(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号

特表2023-507881
(P2023-507881A)

(43)公表日　令和5年2月28日(2023.2.28)

(51)Int.Cl.	F I	テーマコード(参考)
Ｇ０６Ｆ　40/279　(2020.01)	Ｇ０６Ｆ　40/279	５Ｂ０９１

審査請求 未請求　予備審査請求 未請求　（全 31 頁）

(21)出願番号　特願2022-525481(P2022-525481)
(86)(22)出願日　令和2年12月30日(2020.12.30)
(85)翻訳文提出日　令和4年6月20日(2022.6.20)
(86)国際出願番号　PCT/IB2020/062535
(87)国際公開番号　WO2021/137166
(87)国際公開日　令和3年7月8日(2021.7.8)
(31)優先権主張番号　201941054421
(32)優先日　令和1年12月30日(2019.12.30)
(33)優先権主張国・地域又は機関
インド(IN)

(71)出願人　520412811
エルアンドティー　テクノロジー　サービシズ　リミテッド
インド国，チェンナイ　６０００８９，ラーマプラーム，マウント　プーナマリー　ロード，ディーエルエフ　アイティー　エスイーゼット　パーク，ブロック　３，２番　フロア，１／１２４
(74)代理人　100093861
弁理士　大賀　眞司
(74)代理人　100129218
弁理士　百本　宏之

最終頁に続く

(54)【発明の名称】ドメインベースのテキスト抽出方法およびシステム

(57)【要約】
本開示は、入力ファイルのコンテンツから情報を抽出するための方法およびシステムに関する。本方法は、入力ファイルからのテキストデータを識別することと、複数のテキストエンティティから関連テキストエンティティを識別するためにユーザからテキスト入力を受け取ることと、テキスト入力に対応する検索パターンを自動的に生成することと含んでよい。本方法は、複数のテキストエンティティの各々に関連付けられたパターンを決定することと、テキスト入力に対応する検索パターンを複数のテキストエンティティに関連付けられたパターンと対応付けることとをさらに含んでよい。本方法は、対応付けに基づいて複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンを識別することと、１つまたは複数の合致パターンに対応する関連テキストエンティティを複数のテキストエンティティから抽出することとをさらに含んでよい。
【選択図】図２

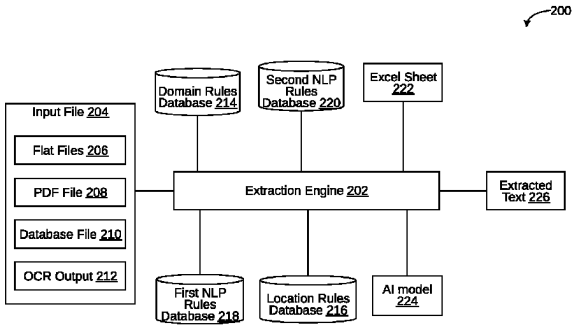


FIG. 2

【特許請求の範囲】
【請求項１】
入力ファイルのコンテンツから情報を抽出する方法であって、
前記入力ファイルからテキストデータを識別することであって、識別された前記テキストデータは複数のテキストエンティティを含む、識別することと、
前記複数のテキストエンティティから関連テキストエンティティを識別するために、ユーザからテキスト入力を受け取ることと、
前記テキスト入力に対応する検索パターンを自動的に生成することと、
前記複数のテキストエンティティの各々に関連付けられたパターンを決定することと、
前記テキスト入力に対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けることと、
前記対応付けに基づいて、前記複数のテキストエンティティに関連付けられた前記パターンから１つまたは複数の合致パターンを識別することと、
前記１つまたは複数の合致パターンに対応する関連テキストエンティティを前記複数のテキストエンティティから抽出することと
を含む方法。
【請求項２】
抽出された前記関連テキストエンティティを用いて表現を生成することを含む、請求項１に記載の方法。
【請求項３】
前記入力ファイルは、画像ファイルおよびポータブル・ドキュメント・フォーマット（ＰＤＦ）ファイルのうちの少なくとも１つを含む、請求項１に記載の方法。
【請求項４】
前記複数のテキストエンティティの各々に関連付けられたドメインベースの名前を決定することをさらに含み、
前記ドメインベースの名前は、前記複数のテキストエンティティの各々をデータデポジトリに記憶されたドメインベースの名前のリストと対応付けることに基づいて決定される、請求項１に記載の方法。
【請求項５】
前記入力ファイルに関連付けられたタイプを識別することと、
前記入力ファイルに関連付けられた前記タイプに基づいて、前記入力ファイルにおける１つまたは複数の関連テキストエンティティの位置を決定することと、
前記位置に基づいて、前記入力ファイルから前記１つまたは複数の関連テキストエンティティを抽出することと
をさらに含む、請求項１に記載の方法。
【請求項６】
前記位置は、（ｘ，ｙ）座標に基づくものであり、前記位置は、前記データリポジトリに記憶された、前記入力ファイルに関連付けられたタイプに対応するテンプレートに基づいて予め決定される、請求項５に記載の方法。
【請求項７】
前記複数のテキストエンティティに関連付けられた品詞（ＰＯＳ）を識別することと、
ＰＯＳが名詞として識別される１つまたは複数のテキストエンティティを抽出することと
をさらに含み、
前記ＰＯＳは、名詞、代名詞、動詞、副詞、形容詞、接続詞、前置詞、および間投詞のうちの１つである、
請求項１に記載の方法。
【請求項８】
ＰＯＳが識別されない１つまたは複数のユニークなテキストエンティティを識別することと、
前記１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティ

ティタイプを決定することと
をさらに含み、
前記エンティティタイプは、値エンティティおよびパターンエンティティのうちの１つであり、各値エンティティは関連付けられた値を有し、各パターンエンティティは関連付けられたパターンを有する、
請求項７に記載の方法。
【請求項９】
各値エンティティについて、
前記値エンティティの前記関連付けられた値と同じ関連付けられた値を有する１つまたは複数のテキストエンティティを自動的に識別すること
をさらに含む、請求項８に記載の方法。
【請求項１０】
各パターンエンティティについて、
前記パターンエンティティに関連付けられたパターンに対応する検索パターンを自動的に生成することと、
前記パターンエンティティに対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けることと、
前記対応付けに基づいて、前記複数のテキストエンティティに関連付けられた前記パターンから１つまたは複数の合致パターンを決定することと、
前記１つまたは複数の合致パターンに対応する合致テキストエンティティを前記複数のテキストエンティティから抽出することと
をさらに含む、請求項８に記載の方法。
【請求項１１】
分類ベースの機械学習（ＭＬ）モデルを用いて、履歴データに基づいて推奨を生成することを含む、請求項１に記載の方法。
【請求項１２】
入力ファイルのコンテンツから情報を抽出するためのシステムであって、
プロセッサと、
前記プロセッサに通信可能に結合されたメモリと
を含み、
前記メモリは、プロセッサ実行可能命令を記憶し、前記プロセッサ実行可能命令は、前記プロセッサにより実行されたとき、前記プロセッサに、
前記入力ファイルからテキストデータを識別することであって、識別された前記テキストデータは複数のテキストエンティティを含む、識別すること、
前記複数のテキストエンティティから関連テキストエンティティを識別するために、ユーザからテキスト入力を受け取ること、
前記テキスト入力に対応する検索パターンを自動的に生成すること、
前記複数のテキストエンティティの各々に関連付けられたパターンを決定すること、
前記テキスト入力に対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けること、
前記対応付けに基づいて、前記複数のテキストエンティティに関連付けられた前記パターンから１つまたは複数の合致パターンを識別すること、および、
前記１つまたは複数の合致パターンに対応する関連テキストエンティティを前記複数のテキストエンティティから抽出することを行わせる、
システム。
【請求項１３】
前記入力ファイルは、画像ファイルおよびポータブル・ドキュメント・フォーマット（ＰＤＦ）ファイルのうちの少なくとも１つを含む、請求項１２に記載のシステム。
【請求項１４】
前記プロセッサ実行可能命令は、前記プロセッサにより実行されたとき、前記プロセッサ

に、
ドメインベースの抽出であって、前記ドメインベースの抽出を行うことは、
前記複数のテキストエンティティの各々に関連付けられたドメインベースの名前を決定することであって、前記ドメインベースの名前は、前記複数のテキストエンティティの各々をデータデポジトリに記憶されたドメインベースの名前のリストと対応付けることに基づいて決定される、決定すること
を含む、ドメインベースの抽出、
位置ベースの抽出であって、前記位置ベースの抽出を行うことは、
前記入力ファイルに関連付けられたタイプを識別すること、
前記入力ファイルに関連付けられた前記タイプに基づいて、前記入力ファイルにおける１つまたは複数の関連テキストエンティティの位置を決定すること、および、
前記位置に基づいて、前記入力ファイルから前記１つまたは複数の関連テキストエンティティを抽出することであって、前記位置は、前記データリポジトリに記憶された、前記入力ファイルに関連付けられたタイプに対応するテンプレートに基づいて予め決定される、抽出すること、
を含む、位置ベースの抽出、
品詞（ＰＯＳ）ベースの抽出であって、前記品詞（ＰＯＳ）ベースの抽出を行うことは、
前記複数のテキストエンティティ関連付けられたＰＯＳを識別することであって、前記ＰＯＳは、名詞、代名詞、動詞、副詞、形容詞、接続詞、前置詞、および間投詞のうちの１つである、識別すること、および、
ＰＯＳが名詞として識別される１つまたは複数のテキストエンティティを抽出すること、
ＰＯＳが識別されない１つまたは複数のユニークなテキストエンティティを識別すること、
前記１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプを決定することであって、前記エンティティタイプは、値エンティティおよびパターンエンティティのうちの１つであり、各値エンティティは関連付けられた値を有し、各パターンエンティティは関連付けられたパターンを有する、決定すること、
各値エンティティについて、
前記値エンティティの前記関連付けられた値と同じ関連付けられた値を有する１つまたは複数のテキストエンティティを自動的に識別すること、
各パターンエンティティについて、
前記パターンエンティティに関連付けられたパターンに対応する検索パターンを自動的に生成すること、
前記パターンエンティティに対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けること、
前記対応付けに基づいて、前記複数のテキストエンティティに関連付けられた前記パターンから１つまたは複数の合致パターンを決定すること、および、
前記１つまたは複数の合致パターンに対応する合致テキストエンティティを前記複数のテキストエンティティから抽出すること
を含む、品詞（ＰＯＳ）ベースの抽出、ならびに、
機械学習（ＭＬ）ベースの抽出であって、前記ＭＬベースの抽出を行うことは、
分類ベースのＭＬモデルを用いて、履歴データに基づいて推奨を生成すること
を含む、機械学習（ＭＬ）ベースの抽出のうちの少なくとも１つをさらに行わせる、
請求項１２に記載のシステム。
【請求項１５】
入力ファイルからテキストデータを識別することであって、識別された前記テキストデータは複数のテキストエンティティを含む、識別することと、
前記複数のテキストエンティティから関連テキストエンティティを識別するために、ユーザからテキスト入力を受け取ることと、
前記テキスト入力に対応する検索パターンを自動的に生成することと、

