

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2025-507838
(P2025-507838A)

(43)公表日 令和7年3月21日(2025.3.21)

(51)国際特許分類
G 0 6 F 16/383 (2019.01)

F I
G 0 6 F 16/383

テーマコード (参考)
5 B 1 7 5

審査請求 未請求 予備審査請求 未請求 (全25頁)			
(21)出願番号	特願2024-551948(P2024-551948)	(71)出願人	519405569
(86)(22)出願日	令和5年2月27日(2023.2.27)		アルテリックス インコーポレイテッド
(85)翻訳文提出日	令和6年10月30日(2024.10.30)		アメリカ合衆国 9 2 6 1 8 カリフォル
(86)国際出願番号	PCT/US2023/013970		ニア州 アーバイン ラグーナ キャニオン
(87)国際公開番号	WO2023/167824		ロード 1 7 2 0 0
(87)国際公開日	令和5年9月7日(2023.9.7)	(74)代理人	110001243
(31)優先権主張番号	17/685,328		弁理士法人谷・阿部特許事務所
(32)優先日	令和4年3月2日(2022.3.2)	(72)発明者	ジャド ディノ ラード
(33)優先権主張国・地域又は機関			アメリカ合衆国 9 2 6 1 2 カリフォル
	米国(US)		ニア州 アーバイン ミッチェルソン ド
(81)指定国・地域	AP(BW,CV,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 最終頁に続く	(72)発明者	アダム ブレイク
			アメリカ合衆国 9 2 6 1 2 カリフォル
			ニア州 アーバイン ミッチェルソン ド
			最終頁に続く

(54)【発明の名称】 自動化されたキーとバリューペアの抽出

(57)【要約】
分析対象の文書および文書から抽出されるキー名のセットが受信される。文書内に含まれる文字列のセットおよび各文字列の位置が識別される。さらに、文書の文書グラフが生成される。文書グラフには、ノードのセットおよびエッジのセットが含まれる。ノードのセットの各ノードは、文字列のセットの文字列に対応する。エッジのセットの各エッジは、2つ以上のノードを相互に接続する。さらに、文書グラフおよび受信したキー名のセットに基づいて、キーのセットが識別される。さらに、文書からバリューのセットが抽出され、識別されたキーのセットおよび抽出されたバリューのセットに基づいて、キーとバリューのペアのセットが生成される。

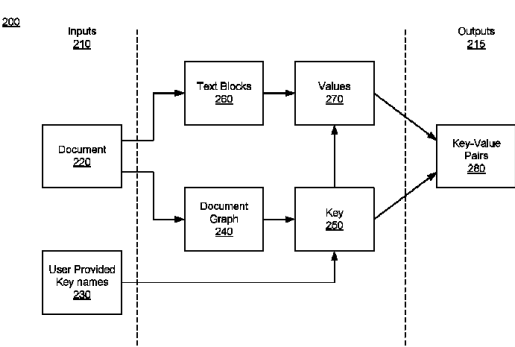


FIG. 2

【特許請求の範囲】**【請求項 1】**

解析対象の文書を受信するステップと、
前記文書から抽出されるキー名のセットを受信するステップと、
前記文書内に含まれる複数の文字列、および前記複数の文字列の各文字列の位置を識別するステップと、
文書グラフを生成するステップであって、前記文書グラフは、
ノードのセットであって、前記ノードの前記セットの各ノードが前記複数の文字列のうちの文字列に対応する、ノードのセット、および
エッジのセットであって、前記エッジのセットの各エッジが2つ以上のノードを互いに接続する、エッジのセット
を含む、ステップと、
前記文書グラフおよび前記受信されたキー名のセットに基づいてキーのセットを識別するステップと、
前記文書からバリューのセットを抽出するステップと、
前記識別されたキーのセットおよび前記抽出されたバリューのセットに基づいてキーとバリューペアのセットを生成するステップと
を含むコンピュータにより実施される方法。

【請求項 2】

前記複数の文字列を識別するステップは、前記受信された文書に対して光学文字認識（OCR）を実施するステップを含む、請求項 1 に記載のコンピュータにより実施される方法。

【請求項 3】

前記文書グラフを生成するステップは、
前記ノードのセットのうちの第2のノードが前記ノードのセットのうちの第1のノードからしきい値距離内にあると決定したことに応答して、前記第1のノードを前記第2のノードに前記エッジのセットのうちのエッジによって接続するステップ
を含む請求項 1 に記載のコンピュータにより実施される方法。

【請求項 4】

前記第1のノードは、前記第2のノードが前記第1のノードからしきい値角度内にあると決定したことにさらに応答して、前記第2のノードに接続される請求項 3 に記載のコンピュータにより実施される方法。

【請求項 5】

前記文書グラフおよび前記受信されたキー名のセットに基づいて前記キーのセットを識別するステップは、

キー名の第1のトークンに基づいて前記文書グラフの前記ノードのセットから第1のノードを識別するステップであって、前記第1のノードに対応する文字列は、前記キー名の前記第1のトークンと一致する、ステップと、

前記キー名の第2のトークンに基づいて前記文書グラフの前記ノードのセットから第2のノードを識別するステップであって、前記第2のノードは、前記文書グラフの前記エッジのセットのうちのエッジによって前記第1のノードに接続され、前記第2のノードに対応する文字列は、前記キー名の第2のトークンと一致する、ステップと
を含む請求項 1 に記載のコンピュータにより実施される方法。

【請求項 6】

前記文書からバリューのセットを抽出するステップは、
前記文書から線除去して、線除去された文書を生成するステップと、
前記線除去された文書にランレングス平滑化アルゴリズム（RLSA）を適用して、前記線除去された文書中の前景画素のセットおよび背景画素のセットを識別するステップと、
前記線除去された文書中の前記前景画素のセットに基づいてバウンディングボックスの

セットを生成するステップと、

前記生成されたバウンディングボックスのセットに基づいて前記バリューのセットを抽出するステップと

を含む請求項 1 に記載のコンピュータにより実施される方法。

【請求項 7】

前記生成されたバウンディングボックスのセットに基づいて前記バリューのセットを抽出するステップは、

前記バウンディングボックスのセットの各バウンディングボックスについて、前記バウンディングボックス内に位置する前記キーのセットのうちのキーの数を決定するステップと、

前記キーの数が 1 よりも大きいことに応答して、前記バウンディングボックスを分割して分割バウンディングボックスのセットを生成するステップであって、各分割バウンディングボックスは、前記分割バウンディングボックス内に位置する前記キーのセットのうちの 1 つのキーを有する、ステップと

を含む請求項 6 に記載のコンピュータにより実施される方法。

【請求項 8】

実行可能コンピュータプログラム命令を記憶する非一時的コンピュータ可読メモリであって、前記命令は、

解析対象の文書を受信することと、

前記文書から抽出されるキー名のセットを受信することと、

前記文書内に含まれる複数の文字列、および前記複数の文字列の各文字列の位置を識別することと、

文書グラフを生成することであって、前記文書グラフは、

ノードのセットであって、前記ノードのセットの各ノードが前記複数の文字列のうちの文字列に対応する、ノードのセット、および

エッジのセットであって、前記エッジのセットの各エッジが 2 つ以上のノードを互いに接続するエッジのセット

を含む、生成することと、

前記文書グラフおよび前記受信されたキー名のセットに基づいてキーのセットを識別することと、

前記文書からバリューのセットを抽出することと、

前記識別されたキーのセットおよび前記抽出されたバリューのセットに基づいてキーとバリューペアのセットを生成することと

を含む動作を実施するために実行可能である非一時的コンピュータ可読メモリ。

【請求項 9】

前記複数の文字列を識別することは、前記受信された文書に対して光学文字認識（OCR）を実施することを含む請求項 8 に記載の非一時的コンピュータ可読メモリ。

【請求項 10】

前記文書グラフを生成することは、

前記ノードのセットのうちの第 2 のノードが前記ノードのセットのうちの第 1 のノードからしきい値距離内にあると決定したことに応答して前記第 1 のノードを前記第 2 のノードに前記エッジのセットのうちのエッジによって接続すること

を含む請求項 8 に記載の非一時的コンピュータ可読メモリ。

【請求項 11】

前記第 1 のノードは、前記第 2 のノードが前記第 1 のノードからしきい値角度内にあると決定したことにさらに応答して前記第 2 のノードに接続される請求項 10 に記載の非一時的コンピュータ可読メモリ。

【請求項 12】

前記文書グラフおよび前記受信されたキー名のセットに基づいて前記キーのセットを識別することは、

10

20

30

40

50

キー名の第1のトークンに基づいて前記文書グラフの前記ノードのセットから第1のノードを識別することであって、前記第1のノードに対応する文字列は、前記キー名の前記第1のトークンと一致する、識別することと、

前記キー名の第2のトークンに基づいて前記文書グラフの前記ノードのセットから第2のノードを識別することであって、前記第2のノードは、前記文書グラフの前記エッジのセットのうちのエッジによって前記第1のノードに接続され、前記第2のノードに対応する文字列は、前記キー名の前記第2のトークンと一致する、識別することとを含む請求項8に記載の非一時的コンピュータ可読メモリ。

