(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6284643号
(P6284643)

(45)発行日　平成30年2月28日(2018.2.28)　(24)登録日　平成30年2月9日(2018.2.9)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210A
G06F 17/30 160F

請求項の数 20　(全 20 頁)

(21)出願番号　特願2016-536850(P2016-536850)
(86)(22)出願日　平成26年12月1日(2014.12.1)
(65)公表番号　特表2016-541069(P2016-541069A)
(43)公表日　平成28年12月28日(2016.12.28)
(86)国際出願番号　PCT/US2014/067918
(87)国際公開番号　WO2015/084724
(87)国際公開日　平成27年6月11日(2015.6.11)
審査請求日　平成29年11月22日(2017.11.22)
(31)優先権主張番号　61/910,739
(32)優先日　平成25年12月2日(2013.12.2)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　516164586
キューベース　リミテッド　ライアビリティ　カンパニー
アメリカ合衆国　ヴァージニア州　２０１９１　レストン　サンライズ　ヴァレー　ドライヴ　１２０１８　スイート　３００
(74)代理人　100086771
弁理士　西島　孝喜
(74)代理人　100088694
弁理士　弟子丸　健
(74)代理人　100094569
弁理士　田中　伸一郎
(74)代理人　100067013
弁理士　大塚　文昭

最終頁に続く

(54)【発明の名称】非構造化テキストにおける特徴の曖昧性除去方法

(57)【特許請求の範囲】
【請求項１】
サーチ質問をエンドユーザ装置から受信することに応答して、
システムのノードにより、１つ以上の抽出された特徴に一致する１つ以上の候補レコードを識別するために共起特徴を含む候補レコードのセットをサーチし、候補レコードに一致する抽出された特徴は、一次特徴であり、前記ノードはインメモリデータベースをホストするメインメモリを含み、前記インメモリデータベースはクラスターの知識ベースを格納し、各クラスターは独特の識別子（独特のＩＤ）を伴う曖昧性除去された一次特徴及び関連付けられた二次特徴を含み、
ノードにより、抽出された特徴の各々を１つ以上のマシン発生トピック識別子（トピックＩＤ）と関連付け、
ノードにより、トピックＩＤの関連度に基づき一次特徴の各々を互いに曖昧性除去し、
ノードにより、トピックＩＤの関連度に基づき各一次特徴に関連した二次特徴のセットを識別し、
ノードにより、トピックＩＤの関連度に基づき二次特徴の関連セットにおける二次特徴の各々から一次特徴の各々を曖昧性除去し、
前記インメモリデータベースから前記知識ベースからデータが検索されるときに、リアルタイムで、ノードにより、各一次特徴を二次特徴の関連セットにリンクして、新たなクラスターを形成し、
ノードのインメモリデータベースの曖昧性除去モジュールにより、曖昧性除去された一

次特徴を伴う既存の知識クラスターへの比較的一致するスコアの指定により前記新たなクラスターの各々が既存の知識ベースクラスターに一致するかどうか決定し、
　一致があるときには、既存の知識ベースクラスターにおける各一致する一次特徴に対応する既存の独特のＩＤを決定しそして前記新たなクラスターを含むように既存の知識ベースクラスターを更新し、及び
　一致がないときには、新たな知識ベースクラスターを生成し、そしてその新たな知識ベースクラスターの一次特徴に新たな独特のＩＤを指定し、及び
　既存の独特のＩＤ及び新たな独特のＩＤの一方を一次特徴として前記エンドユーザ装置へ送出する、
ことを含む方法。
【請求項２】
　ノードにより、抽出された特徴に一致する候補レコードの各々を比較し、及びノードにより、その比較に基づいて前記抽出された特徴の各々に重み付けされた一致スコア結果を指定する、ことを更に含む、請求項１に記載の方法。
【請求項３】
　ノードにより、抽出された特徴の各々を、重み付けされた特徴属性のセットに関連付けることを更に含む、請求項２に記載の方法。
【請求項４】
　ノードにより、１つ以上の重み付けされた特徴属性に基づいて抽出された特徴の各々の関連度を決定することを更に含む、請求項３に記載の方法。
【請求項５】
　１つ以上の抽出された特徴をノードの抽出モジュールにより確認及び抽出し、１つ以上の抽出された特徴において１つ以上の一次特徴を識別し、及び
　ノードの抽出モジュールにより、抽出された特徴の各々をデータベースに記憶する、
ことを更に含む、請求項１に記載の方法。
【請求項６】
　ノードの抽出モジュールにより、各特徴に抽出確度スコアを指定することを更に含む、請求項５に記載の方法。
【請求項７】
　各々の一次特徴は、１つ以上の特徴属性のセットに関連付けられる、請求項１に記載の方法。
【請求項８】
　特徴属性は、トピックＩＤ、ドキュメント識別子（ドキュメントＩＤ）、特徴タイプ、特徴名、信頼性スコア、及び特徴位置より成るグループから選択される、請求項７に記載の方法。
【請求項９】
　各関連特徴は、予め定義されたクラスターハイアラーキーに従って下位順序特徴のセットに関連付けられる、請求項１に記載の方法。
【請求項１０】
　ノードにより、候補レコードのセットの曖昧キーサーチを遂行することを更に含む、請求項１に記載の方法。
【請求項１１】
　ノードのリンクオンザフライモジュールにより、関連トピックＩＤの共起及び１つ以上の特徴属性に基づいて２つ以上のデータソースをリンクすることを更に含む、請求項７に記載の方法。
【請求項１２】
　ノードにより、データソースにおける抽出された特徴が第２データソースにおいて共起するかどうかを、その抽出された特徴を第２データソースにおける特徴と比較することで決定し、及び
　ノードにより、前期比較に基づいてデータソースの各々をリンクする、

ことを更に含む、請求項１に記載の方法。
【請求項１３】
ノードにより、異なるデータソースからの抽出された特徴の共起を分析して、抽出された特徴の曖昧性除去の精度を改善することを更に含む、請求項１に記載の方法。
【請求項１４】
ノードにより、１つ以上の新たなデータソースを連続的に受け取り、
ノードにより、１つ以上の抽出される特徴を連続的に抽出し、
ノードにより、１つ以上の抽出された特徴において候補サーチを連続的に遂行し、
ノードにより、抽出された特徴を連続的に曖昧性除去し、及び
ノードにより、抽出された特徴を１つ以上の新たなクラスターへ連続的にリンクする、
ことを更に含む、請求項１に記載の方法。
【請求項１５】
コンピュータ実行可能なインストラクションが記憶された非一時的コンピュータ読み取り可能な媒体であって、プロセッサによって実行されると、
サーチ質問をエンドユーザ装置から受信することに応答して、
システムのノードにより、１つ以上の抽出された特徴に一致する１つ以上の候補レコードを識別するために共起特徴を含む候補レコードのセットをサーチし、前記ノードはインメモリデータベースをホストするメインメモリを含み、前記インメモリデータベースはクラスターの知識ベースを格納し、各クラスターは独特の識別子（独特のＩＤ）を伴う曖昧性除去された一次特徴及び関連付けられた二次特徴を含み、
ノードにより、抽出された特徴の各々を１つ以上のマシン発生トピック識別子（トピックＩＤ）と関連付け、
ノードにより、トピックＩＤの関連度に基づき一次特徴の各々を互いに曖昧性除去し、
ノードにより、トピックＩＤの関連度に基づき各一次特徴に関連した二次特徴のセットを識別し、
ノードにより、トピックＩＤの関連度に基づき二次特徴の関連セットにおける二次特徴の各々から一次特徴の各々を曖昧性除去し、
前記インメモリデータベースから前記知識ベースからデータが検索されるときに、リアルタイムで、ノードにより、各一次特徴を二次特徴の関連セットにリンクして、新たなクラスターを形成し、
ノードのインメモリデータベースの曖昧性除去モジュールにより、曖昧性除去された一次特徴を伴う既存の知識クラスターへの比較的一致するスコアの指定により前記新たなクラスターの各々が既存の知識ベースクラスターに一致するかどうか決定し、
一致があるときには、既存の知識ベースクラスターにおける各一致する一次特徴に対応する既存の独特のＩＤを決定しそして前記新たなクラスターを含むように既存の知識ベースクラスターを更新し、及び
一致がないときには、新たな知識ベースクラスターを生成し、そしてその新たな知識ベースクラスターの一次特徴に新たな独特のＩＤを指定し、及び
既存の独特のＩＤ及び新たな独特のＩＤの一方を一次特徴として前記エンドユーザ装置へ送出する、
ことを含む機能が実行される、コンピュータ実行可能なインストラクションが記憶された非一時的コンピュータ読み取り可能な媒体。
【請求項１６】
前記インストラクションは、更に、ノードにより、抽出された特徴に一致する候補レコードの各々を比較し、及びノードにより、その比較に基づいて前記抽出された特徴の各々に重み付けされた一致スコア結果を指定する、ことを含む、請求項１５に記載の非一時的コンピュータ読み取り可能な媒体。
【請求項１７】
前記インストラクションは、更に、ノードにより、抽出された特徴の各々を、重み付けされた特徴属性のセットに関連付けることを含む、請求項１６に記載の非一時的コンピュ

