(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6833999号
(P6833999)
(45)発行日　令和3年2月24日(2021.2.24)　(24)登録日　令和3年2月5日(2021.2.5)

(51)Int.Cl.　　F I
G06N 20/00 (2019.01)　G06N 20/00
G06N 5/02 (2006.01)　G06N 5/02
G06F 16/906 (2019.01)　G06F 16/906

請求項の数 36　(全 29 頁)

(21)出願番号　特願2019-527560(P2019-527560)
(86)(22)出願日　平成28年11月23日(2016.11.23)
(65)公表番号　特表2019-536168(P2019-536168A)
(43)公表日　令和1年12月12日(2019.12.12)
(86)国際出願番号　PCT/CA2016/000288
(87)国際公開番号　WO2018/094496
(87)国際公開日　平成30年5月31日(2018.5.31)
審査請求日　令和1年11月20日(2019.11.20)

(73)特許権者　511050240
プライマル　フュージョン　インコーポレイテッド
カナダ国　エヌ２ジー　１ビー９　オンタリオ，キッチナー，キング　ストリート　ウエスト　６０５－３０５
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100091214
弁理士　大貫　進介

最終頁に続く

(54)【発明の名称】機械学習分類器とともに知識表現を使用するためのシステム及び方法

(57)【特許請求の範囲】
【請求項１】
機械学習分類器のためのトレーニングデータを生成する方法であって、
関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、
コンテンツ項目の第１のセットを受け取ることであって、前記第１のセットは、ラベルを有しない１つ以上のコンテンツ項目を含み、前記ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する、受け取ることと、
前記第１のセットのうちの１つ以上のそれぞれのコンテンツ項目についての１つ以上のスコアを決定することであって、それぞれのコンテンツ項目についてのスコアは、前記知識表現と該それぞれのコンテンツ項目の内容とに基づく、決定することと、
前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に関連付けられたスコアに基づいて、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目にラベルを割り当てることにより、前記機械学習分類器のための前記トレーニングデータを生成することと、
を含む、方法。
【請求項２】
前記関心オブジェクトの内容に基づいて、前記知識表現を合成すること
をさらに含む、請求項１に記載の方法。

【請求項３】
前記合成することは、
前記少なくとも１つの概念及び／又は２つ以上の概念間の関係を生成することであって、前記概念及び／又は関係は、前記関心オブジェクトにおいて列挙されていない、生成すること
をさらに含む、請求項２に記載の方法。
【請求項４】
前記知識表現は、前記少なくとも１つの概念に関連付けられた重みを含む、請求項１に記載の方法。
【請求項５】
前記それぞれのコンテンツ項目についての前記スコアは、前記知識表現における前記少なくとも１つの概念と前記それぞれのコンテンツ項目の前記内容との共通部分に基づく、請求項１に記載の方法。
【請求項６】
前記関心オブジェクトは、トピック、ツイート、ウェブページ、ウェブサイト、文書、文書の集合、文書タイトル、メッセージ、広告、及び／又は検索クエリを含む、請求項１に記載の方法。
【請求項７】
前記コンテンツ項目の第１のセットに割り当てられたラベルと前記コンテンツ項目の第１のセットに関連付けられた１つ以上の特徴とに基づいて、１つ以上の分類されていないコンテンツ項目についてのラベルを予測するように、分類器をトレーニングすること
をさらに含む、請求項１に記載の方法。
【請求項８】
コンテンツ項目の第２のセットを受け取ることであって、前記第２のセットは、ラベルを有しない１つ以上のコンテンツ項目を含む、受け取ることと、
前記分類器により、前記第２のセットのうちの前記コンテンツ項目の１つ以上に、前記第２のセットのうちのそれぞれの１つ以上のコンテンツ項目に関連付けられた１つ以上の特徴に基づいて、ラベルを割り当てることと、
をさらに含む、請求項７に記載の方法。
【請求項９】
前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に前記ラベルを割り当てることは、
前記第１のセットのうちのそれぞれのコンテンツ項目についてのスコアが予め定められた閾値を超えていることに基づいて、前記ラベルを割り当てること
を含む、請求項１に記載の方法。
【請求項１０】
前記ラベルは、「推奨する」又は「推奨しない」のうちの１つである、請求項１に記載の方法。
【請求項１１】
前記１つ以上の特徴は、タイトル、長さ、作者、単語頻度、及び／又は逆文書頻度のうちの少なくとも１つを含む、請求項７に記載の方法。
【請求項１２】
前記１つ以上の特徴は、前記知識表現の１つ以上の属性を含む、請求項７に記載の方法。
【請求項１３】
機械学習分類器のためのトレーニングデータを生成するためのシステムであって、
関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、
コンテンツ項目の第１のセットを受け取ることであって、前記第１のセットは、ラベ