【請求項13】

前記文書からバリューのセットを抽出することは、

前記文書から線除去して、線除去された文書を生成することと、

前記線除去された文書にランレングス平滑化アルゴリズム（RLSA）を適用して、前記線除去された文書中の前景画素のセットおよび背景画素のセットを識別することと、

前記線除去された文書中の前記前景画素のセットに基づいてバウンディングボックスのセットを生成することと、

前記生成されたバウンディングボックスのセットに基づいて前記バリューのセットを抽出することと

を含む請求項8に記載の非一時的コンピュータ可読メモリ。

【請求項14】

前記生成されたバウンディングボックスのセットに基づいて前記バリューのセットを抽出することは、

前記バウンディングボックスのセットの各バウンディングボックスについて、前記バウンディングボックス内に位置する前記キーのセットのうちのキーの数を決定することと、

前記キーの数が1よりも大きいことに応答して、前記バウンディングボックスを分割して分割バウンディングボックスのセットを生成することであって、各分割バウンディングボックスは、前記分割バウンディングボックス内に位置する前記キーのセットのうちの1つのキーを有する、生成することと

を含む請求項13に記載の非一時的コンピュータ可読メモリ。

【請求項15】

コンピュータプログラム命令を実行するためのコンピュータプロセッサと、

動作を実行するための、前記コンピュータプロセッサによって実行可能なコンピュータプログラム命令を記憶する非一時的コンピュータ可読メモリであって、前記動作が、

解析対象の文書を受信することと、

前記文書から抽出されるキー名のセットを受信することと、

前記文書内に含まれる複数の文字列、および前記複数の文字列の各文字列の位置を識別することと、

文書グラフを生成することであって、前記文書グラフは、

ノードのセットであって、前記ノードのセットの各ノードが前記複数の文字列のうちの文字列に対応する、ノードのセット、および

エッジのセットであって、前記エッジのセットの各エッジが2つ以上のノードを互いに接続する、エッジのセット

を含む、生成することと、

前記文書グラフおよび前記受信されたキー名のセットに基づいてキーのセットを識別することと、

前記文書からバリューのセットを抽出することと、

前記識別されたキーのセットおよび前記抽出されたバリューのセットに基づいてキーとバリューペアのセットを生成することと

を含む、非一時的コンピュータ可読メモリと

を含むコンピュータシステム。

【請求項16】

10

20

30

40

50

前記複数の文字列を識別することは、前記受信された文書に対して光学文字認識（OCR）を実施することを含む請求項 15 に記載のコンピュータシステム。

【請求項 17】

前記文書グラフを生成することは、

前記ノードのセットのうちの第 2 のノードが前記ノードのセットのうちの第 1 のノードからしきい値距離内にあり、前記第 2 のノードが前記第 1 のノードからしきい値角度内にあると決定したことに応答して前記第 1 のノードを前記第 2 のノードに前記エッジのセットのうちのエッジによって接続すること

を含む請求項 15 に記載のコンピュータシステム。

【請求項 18】

前記文書グラフおよび前記受信されたキー名のセットに基づいて前記キーのセットを識別することは、

キー名の第 1 のトークンに基づいて前記文書グラフの前記ノードのセットから第 1 のノードを識別することであって、前記第 1 のノードに対応する文字列は、前記キー名の前記第 1 のトークンと一致する、識別することと、

前記キー名の第 2 のトークンに基づいて前記文書グラフの前記ノードのセットから第 2 のノードを識別することであって、前記第 2 のノードは、前記文書グラフの前記エッジのセットのうちのエッジによって前記第 1 のノードに接続され、前記第 2 のノードに対応する文字列は、前記キー名の前記第 2 のトークンと一致する、識別することと

を含む請求項 15 に記載のコンピュータシステム。

【請求項 19】

前記文書からバリューのセットを抽出することは、

前記文書から線除去して、線除去された文書を生成することと、

前記線除去された文書にランレングス平滑化アルゴリズム（RLSA）を適用して、前記線除去された文書中の前景画素のセットおよび背景画素のセットを識別することと、

前記線除去された文書中の前記前景画素のセットに基づいてバウンディングボックスのセットを生成することと、

前記生成されたバウンディングボックスのセットに基づいて前記バリューのセットを抽出することと

を含む請求項 15 に記載のコンピュータシステム。

【請求項 20】

前記生成されたバウンディングボックスのセットに基づいて前記バリューのセットを抽出することは、

前記バウンディングボックスのセットの各バウンディングボックスについて、前記バウンディングボックス内に位置する前記キーのセットのうちのキーの数を決定することと、

前記キーの数が 1 よりも大きいことに応答して、前記バウンディングボックスを分割して分割バウンディングボックスのセットを生成することであって、各分割バウンディングボックスは、前記分割バウンディングボックス内に位置する前記キーのセットのうちの 1 つのキーを有する、生成することと

を含む請求項 19 に記載のコンピュータシステム。

【発明の詳細な説明】

【技術分野】

【0001】

記載される実施形態は、一般にデータ解析に関し、さらに詳細には、キーとバリューペアを文書から抽出することに関する。

【背景技術】

【0002】

ビッグデータ解析などのデータ解析プラットフォームの成長は、データ処理を、大量のデータを処理してビジネス的な価値のある情報を抽出するためのツールに拡大させた。例えば、小規模な事業では、専用のコンピューティングリソースおよび人的リソースを利用

10

20

30

40

50

して外部データプロバイダや内部データソース（例えばローカルコンピュータ上のファイルなど）などの様々なソースから膨大な量のデータを集約、処理、および解析する、第三者のデータ解析環境を利用することがある。データ解析で用いられるこのような大きなデータセットを有用な定量的および定性的情報を抽出するように処理するには、通常は、強力なコンピュータデバイスに実装された複雑なソフトウェアツールが必要となる。

【0003】

このようなデータ解析プラットフォームで使用されるデータは、文書のセットに由来することがある。多くの文書は、キーとバリューの間の関連付け（例えば納税申告用紙の「名前」欄など）を含む。これらのキーとバリューペアは、コンピュータが抽出するのがかなり困難である可能性がある。例えば、基本的なOCR抽出は、複雑で面倒な後処理およびテキストマイニングにつながる可能性がある。したがって、ユーザが抽出したいキーの名称のみを提供することによって文書からキーとバリューペアを抽出することができるツールが必要とされている。

【発明の概要】

【0004】

上記その他の問題は、文書からキーとバリューペアを抽出するコンピュータにより実施される方法、コンピュータシステム、およびコンピュータ可読記憶媒体によって対処される。コンピュータにより実施される方法の実施形態は、解析対象の文書、およびその文書から抽出される第1のキーのセットを受信することを含む。この方法は、文書内に含まれる文字列のセット、および各文字列の位置を識別することをさらに含む。この方法は、文書グラフを生成することをさらに含む。文書グラフは、ノードのセットおよびエッジのセットを含む。ノードのセットの各ノードは、文字列のセットのうちの文字列に対応する。エッジのセットの各エッジは、2つ以上のノードを互いに接続する。この方法は、文書グラフおよび受信された第1のキーのセットに基づいてキーの第2のキーのセットを識別することをさらに含む。さらに、この方法は、文書からバリューのセットを抽出することを含む。最後に、この方法は、識別された第2のキーのセットおよび抽出されたバリューのセットに基づいてキーとバリューペアのセットを生成することを含む。

【0005】

コンピュータシステムの実施形態は、コンピュータプログラム命令を実行するためのコンピュータプロセッサを含む。このシステムは、また、実行されたときにコンピュータプロセッサに動作を実施させるコンピュータプログラム命令を記憶する非一時的コンピュータ可読メモリも含む。動作は、解析対象の文書、および文書から抽出される第1のキーのセットを受信することを含む。動作は、文書内に含まれる文字列のセット、および各文字列の位置を識別することをさらに含む。動作は、文書グラフを生成することをさらに含む。文書グラフは、ノードのセットおよびエッジのセットを含む。ノードのセットの各ノードは、文字列のセットのうちの文字列に対応する。エッジのセットの各エッジは、2つ以上のノードを互いに接続する。動作は、文書グラフおよび受信された第1のキーのセットに基づいて第2のキーのセットを識別することをさらに含む。さらに、動作は、文書からバリューのセットを抽出することを含む。最後に、動作は、識別された第2のキーのセットおよび抽出されたバリューのセットに基づいてキーとバリューペアのセットを生成することを含む。

【0006】

非一時的コンピュータ可読メモリの実施形態は、実行可能コンピュータプログラム命令を記憶する。命令は、動作を実行するように実行可能である。動作は、解析対象の文書、および文書から抽出される第1のキーのセットを受信することを含む。動作は、文書内に含まれる文字列のセット、および各文字列の位置を識別することをさらに含む。動作は、文書グラフを生成することをさらに含む。文書グラフは、ノードのセットおよびエッジのセットを含む。ノードのセットの各ノードは、文字列のセットのうちの文字列に対応する。エッジのセットの各エッジは、2つ以上のノードを互いに接続する。動作は、文書グラフおよび受信された第1のキーのセットに基づいて第2のキーのセットを識別することを

10

20

30

40

50

さらに含む。さらに、動作は、文書からバリューのセットを抽出することを含む。最後に、動作は、識別された第2のキーのセットおよび抽出されたバリューのセットに基づいてキーとバリューペアのセットを生成することを含む。

【図面の簡単な説明】

【0007】

【図1】1つまたは複数の実施形態によるデータ解析システムを含むデータ解析環境を示すブロック図である。

【図2】1つまたは複数の実施形態による文書からキーとバリューペアを抽出するためのワークフローを示すブロック図である。

【図3】1つまたは複数の実施形態によるデータ解析アプリケーションを示すブロック図である。

【図4A】1つまたは複数の実施形態によるOCRモジュールの出力を示す文書のマークアップ画像を示す図である。

【図4B】1つまたは複数の実施形態による文書グラフの表現を示す文書のマークアップ画像を示す図である。

【図4C】1つまたは複数の実施形態によるキー識別モジュールの出力を示す文書のマークアップ画像を示す図である。

【図4D】1つまたは複数の実施形態による、カーネルを文書の画像に適用することによって生成される水平方向および垂直方向の線に対応する画像を示す図である。

【図4E】1つまたは複数の実施形態による、文書の画像から文書中の水平方向および垂直方向の線に対応する画像を引くことによって生成される線除去された文書を示す図である。