ータ読み取り可能な媒体。
【請求項１８】
前記インストラクションは、更に、ノードにより、１つ以上の重み付けされた特徴属性に基づいて抽出された特徴の各々の関連度を決定することを含む、請求項１７に記載の非一時的コンピュータ読み取り可能な媒体。
【請求項１９】
前記インストラクションは、更に、
ノードの抽出モジュールにより、１つ以上の抽出された特徴を確認し及び抽出し、その１つ以上の抽出された特徴において１つ以上の一次特徴を識別し、及び
ノードの抽出モジュールにより、抽出された特徴の各々をデータベースに記憶する、ことを含む、請求項１５に記載の非一時的コンピュータ読み取り可能な媒体。
【請求項２０】
前記インストラクションは、更に、ノードの抽出モジュールにより、各特徴に抽出確度スコアを指定することを含む、請求項１９に記載の非一時的コンピュータ読み取り可能な媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般的に、データマネージメントに関するもので、より詳細には、ネットワークを経て受け取ったソースアイテムから資料を抽出しそして記憶するデータマネージメントシステム及び方法に関する。
【背景技術】
【０００２】
ネットワークのようなソースを含む大きなドキュメント集合体におけるエンティティ（即ち、人々、位置、組織）に関する情報のサーチは、しばしば曖昧なものであり、不正確なテキスト処理機能、知識抽出中の特徴の不正確な関連付け、ひいては、不正確なデータ分析を招くことがある。
【０００３】
最新のシステムは、ＰａｇｅＲａｎｋ及びハイパーリンク誘起トピックサーチ（ＨＩＴＳ）アルゴリズムのような多数のアルゴリズムにおけるリンケージベースのクラスター化及びランク付けを使用している。この解決策及び関連解決策の背景にある基本的アイデアは、既存のリンクが、典型的に、関連ページ又はコンセプト間に存在することである。クラスター化ベースの技術の制約は、エンティティを曖昧性除去するのに必要なコンテキスト情報がコンテキストに存在しないことが時々あって、不正確な曖昧性除去結果を招くことである。同様に、同じ又は表面的に同様のコンテキストにおける異なるエンティティに関するドキュメントが不正確にクラスター化されることもある。
【０００４】
他のシステムは、エンティティの１つ以上の外部ディクショナリ（又は知識ベース）を参照することによりエンティティの曖昧性除去を試みる。そのようなシステムでは、エンティティのコンテキストがディクショナリ内の考えられる一致エンティティと比較され、そして最も近い一致が返送される。現在のディクショナリベース技術に関連した制約は、いつでもエンティティの数が増加し、それ故、世界中の全てのエンティティの表現を含むディクショナリがないことから生じる。従って、ドキュメントのコンテキストがディクショナリのエンティティに一致する場合に、この技術では、ディスクショナリ内の最も類似したエンティティだけが識別され、ディクショナリ以外にある正しいエンティティが必ずしも識別されない。
【発明の概要】
【発明が解決しようとする課題】
【０００５】
ほとんどの方法は、曖昧性除去プロセスにおいてエンティティ及びキーフレーズだけを

使用する。それ故、正確なデータ分析を行うことのできる正確なエンティティ曖昧性除去技術が依然要望されている。

【課題を解決するための手段】

【０００６】

ある実施形態では、特徴を曖昧性除去する方法について述べる。この方法は、１つ以上の特徴抽出モジュール、１つ以上の曖昧性除去モジュール、１つ以上のスコア付けモジュール、及び１つ以上のリンキングモジュールのような複数のモジュールを含む。

【０００７】

特徴の曖昧性除去は、特徴の周囲ドキュメントからトピックを抽出し、レイテントディリクレアロケーション（ＭＣ－ＬＤＡ）トピックモデルのマルチコンポーネント拡張を使用することにより、一部分、サポートされる。ここで、各コンポーネントは、既存の知識ベースに記憶されるか又は到来するドキュメントにおいて抽出される各二次的特徴に関してモデリングされる。更に、リンキング又は曖昧性除去プロセスは、ＭＣ－ＬＤＡからのトピック推論としてモデリングされ、これは、ＭＣ－ＬＤＡトレーニングの間に自動重み推定を与え、そして推論中にそれを容易に適用する。

【０００８】

この規範的方法は、ドキュメントリンキングを考慮せずに達成できるものを越えてエンティティ曖昧性除去の精度を改善する。ドキュメントリンケージを考慮し、リンクにより暗示されるドキュメント及びエンティティ関係を考えることで良好な曖昧性除去を行うことができる。

【０００９】

ある実施形態において、方法は、インメモリデータベースをホストするシステムのノードにより、１つ以上の抽出された特徴に一致する１つ以上の候補を識別するために候補レコードのセットをサーチし、候補に一致する抽出された特徴は、一次特徴であり；ノードにより、抽出された特徴の各々を１つ以上のマシン発生トピック識別子（トピックＩＤ）と関連付け；ノードにより、トピックＩＤの関連度に基づき一次特徴の各々を互いに曖昧性除去し；ノードにより、トピックＩＤの関連度に基づき各一次特徴に関連した二次特徴のセットを識別し；ノードにより、トピックＩＤの関連度に基づき二次特徴の関連セットにおける二次特徴の各々から一次特徴の各々を曖昧性除去し；ノードにより、各一次特徴を二次特徴の関連セットにリンクして新たなクラスターを形成し；ノードにより、新たなクラスターが既存の知識ベースクラスターに一致するかどうか決定し、一致があるときには、インメモリデータベースサーバーコンピュータの曖昧性除去モジュールにより、知識ベースクラスターにおける各一致する一次特徴に対応する既存の独特の識別子（独特のＩＤ）を決定しそしてその新たなクラスターを含むように知識ベースクラスターを更新し、及び一致がないときには、ノードにより、新たな知識ベースクラスターを生成し、そしてその新たな知識ベースクラスターの一次特徴に新たな独特のＩＤを指定し；及びノードにより、既存の独特のＩＤ及び新たな独特のＩＤの一方を一次特徴として送出する；ことを含む。

【００１０】

別の実施形態において、非一時的コンピュータ読み取り可能な媒体に記憶されるコンピュータ実行可能なインストラクションは、インメモリデータベースをホストするシステムのノードにより、１つ以上の抽出された特徴に一致する１つ以上の候補を識別するために候補レコードのセットをサーチし、候補に一致する抽出された特徴は、一次特徴であり；ノードにより、抽出された特徴の各々を１つ以上のマシン発生トピック識別子（トピックＩＤ）と関連付け；ノードにより、トピックＩＤの関連度に基づき一次特徴の各々を互いに曖昧性除去し；ノードにより、トピックＩＤの関連度に基づき各一次特徴に関連した二次特徴のセットを識別し；ノードにより、トピックＩＤの関連度に基づき二次特徴の関連セットにおける二次特徴の各々から一次特徴の各々を曖昧性除去し；ノードにより、各一次特徴を二次特徴の関連セットにリンクして新たなクラスターを形成し；ノードにより、新たなクラスターが既存の知識ベースクラスターに一致するかどうか決定し、一致がある

ときには、ノードにより、知識ベースクラスターにおける各一致する一次特徴に対応する既存の独特の識別子（独特のＩＤ）を決定しそしてその新たなクラスターを含むように知識ベースクラスターを更新し、及び一致がないときには、新たな知識ベースクラスターを生成し、そしてその新たな知識ベースクラスターの一次特徴に新たな独特のＩＤを指定し；及びノードにより、既存の独特のＩＤ及び新たな独特のＩＤの一方を一次特徴として送出する；ことを含む。

【００１１】

一実施形態の付加的な特徴及び効果は以下の説明に述べられ、そして一部分はその説明から明らかとなろう。本発明の目的及び他の効果は、以下の説明の規範的実施形態、特許請求の範囲及び添付図面に特に指摘された構造により実現され且つ達成される。