前記複数のテキストエンティティの各々に関連付けられたパターンを決定することと、
前記テキスト入力に対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けることと、
前記対応付けに基づいて、前記複数のテキストエンティティに関連付けられた前記パターンから１つまたは複数の合致パターンを識別することと、
前記１つまたは複数の合致パターンに対応する関連テキストエンティティを前記複数のテキストエンティティから抽出することと
を含むステップを１つまたは複数のプロセッサを含むコンピュータに行わせるコンピュータ実行可能命令のセットが記憶された非一時的コンピュータ可読記憶媒体。
【請求項１６】
前記コンピュータ実行可能命令のセットは、前記１つまたは複数のプロセッサを含む前記コンピュータに、
ドメインベースの抽出であって、前記ドメインベースの抽出を行うことは、
前記複数のテキストエンティティの各々に関連付けられたドメインベースの名前を決定することであって、前記ドメインベースの名前は、前記複数のテキストエンティティの各々をデータデポジトリに記憶されたドメインベースの名前のリストと対応付けることに基づいて決定される、決定すること
を含む、ドメインベースの抽出、
位置ベースの抽出であって、前記位置ベースの抽出を行うことは、
前記入力ファイルに関連付けられたタイプを識別すること、
前記入力ファイルに関連付けられた前記タイプに基づいて、前記入力ファイルにおける１つまたは複数の関連テキストエンティティの位置を決定すること、および、
前記位置に基づいて、前記入力ファイルから前記１つまたは複数の関連テキストエンティティを抽出することであって、前記位置は、前記データリポジトリに記憶された、前記入力ファイルに関連付けられたタイプに対応するテンプレートに基づいて予め決定される、抽出すること、
を含む、位置ベースの抽出、
品詞（ＰＯＳ）ベースの抽出であって、前記品詞（ＰＯＳ）ベースの抽出を行うことは、
前記複数のテキストエンティティ関連付けられたＰＯＳを識別することであって、前記ＰＯＳは、名詞、代名詞、動詞、副詞、形容詞、接続詞、前置詞、および間投詞のうちの１つである、識別すること、および、
ＰＯＳが名詞として識別される１つまたは複数のテキストエンティティを抽出すること、
ＰＯＳが識別されない１つまたは複数のユニークなテキストエンティティを識別すること、
前記１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプを決定することであって、前記エンティティタイプは、値エンティティおよびパターンエンティティのうちの１つであり、各値エンティティは関連付けられた値を有し、各パターンエンティティは関連付けられたパターンを有する、決定すること、
各値エンティティについて、
前記値エンティティの前記関連付けられた値と同じ関連付けられた値を有する１つまたは複数のテキストエンティティを自動的に識別すること、
各パターンエンティティについて、
前記パターンエンティティに関連付けられたパターンに対応する検索パターンを自動的に生成すること、
前記パターンエンティティに対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けること、
前記対応付けに基づいて、前記複数のテキストエンティティに関連付けられた前記パターンから１つまたは複数の合致パターンを決定すること、および、
前記１つまたは複数の合致パターンに対応する合致テキストエンティティを前記複数のテキストエンティティから抽出すること

を含む、品詞（ＰＯＳ）ベースの抽出、ならびに、
機械学習（ＭＬ）ベースの抽出であって、前記ＭＬベースの抽出を行うことは、
分類ベースのＭＬモデルを用いて、履歴データに基づいて推奨を生成すること
を含む、機械学習（ＭＬ）ベースの抽出のうちの少なくとも１つを行わせる、
請求項１５に記載の非一時的コンピュータ可読記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、概してデータ抽出に関し、より詳細には、１つまたは複数のデータ抽出アプローチを用いて入力ファイルのコンテンツからテキスト情報を抽出するための方法およびシステムに関する。
【背景技術】
【０００２】
最近、テキスト抽出技法が重要性を帯びてきている。例えば、光学文字認識（Optical Character Recognition; ＯＣＲ）などの抽出技法は、ユーザが画像またはポータブル・ドキュメント・フォーマット（Portable Document Format; ＰＤＦ）ファイルなどのファイルからテキストデータを抽出することを可能とし得る。さらに、関連情報を抽出し、抽出された関連情報を用いて表現を生成することが望ましい場合がある。
【発明の概要】
【発明が解決しようとする課題】
【０００３】
いくつかの利用可能な技法によれば、入力ファイルがタグを含む場合、または識別されたテキストにおいてパターンを識別することができる場合に、入力ファイルからテキスト情報を抽出し、抽出された情報を用いて表現を生成することが可能であり得る。しかしながら、タグが存在しない場合があるまたはパターンを識別することができない複雑な文書においては、関連情報を抽出したり、表現を生成したりすることが難しい。
【課題を解決するための手段】
【０００４】
本発明による方法およびシステムは、入力ファイルのコンテンツから情報を抽出するものであって、入力ファイルからテキストデータを識別することと、複数のテキストエンティティから関連テキストエンティティを識別するために、ユーザからテキスト入力を受け取ることと、テキスト入力に対応する検索パターンを自動的に生成することと、複数のテキストエンティティの各々に関連付けられたパターンを決定することと、テキスト入力に対応する前記検索パターンを、前記複数のテキストエンティティに関連付けられたパターンと対応付けることと、対応付けに基づいて、複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンを識別することと、１つまたは複数の合致パターンに対応する関連テキストエンティティを複数のテキストエンティティから抽出することとを含む。識別されたテキストデータは、複数のテキストエンティティを含む。
【０００５】
その他、本願が開示する課題とその解決方法は、発明を実施するための形態の欄、および図面の記載によって明らかにされる。
【０００６】
本開示に組み込まれてその一部を構成する添付の図面は、例示的実施形態を示し、説明と共に開示の原理を解説する役割を果たす。
【図面の簡単な説明】
【０００７】
【図１】本開示のいくつかの実施形態に係る、入力ファイルのコンテンツから情報を抽出するための例示的システムの機能ブロック図である。
【図２】本開示のいくつかの実施形態に係る、自然言語処理（Natural Language Processing; ＮＬＰ）メタデータ抽出フレームワークの機能ブロック図である。

【図３】本開示のいくつかの実施形態に係る、スペルチェックシステムのブロック図である。
【図４】本開示のいくつかの実施形態に係る、推奨システムのブロック図である。
【図５】本開示のいくつかの実施形態に係る、メタデータ更新システムのブロック図である。
【図６】本開示の様々な実施形態に係る、入力ファイルのコンテンツから情報を抽出する方法のフローチャートである。
【図７】本開示の様々な実施形態に係る、入力ファイルのコンテンツから情報を抽出する方法のフローチャートである。
【図８】本開示の様々な実施形態に係る、入力ファイルのコンテンツから情報を抽出する方法のフローチャートである。
【図９】本開示のいくつかの実施形態に係る、情報を抽出する対象であるノイジーなエンティティを有する例示的な入力ファイルのスナップショットである。
【図１０】本開示の様々な実施形態に係る、情報を抽出する対象である別の例示的な入力ファイルのスナップショットである。
【発明を実施するための形態】
【０００８】
添付の図面を参照して、例示的実施形態を説明する。都合の良い場合には、同じまたは同様の部分を示すために、図面の全体を通じて同じ参照番号が用いられる。開示の原理の例および特徴を本明細書で説明するが、開示の実施形態の趣旨および範囲から逸脱しない限りにおいて、修正、改変、および他の実装が可能である。以下の詳細な説明は単に例示的なものとみなされることが意図されており、真の範囲および趣旨は、以下の特許請求の範囲によって示される。
【０００９】
ここで図１を参照すると、本開示のいくつかの実施形態に係る、入力ファイルのコンテンツから情報を抽出するための例示的システム１００の機能ブロック図が示されている。システム１００は、テキスト識別デバイス１０２を含んでよい。いくつかの実施形態において、テキスト識別デバイス１０２は、入力ファイルのコンテンツからテキストデータを識別してよい。例えば、入力ファイルは、画像またはポータブル・ドキュメント・フォーマット（Portable Document Format; ＰＤＦ）ファイルであってよい。識別されたテキストデータは、複数のテキストエンティティを含んでよい。したがって、いくつかの実施形態において、テキスト識別デバイス１０２は、入力ファイルからテキストデータを識別するために、光学文字認識（Optical Character Recognition; ＯＣＲ）技法を用いてよい。代替的な実施形態において、テキスト識別デバイス１０２は、入力ファイルからテキストデータを識別するために、当技術分野において知られている任意の他の技法を用いてよい。
【００１０】
システム１００は、テキスト識別デバイス１０２と通信可能な情報抽出デバイス１０４をさらに含んでよい。例えば、情報抽出デバイス１０４は、テキスト識別デバイス１０２によって入力ファイルから識別された複数のテキストエンティティから、関連情報を抽出するように構成されてよい。いくつかの実施形態において、情報を抽出するべく、情報抽出デバイス１０４は、複数のテキストエンティティから関連テキストエンティティを識別するためにユーザからテキスト入力を受け取り、テキスト入力に対応する検索パターンを自動的に生成してよい。情報抽出デバイス１０４はさらに、複数のテキストエンティティの各々に関連付けられたパターンを決定し、テキスト入力に対応する検索パターンを複数のテキストエンティティに関連付けられたパターンと対応付けてよい。情報抽出デバイス１０４はさらに、対応付けに基づいて複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンを識別し、１つまたは複数の合致パターンに対応する関連テキストエンティティを複数のテキストエンティティから抽出してよい。
【００１１】

システム１００は、ディスプレイ１１４をさらに含んでよい。いくつかの実施形態において、情報抽出デバイス１０４は、１つまたは複数のプロセッサ１１０およびコンピュータ可読媒体（例えばメモリ）１１２を含んでよい。コンピュータ可読記憶媒体１１２は、１つまたは複数のプロセッサ１１０により実行されたとき、本開示の態様により、入力ファイルのコンテンツに基づいて表現を生成することを１つまたは複数のプロセッサ１１０に行わせる命令を記憶してよい。コンピュータ可読記憶媒体１１２はまた、システム１００によって取り込まれ、処理され、かつ／または必要とされ得る様々なデータ（例えば、識別されたテキストデータ、値エンティティ、データ、パターンエンティティデータなど）を記憶してよい。システム１００は、ディスプレイ１１４を介してアクセス可能なユーザインターフェース１１６を介してユーザとやり取りしてよい。システム１００はまた、様々なデータを送るまたは受け取るために、通信ネットワーク１０８を介して１つまたは複数の外部デバイス１０６とやり取りしてよい。外部デバイス１０６は、リモートサーバ、デジタルデバイス、または別のコンピューティングシステムを含んでよいが、これらに限定されないものであってよい。

【００１２】

ここで図２を参照すると、本開示のいくつかの実施形態に係る自然言語処理（Natural Language Processing; ＮＬＰ）メタデータ抽出フレームワーク２００の機能ブロック図が示されている。ＮＬＰメタデータ抽出フレームワーク２００は、システム１００に、特に情報抽出デバイス１０４に実装されてよい。ＮＬＰメタデータ抽出フレームワーク２００は、入力ファイルのコンテンツから情報を抽出するための様々な機能を行う様々なモジュールを含んでよい。いくつかの実施形態において、抽出エンジン２０２が、入力ファイル２０４のコンテンツからテキストデータを抽出／識別してよい。入力ファイル２０４は、フラットファイル２０６、またはＰＤＦファイル２０８、またはデータベースファイル２１０、またはＯＣＲ出力２１２であってよい。

【００１３】

抽出エンジン２０２はさらに、ドメインルールデータベース２１４から１つまたは複数のドメインルールを受け取ってよい。さらに、いくつかの実施形態において、ドメインルールデータベース２１４は、（従来の手続き型コードによるものではなく）「ｉｆ－ｔｈｅｎ」ルールに基づくものであってよい。ドメインルールは、複数のテキストエンティティの各々に関連付けられたドメインベースの名前を決定するために用いられてよいことに留意されたい。ドメインベースの名前は、複数のテキストエンティティの各々を、データデポジトリ（図１では不図示）に記憶されたドメインベースの名前のリストと対応付けることに基づいて決定されてよい。例えば、識別されたデータが、ある国のある都市のＰＩＮコードに属するテキストエンティティを含む場合、ドメインベースのデータベースは、例えばＰＩＮコードをデータリポジトリに記憶されたドメインベースの名前のリストと対応付けることにより、ドメイン名、すなわちそのＰＩＮコードを有する都市の名前を決定してよい。

【００１４】

抽出エンジン２０２はさらに、位置ルールデータベース２１６から１つまたは複数の位置ルールを受け取ってよい。いくつかの実施形態において、位置ルールは、ＯＣＲ出力に基づいて、入力ファイルにおけるテキストデータからのエンティティに関連付けられた位置を決定するために用いられてよいことに留意されたい。位置は、入力ファイルに関連付けられたタイプを最初に識別し、入力ファイルに関連付けられたタイプに基づいて入力ファイルにおける１つまたは複数の関連テキストエンティティの位置を決定したときに、決定されてよい。関連テキストエンティティの位置が決定されると、関連テキストエンティティが位置に基づいて入力ファイルから抽出されてよい。１つまたは複数の位置ルールは、表を含む入力データ、すなわち表形式で配置された入力データの場合にはより関連性が高いものであり得るが、１つまたは複数の位置ルールは、任意の他のタイプの入力データ、例えばフリーテキストを有する文書にも同様に適用可能であってよいことに留意されたい。