【図4F】1つまたは複数の実施形態によるバリュー識別モジュールによって生成されるバウンディングボックスを示す文書のマークアップ画像を示す図である。

【図4G】1つまたは複数の実施形態による識別されたキーおよびバリューを有するバウンディングボックスを示す文書のマークアップ画像を示す図である。

【図5】1つまたは複数の実施形態による解析ワークフローを用いてキーとバリューペアを抽出するプロセスを示す流れ図である。

【図6】一実施形態による図1のデータ解析システムとして使用される代表的なコンピュータシステムの機能図を示すハイレベルブロック図である。

【0008】

図面は、説明のみを目的として様々な実施形態を示している。当業者なら、本明細書に記載される実施形態の原理を逸脱することなく本明細書に示される構造および方法の代替の実施形態が利用され得ることを以下の論考から容易に理解するであろう。様々な図面中の同じ参照番号および名称は、同じ要素を示している。

【発明を実施するための形態】

【0009】

図1は、一実施形態によるデータ解析システム110を含むデータ解析環境100を示すブロック図である。環境100は、ネットワーク130を介してデータ解析システム110に接続された複数のデータソース120をさらに含む。図示の環境100は、複数のデータソース120に結合された1つのデータ解析システム110しか含んでいないが、実施形態は、複数のデータ解析システムと1つのデータソースとを有することもできる。さらに、データソース120はネットワークを介してデータ解析システム110に接続されるものとして示されているが、1つまたは複数のデータソースがデータ解析システムにローカルに記憶されていてもよい。あるいは、ユーザは、直接（ネットワーク130を介して、またはデータ解析システムのローカルインタフェースを介して）データをデータ解析システムに提供してもよい。

【0010】

データ解析システム110は、大量のデータを処理するために利用されるコンピュータに基づくシステムである。いくつかの実施形態では、データは、ネットワーク130を介

10

20

30

40

50

して複数のデータソース120から収集、集約、またはその他の形でアクセスされる。データ解析システム110は、幅広い様々なデータソースからのデータ処理タスクに利用されるスケーラブルなソフトウェアツールおよびハードウェアリソースを実装することができる。データ処理タスク（タスクまたはデータ解析タスクとも呼ばれる）は、データを処理するための産出動作である。データ処理タスクは、例えば、データへのアクセス、データの準備、データの混合、データのパッケージ化、データの解析、データに対するその他のタイプの動作、またはそれらの何らかの組合せを含む。データ解析システム110は、複数のタスクを並列に実施することができる。データ解析システム110は、それが利用可能なコンピュータリソース（例えばメモリ）を有するときには、実施される準備ができていて複数のタスクから実施するタスクを選ぶことができる。

10

【0011】

図1の実施形態では、データ解析システム110は、メモリ140と、データ解析アプリケーション150とを含む。メモリ140は、データ解析動作で利用されるデータを記憶し、データ記憶領域へのアクセスを提供する。いくつかの実施形態では、メモリ140は、ランダムアクセスメモリ（RAM）、ハードディスクドライブ（HDD）、ソリッドステートドライブ（SSD）、他の形態のコンピュータメモリ、またはそれらの何らかの組合せを含む。いくつかの実施形態では、メモリ140は、比較的小さなデータ記憶領域への比較的高速のアクセスを提供する1次メモリ140と、比較的大きなデータ記憶領域への比較的低速のアクセスを提供する2次メモリ150とを含む。

20

【0012】

データ解析アプリケーション150は、図2に示すワークフローなどの解析ワークフロー（「ワークフロー」とも呼ばれる）を用いてデータ解析タスクの実行をサポートするソフトウェアアプリケーションである。ワークフローは、データ解析ツール（「ツール」とも呼ばれる）またはデータ解析モジュール（「モジュール」とも呼ばれる）のシーケンスであり、これを通して、データ解析タスクは開始（例えばデータファイルの入力）から完了（例えばタスクの結果の出力）まで進行する。シーケンス中の各ツールは、入力、出力、準備、結合、予測的、空間的、調査、構文解析、および変換など、特定の処理動作またはデータ解析機能を実施する。ワークフローの実行は、ワークフロー中のツールの動作を含み、タスクの結果を表す結果アイテムを生成する。結果アイテムは、一例では、タスクの結果の視覚表現を提供するチャートである。

30

【0013】

データ解析アプリケーション150は、ワークフローの作成を容易にする環境を提供する。例えば、データ解析アプリケーション150は、データ解析アプリケーション150のGUIによってサポートされる視覚ワークフロー環境を提供する。視覚ワークフロー環境は、ワークフローを設計するためにソフトウェアコードを提供する必要をなくし、ワークフローによって実装される複雑な数式を識別する必要をなくすドラッグアンドドロップツールのセットを使用可能にする。別の実施形態では、ワークフロー200は、拡張可能マークアップ言語（XML）文書などの文書に作成され、記述される。

【0014】

この環境は、図2に示されるワークフローなどの論理制御によるワークフローの作成も容易にすることがある。論理制御によるワークフローは、1つまたは複数の論理制御装置と一体化されたツールのシーケンスを含むワークフローである。論理制御装置は、論理機能を有し、入力に基づいてツールまたはモジュールをトリガするかどうか（および／またはしないかどうか）を決定するように適応される。データ解析アプリケーション150は、このツールのシーケンスを使用して、データファイルを処理し、論理制御装置を使用して、データファイルの処理中にシーケンス中のツールの制御動作を制御する。いくつかの実施形態では、論理制御装置は、ワークフロー中の少なくとも2つのツール、すなわち第1のツールおよびシーケンス中で第1のツールの後にある第2のツールと関連付けられる。データ解析アプリケーション150は、第1のツールにロギング機能を適用して、データファイルの処理において第1のツールのデータ解析機能の実施中に起きる事象をロギン

40

50

グする。第1のツールによって生成されるログデータは、データファイルの処理中に第1のツールによって実施されるアクティビティおよび／または第1のツールによって検出される事象を記述する。事象は、例えば、ツールのデータ解析機能の実施中に検出されるエラー、ツールのデータ解析機能の実施中に生成される警告、ツールによって処理されるレコードの数、メタデータファイルの出力、およびツールのデータ解析機能の実施の完了などであり得る。第1のツールは、解析データも例えばメタデータファイルなどで出力する。解析データは、データファイルの処理においてデータ解析機能の実施により第1のツールによって生成される。

【0015】

データ解析アプリケーション150は、第1のツールによって生成されるログデータおよび／または解析データを論理制御装置への入力として使用する。例えば、論理制御装置は、ログデータを解析して、ログデータに記述されている事象を検出し、検出された事象に基づいて第2のツールをトリガするかどうかを決定することができる。いくつかの実施形態では、論理制御装置は、論理機能を有するデータ解析ツールである。論理制御装置のツールは、ログに対してデータ解析機能を実施して、第1のツールによってロギングされた事象を検出する。この論理機能は、論理制御装置のツールの出力に基づいてコマンドを出力する。第2のツールは、このコマンドを受信して、このコマンドに従ってデータを処理するように適応される。

【0016】

例えば、第1のツールのデータ解析機能がうまく完了したことを検出したことに応答して、論理制御装置は、第2のツールをトリガして動作させる。いくつかの実施形態では、論理制御装置は、第1のツールからメタデータファイルを受信し、第2のツールをトリガするという決定に応答してこのメタデータファイルを第2のツールに送信する。第2のツールは、論理制御装置からメタデータファイルを受信すると、このメタデータファイルを処理する。論理制御装置は、第1のツールによって生成されるログに基づいてワークフロー中の他のツールを選択的に実行することもある。第2のツールに続くツールがある場合には、そのツールも動作してそれらのデータ解析機能を実施し、データファイルの処理を終了する。ツールのシーケンスによるデータファイルの処理の結果を表す結果アイテムが生成される。データ解析アプリケーション150は、この結果アイテムをユーザに対する表示のために提供する。

【0017】

別の例では、論理制御装置は、第1のツールの動作中に生じたエラーを検出したことに応答して第2のツールの動作をトリガしない。いくつかの実施形態では、第2のツールの動作をトリガしないと決定したことに応答して、論理制御装置は、第3のツールの動作をトリガする。第3のツールは、シーケンス中になくてもよい。第3のツールは、補助結果アイテムを生成するように適応される。一例では、補助結果アイテムは、論理制御装置が第2のツールのデータ解析機能をトリガしないと決定した根拠になった、第1のツールの動作中に生じた事象を指定するメッセージである。第3のツールは、論理制御装置からのその事象の情報に基づいてこのメッセージを生成し、ユーザに関連付けられたクライアントデバイスにこのメッセージを送信することができる。いくつかの実施形態では、データ解析アプリケーション150は、この事象に対処するアクションをユーザから受け取る。データ解析アプリケーション150は、ユーザのアクションに基づいてイベントを解決することができる。ユーザのアクションを解決したことに応答して、データ解析アプリケーション150は、第2のツールをトリガしてデータファイルの処理を終了するように論理制御装置に指示することができる。

【0018】

これらの例は、説明のために提供されるものである。論理制御装置は、他のタイプのロギングされた事象を検出することもできる。また、論理制御装置は、他のツールの実行を制御することもできる。例えば、論理制御装置は、第2のツールをトリガすることを決定することに加えて、1つまたは複数の他のツールをトリガする（またはしない）と決定す