【００１２】

以上の一般的な説明及び以下の詳細な説明は、どちらも、規範的な説明に過ぎず、請求の範囲に述べる本発明の更なる説明を与えるものであることを理解されたい。

【００１３】

本開示は、添付図面を参照することにより良く理解することができる。添付図面は、本明細書の一部分を構成するもので、本発明の実施形態を示し、そして明細書と共に本発明を説明するものである。図面中のコンポーネントは、必ずしも、正しい縮尺ではなく、むしろ、本開示の原理を例示する際に強調されている。図中、参照番号は、異なる図面全体を通して対応部分を示している。

【図面の簡単な説明】

【００１４】

【図１】規範的実施形態により、非構造化テキストにおける特徴を曖昧性除去する方法のフローチャートである。

【図２】規範的実施形態により、特徴を曖昧性除去する方法に使用される曖昧性除去モジュールにより遂行されるステップのフローチャートである。

【図３】規範的実施形態により、特徴を曖昧性除去する方法に使用されるリンクオンザフライモジュールにより遂行されるステップのフローチャートである。

【図４】規範的実施形態により、特徴を曖昧性除去する方法の実施に使用されるシステムを示す図である。

【図５】規範的実施形態によるマルチコンポーネント条件独立のレイテントディリクレアロケーション（ＭＣ－ＬＤＡ）トピックモデルのグラフ表示である。

【図６】規範的実施形態によるマルチコンポーネント条件独立のレイテントディリクレアロケーション（ＭＣ－ＬＤＡ）トピックモデルのギブスサンプリング方程式の実施形態を示す。

【図７】規範的実施形態によるマルチコンポーネント条件独立のレイテントディリクレアロケーション（ＭＣ－ＬＤＡ）トピックモデルにおけるトレーニング及び推論のための確率変化推論アルゴリズムの実施形態を示す。

【図８】規範的実施形態によるマルチコンポーネント条件独立のレイテントディリクレアロケーション（ＭＣ－ＬＤＡ）トピックモデルのためのサンプルトピックを示すテーブルである。

【発明を実施するための形態】

【００１５】

定義

ここで使用する次の用語は、次のような定義を有する。

【００１６】

「ドキュメント」は、出発点及び終了点を有する情報の個別の電子的表現を指す。

【００１７】

「マルチドキュメント」は、トークン、異なる形式の名前付きエンティティ、及び個別の「bag-of-surface-forms」コンポーネントに編成されるキーフレーズを伴うドキュメントを指す。

【００１８】
「データベース」は、１つ以上の集合体を記憶するのに適し且つ１つ以上の質問を処理するのに適したクラスター及びモジュールの組み合わせを含むシステムを指す。
【００１９】
「コーパス」は、１つ以上のドキュメントの集合体を指す。
【００２０】
「生のコーパス」又は「ドキュメントストリーム」は、新たなドキュメントがネットワークにアップロードされるときに恒常的に供給されるコーパスを指す。
【００２１】
「特徴(Features)」は、ドキュメントから少なくとも一部分導出される情報である。
【００２２】
「特徴属性」は、特徴に関連したメタデータ、例えば、とりわけ、ドキュメントにおける特徴の位置、信頼スコアを指す。
【００２３】
「クラスター」は、特徴の集合体を指す。
【００２４】
「エンティティ知識ベース」は、特徴／エンティティを含むベースを指す。
【００２５】
「リンクオンザフライモジュール」又は「リンクＯＴＦ」は、生のコーパスが更新されるにつれてデータを更新するリンキングモジュールを指す。
【００２６】
「メモリ」は、充分に高い速度で情報を記憶し且つその情報を検索するのに適したハードウェアコンポーネントを指す。
【００２７】
「モジュール」は、１つ以上の定義されたタスクを実行するのに適したコンピュータソフトウェアコンポーネントを指す。
【００２８】
「センチメント(Sentiment)」は、ドキュメント、ドキュメントの一部分、又は特徴に関連した客観的評価を指す。
【００２９】
「トピック」は、コーパスから少なくとも一部分導出されるセマティック情報のセットを指す。
【００３０】
「トピック識別子」又は「トピックＩＤ」は、トピックの特定インスタンスを指す識別子である。
【００３１】
「トピック集合体」は、コーパスから導出されるトピックの特定セットを指し、各トピックは、独特の識別子（独特のＩＤ）を有する。
【００３２】
「トピック分類」は、特定のトピック識別子をドキュメントの特徴として指定することを指す。
【００３３】
「質問」は、１つ以上の適当なデータベースから情報を検索するための要求を指す。
【００３４】
詳細な説明
添付図面に各々示された好ましい実施形態を以下に詳細に説明する。上述した実施形態は、例示に過ぎない。当業者であれば、ここに述べる特定の実施例について、本発明の範囲内で、多数の別のコンポーネント及び実施形態に置き換えできることが認識されよう。
【００３５】
本開示は、非構造化テキストにおける特徴を曖昧性除去する方法について述べる。規範

的な実施形態は、本開示に従って特徴を曖昧性除去する慣習について述べるが、ここに述べるシステム及び方法は、本開示の範囲内で適当に使用するように構成できることが意図される。

【００３６】

既存の知識ベースは、曖昧さのない特徴及びそれに関連した特徴を含み、これは、信頼性の低いテキスト分析を招く。本開示の観点は、特徴及びエンティティの曖昧性除去精度を高め、それ故、テキスト分析の精度を高める。

【００３７】

一実施形態によれば、特徴を曖昧性除去するここに開示する方法は、初期データコーパスに使用されて、ドキュメント取り込み及び特徴抽出を遂行し、初期コーパスに含まれた各ドキュメントに対してトピック分類及び他のテキスト分析を行えるようにする。各特徴は、とりわけ、ドキュメントの名前、タイプ、位置情報、及び信頼性スコアとして識別されそして記録される。

【００３８】

図１は、非構造化テキストにおける特徴を曖昧性除去する複数のステップを示す方法１００のフローチャートである。一実施形態によれば、特徴曖昧性除去方法１００は、既存の知識ベースにおいて新たなドキュメント入力ステップ１０２が実行されるときに開始する。ドキュメントに対して特徴抽出ステップ１０４が遂行される。一実施形態によれば、特徴は、とりわけ、トピック識別子（トピックＩＤ）、ドキュメント識別子（ドキュメントＩＤ）、特徴のタイプ、特徴の名前、信頼性スコア及び特徴の位置のような異なる特徴属性に関連している。

【００３９】

種々の実施形態によれば、ステップ１０２のドキュメント入力は、大量コーパス又は生のコーパス（インターネット又はネットワーク接続のコーパスのような）から供給され、これは、次いで、毎秒供給される。

【００４０】

異なる実施形態によれば、特徴抽出ステップ１０４の間に、ドキュメント入力ステップ１０２の非構造化テキストを分析するために１つ以上の特徴確認及び抽出アルゴリズムが使用される。抽出された各特徴にスコアが指定される。そのスコアは、正しい属性で正しく抽出される特徴の確度レベルを指示する。

【００４１】

加えて、特徴抽出ステップ１０４の間に、ステップ１０２のドキュメント入力から１つ以上の一次特徴が識別される。各一次特徴は、特徴属性のセット及び１つ以上の二次特徴に関連付けられる。各二次特徴は、特徴属性のセットに関連付けられる。ある実施形態では、１つ以上の二次特徴は、特徴属性のそれ自身のセットを各々有する１つ以上の三次特徴を有する。

【００４２】

特徴属性を考慮して、ステップ１０２のドキュメント入力内の各特徴の相対的重み又は関連度が決定される。加えて、重み付けされたスコア付けモデルを使用して、特徴と特徴との間の関連付けの関連度が決定される。

【００４３】

特徴抽出ステップ１０４に続いて、ステップ１０２のドキュメント入力から抽出された特徴及びそれに関連した全ての情報は、ステップ１０６においてＭｅｍＤＢに特徴を含ませる間に、特徴曖昧性除去要求ステップ１０８の一部分として、インメモリデータベース（ＭｅｍＤＢ）にロードされる。

【００４４】

一実施形態では、ＭｅｍＤＢは、図１から８に関連して述べるステップを実行する１つ以上のプロセッサを有する曖昧性除去コンピュータサーバー環境の一部分を形成する。ある実施形態では、ＭｅｍＤＢは、１つ以上のサーチコントローラ、複数のサーチノード、圧縮データの集合体、及び曖昧性除去サブモジュールを含むコンピュータモジュールであ

る。１つのサーチコントローラが１つ以上のサーチノードに選択的に関連付けられる。各サーチノードは、圧縮データの集合体を通して曖昧キーサーチを独立して遂行し、そしてスコア付けされた結果のセットをそれに関連したサーチコントローラへ返送することができる。