【００１５】
例として、入力ファイルのタイプは、「Ａａｄｈａｒ」カードのものであってよい。「Ａａｄｈａｒ」カードのような文書においては、データが特定の形式で存在していてよいことをさらに理解されたい。したがって、関連テキストエンティティの位置は、データリポジトリに記憶された、入力ファイルに関連付けられたタイプに対応するテンプレートに基づいて予め決定されてよい。例えば、（対象の人物の）テキストエンティティ「名前」および「住所」の位置は、文書の中程であってよく、この情報は、システム１００のデータベースにおいて取得可能であり記憶されていてよい。特に、位置は、（ｘ，ｙ）座標に基づくものであってよい。上記の例において、テキストエンティティ「名前」および「住所」の位置は、（ｘ，ｙ）座標（５，１０）～（１５，１５）によって画定される領域内であってよい。したがって、文書のタイプが「Ａａｄｈａｒ」カードとして識別されると、システム１００は、上記位置にアクセスして、その位置から関連テキストエンティティ（すなわち「名前」および「住所」）を抽出してよい。
【００１６】
抽出エンジン２０２はさらに、第１のＮＬＰルールデータベース２１８から１つまたは複数の第１のＮＬＰルールを受け取ってよい。第１のＮＬＰルールデータベース２１８は、正解値／属性について提供されるＮＬＰルールに基づいて正規表現（Regular Expression；Ｒｅｇｅｘ）を作成してよいことに留意されたい。当業者には理解されるように、Ｒｅｇｅｘは、検索パターンを記述するための特別な文字列であってよい。
【００１７】
抽出エンジン２０２はさらに、第２のＮＬＰルールデータベース２２０から１つまたは複数の第２のＮＬＰルールを受け取ってよい。いくつかの実施形態において、第２のＮＬＰルールデータベース２２０は、正解値／属性について品詞（part of speech; ＰＯＳ）を定義してよいことに留意されたい。例えば、複数のテキストエンティティに関連付けられたＰＯＳが識別されてよい。当業者には理解されるように、ＰＯＳは、名詞、代名詞、動詞、副詞、形容詞、接続詞、前置詞、および間投詞のうちの１つであってよい。ＰＯＳを識別すると、ＰＯＳが名詞である複数のテキストエンティティから１つまたは複数のユニークなテキストエンティティが選択されてよい。ＰＯＳが名詞として識別された１つまたは複数の（選択された）ユニークなテキストエンティティは、関連データとして抽出されてよい。
【００１８】
例えば、給与明細文書（入力データ）は、対象の人物の名前を含む複数のテキストエンティティを含んでよい。名前を抽出することが、抽出エンジン２０２の目的であってよい。さらに、名前は、名詞のＰＯＳカテゴリに属するものであることが理解されよう。したがって、ＰＯＳが名詞であるテキストエンティティが選択されることになる。したがって、必要とされる名前が、選択されたエンティティから抽出されることになる。
【００１９】
いくつかの実施形態において、複数のテキストエンティティに関連付けられたＰＯＳを識別する試行が行われると、ＰＯＳが識別されない１つまたは複数のユニークなテキストエンティティが識別されてよい。１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプが決定されてよい。エンティティタイプは、値エンティティまたはパターンエンティティであってよい。各値エンティティは関連付けられた値を有していてよく、各パターンエンティティは関連付けられたパターンを有する。
【００２０】
例えば、識別されたテキストエンティティは、電子メールＩＤまたは油田名（例えばそれぞれ「ｊｏｈｎ＠ｇｍａｉｌ．ｃｏｍ」または「ＡＢ－１２３」）のようなコードネームを示すテキスト文字の組み合わせから構成されていてよい。したがって、そのようなテキストエンティティについては、ＰＯＳが識別されない場合がある。したがって、そのようなテキストエンティティは、値エンティティまたはパターンエンティティのいずれかとして識別されてよい。したがって、電子メールＩＤ（「ｊｏｈｎ＠ｇｍａｉｌ．ｃｏｍ」）

は値エンティティとして決定されてよく、値は電子メールＩＤである。これは、電子メールＩＤのパターンが通常は典型的かつ既知のパターンであり、全ての電子メールＩＤが同じまたは類似のパターンを共有し得るためであり得る。一方、油田名に対応する「ＡＢ－１２３」のようなテキストエンティティは、パターンエンティティとして決定されてよい。これは、油田名が同じパターンを共有し得るが、そのようなパターンは一般的に知られていない場合がある、すなわちＮＬＰメタデータ抽出フレームワーク２００がそのようなパターンを予め記憶していない場合があるためである。値エンティティを識別すると、値エンティティの関連付けられた値と同じ関連付けられた値を有する入力ファイルにおけるテキストエンティティが、自動的に識別されてよい。換言すると、全ての電子メールＩＤテキストエンティティが、識別および抽出されてよい。

【００２１】

パターンエンティティを識別すると、パターンエンティティに関連付けられたパターン（Ｒｅｇｅｘ）に対応する検索パターンが、自動的に生成されてよい。さらに、入力ファイルにおける複数のテキストエンティティの各々に関連付けられたパターンが決定されてよい。パターンエンティティに対応する検索パターン（Ｒｅｇｅｘ）は、複数のテキストエンティティに関連付けられたパターンと対応付けられてよく、対応付けに基づいて、複数のテキストエンティティに関連付けられたパターンからの１つまたは複数の合致パターンが決定されてよい。１つまたは複数の合致パターンに対応する合致テキストエンティティが、複数のテキストエンティティから抽出されてよい。

【００２２】

例えば、油田に対応するテキストエンティティ「ＡＢ－１２３」について、検索パターン（Ｒｅｇｅｘ）「ａｐｄ」（ここで、「ａ」は１つまたは複数のアルファベットを意味し、「ｐ」は１つまたは複数の句読点を意味し、「ｄ」は１つまたは複数の数字を意味する）が自動的に生成されてよい。さらに、入力ファイルにおける他のテキストエンティティに関連付けられたパターンが決定され、検索パターン「ａｐｄ」が他のテキストエンティティに関連付けられたパターンと対応付けられてよく、合致パターンが識別されてよい。その後、合致パターンに対応する全てのテキストエンティティが抽出される。これにより、入力ファイルにおける全ての油田名が抽出されてよい。

【００２３】

いくつかの実施形態において、入力ファイルからのテキストデータが識別されると、第１のＮＬＰルールは、複数のテキストエンティティから関連テキストエンティティを識別するためにユーザからテキスト入力を受け取り、テキスト入力に対応する検索パターンを自動的に生成することを行わせてよい。第１のＮＬＰルールはさらに、複数のテキストエンティティの各々に関連付けられたパターンを決定し、テキスト入力に対応する検索パターンを複数のテキストエンティティに関連付けられたパターンと対応付けることを行わせてよい。第１のＮＬＰルールはさらに、対応付けに基づいて複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンを識別し、１つまたは複数の合致パターンに対応する関連テキストエンティティを複数のテキストエンティティから抽出することを行わせてよい。

【００２４】

例えば、ユーザは、入力ファイルから抽出された複数のテキストエンティティから関連テキストエンティティを識別するために、テキスト入力「ＡＢ－１２３」を提供してよい。検索パターン（Ｒｅｇｅｘ）「ａｐｄ」が、入力されたテキストエンティティ「ＡＢ－１２３」について識別されてよい。さらに、入力ファイルにおける他のテキストエンティティに関連付けられたパターンが決定されてよく、検索パターン（Ｒｅｇｅｘ）「ａｐｄ」が他のテキストエンティティに関連付けられたパターンと対応付けられてよく、合致パターンが識別される。その後、合致パターンに対応する全てのテキストエンティティが抽出されてよい。これにより、入力ファイルにおける全ての油田名が抽出されてよい。

【００２５】

いくつかの実施形態において、ユーザからのテキスト入力は、「Ｍｉｃｒｏｓｏｆｔ　Ｅ

ｘｃｅｌ（登録商標）」シートにおけるエントリの形態で受け取られてよいことに留意されたい。テキスト入力に対応する検索パターン（Ｒｅｇｅｘ）を自動的に生成し得る１つまたは複数のルールが定義されてよい。その後、１つまたは複数のルールは、入力ファイルの複数のテキストエンティティの各々に関連付けられたパターンを決定し、テキスト入力に対応する検索パターンを複数のテキストエンティティに関連付けられたパターンと対応付け、対応付けに基づいて複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンを識別し、１つまたは複数の合致パターンに対応する関連テキストエンティティを複数のテキストエンティティから抽出してよい。

【００２６】

したがって、いくつかの実施形態において、「Ｍｉｃｒｏｓｏｆｔ　Ｅｘｃｅｌ（登録商標）」データは、１つまたは複数の「Ｍｉｃｒｏｓｏｆｔ　Ｅｘｃｅｌ（登録商標）」シート２２２から受け取られてよい。様々なルール（ドメインルール、ＮＬＰルール、位置ルール等）が適用されることで、抽出されたデータ２２６が得られてよい。いくつかの実施形態において、「Ｍｉｃｒｏｓｏｆｔ　Ｅｘｃｅｌ（登録商標）」マクロが抽出に用いられてよい。さらに、いくつかの実施形態において、同マクロがＰｙｔｈｏｎ（言語；登録商標）に埋め込まれてよい。ここで、抽出されたデータ２２６は、複数のテキストエンティティから抽出された関連テキストエンティティ、または、様々なルールを用いて抽出された関連テキストエンティティに基づいて生成され得る表現を指すものであってよい。いくつかの追加的な実施形態において、上記のアプローチ、すなわちドメインベース、位置ベース、ＰＯＳベース、およびＲｅｇｅｘベースのアプローチがテキスト抽出を提供することに失敗した場合、機械学習（Machine Learning; ＭＬ）ベースのアプローチが用いられてよい。ＭＬベースのアプローチには、人工知能（ＡＩ）モデル（すなわちＭＬモデル）２２４が用いられてよい。

【００２７】

ここで図３を参照すると、本開示のいくつかの実施形態に係るスペルチェックシステム３００のブロック図が示されている。いくつかの実施形態において、抽出モジュール３０２（テキスト識別デバイス１０２に対応する）により抽出されたテキストエンティティが、スペルチェックシステム３００により受け取られてよい。抽出モジュール３０２により生成されたこれらの表現は、スペルチェックシステム３００への入力テキスト３０４として働いてよい。入力テキスト３０４は、前処理モジュール３０６により前処理されてよい。

【００２８】

前処理が行われると、コサイン類似度モジュール３０８により、前処理されたデータに対してコサイン類似度解析が行われてよい。一例として、コサイン類似度モジュール３０８は、閾値を用いてコサイン類似度解析を行ってよい。最小編集距離モジュール３１０が、コサイン類似度モジュール３０８から受け取られたデータに対して最小編集距離解析を行ってよい。いくつかの実施形態において、最小編集距離解析は、フィルタリング済みの単語に対して行われてよい。

【００２９】

最大コサイン類似度モジュール３１２が、最小編集距離モジュール３１０から受け取られたデータに対して最大コサイン類似度解析を行ってよい。最大コサイン類似度解析は、最小編集距離単語内の単語に基づくものであってよいことに留意されたい。置換モジュール３１４は、最大コサイン類似度モジュール３１２により行われた解析に基づいて、不正確な単語を修正された単語で置換してよい。

【００３０】

ここで図４を参照すると、本開示のいくつかの実施形態に係る推奨システム４００のブロック図が示されている。例えば、推奨システム４００は、エンティティタイプ、すなわち値エンティティまたはパターンエンティティを識別するために用いられてよい。いくつかの実施形態において、推奨システム４００は、分類ベースの機械学習（ＭＬ）モデル４０２を用いてよい。推奨システム４００は、入力データ（ＮＬＰ抽出・スペルチェックデータ）４０６を受け取ってよい。入力データは、ＮＬＰおよびスペルチェックが行われたデ

ータであってよいことに留意されたい。推奨システム４００は、設定（ｃｏｎｆｉｇ）ファイル４０８を含んでよい。ｃｏｎｆｉｇファイル４０８は、入力データについて値が推奨され得るか、またはパターンが推奨され得るかを決定することをトリガしてよい。

【００３１】

推奨システム４００は、入力データをＭＬモデル４０２に供給してよい。ＭＬモデル４０２は、アーカイブデータベース４０４からの履歴データを用いてよい。履歴データに基づいて、ＭＬモデル４０２は、予測データ４１０を提供してもよく、または推奨データ４１２を提供してもよい。したがって、推奨値４１４が予測データ４１０に基づいて生成されてもよく、または推奨パターン４１６が推奨データ４１２に基づいて生成されてもよい。