10

20

30

40

50

ることもある。この1つまたは複数の他のツールは、例えばそれらがツールのシーケンス中に配置されているのと同じ順序で第2のツールに接続されることもある。同様に、論理制御装置は、第2のツールをトリガしないと決定することに加えて、1つまたは複数の他のツールをトリガする（またはしない）と決定することもある。さらに、論理制御装置は、第1のツールに加えて1つまたは複数の他のツールによって生成されるログを入力として受信して、その入力を使用することもできる。

【0019】

データ解析アプリケーション150によって生成される論理制御によるワークフローは、それによりユーザがデータ解析プロセスを監視および制御することができるので有利である。このような論理制御によるワークフローを使用することにより、否定的な事象（すなわち論理制御装置が第2のツールをトリガしないと決定する根拠になる事象）の後、第1のツールでデータ解析プロセスを一時停止させることができる。否定的な事象は、第1のツールによって生成されるデータを、更なる処理に不適切なものにする可能性がある。例えば、第1のツールは、エラーが発生したときにデータファイルの処理を停止して、第1のツールによって生成されるデータが不完全になるようにすることができる。したがって、データ解析プロセスの一時停止により、他のツールが不適切なデータを処理するためにコンピューティングリソースを浪費することが回避される。また、ユーザは、第3のツールをワークフローに追加して、彼らが補助結果アイテムを通して否定的な事象を通知されることが可能になるようにすることができる。このようにして、ユーザは、否定的な事象を解決し、データ解析プロセスを再開する機会を得る。従来のデータ解析技術と比較すると、この論理制御によるワークフローでは、データ解析プロセスにユーザが寄与することが容易になり、不適切なデータの処理に浪費されていたはずのコンピューティングリソースを節約する。

【0020】

いくつかの場合には、データ解析アプリケーション150は、データ解析ツールおよびマクロへのネットワークによる、またはクラウドベースのアクセスをサポートするソフトウェアを、複数のエンドユーザに提供する。例えば、データ解析アプリケーション150は、ダッシュボード、ウェブアプリケーション、モバイルアプリケーションなどの方式でワークフローの作成をサポートする。別の例では、データ解析アプリケーション150は、モバイルアプリケーションストアまたはその他のクラウドベースのサービスと同様にユーザが解析を共有し、閲覧し、消費することを可能にする。解析データ、マクロ、およびワークフローは、例えばデータ解析システム110の他のユーザからアクセスされることが可能な、より小規模でカスタマイズ可能な解析アプリケーション（すなわちアプリ）としてパッケージ化されて実行されることが可能である。いくつかの場合には、公開された解析アプリへのアクセスがデータ解析システム110によって管理される、すなわちアクセスが認可または取り消され、それによりアクセス制御およびセキュリティ機能が提供されることが可能である。データ解析アプリケーション150は、作成、デプロイ、公開、反復、および更新など、解析アプリに関連付けられた機能を実施することができる。

【0021】

さらに、データ解析アプリケーション150は、解析結果のアクセス、準備、混合、解析、および出力の能力など、データ解析に関連する様々なステージで実施される機能をサポートすることができる。いくつかの場合には、データ解析アプリケーション150は、様々なデータソースにアクセスして、例えばデータのストリーム中の生データを取り出すことができる。データ解析アプリケーション150によって収集されるデータストリームは、複数のレコードの生データを含むことができ、ここで、この生データは、様々なフォーマットおよび構造である。少なくとも1つのデータストリームを受信した後で、データ解析アプリケーション150は、データストリームのレコードがデータ解析動作への入力として使用されることを可能にする動作を実施することができる。さらに、予測解析（例えば予測モデリング、クラスタリング、データ調査など）など、レコードの統計的、定性的、または定量的な処理に関連する解析機能が、データ解析アプリケーション150によ

って実施されることが可能である。

【0022】

上述のように、データソース120は、データ解析システム110に電子データを提供する。データソース120は、コンピュータ、データベース、ネットワークAPI（アプリケーションプログラミングインタフェース）、またはクラウドストレージシステムであり得る。データソース120は、別のソースからデータを取り出すことができるコンピュータシステムであることもある。データソース120は、データ解析システム110から遠隔にあって、ネットワーク130を介してデータを提供することもある。さらに、一部または全てのデータソース120は、データ解析システムに直接結合され、データをネットワーク130に通すことなく提供することもある。データソース120によって提供されるデータは、通常はデータレコードに編成され、各データレコードが1つまたは複数の値を含む。例えば、データソースによって提供されるデータレコードは、カンマ区切り値の連続を含むことがある。このデータは、データ解析システムを用いる企業との関連性の情報を記述する。例えば、データソース120のデータは、ウェブサイト上でアクセス可能なコンテンツおよび／またはソーシャルメディアアプリケーションとのコンピュータに基づく対話（例えばクリックトラッキングデータ）を記述することができる。

10

【0023】

ネットワーク130は、データ解析システム110とデータソース120の間の通信経路を表す。一実施形態では、ネットワーク130は、インターネットであり、標準的な通信技術および／またはプロトコルを使用する。ネットワーク130上で交換されるデータは、ハイパーテキストマークアップ言語（HTML）、拡張可能マークアップ言語（XML）などの技術および／またはフォーマットを用いて表されることが可能である。別の実施形態では、これらのエンティティは、上述のものの代わりに、または上述のものに加えて、カスタムおよび／または専用のデータ通信技術を使用することもできる。

20

【0024】

図2は、一実施形態によるデータを処理するためにデータ解析システム110を使用して実施されるワークフロー200を示すブロック図である。ワークフロー200は、データ解析システム110のコンピュータデバイスによって実行される。ただし、他の実施形態では、ワークフロー200は、ネットワーク（例えばネットワーク130）を介してデータ解析システム110に通信可能に接続され得る別のコンピュータデバイスにデプロイされる。

30

【0025】

図示のように、図2のワークフロー200は、入力210のセットを受信することと、出力215のセットを生成および／または提示することを含む。いくつかの実施形態では、入力のセットは、入力ツールのセットを用いて（例えばデータ解析システム110のユーザから、データ解析システム110と対話する別のプロセスによって、または記憶デバイスから）受信される。さらに、出力のセットは、ブラウジングツールを介してデータ解析システム110のユーザに対して提示されることもある。

【0026】

データ解析ツール110は、文書220およびユーザから提供されるキー名230のセットを入力として受信する。いくつかの実施形態では、文書は、指定された文書フォーマット（例えばポータブルドキュメントフォーマット（PDF））で受信される。データ解析ツール110は、次いで、ユーザから提供されるキー名230のセットに基づいて文書220からキーとバリューペア280のセットを抽出する。

40

【0027】

いくつかの実施形態では、キーとバリューペア280のセットを抽出するために、データ解析システム110は、受信される文書220を解析することによって文書グラフ240を生成する。文書グラフ240は、ノードのセットおよびエッジのセットを含むことがある。ノードのセットの各ノードは、文書220中で識別される文字列に対応することがある。いくつかの実施形態では、各ノードは、それが対応する文字列と、その文字列が文

50

書 220 内で位置する関連付けられた位置とを記憶する。さらに、エッジのセットの各エッジは、2 つ以上のノードを互いに接続する。

【0028】

さらに、生成された文書グラフ 240 およびユーザから提供されるキー名 230 のセットに基づいて、データ解析システム 110 は、文書内のキー 250 のセットを識別する。いくつかの実施形態では、データ解析システム 110 は、文書グラフ 240 を探索して、文書グラフ 240 内でユーザから提供されるキー名 230 の各々を特定する。データ解析ツールは、さらに、文書グラフ 240 から文書内の各キー 250 の位置を抽出することがある。

【0029】

データ解析システム 110 は、文書 220 内のテキストブロック 260 のセットを識別し、識別されたテキストブロック 260 のセットと識別されたキー 250 のセットとに基づいて文書からバリュー 270 のセットを抽出する。データ解析システム 110 は、次いで、識別されたキー 250 のセットのうちのキーと識別されたバリュー 270 のセットのうちの対応するバリューとをマッチングして、キーとバリューペア 280 のセットを生成する。

【0030】

図 3 は、1 つまたは複数の実施形態によるデータ解析アプリケーション 300 を示すブロック図である。データ解析アプリケーション 300 は、図 1 のデータ解析アプリケーション 150 の実施形態である。図 3 の実施形態では、データ解析アプリケーション 300 は、データ受信機モジュール 310 と、光学文字認識（OCR）モジュール 320 と、グラフ生成モジュール 330 と、キー識別モジュール 340 と、バリュー識別モジュール 350 と、ユーザインタフェースモジュール 360 とを含む。当業者なら、他の実施形態ではここに述べるものとは異なる、および／またはここに述べるもの以外の構成要素を有することができること、ならびに機能がそれらの構成要素間で異なる様式で分散されることが可能であることを理解するであろう。例えば、データ解析アプリケーション 300 の一部または全ての機能が、クライアントデバイスまたは第三者サーバで実施されてもよい。

【0031】

データ受信機モジュール 310 は、例えばデータソース 120 などのデータソースからデータファイルを受信し、そのデータファイルをデータ解析アプリケーション 300 内の他のモジュールに提供する。一実施形態では、データ受信機モジュール 310 は、データ解析アプリケーション内のモジュール間のインタフェースとして機能する。別の実施形態では、データ受信機モジュール 310 がなく、データ受信機モジュール 310 の機能は、データ解析アプリケーションの各モジュールによって実施される。