【００４５】

特徴の曖昧性除去ステップ１０８は、ＭｅｍＤＢ内の曖昧性除去サブモジュールにより遂行される。特徴の曖昧性除去１０８プロセスは、マシンで発生されるトピックＩＤを含み、これは、特徴、ドキュメント、又はコーパスを分類するのに使用される。個々の特徴及び特定のトピックＩＤの関連度は、曖昧性除去アルゴリズムを使用して決定される。あるドキュメントにおいて、そのドキュメント内の特徴の異なる発生のコンテキストに基づき、１つ以上のトピックＩＤに同じ特徴が関連付けられる。

【００４６】

あるドキュメントから抽出された特徴（同じトピック、接近用語及びエンティティ、キーフレーズ、イベント及びファクト）のセットは、異なるドキュメントにわたる２つ以上の特徴が単一の特徴である場合、又はそれらが別々の特徴である場合に、ある精度レベルで定義する曖昧性除去アルゴリズムを使用して、他のドキュメントからの特徴のセットと比較される。ある例では、データベースにおけるドキュメントの集合体にわたる２つ以上の特徴の共起を分析して、特徴曖昧性除去プロセス１０８の精度を改善する。ある実施形態では、全体的スコア付けアルゴリズムを使用して、特徴が同じである確率を決定する。

【００４７】

ある実施形態では、特徴曖昧性除去プロセス１０８の一部分として、ＭｅｍＤＢ内に知識ベースが発生される。この知識ベースは、関連する曖昧性除去された一次特徴及びそれに関連する二次特徴のクラスターを一時的に記憶するのに使用される。新たなドキュメントがＭｅｍＤＢにロードされたときに、曖昧性除去された新たな特徴セットを既存の知識ベースと比較し、特徴と特徴との関係を決定し、そして新たな特徴と既に抽出された特徴との間に一致があるかどうか決定する。

【００４８】

比較された特徴が一致する場合には、知識ベースが更新され、一致する特徴の特徴ＩＤがユーザ及び／又は要求側アプリケーション又はプロセスへ返送され、そして更に、一致の頻度に基づいて目立った手段を特徴ＩＤと共に取り付けることができ、これは、所与のコーパスにおいてその人気指数を捕らえるものである。比較された特徴が既に抽出された特徴のいずれとも一致しない場合には、曖昧性除去されたエンティティ又は特徴に独特の特徴ＩＤが指定され、その独特の特徴ＩＤは、特徴を定義するクラスターに関連付けられそしてＭｅｍＤＢの知識ベース内に記憶される。その後、ステップ１１０において、曖昧性除去された特徴の特徴ＩＤがシステムインターフェイスを通してソースへ返送される。ある実施形態では、曖昧性除去された特徴の特徴ＩＤは、二次特徴、特徴のクラスター、関連特徴属性、又は他の要求データを含む。特徴曖昧性除去ステップ１０８に対して使用される曖昧性除去サブモジュールを、図２について以下に詳細に述べる。

【００４９】

曖昧性除去サブモジュール

図２は、一実施形態により、方法１００（図１）の特徴曖昧性除去ステップ１０８の非構造化テキストに使用される曖昧性除去サブモジュールにより遂行されるプロセス２００のフローチャートである。曖昧性除去プロセス２００は、図１のステップ１０６においてＭｅｍＤＢに特徴を含ませた後に始まる。ステップ２０２において与えられる抽出された特徴は、ステップ２０４において候補サーチを遂行するのに使用され、抽出された特徴についてのサーチは、共起特徴を含めて全ての候補レコードを通して遂行される。

【００５０】

種々の実施形態によれば、候補は、特徴の曖昧性除去プロセス１０８に使用される関連二次特徴のセットを伴う一次特徴である。

【００５１】

曖昧性除去結果は、トピックＩＤの共起とトピックＩＤの中の関連度とにより改善される。トピックＩＤの関連度は、異なるトピックモデルにわたるものであっても、トピックＩＤが指定された大きなコーパスから発見することができる。関連トピックＩＤをレコードリンケージステップ２０６の間に使用して、厳密なトピックＩＤを含まないが１つ以上の関連トピックＩＤを含むドキュメントへのリンケージを与えることができる。この解決策は、レコードリンケージステップ２０６に含まれるべき関連特徴のリコールを改善し、そしてあるケースでは、曖昧性除去結果を改善する。

【００５２】

潜在的に関連するドキュメントのセットが識別され、そしてそれらのドキュメント内の関連する一次及び二次特徴が抽出されると、特徴の属性、同じドキュメント（意義のあるコンテキスト）の特徴と特徴との間の関係、特徴の相対的重み、及び他の変数をレコードリンケージプロセス２０６の間に使用して、それらのドキュメントにわたる一次及び二次特徴を曖昧性除去する。次いで、各レコードを他のレコードにリンクして、曖昧性除去された一次特徴及びそれらの関連する二次特徴のクラスターを決定する。レコードリンケージ２０６に使用されるアルゴリズムは、マイニング非構造化データセットのスペルエラー又は翻字及び他の課題を克服することができる。

【００５３】

クラスター比較ステップ２０８は、比較的一致するスコアを、曖昧性除去された特徴のクラスターに指定することを含み、異なるアプリケーションに対して異なる受け容れスレッシュホールドが定義される。定義された精度レベルは、どのスコアが肯定的一致サーチと考えられそしてどのスコアが否定的一致サーチと考えられるか決定する（ステップ２１０）。各新たなクラスターは、独特のＩＤが与えられ、そして知識ベースに一時的に記憶される。各新たなクラスターは、曖昧性除去された新たな一次特徴及び二次特徴のセットを含む。新たなクラスターが、知識ベースに既に記憶されているクラスターに一致する場合には、システムは、ステップ２１２において知識ベースを更新し、そしてユーザ及び／又は要求側アプリケーション又はプロセスへの一致特徴ＩＤの返送がステップ２１４において遂行される。知識ベースの更新２１２は、１つの一次特徴への付加的な二次特徴の関連付け、或いは一次又は二次特徴に以前に関連付けされていない特徴属性の追加を意味する。

【００５４】

評価されているクラスターに、肯定的一致サーチ２１０のスレッシュホールドより低いスコアが指定された場合には、システムは、ステップ２１６において、クラスターの一次特徴に独特のＩＤ指定を遂行し、そしてステップ２１２において、知識ベースを更新する。その後、システムは、一致ＩＤ返送プロセス２１４を遂行する。レコードリンケージステップ２０６は、図３を参照して更に詳細に説明する。

【００５５】

リンクオンザフライサブモジュール

図３は、一実施形態により、特徴を曖昧性除去する方法１００に使用されるリンクオンザフライ（リンクＯＴＦ）サブモジュールにより遂行されるプロセス３００のフローチャートである。リンクＯＴＦプロセス３００は、情報のフィードを定常的に評価し、スコア付けし、リンクし、そしてクラスター化することができる。リンクＯＴＦサブモジュールは、複数のアルゴリズムを使用してレコードリンケージ２０６を遂行する。ステップ２０４の候補サーチ結果は、リンクＯＴＦモジュール３００へ定常的にフィードされる。データの入力に続いて、一致スコア付けアルゴリズムが適用され（ステップ３０２）、ここでは、１つ以上の一致スコア付けアルゴリズムがＭｅｍＤＢの複数のサーチノードに同時に適用される一方、とりわけ、ストリンク編集距離、表音及び意味のような複数の特徴属性を考慮して、関連する結果を評価及びスコア付けするために曖昧キーサーチを遂行する。

【００５６】

その後、一致スコア付けアルゴリズム適用ステップ３０２の間に識別された全ての候補レコードを互いに比較するために、リンキングアルゴリズムの適用ステップ３０４が追加

される。リンキングアルゴリズムの適用３０４は、ＭｅｍＤＢの複数のサーチノードの内部で遂行される曖昧キーサーチのスコア付けされた結果をフィルタリング及び評価できる１つ以上の分析リンキングアルゴリズムの使用を含む。ある例では、ＭｅｍＤＢにおける識別された候補レコードの集合体にわたる２つ以上の特徴の共起を分析して、プロセスの精度を改善する。リンキングアルゴリズムの適用３０４には、異なる特徴属性に関連した異なる重み付けモデル及び信頼性スコアが考慮される。

【００５７】

リンキングアルゴリズムの適用ステップ３０４の後に、リンクされた結果が関連特徴のクラスターに配置され、そしてステップ３０６において、リンクされたレコードのクラスターの返送の一部分として返送される。

【００５８】

図４は、図１を参照して上述した非構造化テキストにおいて特徴を曖昧性除去するシステム４００の一実施形態を例示する図である。このシステム４００は、インメモリデータベースをホストし、そして１つ以上のノードを含む。