【００３２】

ここで図５を参照すると、本開示のいくつかの実施形態に係るメタデータ更新システム５００のブロック図が示されている。いくつかの実施形態において、メタデータ更新システム５００は、ＮＬＰ抽出済みデータ５０２を受け取ってよい。メタデータ更新システム５００は、システム日時モジュール５０４を含んでよい。メタデータ更新システム５００は、信頼度モジュール５０６および確率モジュール５０８をさらに含んでよい。信頼度モジュール５０６は、抽出されたデータおよびＮＬＰルールにより生成された表現の正確度に関連付けられた信頼度スコアを決定してよい。確率モジュール５０８は、信頼度スコアに基づいて、抽出されたデータおよびＮＬＰルールにより生成された表現の精度に関連付けられた正確度スコアを決定してよい。換言すると、確率モジュール５０８は、抽出されたデータまたはＮＬＰルールを用いて生成された表現がどれだけ正確／有用であるかについての確率を決定してよい。

【００３３】

確率モジュール５０８により決定される確率が十分に高い（すなわち閾値よりも大きい）場合、抽出されたデータおよびＮＬＰルールを用いて生成された表現、すなわちルールベース値５１４が提供されてよい。一方、確率モジュール５０８により決定される確率が十分に高くない（すなわち閾値よりも小さい）場合、抽出されたデータおよびＭＬモデルを用いて生成された表現、すなわち機械学習（ＭＬ）値５１６が提供されてよい。前述の通り、ＭＬ値は、アーカイブデータベース５１０に記憶され得る履歴データを用いてＭＬモデルにより生成されてよい。抽出されたデータおよび生成された表現は、最終データベース５１２に記憶されてよい。

【００３４】

ここで図６を参照すると、本開示の一実施形態に係る、入力ファイルのコンテンツから情報を抽出する方法６００のフローチャートが示されている。理解されるように、方法６００は、テキスト抽出のＲｅｇｅｘベースの方法を提供してよい。ステップ６０２において、入力ファイルからテキストデータが識別されてよい。識別されたテキストデータは、複数のテキストエンティティを含んでよい。いくつかの例において、入力ファイルは、画像ファイルおよびポータブル・ドキュメント・フォーマット（ＰＤＦ）ファイルのうちの少なくとも１つを含んでよいことに留意されたい。

【００３５】

ステップ６０４において、複数のテキストエンティティから関連テキストエンティティを識別するために、テキスト入力がユーザから受け取られてよい。例えば、ユーザは、テキストを「Ｍｉｃｒｏｓｏｆｔ　Ｅｘｃｅｌ（登録商標）」シートにおけるエントリとして提供してよい。ステップ６０６において、テキスト入力に対応する検索パターン（Ｒｅｇｅｘ）が、自動的に生成されてよい。ステップ６０８において、複数のテキストエンティティの各々に関連付けられたパターンが決定されてよい。ステップ６１０において、テキスト入力に対応する検索パターンが、複数のテキストエンティティに関連付けられたパターンと対応付けられてよい。ステップ６１２において、複数のテキストエンティティに関連付けられたパターンからの１つまたは複数の合致パターンが、対応付けに基づいて識別されてよい。ステップ６１４において、１つまたは複数の合致パターンに対応する関連テキストエンティティが、複数のテキストエンティティから抽出されてよい。

【００３６】
加えて、いくつかの実施形態においては、複数のテキストエンティティから１つまたは複数の合致パターンに対応する関連テキストエンティティを抽出すると、抽出された関連テキストエンティティを用いて、表現が生成されてよい。
【００３７】
さらに、例えば方法６００では関連エンティティを抽出することができない場合に、関連エンティティを抽出するために機械学習（ＭＬ）モデルが用いられてよい。換言すると、データ（関連テキストエンティティ）を抽出することができない場合、ＭＬベースのアプローチが開始してよい。
【００３８】
ここで図７を参照すると、本開示の別の実施形態に係る、入力ファイルのコンテンツから情報を抽出する方法７００のフローチャートが示されている。ステップ７０２において、入力ファイルからテキストデータが識別されてよい。識別されたテキストデータは、複数のテキストエンティティを含んでよい。入力ファイルは、画像ファイルおよびポータブル・ドキュメント・フォーマット（ＰＤＦ）ファイルのうちの少なくとも１つを含んでよいことに留意されたい。
【００３９】
ステップ７０４において、複数のテキストエンティティの各々に関連付けられたドメインベースの名前が決定されてよい。ドメインベースの名前は、複数のテキストエンティティの各々を、データデポジトリに記憶されたドメインベースの名前のリストと対応付けることに基づいて決定されてよい。ステップ７０６において、入力ファイルに関連付けられたタイプが識別されてよい。ステップ７０８において、入力ファイルに関連付けられたタイプに基づいて、入力ファイルにおける１つまたは複数の関連テキストエンティティの位置が決定されてよい。ステップ７１０において、位置に基づいて、１つまたは複数の関連テキストエンティティが入力ファイルから抽出されてよい。
【００４０】
ステップ７１２において、複数のテキストエンティティに関連付けられた品詞（ＰＯＳ）が識別されてよい。ＰＯＳは、名詞、代名詞、動詞、副詞、形容詞、接続詞、前置詞、および間投詞のうちの１つであってよいことを理解されたい。ステップ７１４において、ＰＯＳが名詞として識別される１つまたは複数のテキストエンティティが決定されてよい。
【００４１】
ステップ７１６において、ＰＯＳが識別されない１つまたは複数のユニークなテキストエンティティが選択されてよい。ステップ７１８において、１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプが決定されてよい。エンティティタイプは、値エンティティおよびパターンエンティティのうちの１つであってよい。さらに、各値エンティティは関連付けられた値を有していてよく、各パターンエンティティは関連付けられたパターンを有する。図４に関連して説明したように、１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプを決定するために、（ＭＬベースの）推奨システム４００が用いられてよい。
【００４２】
ステップ７２０において、各値エンティティについて、値エンティティの関連付けられた値と同じ関連付けられた値を有する１つまたは複数のテキストエンティティが、自動的に識別されてよい。各パターンエンティティについて、ステップ７２２～７２８が行われてよい。したがって、ステップ７２２において、パターンエンティティに関連付けられたパターンに対応する検索パターン（Ｒｅｇｅｘ）が、自動的に生成されてよい。ステップ７２４において、パターンエンティティに対応する検索パターンが、複数のテキストエンティティに関連付けられたパターンと対応付けられてよい。ステップ７２６において、対応付けに基づいて、複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンが決定されてよい。ステップ７２８において、１つまたは複数の合致パターンに対応する合致テキストエンティティが、複数のテキストエンティティから抽出

されてよい。
【００４３】
方法７００は、方法６００と組み合わせて行われてよいことに留意されたい。例示的シナリオにおいて、方法６００は、方法７００のステップ７１０が完了した後、方法７００のステップ７１２が始まる前に開始してよい。
【００４４】
ここで図８を参照すると、本開示の別の実施形態に係る、入力ファイルのコンテンツから情報を抽出する方法８００のフローチャートが示されている。ステップ８０２において、入力ファイルのコンテンツからテキストデータが抽出されてよい。抽出されたテキストデータは、複数のテキストエンティティを含んでよい。
【００４５】
ステップ８０４において、複数のテキストエンティティの各々に関連付けられたドメインベースの名前が決定されてよい。ドメインベースの名前は、複数のテキストエンティティの各々を、データデポジトリに記憶されたドメインベースの名前のリストと対応付けることに基づいて決定されてよい。例えば、ＰＩＮコードが取得可能な場合、例えばＰＩＮコードをナレッジベースに記憶された（ＰＩＮコードおよび関連付けられた都市名の）リストと対応付けることにより、ドメインベースのアプローチを用いてＰＩＮコードから都市が決定されてよい。同様に、都市名（例えばデリー）が識別された場合、国名（例えばインド）が決定されてよい。いくつかの実施形態においては、それに応じて、第２の複数のエンティティの各々にタグが割り当てられてよい。
【００４６】
ステップ８０６において、ドメインベースの名前を決定することに成功したか否かを確認するためのチェックが行われてよい。ドメインベースの名前を決定することに成功した場合、本方法はステップ８２２（「Ｙｅｓ」の経路）に進んでよく、このとき、識別されたテキストエンティティが抽出されてよい。さらに、抽出されたテキストエンティティを用いて、表現が生成されてよい。ステップ８０６において、ドメインベースの名前が決定されない場合、本方法はステップ８０８（「Ｎｏ」の経路）に進んでよい。換言すると、ドメインベースの名前を識別することができない場合、本方法は、次の代替的アプローチの実行に進んでよい。
【００４７】
ステップ８０８において、複数のテキストエンティティの各々の位置に基づいて、所定のフィールドの値が決定されてよい。位置は、（ｘおよびｙ）座標の形態であってよい。一例として、位置は、データリポジトリに記憶されたテンプレートに基づいて予め決定されてよい。例えば、いくつかのシナリオにおいて、特定の名前の値がその名前の横または特定の名前ヘッダの下に記載されていてよく、これは（ｘおよびｙ）座標に基づいて識別可能であってよい。例えば、「Ａａｄｈａｒカード」、または「ＰＡＮカード」、または運転免許証において、エンティティ、例えばユーザの名前が特定の位置において取得可能であってよく、そこから名前を識別することができる。位置の技法は、表の場合に特に有用であり得る。これは、表におけるデータは構造化された形式で利用可能であり、エンティティの位置が容易かつ正確に特定され得るためである。したがって、入力ファイルに関連付けられたタイプ、すなわち入力ファイルが「Ａａｄｈａｒカード」であるか、または「ＰＡＮカード」であるか、または運転免許証であるかが、まず識別されてよい。さらに、入力ファイルに関連付けられたタイプに基づいて、入力ファイルにおける１つまたは複数の関連テキストエンティティの位置が決定されてよい。
【００４８】
ステップ８１０において、入力ファイルにおける１つまたは複数の関連テキストエンティティの位置を決定することに成功したか否かを確認するためのチェックが行われてよい。位置を決定することに成功した場合、方法８００は再びステップ８２２（「Ｙｅｓ」の経路）に進んでよく、このとき、入力ファイルからの１つまたは複数の関連テキストエンティティが位置に基づいて抽出されてよく、抽出されたテキストエンティティを用いて表現

が生成されてよい。位置が決定されないと認められる場合、ステップ８１０において、方法８００はステップ８１２に進んでよい。

【００４９】

ステップ８１２において、関連テキストエンティティを抽出するために、入力ファイルに対してＲｅｇｅｘベースのアプローチが試行されてよい。この目的で、複数のテキストエンティティから関連テキストエンティティを識別するためにユーザからテキスト入力が受け取られてよく、テキスト入力に対応する検索パターンが自動的に生成されてよい。換言すると、Ｒｅｇｅｘが生成されてよい。さらに、複数のテキストエンティティの各々に関連付けられたパターンが生成されてよく、テキスト入力に対応する検索パターン（Ｒｅｇｅｘ）が、複数のテキストエンティティに関連付けられたパターンと対応付けられてよく、対応付けに基づいて、複数のテキストエンティティに関連付けられたパターンからの１つまたは複数の合致パターンが識別されてよい。いくつかの実施形態において、方法８００は、Ｒｅｇｅｘに対応する、抽出されたテキストデータから関連エンティティを決定する確率を算出することを含んでよい。

【００５０】

ステップ８１４において、Ｒｅｇｅｘベースのアプローチを適用することに成功したかを決定するためのチェックが行われてよい。Ｒｅｇｅｘベースのアプローチを適用することに成功した場合、方法８００はステップ８２２（「Ｙｅｓ」の経路）に進んでよく、このとき、１つまたは複数の合致パターンに対応する関連テキストエンティティが複数のテキストエンティティから抽出されてよく、抽出されたテキストエンティティを用いて表現が生成されてよい。Ｒｅｇｅｘベースのアプローチが成功しない場合、方法８００はステップ８１６（「Ｎｏ」の経路）に進んでよい。