【0032】

いくつかの実施形態では、データ受信機モジュール 310 は、解析対象の文書 220 と、文書を解析するためのパラメータのセットとを受信するように構成される。文書を解析するためのパラメータのセットは、文書 220 からキーとバリューペア 280 を抽出するためのユーザから提供されるキー名 230 のセットを含むことがある。いくつかの実施形態では、データ受信機モジュール 310 は、既定のフォーマット（PDF フォーマットなど）を有する文書を受信するように構成される。データ受信機モジュール 310 は、文書 220 が正しいフォーマットを有するかどうかを識別するように構成されることがあり、文書が既定のフォーマットを有していない場合にはエラーを出力することがある。

【0033】

OCR モジュール 320 は、文書 220 を解析し、文書 220 に含まれる文字列を識別する。いくつかの実施形態では、OCR モジュール 320 は、さらに、識別された文字列の各々の位置または座標を決定する。文字列の位置または座標は、文書 220 内の文字列の周りのバウンディングボックスのセントロイドに対応することもある。いくつかの実施形態では、OCR モジュール 320 は、文書の画像（または文書のページに対応する 1 つもしくは複数の画像）を入力として受信する。OCR モジュール 320 は、入力画像内に

10

20

30

40

50

含まれる文字列を正しく検出する可能性を高めるために入力画像を前処理することもある。例えば、OCRモジュールは、画像を所定の解像度（例えば300ドットパーインチ（dpi））に変換することがあり、この画像をグレースケールに変換し、適応的ガウシアンしきい値処理、膨張、収縮、中央値ぼかし、および単純しきい値処理を実施する。

【0034】

図4Aは、1つまたは複数の実施形態によるOCRモジュールの出力を示す文書のマークアップ画像を示している。図4Aに示されるように、OCRモジュールは、文字列を含む文書の領域を識別し、文字列内の文字を認識する。図4Aの例では、各文字列は、枠で囲まれている。

【0035】

図3に戻って、グラフ生成モジュール330は、OCRモジュール320の出力から文書220の文書グラフ240を生成する。具体的には、OCRモジュール320の出力に基づいてグラフ生成モジュール330によって生成される文書グラフ240は、エッジのセットによって相互接続されたノードのセットを含む。いくつかの実施形態では、グラフ生成モジュール330は、OCRモジュール320によって文書から識別された各文字列についてノードを追加する。いくつかの実施形態では、各ノードは、それが対応する文字列と、文書内でその文字列が位置する関連付けられた位置とを記憶する。

【0036】

いくつかの実施形態では、文書グラフ240中の各ノードについて、グラフ生成モジュール330は、そのノードに接続される近隣ノードのセットを識別する。グラフ生成モデル330は、近隣ノードがそのノードからしきい値距離内にある場合に、そのノードを近隣ノードに接続するためのエッジを追加することもある。いくつかの実施形態では、モデル生成モデル330は、近隣ノードがそのノードからしきい値距離内にあり、ノードの位置と近隣ノードの位置とが互いにある角度範囲内にある場合に、そのノードを近隣ノードに接続するためにエッジを追加する。

【0037】

図4Bは、1つまたは複数の実施形態によるグラフ生成モジュールの出力を示す文書のマークアップ画像を示している。図4Bの例では、図4Aに示される文書の画像中のOCRモジュール320によって識別される各文字列が、文書グラフ中のノード421に割り当てられる。さらに、各ノード421は、エッジ423によって1つまたは複数のノードに接続される。例えば、文字列「Work」に対応するノード421Aは、エッジ423Aによって文字列「Request」421Bに対応するノード421Bに接続される。グラフ生成モジュール330は、ノード421Bがノード421Aからしきい値距離内にあり、ノード421Aからしきい値角度内にあると決定したことに応答して、ノード421Aをノード421Bに接続することがある。同様に、ノード421Aは、エッジ423Bによって、文字列「Requesting」に対応するノード421Dに接続される。ただし、ノード421Aは、文字列「Number EWR-001-1001」に対応するノード421Cには接続されない。図4Bの例では、ノード421Cがノード421Aからしきい値角度内にないので、ノード421Aはノード421Cに接続されない。

【0038】

図3に戻って、キー識別モジュール340は、文書220からキー250のセットを識別して、文書から抽出されるキーとバリューペア280を形成する。キー識別モジュール340は、ユーザから提供されるキー名230のセットを受信し、文書グラフ240をトラバースして、文書220内のユーザから提供されるキー名230を特定する。

【0039】

いくつかの実施形態では、ユーザから提供されるキー名ごとに、キー識別モジュール340は、ユーザから提供されるキー名が複数の単語を含むかどうかを決定する。ユーザから提供されるキー名が複数の単語を含む場合には、キー識別モジュールは、ユーザから提供されるキー名を単語のセット（またはトークンのセット）に分割する。キー識別モジュール340は、次いで、文書グラフ240内のユーザから提供されるキー名の第1の単語

10

20

30

40

50

を探索することによってユーザから提供されるキー名を探索し、文書グラフをトラバースしてユーザから提供されるキー名の後続の単語を識別することによって、ユーザから提供されるキー名の単語のセットに対応するノードの1つまたは複数のチェーンを識別することがある。すなわち、キー識別モジュール340は、OCRモジュール320によって識別される各ノード列をユーザから提供されるキー名の第1の単語と比較することによってユーザから提供されるキー名の第1の単語に一致するノードのセットを識別することがある。ユーザから提供されるキー名の第1の単語に一致する識別された各ノードについて、キー識別モジュール340は、識別されたノードにエッジを介して接続された第2のノードがユーザから提供されるキー名の第2の単語と一致するかどうかを決定する。識別されたノードがユーザから提供されるキー名の第2の単語と一致する第2のノードに接続されていない場合には、キー識別モジュール340は、識別されたノードがユーザから提供されるキー名に対応しないと決定する。あるいは、識別されたノードがユーザから提供されるキー名の第2の単語と一致する第2のノードを有する場合には、キー識別モジュール340は、その識別されたノードをユーザから提供されるキー名に対応する候補キーとして保持する。ユーザから提供されるキー名がさらなる単語を含む場合には、キー識別モジュール340は、ユーザから提供されるキー名の全ての単語が見つかるまで、文書グラフをトラバースし続ける。例えば、ユーザから提供されるキー名が第3の単語を含む場合に、キー識別モジュール340は、エッジを介して第2のノードに接続された第3のノードがユーザから提供されるキー名の第3の単語と一致するかどうかを決定する、などである。

10

20

【0040】

いくつかの実施形態では、ノードのチェーン（すなわちエッジのセットを介して接続されたノードのセット）が、ユーザから提供されるキー名と一致し、キー識別モジュール340は、このノードのチェーン中のノードを1つのノードにマージする。キー識別モジュール340は、さらに、マージされたノードの新たな位置を決定することもある。キー識別モジュール340は、次いで、ノードのチェーン中のノードをマージされたノードで置換することがある。

【0041】

いくつかの実施形態では、キー識別モジュール340は、ファジーマッチングを使用して、単語が文書グラフ中のノードと一致するかどうかを決定する。例えば、キー識別は、単語をノードに対応する文字列と比較することがあり、この比較に基づいてレーベンシュタイン比（LR）を決定することがある。LRがしきい値（例えば90%）を超える場合には、キー識別モジュール340は、その単語がノードに対応する文字列と一致すると決定する。あるいは、キー識別モジュール340は、考慮中の単語について最高のLR値を有する上位の既定数のノード（例えば上位10個のノード）を選択する。

30

【0042】

図4Cは、1つまたは複数の実施形態によるキー識別モジュールの出力を示す文書のマークアップ画像を示している。具体的には、図4Cは、キー識別モジュールによって識別されるキーのセットを示している。キー識別モジュールによって識別される各キーは、枠で囲まれている。

【0043】

40

図3に戻って、バリュー識別モジュール350は、文書220からバリュー270のセットを識別して、文書から抽出するキーとバリューペア280のセットを形成する。バリュー識別モジュール350は、文書220からテキストブロック260（すなわちテキストのブロック）を識別し、テキストブロック260をキー識別モジュール340によって識別されるキー250とマッチングする。バリュー識別モジュール350は、次いで、テキストブロック260からバリューを抽出する。

【0044】

いくつかの実施形態では、バリュー識別モジュール350は、文書から線除去（*deline*）する。例えば、バリュー識別モジュール350は、文書の画像から（または文書の各ページに対応する画像から）水平方向および垂直方向の線を除去する。文書から線除

50

去するために、文書の画像にカーネルが適用されて、水平方向および垂直方向の線に対応する画像を生成する。いくつかの実施形態では、カーネルは、文書のサイズによって決定される。さらに、いくつかの実施形態では、第1のカーネルが適用されて、水平方向の線に対応する第1の画像を生成し、第2のカーネルが適用されて、垂直方向の線に対応する第2の画像を生成する。水平方向および垂直方向の線に対応する生成された画像（または複数の画像）が文書の画像から引かれて、文書の線除去された画像を生成する。

【0045】

図4Dは、1つまたは複数の実施形態による、カーネルを文書の画像に適用することによって生成される水平方向および垂直方向の線に対応する画像を示している。図4Eは、1つまたは複数の実施形態による、文書の画像から文書中の水平方向および垂直方向の線に対応する画像を引くことによって生成される線除去された文書を示している。