【００５９】

一実施形態によれば、システム４００は、１つ以上のドキュメント内の特徴を曖昧性除去するため複数の特殊目的コンピュータモジュール４０１、４０２、４１１、４１２及び４１４（以下に述べる）のコンピュータインストラクションを実行する１つ以上のプロセッサを備えている。図４に示すように、ドキュメント入力モジュール４０１、４０２は、インターネットベースのソース及び／又はドキュメントの生のコーパスからドキュメントを受け取る。多数の新たなドキュメントがネットワーク接続４０４を通してドキュメント入力モジュール４０２へアップロードされる。それ故、ソースは、常時、新たな知識を得て、ユーザワークステーション４０６により更新され、そのような新たな知識は、スタティックな仕方で予めリンクされない。従って、評価されるべきドキュメントの数は、無限に増加する。

【００６０】

この評価は、ＭｅｍＤＢコンピュータ４０８を経て達成される。ＭｅｍＤＢ４０８は、高速の曖昧性除去プロセスを促進し、曖昧性除去プロセスをオンザフライで促進し、これは、ＭｅｍＤＢ４０８に貢献しようとする最新情報の受信を促進する。特徴をリンクするための種々の方法が使用され、これは、重み付けされたモデルを本質的に使用して、どのエンティティタイプが最も重要であるか決定し、どれがより大きな重みを有するか決定し、そして信頼性スコアに基づき、正しい特徴の抽出及び曖昧性除去がどれほどの信頼性で遂行されたか決定し、且つ正しい特徴が結果の特徴クラスターに向かうことを決定する。図４に示すように、より多くのシステムノードが並列に機能するほど、プロセスは、より効率的となる。

【００６１】

種々の実施形態によれば、新たなドキュメントがドキュメント入力モジュール４０１、４０２を経てネットワーク接続４０４を通してシステム４００に到着するとき、特徴抽出が抽出モジュール４１１を経て遂行され、次いで、特徴の曖昧性除去が新たなドキュメントにおいてＭｅｍＤＢ４０８の特徴曖昧性除去サブモジュール４１４を経て遂行される。ある実施形態では、新たなドキュメントの特徴曖昧性除去が遂行された後に、抽出された新たな特徴４１０は、リンクＯＴＦサブモジュール４１２を通過するためにＭｅｍＤＢに含まれ、ここで、特徴は、比較され及びリンクされ、そして曖昧性除去された特徴１１０の特徴ＩＤが質問からの結果としてユーザに返送される。特徴ＩＤに加えて、曖昧性除去された特徴を定義する結果の特徴クラスターが任意に返送されてもよい。

【００６２】

ＭｅｍＤＢコンピュータ４０８は、装置メインメモリにデータレコードを記憶するように構成されたデータベースマネージメントシステム（ＤＢＭＳ）（図示せず）により制御されるレコードにデータを記憶するデータベースであり、これは、データを「ディスク」メモリに記憶する従来のデータベース及びＤＢＭＳモジュールと対照的である。従来のデ

ィスクストレージは、装置のハードディスクへの読み取り及び書き込みコマンドをプロセッサ（ＣＰＵ）が実行することを要求し、従って、ＣＰＵがデータのためのメモリ位置を位置付け（即ち、シークし）及び検索するインストラクションを実行した後に、そのメモリ位置におけるデータとのある形式のオペレーションを遂行することを要求する。インメモリデータベースシステムは、メインメモリに入れられて適宜にアドレスされるデータにアクセスし、従って、ＣＰＵにより遂行されるインストラクションの数を軽減し、そしてハードディスクのデータをＣＰＵがシークするのに関連したシークタイムを排除する。

【００６３】

インメモリデータベースは、ノードの各リソース（例えば、メモリ、ディスク、プロセッサ）をアグリゲートするように構成された１つ以上のノードを含むコンピューティングシステムである分散型コンピューティングアーキテクチャーにおいて実施される。ここに開示されるように、インメモリデータベースをホストするコンピューティングシステムの実施形態は、１つ以上のノードの間でデータベースのデータレコードを分散しそして記憶する。ある実施形態では、これらのノードは、ノードの「クラスター」へと形成される。ある実施形態では、ノードのこれらクラスターは、データベース情報の部分又は「集合体」を記憶する。

【００６４】

種々の実施形態は、共起トピック、キーフレーズ、接近用語、イベント、ファクト及びトレンド人気指数のような二次特徴を記憶するように構成された進化する効率的にリンク可能な特徴知識ベースを使用するコンピュータ実行の特徴曖昧性除去技術を提供する。ここに開示する実施形態は、知識ベースに記憶された特徴に対して所与の抽出特徴を分析する上で役立つ関連二次特徴の次元に基づいて簡単な概念的距離尺度から精巧なグラフクラスター化解決策まで変化し得る種々様々なリンキングアルゴリズムを経て遂行される。加えて、それらの実施形態は、既存の特徴エントリの二次特徴を更新するだけでなく、知識ベースに追加できる新たな特徴を発見することでそれを拡張もする能力により既存の特徴知識ベースを進化させる解決策を導入することができる。

【００６５】

曖昧性除去解決策の実施形態は、トピックモデリング解決策を使用して、トピック推論としてモデリングされる自動重み付け（全ての二次特徴にわたる）リンキングプロセスを提供する。この自動重み付け型リンキングプロセスをサポートするため、それら実施形態は、多数のコンポーネント（二次特徴）を条件独立としてサポートできるマルチコンポーネントＬＤＡ（ＭＣ－ＬＤＡ）と称される新規なトピックモデリング解決策を構築するように従来のＬＤＡトピックモデリングを拡張する。又、モデリング解決策の実施形態は、トレーニング中にコンポーネントの重みを自動的に学習し、そしてそれを曖昧性除去に関する推論（リンキング）のために使用することができる。曖昧性除去のために導入されるＭＣ－ＬＤＡ解決策は、曖昧性除去精度を高めるために導入できる付加的な数の二次特徴のためにスケーリングすることができる。

【００６６】

図５は、上述した図４のシステム４００によって使用されるマルチコンポーネント条件独立のレイテントディリクレアロケーション（ＭＣ－ＬＤＡ）トピックコンピュータモデリング解決策の実施形態のグラフィック表示である。ここに示す実施形態では、各コンポーネントブロックは、例えば、図５に示すパラメータで初期化される図４のＭｅｍＤＢ４０８を経て実行される、知識ベースにわたる各二次特徴のモデリングを表す。

【００６７】

図６は、上述した図５に使用されるＭＣ－ＬＤＡトピックモデルのギブスサンプリング方程式の実施形態を示す。このサンプリング解決策の実施形態は、個々のコンポーネント（二次特徴）の重みを自動的に且つ効率的にトレーニングする上で図４のシステム４００の助けとなる。

【００６８】

図７は、例えば、図７に示すパラメータで初期化される図４のシステム４００のＭｅｍ

ＤＢ４０８を経て実行される、図５－６のＭＣ－ＬＤＡトピックモデルにおけるトレーニング及び推論のための確率論的変化推論アルゴリズムのコンピュータ実行の実施形態を示す。この推論方法の実施形態は、全ての二次特徴（当該ドキュメントから抽出された）を入力として取り上げそして重み付けされたトピックを出力として与えることにより、リンキング／曖昧性除去プロセスをトピック推論としてモデリングするように容易に適用される。これらの重み付けされたトピックは、次いで、記憶された特徴知識ベースエントリに対して類似性スコアを計算するのに使用できる。

【００６９】

図８は、ＭＣ－ＬＤＡトピックモデルに対するサンプルトピックを示すテーブルである。図８は、一実施形態により、例えば、図４のシステム４００のＭｅｍＤＢ４０８を経て実行される、モデルの各コンポーネントに対するトップスコア付け表面フォームを示す。

【００７０】

例＃１は、当該特徴（一次特徴）がフットボール選手のＪｏｈｎ　Ｄｏｅでありそしてユーザがｊｏｈｎ　Ｄｏｅについて言及するニュースの監視を希望する場合に、非構造化テキストにおける特徴を曖昧性除去する方法１００を適用するものである。ある実施形態によれば、Ｊｏｈｎ　Ｄｏｅについて述べるドキュメント入力１０２がネットワークにアップロードされる。ドキュメント入力１０２の特徴が抽出されて、ＭｅｍＤＢ４０８に含まれ、曖昧性除去されて、一次特徴（Ｊｏｈｎ　Ｄｏｅ）に関連した二次特徴のクラスターにリンクされ、そして同様の特徴の既存のクラスターと比較される。方法１００は、異なる特徴ＩＤ及び特徴ＩＤの関連クラスターを出力し、これは、Ｊｏｈｎ　Ｄｏｅに対する全ての関連二次特徴、例えば、エンジニアのＪｏｈｎ　Ｄｏｅ；教師のＪｏｈｎ　Ｄｏｅ；及びフットボール選手のＪｏｈｎ　Ｄｏｅ；を含む。同様の二次特徴を伴う他の一次特徴、例えば、ニックネーム又は省略名が考えられる。フットボール選手のＪｏｈｎ　Ｄｏｅと同じチームから、同じ年齢及び経験の「ＪＤ」フットボール選手は、同じ一次特徴と考えられる。それ故、フットボール選手のＪｏｈｎ　Ｄｏｅに関連した全てのドキュメントは、容易にアクセスすることができる。

