者 ナクヴィ, シャミム

アメリカ合衆国 ニュージャージー州 07960, モリスタウン, スプリング バレー ロード 19

審査官 上島 拓也

(56)参考文献 米国特許出願公開第2018/0212971 (US, A1)

米国特許出願公開第2015/0347768 (US, A1)

特開2015-170169 (JP, A)

(58)調査した分野 (Int.Cl., DB名)

G06F 21/62

G16H 10/00

G06F 21/71

H04L 9/10

G06F 21/60