10

【0046】

次いで、文書の線除去された画像が処理されて、バウンディングボックスのセットを生成する。いくつかの実施形態では、ランレングス平滑化アルゴリズム（RLSA）が線除去された画像に適用されて、バウンディングボックスを生成する。いくつかの実施形態では、RLSAは、前景画素（例えば非空白画素または黒画素）のセットおよび背景画素（例えば空白画素または白画素）のセットを識別する。RLSAは、次いで、背景画素の前景画素に対する近接度に基づいて、背景画素を前景画素に変換する。いくつかの実施形態では、RLSAは、背景画素が線除去された画像の水平方向軸上の両方向に前景画素から既定数の画素以内である場合に背景画素を前景画素に変換する。

20

【0047】

いくつかの実施形態では、RLSAは、しきい値限界値よりも小さい数の隣接する背景画素を有する背景画素を変換する。いくつかの実施形態では、RLSAは、文書の背景画素を0で表し、文書の前景画素を1で表す。次いで、文書の表現中の各0について、隣接する0の数がしきい値限界値未満である場合に、RLSAは0を1で置換する。例えば、しきい値限界値が3である場合には、10001110000001という表現を有する文書の一部は、11111110000001という表現に変換される。いくつかの実施形態では、RLSAは、水平方向に適用される。他の実施形態では、RLSAは、水平方向および垂直方向の両方に適用される。

【0048】

RLSAを実行した後で、バリュー識別モジュール350は、輪郭検出アルゴリズムを実行して、互いに関連付けられたテキストのブロックの周りにバウンディングボックスを生成する。輪郭検出アルゴリズムは、RLSAが適用された後で前景画素として分類された画素のバウンディングボックスを識別する。いくつかの実施形態では、バリュー識別モジュールは、バウンディングボックスに対して輪郭検出アルゴリズムを実行して、隣接するボックス同士をマージする。例えば、バリュー識別モジュール350は、バウンディングボックスによって囲まれる全ての画素を前景画素として変換することがあり、輪郭検出アルゴリズムを再適用して、第1の輪郭検出アルゴリズムが適用された後で前景画素として分類された画素のバウンディングボックスを検出する。

30

【0049】

図4Fは、1つまたは複数の実施形態によるバリュー識別モジュールによって生成されるバウンディングボックスを示す文書のマークアップ画像を示している。具体的には、図4Fは、RLSAによって識別される前景画素のセットと、各々が前景画素の1つのまわりを取り囲むバウンディングボックスのセットとを示している。

40

【0050】

バウンディングボックスを用いて、次いで、バリュー識別モジュール350は、キー識別モジュール340によって識別された各キーについての1つまたは複数のバリューを識別する。いくつかの実施形態では、キー識別モジュール340によって識別される各キーについて、バリュー識別モジュール350は、キーの位置に基づいてバウンディングボックスを識別し、識別されたバウンディングボックスによって囲まれるテキストを抽出する

50

ことによって、1つのバリューを識別する。

【0051】

いくつかの実施形態では、所与のキーについて、バリュー識別モジュール350は、そのキーを囲むバウンディングボックスを識別する。あるいは、バリュー識別モジュール350は、そのキーに最も近いバウンディングボックスを識別する。いくつかの実施形態では、バリュー識別モジュール350は、文書グラフをトラバースして、そのキーに関連付けるバウンディングボックスを識別する。いくつかの実施形態では、バリュー識別モジュール350は、任意の所与のバウンディングボックスが複数のキーに対して最も近いバウンディングボックスであるかどうかを決定する。バリュー識別モジュール350が、バウンディングボックスが複数のキーに対して最も近いバウンディングボックスであると決定した場合には、バリュー識別モジュール350は、そのバウンディングボックスを分割し、各キーをそのバウンディングボックスの一部と関連付ける。

10

【0052】

あるいは、いくつかの実施形態では、バリュー識別モジュール350は、各キーを処理してそのキーに関連付けられるバウンディングボックスを識別する代わりに、各バウンディングボックスを処理して、そのバウンディングボックスを1つまたは複数のキーと関連付けるかどうかを識別する。例えば、各バウンディングボックスについて、バリュー識別モジュール350は、そのバウンディングボックスによって囲まれるキーを識別することがある。この実施形態では、バリュー識別モジュール350は、3つの異なるシナリオを考慮し得る。第1に、バウンディングボックスは、1つのキーおよび関連付けられたバリューを含むことがある。第2に、バウンディングボックスは、複数のキーおよび関連付けられたバリューを含むことがある。第3に、バウンディングボックスは、バリューを含み、キーは含まないことがある。

20

【0053】

バウンディングボックスが1つのキーを含むとバリュー識別モジュール350が決定した場合には、バリュー識別モジュール350は、そのバウンディングボックスからバリューを抽出し、キーおよび抽出されたバリューに基づいてキーとバリューペアを生成する。例えば、バリューは、バウンディングボックスによって囲まれるテキストからキーのテキストを除去することによって抽出されることがある。

【0054】

さらに、バウンディングボックスが複数のキーを含むとバリュー識別モジュール350が決定した場合には、バリュー識別モジュール350は、バウンディングボックスによって囲まれるキーの数に基づいてバウンディングボックスを分割する。いくつかの実施形態では、バリュー識別モジュール350は、キーとバリューの間の境界を最適化しようと試みることによってバウンディングボックスを分割する。例えば、バウンディングボックスが一行に積み重ねられたn個のキーを含む場合には、バリュー識別モジュール350は、そのバウンディングボックスを、それぞれが1つのキーおよび1つのバリューを含むn個のさらに小さいバウンディングボックスに水平方向に分割する。バリュー識別モジュール350は、次いで、バウンディングボックスが1つのキーおよび関連付けられたバリューを含む場合に対応するステップを、複数のキーを含むバウンディングボックスを分割することによって得られたバウンディングボックスの各々に適用する。

30

40

【0055】

さらに、バウンディングボックスがキーを含まないとバリュー識別モジュール350が決定した場合には、バリュー識別モジュールは、文書グラフを用いてバウンディングボックスをキーと関連付ける。

【0056】

図4Gは、1つまたは複数の実施形態による識別されたキーおよびバリューを有するバウンディングボックスを示す文書のマークアップ画像を示している。図4Gの例では、文書は、1つのキーおよび1つのバリューを含むバウンディングボックス471のセットを含む。例えば、バウンディングボックス471Aは、キー「Type」およびバリュー「

50

C h a n g e」を含む。バウンディングボックス 4 7 1 A を処理するとき、バリュー識別モジュールは、1 つのキーのみがバウンディングボックス 4 7 1 A によって囲まれていると決定し、バウンディングボックスからバリューを抽出して、バウンディングボックス内に位置するキーについてのキーとバリューペアを生成する。例えば、バリュー識別モジュールは、バウンディングボックス 4 7 1 A 内に位置する文字列からキーを除去することによって、バウンディングボックスからバリューを抽出する。さらに、図 4 G の例では、文書は、バリューを含むがキーは含まないバウンディングボックス 4 7 3 のセットを含む。例えば、バウンディングボックス 4 7 3 A は、バリュー「C A R B U R E T O R」を含むが、キーは含まない。ここで、バリュー識別モジュールは、そのバリューの最も近くに位置するキーを識別し、その識別されたキーをバウンディングボックスによって囲まれるバリューと関連付ける。この場合には、バリュー識別モジュールは、キー「N a m e」がバウンディングボックス 4 7 3 A に最も近いと識別子、キー「N a m e」をバリュー「C A R B U R E T O R」と関連付けることによってキーとバリューペアを生成する。

10

【0057】

図 5 は、1 つまたは複数の実施形態による解析ワークフローを用いてキーとバリューペアを抽出するプロセスを示す流れ図である。いくつかの実施形態では、このプロセスはデータ解析アプリケーション 1 5 0 によって実施されるが、プロセス中の動作の一部または全ては、他の実施形態では他のエンティティによって実施されることもあり得る。いくつかの実施形態では、プロセス中の動作は、異なる順序で実施され、異なる、および／または追加のステップを含むこともできる。

20

【0058】

いくつかの実施形態では、データ解析アプリケーション 1 5 0 は、ワークフローにアクセスする (5 1 0)。解析ワークフローは、ツールのシーケンスを含む。シーケンス中の各ツールは、入力、出力、準備、結合、予測的、空間的、調査、構文解析、および変換などのデータ解析機能を実施するように適応される。解析ワークフローは、論理制御によるワークフローであり、1 つまたは複数の論理制御装置をさらに含む。

【0059】

データ解析アプリケーション 1 5 0 のデータ受信モジュール 3 1 0 は、解析対象の文書を受信する (5 1 0)。解析対象の文書は、データファイル (例えばデータ解析アプリケーションのユーザによってアップロードされる) として受信されることもある。データファイルは、サポートされる文書フォーマットのセットのうちの 1 つ (例えば P D F フォーマット) として受信されることがある。さらに、データ解析アプリケーション 1 5 0 のデータ受信モジュール 3 1 0 は、文書から抽出されるユーザから提供されるキー名のセットを受信する (5 1 5)。いくつかの実施形態では、キーは、データ解析アプリケーションのユーザインタフェースモジュール 3 6 0 によって提供されるグラフィカルユーザインタフェースを介してデータ解析アプリケーション 1 5 0 のユーザから受け取られる。

30

【0060】

データ解析アプリケーション 1 5 0 は、文書を O C R モジュール 3 2 0 に提供する。データ解析アプリケーション 1 5 0 の O C R モジュール 3 2 0 は、文書内の文字列 (またはテキスト列) のセットを識別する (5 2 0)。さらに、O C R モジュール 3 2 0 は、識別された各文字列の位置を識別することがある。いくつかの実施形態では、O C R モジュールは、文書に対応する画像から文字列のセットを抽出する。さらに、O C R モジュール 3 2 0 は、文書から文字列のセットを抽出する前に、文書または文書の画像に対して前処理を実施することもある。