【００７１】

例＃２は、一次特徴が画像である場合に、非構造化ドキュメントにおける特徴を曖昧性除去する方法１００を適用するものである。ある実施形態によれば、方法１００は、特徴の抽出１０４を含み、ここで、特徴は、とりわけ、縁及び形状のような一般的な属性であるか、或いはとりわけ、タンク、個人及び時計のような特定の属性である。例えば、新たな画像が入力され、ここで、画像は、特定の形状（例えば、方形、個人又は車の形状）のような二次特徴を有し、二次特徴が抽出されてＭｅｍＤＢ４０８に含まれ、ここで、同様の二次特徴を有する他の全ての画像の間で一致が見出される。ここに示す実施形態によれば、特徴は、画像のみを含み、即ちテキストは、特徴として含まれない。

【００７２】

例＃３は、一次特徴がイベントである場合に、非構造化テキストにおける特徴を曖昧性除去する方法１００を適用するものである。ある実施形態によれば、質問がなされたときに、方法１００は、ユーザが、とりわけ、地震、火災、又は伝染病の発生のようなイベントに関連した結果を受け取ることができるようにする。方法１００は、特徴の抽出１０４及び特徴の曖昧性除去１０８を遂行して、イベントに関連した特徴を見出すと共に、曖昧性除去された特徴１１０の特徴ＩＤを与える。

【００７３】

例＃４は、１つ以上のイベントの発生の予想がなされる場合の方法１００の実施形態である。ある実施形態によれば、ユーザは、オペレーションの前に当該特徴及びイベントを前もって指示し、それ故、当該イベントに関連した異なる特徴間のリンクが前もって確立される。関連特徴が高い発生数でネットワークに現われるとき、方法１００は、関連特徴の発生数増加に基づいて、当該事象が発生することを予想する。切迫したイベントが検出されると、ユーザに警報が送られる。例えば、タイからの保健省に対して仕事をするユーザは、デング熱の伝染病発生についての警報を受け取ることを選択する。例えば、ソーシ

ャルネットワークからの他のユーザ４０６がデング熱の兆候又は包括を含めたコメントを病院へアップロードするとき、方法１００は、ソーシャルネットワークからの全ての関連コメントを曖昧性除去し、そして関連情報を含めたユーザ４０６の数を考慮して、デング熱の伝染病発生が生じることを予想し、保健省の職員に警報する。それ故、保健省の職員は、付加的な形跡を得て、影響のある共同体への更なる対策を取り、伝染病が広がらないようにする。

【００７４】

例＃５は、一次特徴が地理的な場所の名前である場合の方法１００の適用である。一実施形態によれば、方法１００は、都市の名前を曖昧性除去するのに使用され、曖昧性除去サブモジュールにおいて二次特徴に異なるスコア付け重みが関連付けられる。例えば、方法１００は、Ｐａｒｉｓ、ＴｅｘａｓをＰａｒｉｓ、Ｆｒａｎｃｅから曖昧性除去するのに使用される。

【００７５】

例＃６は、一次特徴が、とりわけ、個人、イベント、又は会社に関連した感情であり、その感情が、とりわけ、個人、イベント、又は会社に関する肯定的又は否定的コメントであって、ソーシャルネットワークを含む適当なソースから供給される場合に、非構造化テキストにおける特徴を曖昧性除去する方法１００を適用するものである。ある実施形態によれば、方法１００は、会社が一般大衆の中で有している容認性を確認するために使用される。

【００７６】

例＃７は、特徴の信頼性スコアを高めるために人間の確認を含む方法１００の実施形態である。ある実施形態によれば、リンクＯＴＦプロセス３００（図４）は、ユーザにより支援され、ユーザは、曖昧性除去された特徴が正しく曖昧性除去されたかどうか指示し、そして２つの異なるクラスターが１つでなければならないかどうか指示し、これは、ユーザが知っている２つの異なる一次特徴が同じであるときに方法１００が（全ての特徴及びトピック共起情報を考慮して）何を指示するかを意味する。それ故、そのクラスターに関連した信頼性スコアが高くなり、従って、特徴が正しく曖昧性除去されたという確率が高くなる。

【００７７】

例＃８は、曖昧性除去プロセス２００及びリンクＯＴＦプロセス３００を使用する方法１００の実施形態である。この例では、リンキングアルゴリズムの適用３０４に使用されるリンキングアルゴリズムは、１０００ｍｓの期間内に０．８５より高い信頼性スコアを与えるように構成される。

【００７８】

例＃９は、曖昧性除去プロセス２００及びリンクＯＴＦプロセス３００を使用する方法１００の実施形態である。この例では、リンキングアルゴリズムの適用３０４に使用されるリンキングアルゴリズムは、３００ｍｓ以下の期間内に０．８０より高い信頼性スコアを与えるように構成される。この例に使用されるアルゴリズムは、例＃８に使用されるアルゴリズムに比して短い期間内に応答を与えるが、一般的に、低い信頼性スコアを返送する。

【００７９】

例＃１０は、曖昧性除去プロセス２００及びリンクＯＴＦプロセス３００を使用する方法１００の実施形態である。この例では、リンキングアルゴリズムの適用３０４に使用されるリンキングアルゴリズムは、一般的に３０００ｍｓを越える期間内に０．９０より高い信頼性スコアを与えるように構成される。この例に使用されるアルゴリズムは、例＃８に使用されるアルゴリズムにより返送されるものより一般的に大きな信頼性スコアをもつ応答を与えるが、著しく長い期間を一般的に要求する。

【００８０】

例＃１１は、複数のソースからのドキュメントの大きなコーパスにおいてｅディスカバリーを遂行するために非構造化テキストにおける特徴を曖昧性除去する方法１００の一例

である。複数のリソースからのドキュメントの大きなコーパスが与えられると、それらドキュメントにおける全ての特徴を曖昧性除去するための方法１００の適用は、コーパスにおいて全ての特徴を発見できるようにする。発見された特徴の集合体は、特徴に関連した全てのドキュメントの発見及び関連特徴の発見に更に使用することができる。

【００８１】

以上の方法の説明及びプロセスフロー図は、単なる例示として示されたもので、種々の実施形態のステップを、提示した順序で遂行しなければならないことを要求し又は意味することは意図されない。当業者に明らかなように、前記実施形態におけるステップは、任意の順序で遂行されてもよい。「次いで(then)」、「次に(next)」、等のワードは、ステップの順序を限定するものではなく、これらのワードは、単に、方法の説明を通して読者を誘導するのに使用されるだけである。プロセスフロー図は、オペレーションを一連のプロセスとして示すが、多数のオペレーションを並列に又は同時に遂行することもできる。加えて、オペレーションの順序は、再構成してもよい。プロセスは、方法、機能、手順、サブルーチン、サブプログラム、等に対応する。プロセスが機能に対応するとき、その終了は、コーリング機能又はメイン機能への機能の復帰に対応する。

【００８２】

ここに開示する実施形態に関連して述べた種々の例示的論理ブロック、モジュール、回路及びアルゴリズムステップは、電子的ハードウェア、コンピュータソフトウェア又はその両方の組み合わせとして具現化されてもよい。ハードウェア及びソフトウェアのこの互換性を明確に示すために、種々の例示的コンポーネント、ブロック、モジュール、回路、及びステップは、それらの機能に関して一般的に説明された。そのような機能がハードウェアとして具現化されるかソフトウェアとして具現化されるかは、システム全体に課せられる特定アプリケーション及び設計上の制約に依存する。当業者であれば、ここに述べた機能を特定アプリケーションごとに色々な仕方で具現化できるが、そのような具現化の判断は、本発明の範囲から逸脱すると解釈されてはならない。