ルを有しない１つ以上のコンテンツ項目を含み、前記ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する、受け取ることと、
前記第１のセットのうちの１つ以上のそれぞれのコンテンツ項目についての１つ以上のスコアを決定することであって、それぞれのコンテンツ項目についてのスコアは、前記知識表現と該それぞれのコンテンツ項目の内容とに基づく、決定することと、
前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に関連付けられたスコアに基づいて、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目にラベルを割り当てることにより、前記機械学習分類器のための前記トレーニングデータを生成することと、
を含む方法を実行するよう構成されている少なくとも１つのプロセッサ
を有する、システム。
【請求項１４】
前記方法は、
前記関心オブジェクトの内容に基づいて、前記知識表現を合成すること
をさらに含む、請求項１３に記載のシステム。
【請求項１５】
前記合成することは、
前記少なくとも１つの概念及び／又は２つ以上の概念間の関係を生成することであって、前記概念及び／又は関係は、前記関心オブジェクトにおいて列挙されていない、生成すること
をさらに含む、請求項１４に記載のシステム。
【請求項１６】
前記知識表現は、前記少なくとも１つの概念に関連付けられた重みを含む、請求項１３に記載のシステム。
【請求項１７】
前記それぞれのコンテンツ項目についての前記スコアは、前記知識表現における前記少なくとも１つの概念と前記それぞれのコンテンツ項目の前記内容との共通部分に基づく、請求項１３に記載のシステム。
【請求項１８】
前記関心オブジェクトは、トピック、ツイート、ウェブページ、ウェブサイト、文書、文書の集合、文書タイトル、メッセージ、広告、及び／又は検索クエリを含む、請求項１３に記載のシステム。
【請求項１９】
前記方法は、
前記コンテンツ項目の第１のセットに割り当てられたラベルと前記コンテンツ項目の第１のセットに関連付けられた１つ以上の特徴とに基づいて、１つ以上の分類されていないコンテンツ項目についてのラベルを予測するように、分類器をトレーニングすること
をさらに含む、請求項１３に記載のシステム。
【請求項２０】
前記方法は、
コンテンツ項目の第２のセットを受け取ることであって、前記第２のセットは、ラベルを有しない１つ以上のコンテンツ項目を含む、受け取ることと、
前記分類器により、前記第２のセットのうちの前記コンテンツ項目の１つ以上に、前記第２のセットのうちのそれぞれの１つ以上のコンテンツ項目に関連付けられた１つ以上の特徴に基づいて、ラベルを割り当てることと、
をさらに含む、請求項１９に記載のシステム。
【請求項２１】
前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に前記ラベルを割り当てることは、
前記第１のセットのうちのそれぞれのコンテンツ項目についてのスコアが予め定められ

た閾値を超えていることに基づいて、前記ラベルを割り当てること
を含む、請求項１３に記載のシステム。
【請求項２２】
前記ラベルは、「推奨する」又は「推奨しない」のうちの１つである、請求項１３に記載のシステム。
【請求項２３】
前記１つ以上の特徴は、タイトル、長さ、作者、単語頻度、及び／又は逆文書頻度のうちの少なくとも１つを含む、請求項１９に記載のシステム。
【請求項２４】
前記１つ以上の特徴は、前記知識表現の１つ以上の属性を含む、請求項１９に記載のシステム。
【請求項２５】
プロセッサ実行可能な命令を記憶している少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体であって、前記命令は、少なくとも１つのプロセッサにより実行されたときに、前記少なくとも１つのプロセッサに、機械学習分類器のためのトレーニングデータを生成する方法を実行させ、前記方法は、
関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、
コンテンツ項目の第１のセットを受け取ることであって、前記第１のセットは、ラベルを有しない１つ以上のコンテンツ項目を含み、前記ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する、受け取ることと、
前記第１のセットのうちの１つ以上のそれぞれのコンテンツ項目についての１つ以上のスコアを決定することであって、それぞれのコンテンツ項目についてのスコアは、前記知識表現と該それぞれのコンテンツ項目の内容とに基づく、決定することと、
前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に関連付けられたスコアに基づいて、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目にラベルを割り当てることにより、前記機械学習分類器のための前記トレーニングデータを生成することと、
を含む、少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。
【請求項２６】
前記方法は、
前記関心オブジェクトの内容に基づいて、前記知識表現を合成すること
をさらに含む、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。
【請求項２７】
前記合成することは、
前記少なくとも１つの概念及び／又は２つ以上の概念間の関係を生成することであって、前記概念及び／又は関係は、前記関心オブジェクトにおいて列挙されていない、生成すること
をさらに含む、請求項２６に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。
【請求項２８】
前記知識表現は、前記少なくとも１つの概念に関連付けられた重みを含む、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。
【請求項２９】
前記それぞれのコンテンツ項目についての前記スコアは、前記知識表現における前記少なくとも１つの概念と前記それぞれのコンテンツ項目の前記内容との共通部分に基づく、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。
【請求項３０】