40

【0061】

データ解析アプリケーション 1 5 0 は、O C R モジュール 3 2 0 の出力をグラフ生成モジュール 3 3 0 に提供することがある。データ解析アプリケーション 1 5 0 のグラフ生成モジュール 3 3 0 は、O C R モジュール 3 2 0 によって抽出された文字列のセットに基づいて文書の文書グラフを生成する (5 2 5)。文書グラフは、ノードのセット (それぞれが O C R モジュール 3 2 0 によって識別された文字列のセットの文字列に対応する) と、

50

エッジのセット（各エッジが2つ以上のノードを互いに接続する）とを含むことがある。

【0062】

データ解析アプリケーション150は、次いで、グラフ生成モジュール330によって生成された文書グラフおよびユーザから提供されるキー名を、キー識別モジュール340に提供することがある。データ解析アプリケーション150のキー識別モジュール340は、文書中のキーのセットを識別する（530）。キー識別モジュール340は、各ユーザから提供されるキー名を、文書グラフ中の1つまたは複数のノードまたはノードチェーン（すなわちエッジのセットを介して接続されたノードのセット）とマッチングすることがある。

【0063】

データ解析アプリケーション150は、次いで、文書およびキー識別モジュール340によって識別されたキーを、バリュー識別モジュール350に提供することがある。いくつかの実施形態では、データ解析アプリケーション150は、さらに文書グラフをバリュー識別モジュール350に提供する。データ解析アプリケーション150のバリュー識別モジュール350は、文書内のテキストのブロックのセットを識別する（535）。特に、バリュー識別モジュール350は、それぞれが文書内の1つまたは複数の文字列を囲むバウンディングボックスのセットを生成する。バウンディングボックスを生成するために、バリュー識別モジュール350は、文書から線除去し、RLSAを適用し、輪郭検出アルゴリズムを適用することがある。

【0064】

さらに、データ解析アプリケーション150のバリュー識別モジュール350は、識別されたテキストのブロックから各キーのバリューを抽出する（540）。いくつかの実施形態では、バリューを抽出するために、バリュー識別モジュール350は、識別されたバウンディングボックスの各々を処理し、1つまたは複数のキーがそれらのバウンディングボックスの各々の中に位置するかどうかを決定する。複数のキーがバウンディングボックス内に位置する場合には、バリュー識別モデル350は、バウンディングボックスを分割して、その結果得られたバウンディングボックスの各々の中にキーが1つだけ位置するようにする。キーを1つだけ含む各バウンディングボックスについて、バリュー識別モジュール350は、バウンディングボックスによって囲まれるテキストから、バウンディングボックス内に位置するキーと関連付けられるバリューを抽出する。さらに、バウンディングボックス内にキーが位置していない場合には、バリュー識別モジュールは、そのバウンディングボックスをキーと関連付けるかどうかを決定する。

【0065】

最後に、キー識別モジュール340によって識別されたキーおよびバリュー識別モジュール350によって識別されたバリューに基づいて、キーとバリューペアのセットが生成される（545）。いくつかの実施形態では、データ解析アプリケーション150は、生成されたキーとバリューペアをデータ解析アプリケーションのユーザに対して出力する。あるいは、データ解析アプリケーション150は、生成されたキーとバリューペアを、ファイルに、またはデータ解析アプリケーションに接続されたデータベースに記憶することがある。

【0066】

図6は、1つまたは複数の実施形態による図1の機械学習サーバ110として使用される代表的なコンピュータシステム600の機能図を示すハイレベルブロック図である。

【0067】

図示のコンピュータシステムは、チップセット604に結合された少なくとも1つのプロセッサ602を含む。プロセッサ602は、複数のプロセッサコアを同じダイ上に含むことができる。チップセット604は、メモリ制御装置ハブ620および入出力（I/O）制御装置ハブ622を含む。メモリ606およびグラフィックスアダプタ612は、メモリ制御装置ハブ620に結合され、ディスプレイ618は、グラフィックスアダプタ612に結合される。記憶デバイス608、キーボード610、ポインティングデバイス6

10

20

30

40

50

14、およびネットワークアダプタ616は、I/O制御装置ハブ622に結合されることがある。いくつかの他の実施形態では、コンピュータシステム600は、追加の構成要素を有することも、これよりも少ない構成要素を有することも、または異なる構成要素を有することもあり、これらの構成要素が異なるかたちで結合されることがある。例えば、コンピュータシステム600の実施形態には、ディスプレイおよび／またはキーボードがなくてもよい。さらに、コンピュータシステム600は、いくつかの実施形態では、ラックマウント式のブレードサーバとして、またはクラウドサーバインスタンスとしてインスタンス化されることがある。

【0068】

メモリ606は、プロセッサ602によって使用される命令およびデータを保持する。いくつかの実施形態では、メモリ606は、ランダムアクセスメモリである。記憶デバイス608は、非一時的コンピュータ可読記憶媒体である。記憶デバイス608は、HDD、SSD、または他のタイプの非一時的コンピュータ可読記憶媒体とすることができる。機械学習サーバ110によって処理および解析されたデータは、メモリ606および／または記憶デバイス608に記憶されることが可能である。

【0069】

ポインティングデバイス614は、マウス、トラックボール、または他のタイプのポインティングデバイスであり得、コンピュータシステム600にデータを入力するためにキーボード610と組み合わせて使用される。グラフィックスアダプタ612は、画像および他の情報をディスプレイ618上に表示する。いくつかの実施形態では、ディスプレイ618は、ユーザの入力および選択を受け取るタッチスクリーン機能を含む。ネットワークアダプタ616は、コンピュータシステム600をネットワーク160に結合する。

【0070】

コンピュータシステム600は、本明細書に記載される機能を提供するコンピュータモジュールを実行するように適応される。本明細書で使用される「モジュール」という用語は、指定された機能を提供するコンピュータプログラム命令および他の論理を指す。モジュールは、ハードウェア、ファームウェア、および／またはソフトウェアとして実装されることが可能である。モジュールは、1つもしくは複数のプロセスを含むことができ、かつ／またはプロセスの一部のみによって提供されることが可能である。モジュールは、通常は、記憶デバイス608に記憶され、メモリ606にロードされ、プロセッサ602によって実行される。

【0071】

構成要素の個々の名称、用語に関する大文字の使用、属性、データ構造、または任意の他のプログラミングもしくは構造上の態様は、必須または重要なものではなく、記載されている実施形態を実装する機構は、異なる名称、フォーマット、またはプロトコルを有していてもよい。さらに、システムは、記載されるようにハードウェアとソフトウェアの組み合わせとして実装されることがあるし、または完全にハードウェア要素で実装されることがある。また、本明細書に記載される様々なシステム構成要素間の機能の特定の分割は単なる例示であり、必須のものではなく、1つのシステム構成要素によって実施されている機能が、1つではなく複数の構成要素によって実施されることがあるし、複数の構成要素によって実施されている機能が、複数ではなく1つの構成要素によって実施されることがある。

【0072】

上記の説明のいくつかの部分は、特徴を情報に対する動作のアルゴリズムおよび象徴的表現として提示している。これらのアルゴリズムの記述および表現は、それらの作業の実質を他の当業者に対して最も効果的に伝えるためにデータ処理技術の当業者によって使用される手段である。これらの動作は、機能的または論理的に説明されているが、コンピュータプログラムによって実施されるものとして理解される。さらに、これらの動作の構成を普遍性を失うことなくモジュールと呼ぶ、または機能名と呼ぶことは、好都合であることがあることも分かっている。

10

20

30

40

50

【0073】

上記の論考から明らかであるように特にそうでないと述べられていない限り、説明全体を通じて、「処理する」、「算出する」、「計算する」、「決定する」、または「表示する」などの用語を利用した論考は、コンピュータシステムのメモリもしくはレジスタ、またはその他のそのような情報記憶、伝送、または表示デバイス内で物理（電子）量として表されるデータを操作および変換する、コンピュータシステムまたは同様の電子コンピューティングデバイスのアクションおよびプロセスを指すものと理解される。

【0074】

本明細書に記載される特定の実施形態は、アルゴリズムの形態で記載されるプロセスステップおよび命令を含む。これらの実施形態のプロセスステップおよび命令は、ソフトウェア、ファームウェア、またはハードウェアとして実施されることが可能であり、ソフトウェアとして実施されるときには、ダウンロードされて、リアルタイムネットワークオペレーティングシステムによって使用される様々なプラットフォームに存在し、そこから操作されることも可能であることに留意されたい。

【0075】

最後に、本明細書で使用される表現は、第一に読みやすく教育的目的のために選択されたものであり、本発明の主題を画定または制限するようには選択されていない可能性がある。したがって、これらの実施形態の開示は、限定ではなく例示を意図している。