【００５１】

ステップ８１６において、品詞（ＰＯＳ）ベースのアプローチが試行されてよい。例えば、複数のテキストエンティティの各々に関連付けられたＰＯＳが決定されてよい。換言すると、ステップ８１６において、エンティティがいずれのＰＯＳと関連付けられ得るか、例えばＰＯＳが名詞、副詞、形容詞等であり得るかを決定するための試行が行われてよい。さらに、ステップ８１６において、１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプを決定するための試行が行われてよい。この目的で、ＰＯＳが識別されない１つまたは複数のユニークなテキストエンティティが選択されてよい。例えば、アルファベット、数字、句読点等の組み合わせを有するユニークなエンティティ（上記で述べた電子メールＩＤまたは油田名など）については、ＰＯＳが識別されない場合がある。そのようなテキストエンティティについて、１つまたは複数のユニークなテキストエンティティの各々に関連付けられたエンティティタイプが決定されてよい。エンティティタイプは、値エンティティおよびパターンエンティティのうちの１つであってよい。各値エンティティは関連付けられた値を有していてよく、各パターンエンティティは関連付けられたパターンを有していてよい。さらに、各値エンティティについて、値エンティティの関連付けられた値と同じ関連付けられた値（例えば電子メールＩＤ）を有する１つまたは複数のテキストエンティティを自動的に識別するための試行が行われてよい。

【００５２】

ステップ８１８において、試行が成功したか（すなわちＰＯＳベースのアプローチが成功したか）を決定するためのチェックが行われてよい。試行が成功したと認められる場合、方法８００はステップ８２２（「Ｙｅｓ」の経路）に進んでよく、このとき、値エンティティの関連付けられた値と同じ関連付けられた値を有する１つまたは複数のテキストエンティティ（関連テキストエンティティ）が抽出されてよく、抽出されたテキストエンティティを用いて表現が生成されてよい。

【００５３】

さらに、各パターンエンティティについて、パターンエンティティに関連付けられたパターンに対応する検索パターンを自動的に生成し、パターンエンティティに対応する検索パ

ターンを複数のテキストエンティティに関連付けられたパターンと対応付け、対応付けに基づいて複数のテキストエンティティに関連付けられたパターンから１つまたは複数の合致パターンを決定するための試行が行われてよい。ステップ８１８において、試行が成功したか（すなわちＰＯＳベースのアプローチが成功したか）を決定するためのチェックが行われてよい。試行が成功したと認められる場合、方法８００はステップ８２２（「Ｙｅｓ」の経路）に進んでよく、このとき、１つまたは複数の合致パターンに対応する合致テキストエンティティが複数のテキストエンティティから抽出されてよく、抽出されたテキストエンティティを用いて表現が生成されてよい。ステップ８１８において、ＰＯＳベースのアプローチが成功しないと認められる場合、方法８００はステップ８２０（「Ｎｏ」の経路）に進んでよい。

【００５４】

ステップ８２０において、機械学習（ＭＬ）モデルを用いて複数のテキストエンティティから関連テキストエンティティ（情報）を抽出するための試行が行われてよい。ＭＬモデルは、関連テキストエンティティのタイプを抽出するようにまず訓練されてよいことが理解されよう。したがって、ＭＬモデルの訓練に基づいて、ＭＬベースの分類が、関連テキストエンティティを抽出するように入力ファイルに適用されてよい。したがって、ステップ８２２において、ＭＬベースの分類に基づいて関連テキストエンティティが抽出されてよく、抽出されたテキストエンティティを用いて表現が生成されてよい。

【００５５】

いくつかの実施形態において、様々なステップ（すなわちステップ８０４、８０８、８１２、８１６、および８２０）は、上記で説明した順序で行われてもよく、または任意の他の順序で行われてもよいことに留意されたい。

【００５６】

さらに、下記のシナリオにおいてＮＬＰベースのアプローチが用いられてよいことに留意されたい。

（ｉ）検索キー（すなわちユーザからのテキスト入力）が利用可能であり、かつそれに関連付けられたテキスト抽出（すなわち関連エンティティ）が入力ファイルにおいて予期される場合。

（ｉｉ）検索キー（すなわちユーザからのテキスト入力）が利用可能でないが、それに関連付けられたテキスト抽出（すなわち関連エンティティ）が入力ファイルにおいて予期される場合。

（ｉｉｉ）検索キー（すなわちユーザからのテキスト入力）が利用可能であるが、それに関連付けられたテキスト抽出（すなわち関連エンティティ）が入力ファイルにおいて予期されない場合。

【００５７】

さらに、各々のアプローチにおいて信頼度スコア（確率スコア）が算出されてよいことに留意されたい。例えば、「ドメインベース」のアプローチについては、検索キー（すなわちユーザからのテキスト入力）が入力ファイルにおける残りのテキストエンティティと厳密に合致する場合に、最大の確率スコア（例えば１）が得られてよい。さらに、「位置ベース」のアプローチでは、確率スコア（ｐ）は、検索キーの合致の確率（ｗ）および位置（ｌ）の識別、すなわちｐ＝ｗ＊ｌに基づいていてよい。Ｒｅｇｅｘベースのアプローチについては、位置識別についての確率は０または１のいずれかであってよく、抽出の確率スコア（ｅ）は１または０であり得る。さらに、抽出される値は複数（ｎ）であり得、それらの値のうちの１つが選択されてよい。ゆえに、選択される値の確率は「１／ｎ」であってよく、信頼度スコアはｅ＊（１／ｎ）であってよい。「位置ベース」および「Ｒｅｇｅｘベース」のアプローチの両方が用いられる場合、各アプローチに５０％の重みが与えられてよい。ゆえに、信頼度スコアは（ｗ＊ｌ）／２＋（ｅ＊（１／ｎ））／２であってよい。ＭＬベースのアプローチについては、確率スコアが混同行列（正解率）に基づいて導出されてよい。

【００５８】

さらに、以下のアプローチ、すなわち純粋な単語合致アプローチ、ルールベースのアプローチ（すなわちドメインベース、位置ベース、ＰＯＳベース、およびＲｅｇｅｘベースのアプローチのうちの１つまたは複数）、およびＭＬベースのアプローチが用いられる場合、複合確率スコア（Ｐ）が算出されてよい。
したがって、
複合確率スコア（Ｐ）＝（Ｐ１＊０．３）＊（Ｐ３＊０．７）、または
複合確率スコア（Ｐ）＝（Ｐ１＊０．３）＊（Ｐ２＊０．７）
式中、
Ｐ１＝純粋な単語合致アプローチの確率スコア、
Ｐ２＝ルールベースのアプローチの確率スコア、および
Ｐ３＝ＭＬベースのアプローチの確率スコア。

【００５９】
＜ケースシナリオ１＞
ここで図９を参照すると、ノイジーなエンティティ９０２が入力ファイル９００に存在する例示的な入力ファイル９００のスナップショットが示されている。例えば、エンティティの第１の部分は汚れまたはかすれ等により判読できない場合があり、エンティティの第２の部分は「Ｕｎｉｔｅｄ」という単語を含む場合がある。したがって、実際のエンティティは「Ｕｎｉｔｅｄ　Ａｉｒｌｉｎｅｓ（登録商標）」または「Ｕｎｉｔｅｄ　Ｓｔａｔｅｓ」等である可能性がある。このエンティティを決定するために、１つまたは複数のアプローチ（すなわちドメインベース、位置ベース、ＰＯＳベース、Ｒｅｇｅｘベース、またはＭＬベース）が実際のエンティティの決定に用いられてよい。例えば、ドメインベース、位置ベース、ＰＯＳベース、およびＲｅｇｅｘベースのアプローチが抽出を提供することに失敗した場合、ＭＬベースのアプローチが開始してよく、分類ベースの機械学習（ＭＬ）モデルを用いて履歴データに基づいてテキスト抽出を提供してよい。ＭＬモデルは、決定論的モデルおよび／または確率論的モデルであってよい。

【００６０】
＜ケースシナリオ２＞
ここで図１０を参照すると、入力ファイル「Ｌｏｃａｌ　Ｏｒｄｅｒ」１０００のスナップショットが示されており、そこから様々なテキストエンティティを抽出することが必要とされ得る。図１０に示すように、入力ファイル「Ｌｏｃａｌ　Ｏｒｄｅｒ」１０００は表形式構造である。

【００６１】
例えば、属性「ＲＯ　Ｎｕｍｂｅｒ」についての値を抽出するために、ＲＯ　ｄａｔｅの属性を対応付けることにより、「ＲＯ　Ｎｕｍｂｅｒ」がドメインディクショナリデータベース（ルックアップテーブル）１００２から抽出されてよい。

【００６２】
属性Ｐａｒｔ　Ｎｕｍｂｅｒを抽出するために、入力ファイル１０００に存在するキー「Ｐａｒｔ　Ｎｕｍｂｅｒ」が識別され、またはユーザ入力として受け取られてよい。その後、位置ベースのアプローチにより、データベースに記憶されたテンプレート情報を用いて、Ｐａｒｔ　Ｎｕｍｂｅｒの可能性のある位置が決定されてよい。例えば、位置ベースのアプローチにより、Ｐａｒｔ　Ｎｕｍｂｅｒが入力ファイル１０００の「右」側に存在し得ることが示唆されてよい。したがって、必要とされるＰａｒｔ　Ｎｕｍｂｅｒ「３２１４５６４３」が、入力ファイル「Ｌｏｃａｌ　Ｏｒｄｅｒ」１０００の表形式構造の右のセル座標から抽出されてよい。

【００６３】
入力ファイル１０００に存在する電子メールＩＤの属性を抽出するために、Ｒｅｇｅｘベースのアプローチが用いられてよい。したがって、電子メールの属性に関連付けられたキー（Ｒｅｇｅｘ）が識別され、またはデータベースから受け取られてよい。したがって、Ｒｅｇｅｘは、電子メールＩＤ「ａｊａｙ．ｔｈａｋｕｒ@ｇｍａｉｌ．ｃｏｍ」に対応する「ａｐａｐａｐａ」（「ａ」は１つまたは複数のアルファベットを意味し、「ｐ」は

１つまたは複数の句読点を意味し、「ｄ」は１つまたは複数の数字を意味する）であってよい。したがって、合致パターン（Ｒｅｇｅｘ）を有するテキストエンティティ、すなわち電子メールＩＤのテキストエンティティが識別されてよく、値「ａｊａｙ．ｔｈａｋｕｒ＠ｇｍａｉｌ．ｃｏｍ」が入力ファイル１０００の表形式構造の対応するセルから抽出されてよい。

【００６４】

入力ファイル１０００から属性Ｎａｍｅを抽出するために、ＰＯＳベースのアプローチ用いられてよい。例えば、まず属性（Ｎａｍｅ）について存在するキーが入力ファイル１０００において識別されてよい。その後、例えば固有表現抽出（Named Entity Recognizer; ＮＥＲ）タガーを用いて、入力ファイル１０００における識別されたテキストエンティティのＰＯＳが決定されてよい。属性Ｎａｍｅについてデータベースにおいて指定されるＮＥＲタガーが検索されてよい。したがって、名詞のＰＯＳを有する属性Ｎａｍｅ（「Ａｊａｙ　Ｔｈａｋｕｒ」）に対応するテキストエンティティが抽出されてよい。したがって、抽出されたテキストエンティティ「Ａｊａｙ　Ｔｈａｋｕｒ」を用いる表現が生成されてよい。

【００６５】

入力ファイルのコンテンツから情報を抽出するための１つまたは複数のテキスト抽出技法が、上記に開示されている。上記の技法は、従来の技法に対して１つまたは複数の利点を提供する。例えば、これらの技法は、テキスト抽出を行うための様々なアプローチ、すなわちドメインベース、位置ベース、ＰＯＳベース、Ｒｅｇｅｘベースのアプローチ、およびＭＬベースのアプローチを提供する。したがって、単一のアプローチ、または任意の順序で適用される複数のアプローチを用いて、テキスト抽出を行うことができる。さらに、これらの技法は、抽出された情報を用いて表現を生成することを可能とする。さらに、これらの技法は、タグが存在しない、または抽出された情報においてパターンを識別することができない複雑な文書においても、情報を抽出し、抽出された情報を用いて表現を生成することを可能とする。