【００８３】

コンピュータソフトウェアで具現化される実施形態は、ソフトウェア、ファームウェア、ミドルウェア、マイクロコード、ハードウェア記述言語、又はその組み合わせで具現化される。コードセグメント又はマシン実行可能なインストラクションは、手順、機能、サブプログラム、プログラム、ルーチン、サブルーチン、モジュール、ソフトウェアパッケージ、クラス、或いはインストラクション、データ構造体又はプログラムステートメントの組合せを表わす。コードセグメントは、情報、データ、アーギュメント、パラメータ又はメモリコンテンツを通し及び／又は受け取ることにより別のコードセグメント又はハードウェア回路に結合される。情報、アーギュメント、パラメータ、データ、等は、メモリ共有、メッセージ通過、トークン通過、ネットワーク送信、等を含む適当な手段を経て通され、転送され又は送信される。

【００８４】

これらのシステム及び方法を実施するのに使用される実際のソフトウェアコード又は特殊な制御ハードウェアは、本発明を限定するものではない。従って、システム及び方法のオペレーション及び振舞いは、ここでの記載に基づいてシステム及び方法を実施するようにソフトウェア及び制御ハードウェアを設計できることを理解して、特定のソフトウェアコードを参照せずに説明した。

【００８５】

ソフトウェアで実施されるときに、機能は、非一時的コンピュータ読み取り可能な又はプロセッサ読み取り可能なストレージ媒体に１つ以上のインストラクション又はコードとして記憶される。ここに開示する方法又はアルゴリズムのステップは、コンピュータ読み取り可能な又はプロセッサ読み取り可能なストレージ媒体に存在するプロセッサ実行可能なソフトウェアモジュールにおいて実施される。非一時的なコンピュータ読み取り可能な又はプロセッサ読み取り可能な媒体は、ある場所から別の場所へのコンピュータプログラムの転送を容易にするコンピュータストレージ媒体及び有形のストレージ媒体の両方を含

む。非一時的なプロセッサ読み取り可能なストレージ媒体は、コンピュータによりアクセスされる利用可能な媒体である。これに限定されないが、一例として、そのような非一時的なプロセッサ読み取り可能な媒体は、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、ＣＤ－ＲＯＭ又は他の光学ディスクストレージ、磁気ディスクストレージ又は他の磁気ストレージ装置、或いはインストラクション又はデータ構造体の形態で望ましいプログラムコードを記憶するのに使用され且つコンピュータ又はプロセッサによりアクセスされる他の有形のストレージ媒体を含む。ここで使用するディスク(disk & disc)とは、コンパクトディスク（ＣＤ）、レーザーディスク（登録商標）、光学ディスク、デジタル多様性ディスク（ＤＶＤ）、フロッピーディスク、及びブルーレイディスクを含み、ここで、ディスク(disk)は、通常、データを磁気的に再生するものであり、一方、ディスク(disc)は、データをレーザで光学的に再生するものである。前記の組み合わせも、コンピュータ読み取り可能な媒体の範囲内に包含される。加えて、方法又はアルゴリズムのオペレーションは、コンピュータプログラム製品に合体される非一時的プロセッサ読み取り可能な媒体及び／又はコンピュータ読み取り可能な媒体にコード及び／又はインストラクションの１つ又は組み合わせ或いはセットとして存在する。

【００８６】

技術の種々のコンポーネントは、分散型ネットワーク及び／又はインターネットの遠隔部分に、或いは専用のセキュア、アンセキュア及び／又は暗号化システム内に配置できることが明らかである。従って、システムのコンポーネントは、１つ以上の装置に結合するか、又はテレコミュニケーションネットワークのような分散型ネットワークの特定ノードに共通配置できることが明らかである。以上の説明から明らかなように、計算効率の理由で、システムのコンポーネントは、システムのオペレーションに影響することなく、分散型ネットワーク内の任意の位置に配置することができる。更に、それらのコンポーネントは、専用マシンに埋め込むこともできる。

【００８７】

更に、エレメントを接続する種々のリンクは、ワイヤード又はワイヤレスリンク又はその組み合わせ、或いは接続されたエレメントへ及びそこからデータを供給及び／又は通信することのできる他の既知の又は今後開発されるエレメントであることが明らかである。ここで使用するモジュールという語は、エレメントに関連した機能を遂行できる既知の又は今後開発されるハードウェア、ソフトウェア、ファームウェア、又はその組み合わせを指す。又、ここで使用する決定、計算及びコンピューティング、並びにその変形の語は、交換可能に使用され、そして任意のタイプの方法、プロセス、数学演算又は技術を包含する。

【００８８】

ここに開示する実施形態の前記説明は、当業者が本発明を実施又は利用できるようにするためになされたものである。これら実施形態に対する種々の変更は、当業者に容易に明らかであり、そしてここに定義する一般的な原理は、本発明の精神又は範囲から逸脱せずに他の実施形態に適用される。従って、本発明は、ここに示す実施形態に限定されるものではなく、特許請求の範囲並びにここに開示した原理及び新規な特徴に一致する最も広い範囲と調和されるべきである。

【００８９】

以上に述べた実施形態は、例示に過ぎない。当業者であれば、ここに述べた特定例に対して置き換えられ且つ依然として本発明の範囲内に入る多数の代替的コンポーネント及び実施形態が認識されよう。