【図面】

【図 1】

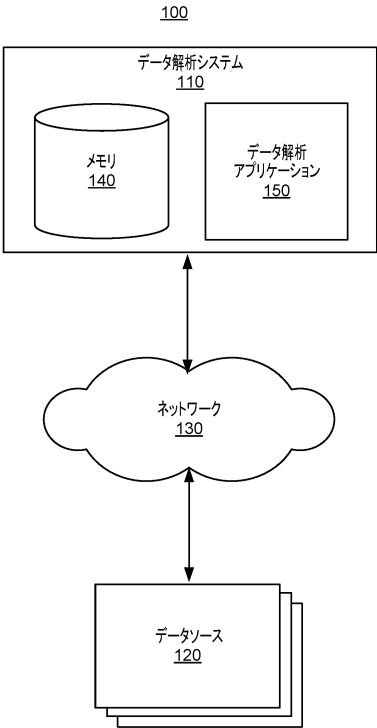


FIG. 1

【図 2】

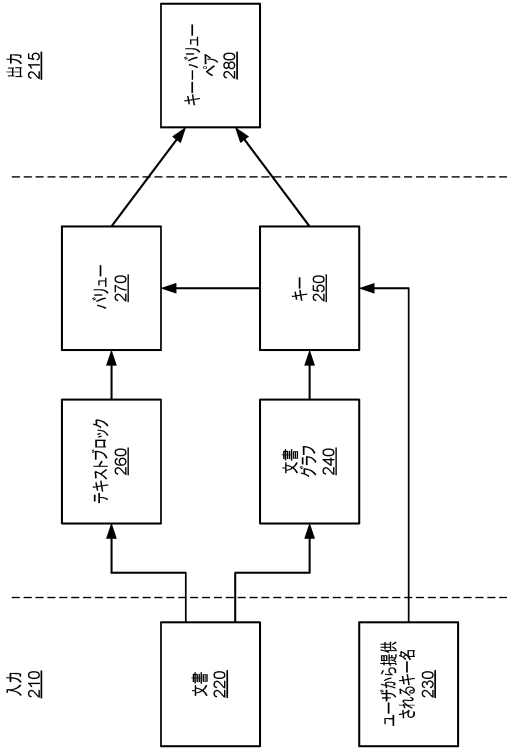


FIG. 2

【図 3】

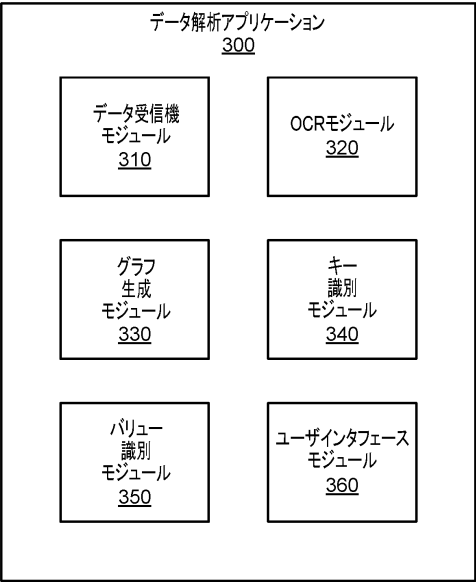


FIG. 3

【図 4 B】

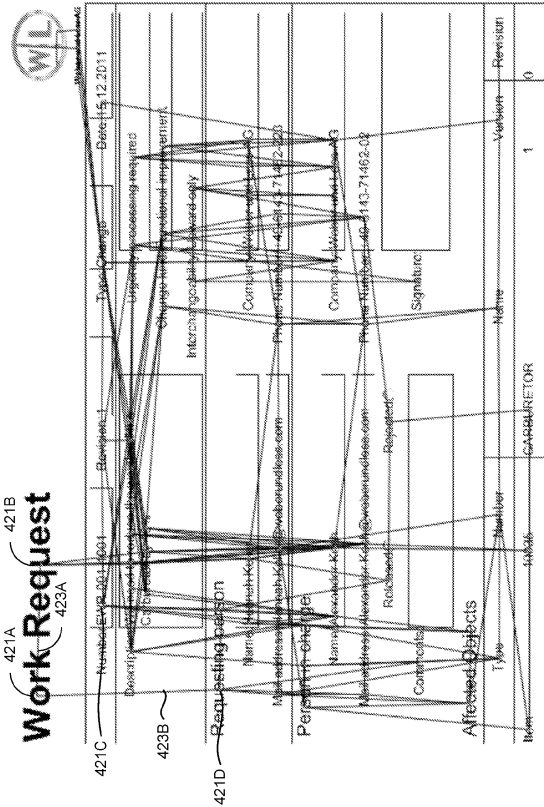


FIG. 4B

【図 4 A】

Work Request

Number: WVR-001-1001 Revision: 1 Date: 15.12.2011

Description: We need to reduce the weight of the Carburetor by 10%

Urgency: Urgency processing required

Change kind: Functional improvement

Interchangeability: Upward only

Requesting person: Name: Hannah Keller Company: Weber und Less AG Phone Number: +49-6143-71462-223

Person in charge: Name: Alexander Koch Company: Weber und Less AG Phone Number: +49-6143-71462-02

Affected Objects: Type: CARBURETOR Number: 10035 Name: CARBURETOR Version: 1 Revision: 0

FIG. 4A

【図 4 C】

Work Request

Number: WVR-001-1001 Revision: 1 Date: 15.12.2011

Description: We need to reduce the weight of the Carburetor by 10%

Urgency: Urgency processing required

Change kind: Functional improvement

Interchangeability: Upward only

Requesting person: Name: Hannah Keller Company: Weber und Less AG Phone Number: +49-6143-71462-223

Person in charge: Name: Alexander Koch Company: Weber und Less AG Phone Number: +49-6143-71462-02

Affected Objects: Type: CARBURETOR Number: 10035 Name: CARBURETOR Version: 1 Revision: 0

FIG. 4C

10

20

30

40

50

【図 5】

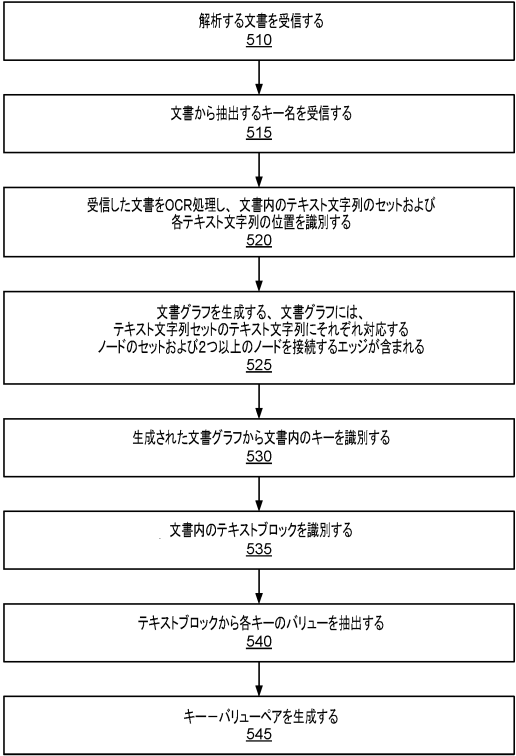


FIG. 5

【図 6】

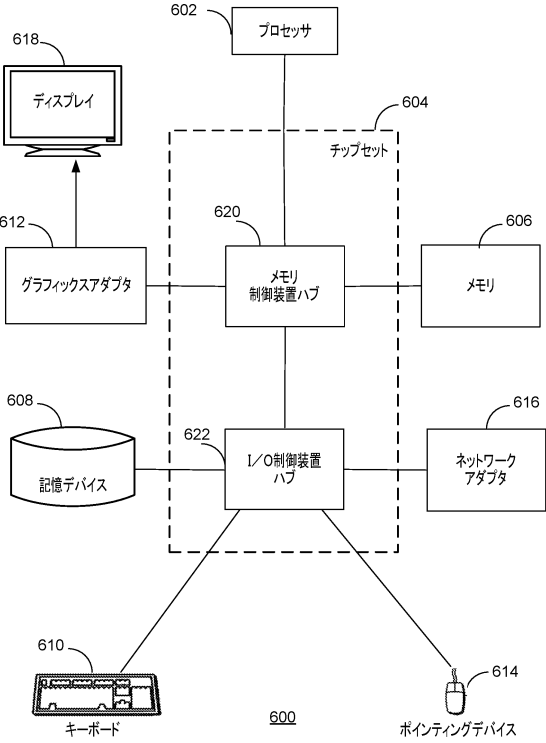


FIG. 6

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 23/13970

A. CLASSIFICATION OF SUBJECT MATTER IPC - INV. G06F 16/00, G06V 30/10 (2023.01) ADD. G06V 30/00 (2023.01) CPC - INV. G06F 16/00, G06V 30/10 ADD. G06V 30/00 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) See Search History document		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched See Search History document		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) See Search History document		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2010/0174732 A1 (Levy) 08 July 2010 (08.07.2010), entire document, especially Para [0016], [0117], [0120]-[0121], [0143], [0146], [0149], [0155], [0159], [0187], [0303]-[0304], [0310], [0675]	1-4, 8-11, 15-17
Y		5-7, 12-14, 18-20
Y	US 9,645,999 B1 (Quid, Inc.) 09 May 2017 (09.05.2017), entire document, especially Col 1, In 54-57; Col 12, In 43-53; Col 14, In 28-38; Col 17, In 31-42; Col 17, In 43-47	5, 12, 18
Y	CN 114,005,123 A (Peking University) 01 February 2022 (01.02.2022), entire document, especially Para [0003], [0083]	6-7, 13-14, 19-20
☐ 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 "I" 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 10 April 2023 (10.04.2023)		Date of mailing of the international search report MAY 30 2023
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300		Authorized officer Kari Rodriguez Telephone No. PCT Helpdesk: 571-272-4300

Form PCT/ISA/210 (second sheet) (July 2022)

フロントページの続き

,MC,ME,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,C
O,CR,CU,CV,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,I
R,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,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,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

ライブ 3 3 4 5 スイート 4 0 0 アルテリックス インコーポレイテッド内

F ターム (参考) 5B175 DA01 FB02 GA03 KA12