【００６６】

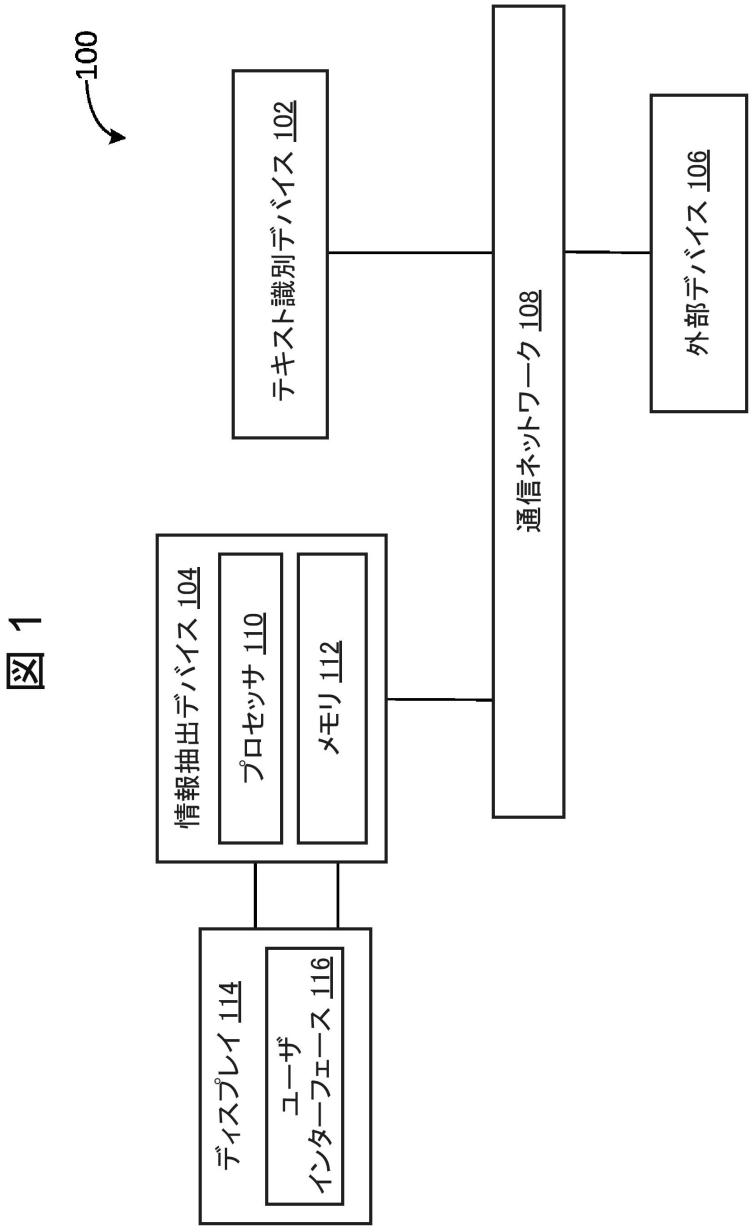
本開示と整合する実施形態の実装において、１つまたは複数のコンピュータ可読記憶媒体が利用されてよい。コンピュータ可読記憶媒体とは、プロセッサにより読み取り可能な情報またはデータが記憶され得る任意のタイプの物理メモリを指す。よって、コンピュータ可読記憶媒体は、本明細書に記載の実施形態と整合するステップまたは段階をプロセッサに実行させるための命令を含む、１つまたは複数のプロセッサが実行するための命令を記憶してよい。「コンピュータ可読媒体」という用語は、有形の物品を含み、搬送波および過渡信号を除外する、すなわち非一時的なものであると理解されるべきである。例としては、ランダムアクセスメモリ（ＲＡＭ）、リードオンリメモリ（ＲＯＭ）、揮発性メモリ、不揮発性メモリ、ハードドライブ、ＣＤ‐ＲＯＭ、ＤＶＤ、フラッシュドライブ、ディスク、および任意の他の既知の物理記憶媒体が挙げられる。

【００６７】

開示および例は単に例示的なものとみなされることが意図されており、開示の実施形態の真の範囲および趣旨は、以下の特許請求の範囲によって示される。

【図１】

図１


100
テキスト識別デバイス 102
通信ネットワーク 108
外部デバイス 106
情報抽出デバイス 104
プロセッサ 110
メモリ 112
ディスプレイ 114
ユーザ
インターフェース 116

【図２】

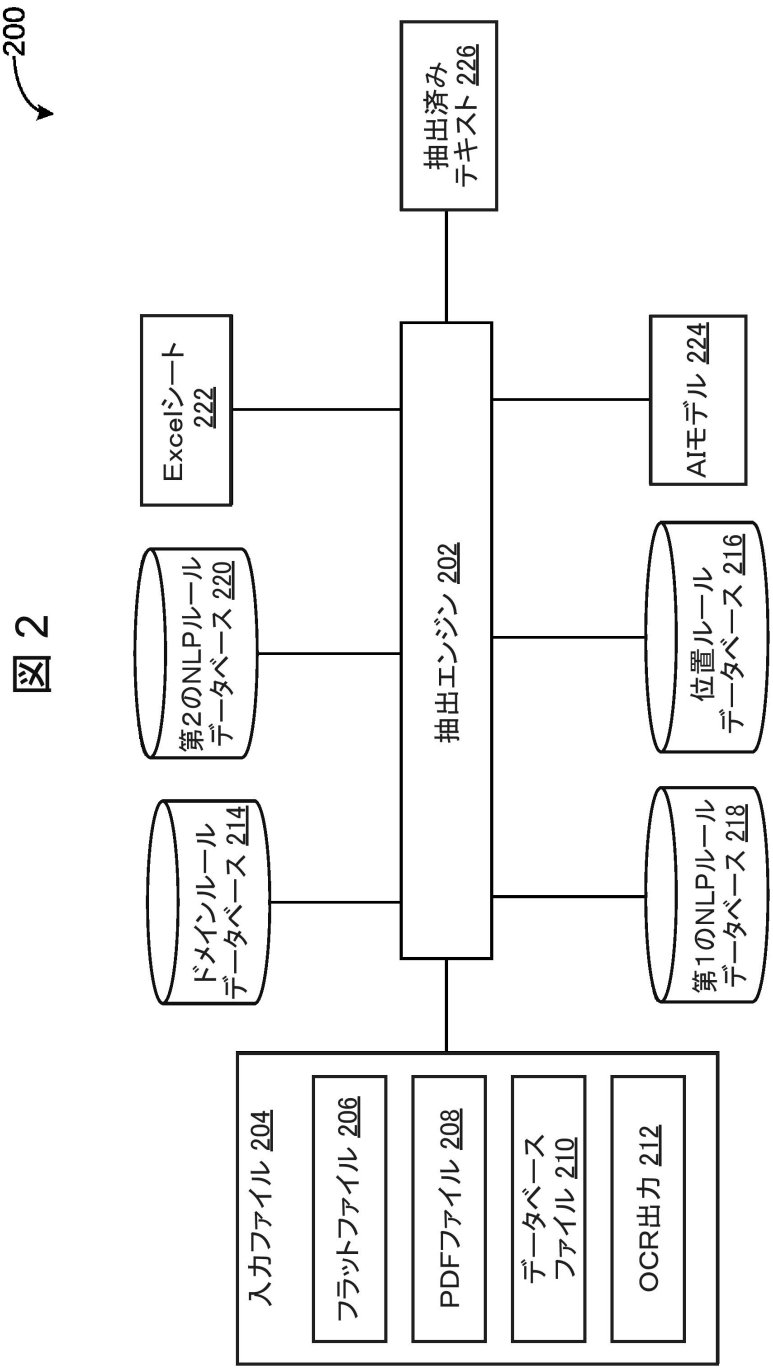
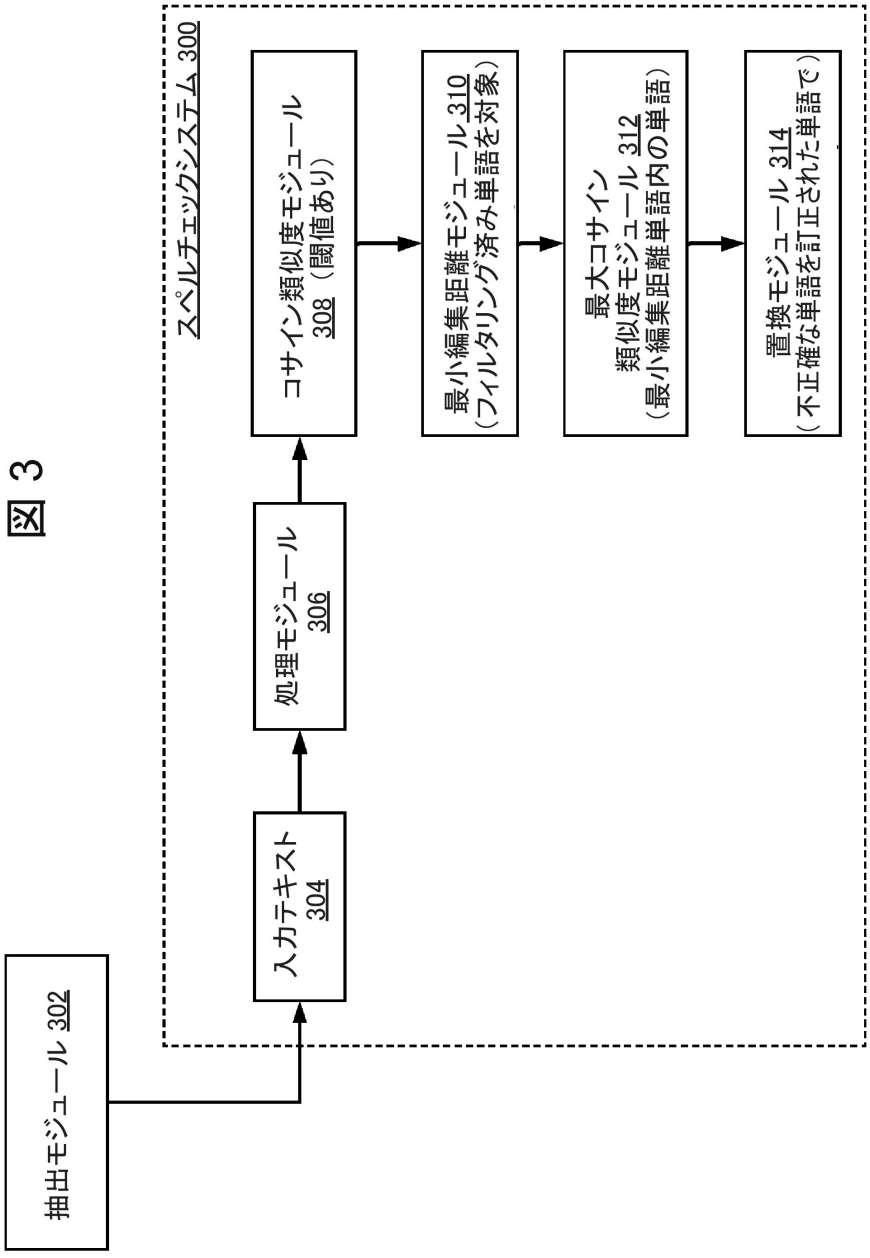

200
抽出済みテキスト 226
Excelシート 222
AIモデル 224
抽出エンジン 202
第2のNLPルールデータベース 220
位置ルールデータベース 216
ドメインルールデータベース 214
第1のNLPルールデータベース 218
入力ファイル 204
フラットファイル 206
PDFファイル 208
データベースファイル 210
OCR出力 212