【符号の説明】

【００９０】

４００：システム
４０１、４０２：ドキュメント入力モジュール
４０４：ネットワーク接続
４０６：ユーザワークステーション

４０８：ＭｅｍＤＢコンピュータ
４１０：抽出された新たな特徴
４１１：抽出ノード
４１２：リンクＯＴＦサブモジュール

【図１】

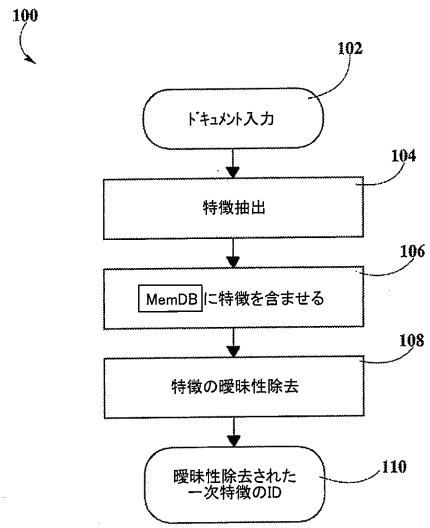


FIG. 1

【図２】

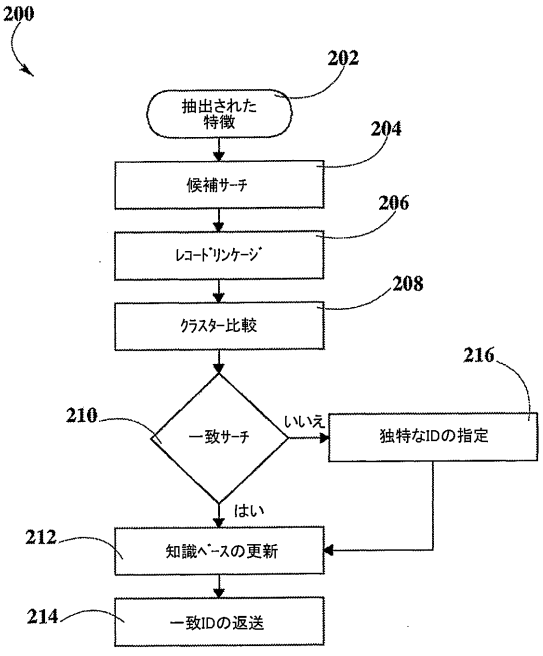


FIG. 2

【図３】

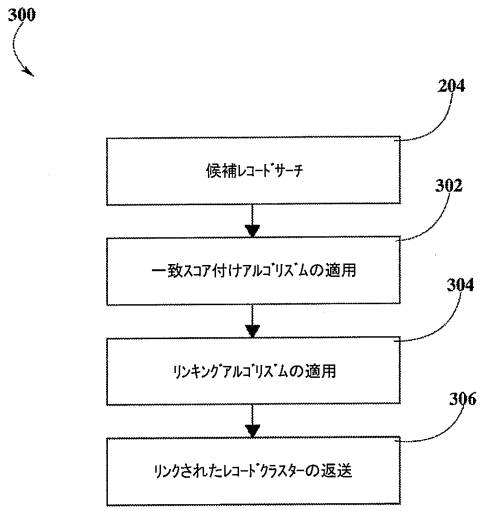


FIG. 3

【図４】

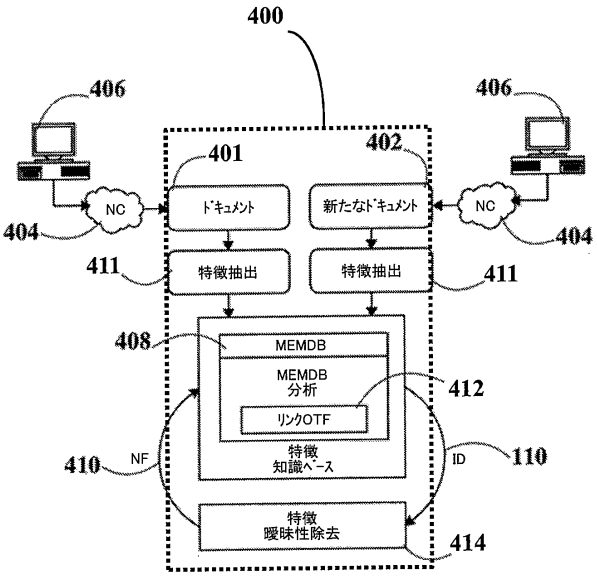
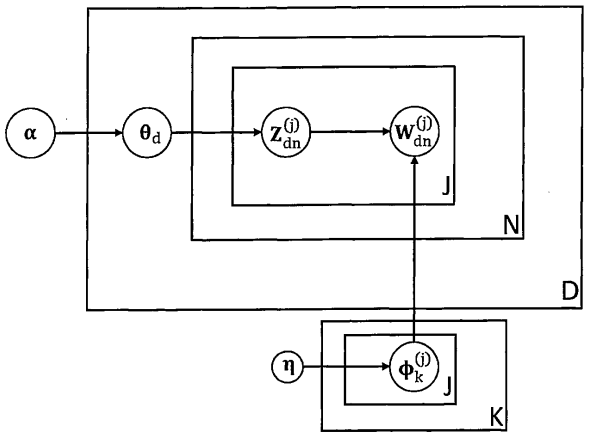


FIG. 4

【図５】



J = マルチドキュメントコンポーネントの数
$V^{(j)}$ = j番目のコンポーネントの語彙における用語の数
D = ドキュメントの数
N = ドキュメントの(コンポーネント)におけるトークンの数(実際にはj及びdの両方に依存する)
K = トピックの数
α = 混合比率でのハイパーパラメータ(対称的である場合にはKベクトル又はスカラー)
$\eta^{(j)}$ = 混合コンポーネントでのハイパーパラメータ(対称的である場合には$V^{(j)}$ベクトル又はスカラー)
θd = ドキュメントdに対するトピック混合比率
$\phi_k^{(j)}$ = k番目のトピックのj番目のコンポーネントに対する混合コンポーネント
$z_{d,n}^{(j)}$ = ドキュメントdのj番目のコンポーネントにおけるn番目のワードに対する
トピックを選択する混合インジケータ
$\omega_{d,n}^{(j)}$ = ドキュメントdのj番目のコンポーネントにおけるn番目のワードに対する項目インジケータ

FIG. 5

【図６】

$$p(z_i^{(j)}=k|w_i^{(j)}=t, z_{-i}^{(j)}, z^{(-j)}, w_{-i}^{(j)}, w^{(-j)}, \alpha, \eta) \propto \frac{n_{d,-i}^{(j,k)}+\alpha}{\sum_{k'=1}^{K}(n_{d,-i}^{(j,k')}+\alpha)} \cdot \frac{n_{k,-i}^{(j,t)}+\beta^{(j)}}{\sum_{t'=1}^{V_j}(n_{k,-i}^{(j,t')}+\beta^{(j)})}$$

$$\mathrm{E}\{\phi_{k,t}^{(j)}|z^{(j)}, w^{(j)}, \beta\} = \frac{n_k^{(j,t)}+\beta^{(j)}}{\sum_{t'=1}^{V_j}(n_k^{(j,t')}+\beta^{(j)})}$$

$$\mathrm{E}\{\theta_{d,k}|z^{(1)}, \ldots, z^{(J)}, \alpha\} = \frac{\sum_{j=1}^{J}(n_d^{(j,k)}+\alpha)}{\sum_{k=1}^{K}\sum_{j=1}^{J}(n_d^{(j,k)}+\alpha)}$$

J = マルチドキュメントコンポーネントの数
$V^{(j)}$ = j番目のコンポーネントの語彙における用語の数
K = トピックの数
α = 混合比率でのハイパーパラメータ(対称的である場合にはKベクトル又はスカラー)
η = 混合コンポーネントでのハイパーパラメータ(実際には、各々、長さV(j)のJ個のベクトル)
θd = ドキュメントdに対するトピック混合比率
$\phi_k^{(j)}$ = k番目のトピックのj番目のコンポーネントに対する混合コンポーネント
$z_{d,n}^{(j)}$ = ドキュメントdのj番目のコンポーネントにおけるn番目のワードに対するトピックを選択する混合インジケータ
$\omega_{d,n}^{(j)}$ = ドキュメントdのj番目のコンポーネントにおけるn番目のワードに対する項目インジケータ
$n_d^{(j,k)}$ = トピックkに指定されたドキュメントdのコンポーネントjにおける用語の数
$n_k^{(j,t)}$ = コーパスのコンポーネントjにおけるトピックkに用語tが指定された回数

FIG. 6

【図７】

アルゴリズム1 MC-TMのための確率論的変化推論

Set $\epsilon = 0.00001$
Define $\rho_t := (\tau_0 + t)^{-\kappa}$
Initialize λ randomly.
for $t = 0$ to ∞ **do**
  E-Step:
  Initialize $\gamma_{tk} = 1$. (The constant 1 is arbitrary.)
  repeat
    Set $\phi_{twk}^{(j)} \propto e^{\mathrm{E}_q[\log \theta_{tk}] + \mathrm{E}_q[\log \beta_{kw}^{(j)}]}$
    Set $\gamma_{tk} = \alpha + \sum_{j=1}^{J} \sum_{w=1}^{W^{(j)}} \phi_{twk}^{(j)} n_{tw}^{(j)}$
  until $\frac{1}{K} \sum_k |\Delta\gamma_{tk}| < \epsilon$
  M-Step:
  Compute $\tilde{\lambda}_{k,w}^{(j)} = \eta^{(j)} + D \cdot n_{t,w}^{(j)} \cdot \phi_{t,w,k}^{(j)}$
  Set $\lambda = (1 - \rho_t)\lambda + \rho_t \tilde{\lambda}$
end-for

α = ドキュメントトピック混合比率でのハイパーパラメータ
η = トピック用語混合比率でのハイパーパラメータ
$\phi_{d,n}$ = ドキュメントdにおけるトピック指定の多項式分布のパラメータ
γ_d = ドキュメントdに対するドキュメントトピック比率でのディリクレ分布のパラメータ
λ_k = トピックkに対するトピック用語比率でのディリクレ分布のパラメータ

FIG. 7

【図８】

トークン	album, music, song, released, songs, single, albums, love, tour, first . . .
キーフレーズ	album, song, music, rock, debut_album, studio_album, solo_album, edit, music_video, guitar . . .
組織	mtv, beatles, rolling_stone, youtube, emi, warner_bros, rca, official_charts_company, rovi_corporation, atlantic_records . . .
個人	britney_spears, bruce_springsteen, robin_thicke, bob_dylan, mariah_carey, john_lennon, jennifer_lopez, stevie_nicks, melissa_etheridge, dylan . . .
位置	us, united_states, uk, united_kingdom, united_states_of_america, u_s, australia, glastonbury . . .

FIG. 8

フロントページの続き

(74)代理人 100109070
弁理士 須田 洋之
(74)代理人 100109335
弁理士 上杉 浩
(74)代理人 100120525
弁理士 近藤 直樹
(74)代理人 100139712
弁理士 那須 威夫
(72)発明者 ライトナー スコット
アメリカ合衆国 ヴァージニア州 ２０１７５ リーズバーグ レッドヒル マナー コート ２２５９６
(72)発明者 ウェックザー フランツ
アメリカ合衆国 オハイオ州 ４５３７０ スプリング ヴァレー イースト センターヴィル ロード ３９４２
(72)発明者 ボッヅ サンジェイ
アメリカ合衆国 オハイオ州 ４５４４０ デイトン サマーセット パス ４４０８
(72)発明者 デイヴ ラケシュ
アメリカ合衆国 オハイオ州 ４５４４０ デイトン アマースト ベンド ３７６
(72)発明者 フラッグ ロバート
アメリカ合衆国 メイン州 ０４１０１ ポートランド イースタン プロムナード ６ アパートメント １

審査官 齊藤 貴孝

(56)参考文献 米国特許出願公開第２００２／０１６５８４７（ＵＳ，Ａ１）
米国特許出願公開第２０１３／０２９０６６５（ＵＳ，Ａ１）
特開２０１３－２３９１６２（ＪＰ，Ａ）
特開２００３－１５０４４２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／３０