図２

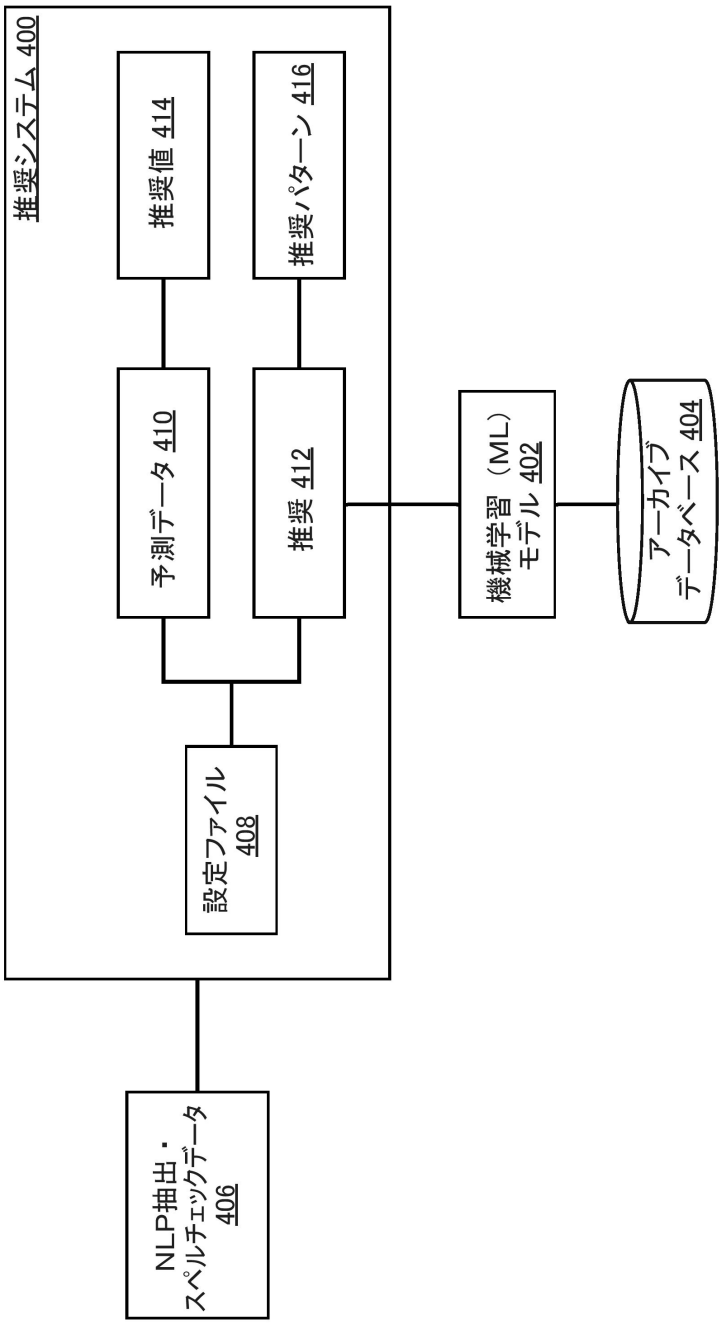
【図３】

図３


抽出モジュール 302
スペルチェックシステム 300
入力テキスト 304
処理モジュール 306
コサイン類似度モジュール 308（閾値あり）
最小編集距離モジュール 310（フィルタリング済み単語を対象）
最大コサイン類似度モジュール 312（最小編集距離単語内の単語）
置換モジュール 314（不正確な単語を訂正された単語で）

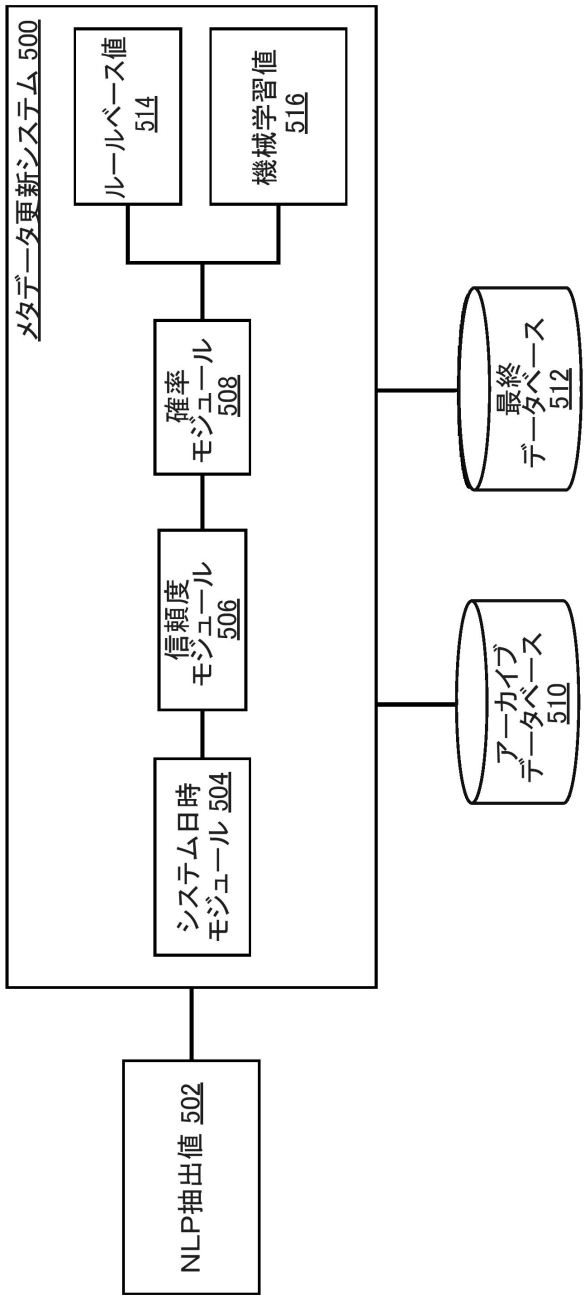
【図４】

図４


推奨システム 400
推奨値 414
推奨パターン 416
予測データ 410
推奨 412
設定ファイル 408
機械学習（ML）モデル 402
アーカイブデータベース 404
NLP抽出・スペルチェックデータ 406

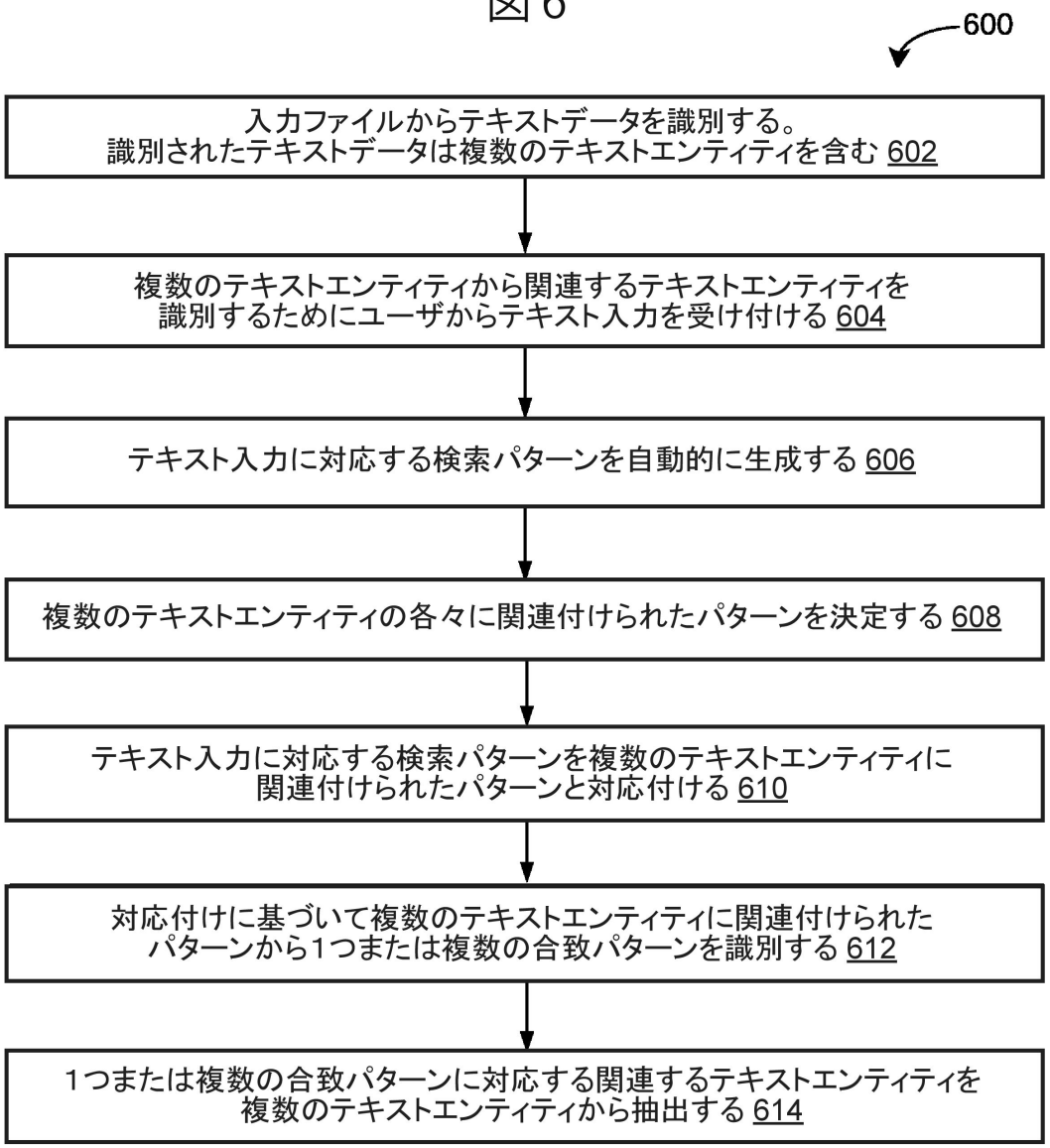
【図５】

図５


メタデータ更新システム 500
ルールベース値 514
機械学習値 516
確率モジュール 508
信頼度モジュール 506
システム日時モジュール 504
NLP抽出値 502
最終データベース 512
アーカイブデータベース 510

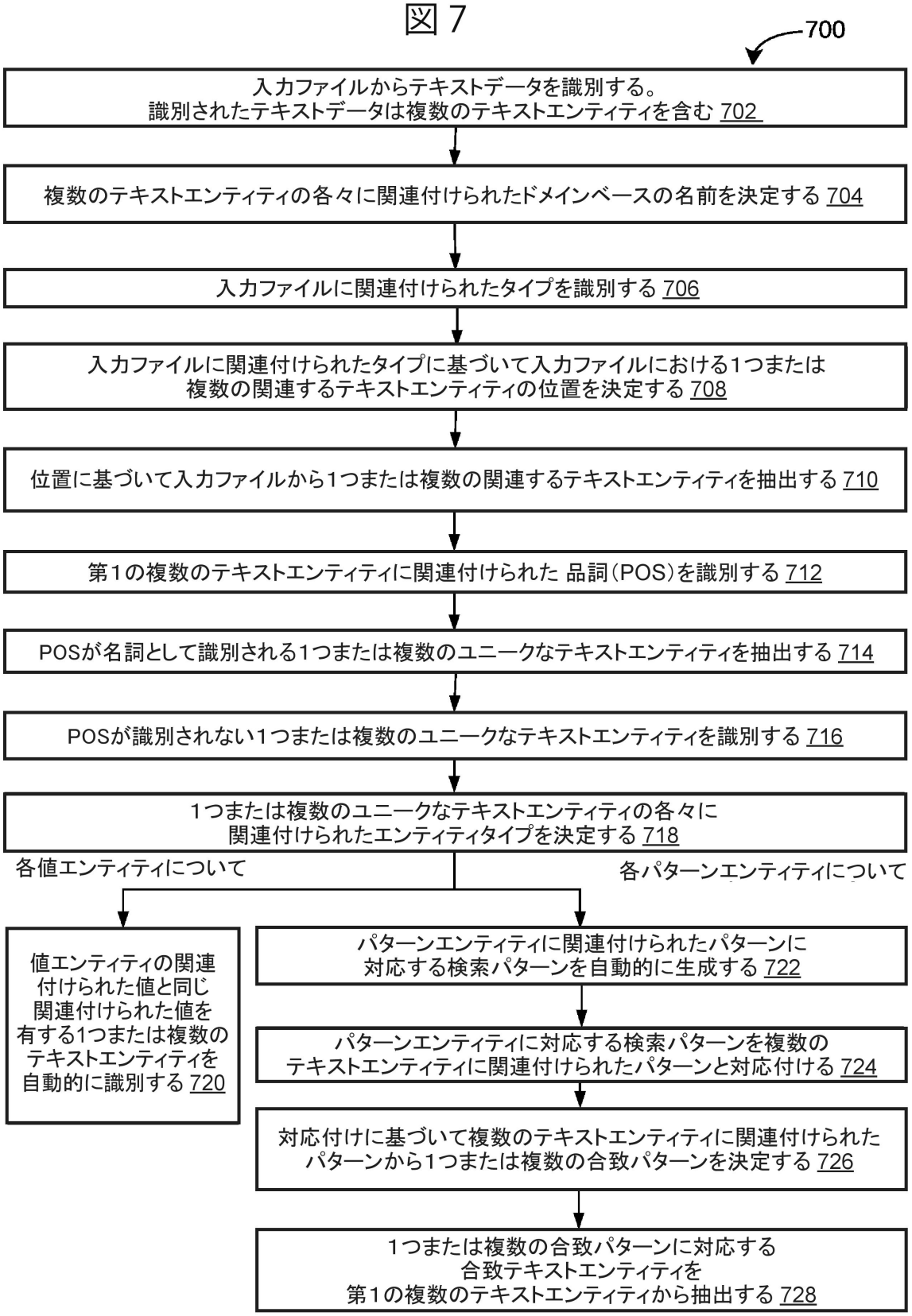
【図６】

図６


600
入力ファイルからテキストデータを識別する。
識別されたテキストデータは複数のテキストエンティティを含む 602
複数のテキストエンティティから関連するテキストエンティティを
識別するためにユーザからテキスト入力を受け付ける 604
テキスト入力に対応する検索パターンを自動的に生成する 606
複数のテキストエンティティの各々に関連付けられたパターンを決定する 608
テキスト入力に対応する検索パターンを複数のテキストエンティティに
関連付けられたパターンと対応付ける 610
対応付けに基づいて複数のテキストエンティティに関連付けられた
パターンから1つまたは複数の合致パターンを識別する 612
1つまたは複数の合致パターンに対応する関連するテキストエンティティを
複数のテキストエンティティから抽出する 614

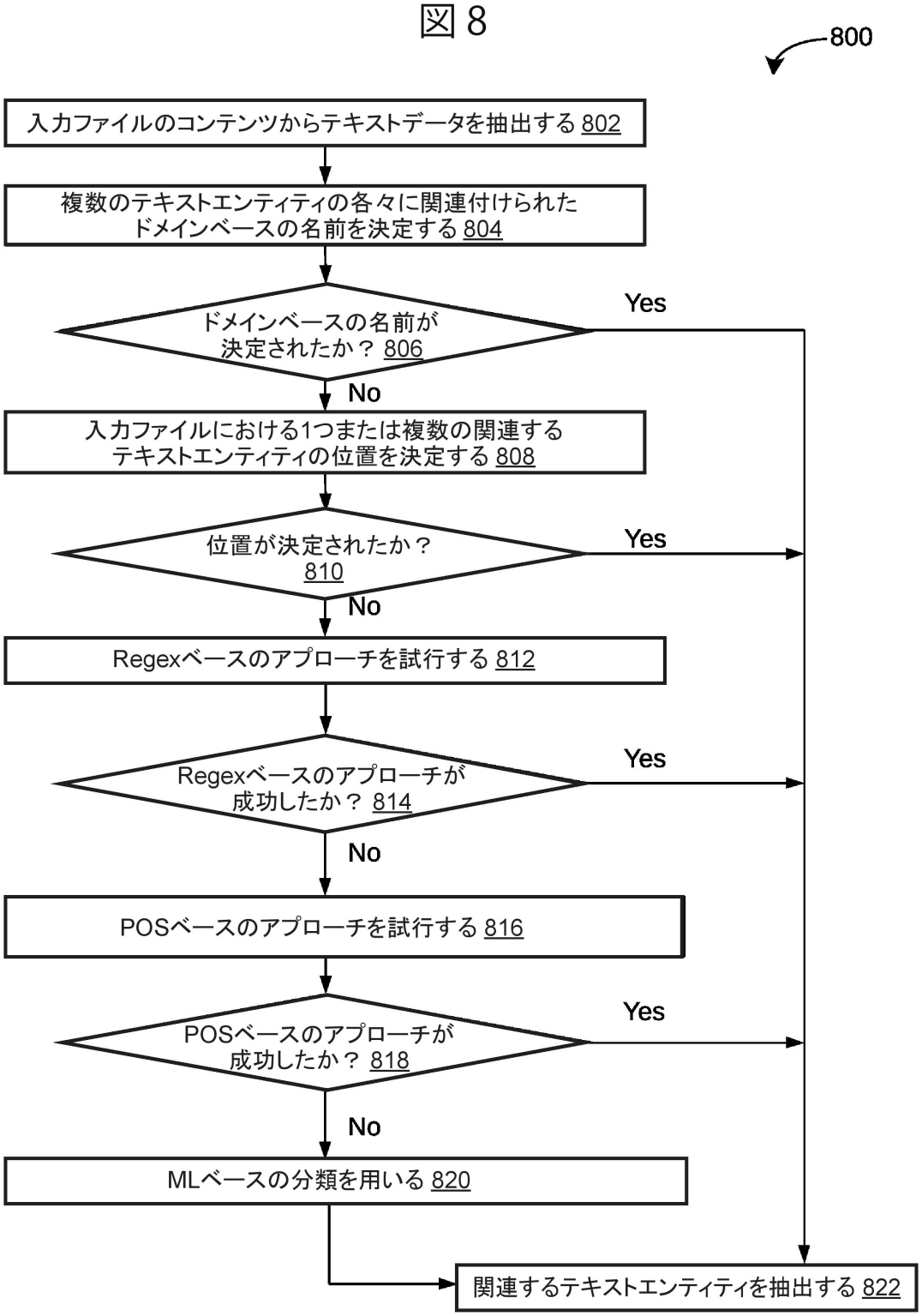
【図７】

図７


700
入力ファイルからテキストデータを識別する。
識別されたテキストデータは複数のテキストエンティティを含む 702
複数のテキストエンティティの各々に関連付けられたドメインベースの名前を決定する 704
入力ファイルに関連付けられたタイプを識別する 706
入力ファイルに関連付けられたタイプに基づいて入力ファイルにおける1つまたは
複数の関連するテキストエンティティの位置を決定する 708
位置に基づいて入力ファイルから1つまたは複数の関連するテキストエンティティを抽出する 710
第1の複数のテキストエンティティに関連付けられた 品詞(POS)を識別する 712
POSが名詞として識別される1つまたは複数のユニークなテキストエンティティを抽出する 714
POSが識別されない1つまたは複数のユニークなテキストエンティティを識別する 716
1つまたは複数のユニークなテキストエンティティの各々に
関連付けられたエンティティタイプを決定する 718
各値エンティティについて
各パターンエンティティについて
値エンティティの関連
付けられた値と同じ
関連付けられた値を
有する1つまたは複数の
テキストエンティティを
自動的に識別する 720
パターンエンティティに関連付けられたパターンに
対応する検索パターンを自動的に生成する 722
パターンエンティティに対応する検索パターンを複数の
テキストエンティティに関連付けられたパターンと対応付ける 724
対応付けに基づいて複数のテキストエンティティに関連付けられた
パターンから1つまたは複数の合致パターンを決定する 726
1つまたは複数の合致パターンに対応する
合致テキストエンティティを
第1の複数のテキストエンティティから抽出する 728

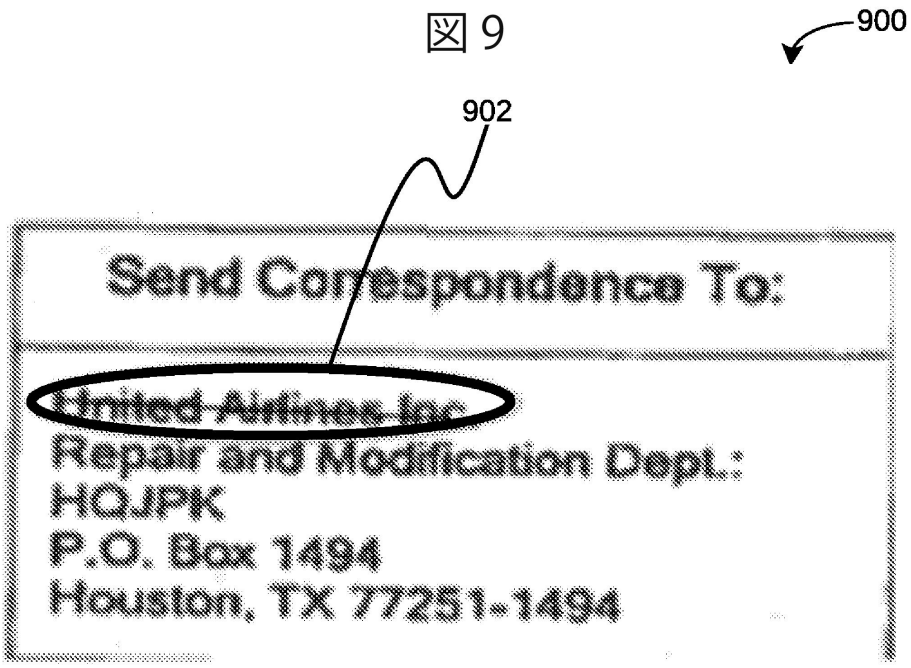
【図８】

図８


800
入力ファイルのコンテンツからテキストデータを抽出する 802
複数のテキストエンティティの各々に関連付けられた
ドメインベースの名前を決定する 804
ドメインベースの名前が
決定されたか？ 806
Yes
No
入力ファイルにおける1つまたは複数の関連する
テキストエンティティの位置を決定する 808
位置が決定されたか？
810
Yes
No
Regexベースのアプローチを試行する 812
Regexベースのアプローチが
成功したか？ 814
Yes
No
POSベースのアプローチを試行する 816
POSベースのアプローチが
成功したか？ 818
Yes
No
MLベースの分類を用いる 820
関連するテキストエンティティを抽出する 822

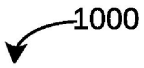
【図９】

図９


900
902
Send Correspondence To:
United Airlines Inc.
Repair and Modification Dept.:
HQJPK
P.O. Box 1494
Houston, TX 77251-1494

【図１０】

図10



Local Purchase Order

		R.O. Number	156789214200	Account ID: 125673
		Date Required	12/02/2020	Special Instruction: None
		R.O. Date	14/02/2020	Warranty: None
Mode of Payment	Serial No.	234761	Customer Info	
Online	Part No.	32145643	Name:Ajay Thakur	
	IMEI No.	867345823421	Email:ajay.thakur@gmail.com	
	Tracking Tag:	1345276	Phone No. 8567434651	
			Fax No:3214-292-2123	

This document constitutes an agreement between the vendor and the buyer. See terms and conditions of this purchase listed on the reverse side

Item No.	Specification	Unit	Quantity	Unit Price	Total Value
1	Mobile Display Capacitive Touch Screen-P5123Y1	1	1	5500	5500
2					
3					
4					
				Tax	50.65
				Total Value	5550.65

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB2020/062535

A. CLASSIFICATION OF SUBJECT MATTER
G06F16/00,G06K9/46,G06K9/72,G06F40/00 Version=2021.01

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F, G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Databases: TotalPatent One, IPO Internal Database
Keywords:extract, data, pattern, document, search

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 20190286898 A1 (FIRST AMERICAN FINANCIAL CORPORATION) 19 September 2019 (19/09/2019) (abstract, paragraphs [0005]-[0007], [0029]-[0037], [0043]-[0048], figures 2, 4 and 5	1-16

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21-04-2021	21-04-2021
Name and mailing address of the ISA/ Indian Patent Office Plot No.32, Sector 14,Dwarka,New Delhi-110075 Facsimile No.	Authorized officer Prashant Singh Telephone No. +91-1125300200

Form PCT/ISA/210 (second sheet) (July 2019)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2020/062535

Citation	Pub.Date	Family	Pub.Date
US 2019286898 A1	19-09-2019	AU 2015203150 A1	21-01-2016
		CA 2895917 A1	30-12-2015

Form PCT/ISA/210 (patent family annex) (July 2019)

フロントページの続き

(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(72)発明者　シン，　マドゥスダン
インド　５６０１００　バンガロール，　エレクトロニック　シティー－１，　ファースト　クロス，　アジメラ　ストーン　パーク，　ハウス　ナンバー　ビー－６０３
(72)発明者　ハルダー，　カウシク
インド　７０００１０　コルカタ，　ベレガタ，　１５２－ラジャ　ラジェンドラ　ラル　ミトラ　ロード，　アイディアル　アパートメント，　ブロック－エー，　フラット－４エー
(72)発明者　バナパリ　ベンカタ　ラメッシュ　ラユル，　ニルマル
インド　５３１０３５　ヴィシャーカパトナム，　サババラム　マンダル，　ビハインド　ガバメント　ジュニア　カレッジ，　ゴポルナガー，　ハウス　ナンバー　６－３７
(72)発明者　ゴーシュ　ダスティダル，　アリトラ
インド　７０００８２　コルカタ，ハリデブプール，１８１　バナマリ　バネルジー　ロード，フラット　ナンバー　０６
(72)発明者　シャ，　アジェイ
インド　７３６１８２，　アリプルドゥアル，　ジャイガオン　ディストリクト，　エヌ・エス　ロード，　ビサイド　ポスト　オフィス，　シャンブー　シャ内，　ハウス　ナンバー　２９８

Ｆターム(参考) 5B091 AA15