前記関心オブジェクトは、トピック、ツイート、ウェブページ、ウェブサイト、文書、文書の集合、文書タイトル、メッセージ、広告、及び／又は検索クエリを含む、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【請求項３１】

前記方法は、

前記コンテンツ項目の第１のセットに割り当てられたラベルと前記コンテンツ項目の第１のセットに関連付けられた１つ以上の特徴とに基づいて、１つ以上の分類されていないコンテンツ項目についてのラベルを予測するように、分類器をトレーニングすること

をさらに含む、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【請求項３２】

前記方法は、

コンテンツ項目の第２のセットを受け取ることであって、前記第２のセットは、ラベルを有しない１つ以上のコンテンツ項目を含む、受け取ることと、

前記分類器により、前記第２のセットのうちの前記コンテンツ項目の１つ以上に、前記第２のセットのうちのそれぞれの１つ以上のコンテンツ項目に関連付けられた１つ以上の特徴に基づいて、ラベルを割り当てることと、

をさらに含む、請求項３１に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【請求項３３】

前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に前記ラベルを割り当てることは、

前記第１のセットのうちのそれぞれのコンテンツ項目についてのスコアが予め定められた閾値を超えていることに基づいて、前記ラベルを割り当てること

を含む、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【請求項３４】

前記ラベルは、「推奨する」又は「推奨しない」のうちの１つである、請求項２５に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【請求項３５】

前記１つ以上の特徴は、タイトル、長さ、作者、単語頻度、及び／又は逆文書頻度のうちの少なくとも１つを含む、請求項３１に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【請求項３６】

前記１つ以上の特徴は、前記知識表現の１つ以上の属性を含む、請求項３１に記載の少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

本開示に記載の教示は、機械学習及び情報検索の分野に関する。詳細には、本出願において開示される教示は、情報を分類し、情報を分類するための分類器モデルをトレーニングし、且つ／又は、分類器の結果を使用して知識表現を洗練させるために、知識表現として表現される１つ以上のデータセットに関連する情報を使用するための方法の、デジタル情報システム環境における展開に関する。

【背景技術】

【０００２】

情報技術は、テキスト、オーディオ、ビデオ、及び任意の適切な他のタイプの情報等の様々なタイプの情報をユーザに提供するためにしばしば使用される。いくつかの場合において、情報は、ユーザが行ったアクションに応じて、ユーザに提供される。例えば、情報は、ユーザにより入力された検索クエリに応じて、又は、ユーザが電子メールアラート又

は電子ニュースレター等のコンテンツをサブスクライブしたことに応じて、ユーザに提供されることがある。他の場合において、情報は、ユーザがそのような情報を明確に要求していなくても、ユーザに提供又は「プッシュ」される。例えば、広告又は勧誘が、ユーザに時折提示されることがある。

【０００３】

情報技術を介してユーザに提供され得る膨大なコンテンツ及び情報が存在する。実際、インターネット、ワールドワイドウェブ（ＷＷＷ）、及び任意の他の適切な情報提供ソースを介して利用可能な膨大な量の情報を理由として、且つ、利用可能な情報が、独立して所有及び運営される膨大な数のネットワーク及びサーバにわたって分散されるので、ユーザにとって関心がある情報を特定することは課題を伴う。関心がある情報が大規模なプライベートネットワークに分散されている場合にも同様の課題がある。

【発明の概要】

【０００４】

本開示において提示される発明のコンセプトが、それぞれが１つ以上のコンセプトを示す複数の異なる実施形態において示されているが、概して、これらのコンセプトは、相互に排他的ではなく、示されていない場合にも組み合わせて使用され得ることを理解されたい。

【０００５】

いくつかの実施形態は、機械学習分類器のためのトレーニングデータを生成する方法であって、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、コンテンツ項目の第１のセットを受け取ることであって、前記第１のセットは、ラベルを有しない１つ以上のコンテンツ項目を含み、前記ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する、受け取ることと、前記第１のセットのうちの１つ以上のそれぞれのコンテンツ項目についての１つ以上のスコアを決定することであって、それぞれのコンテンツ項目についてのスコアは、前記知識表現と該それぞれのコンテンツ項目の内容とに基づく、決定することと、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に関連付けられたスコアに基づいて、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目にラベルを割り当てることにより、前記機械学習分類器のための前記トレーニングデータを生成することと、を含む、方法を提供する。

【０００６】

いくつかの実施形態は、機械学習分類器のためのトレーニングデータを生成するためのシステムであって、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、コンテンツ項目の第１のセットを受け取ることであって、前記第１のセットは、ラベルを有しない１つ以上のコンテンツ項目を含み、前記ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する、受け取ることと、前記第１のセットのうちの１つ以上のそれぞれのコンテンツ項目についての１つ以上のスコアを決定することであって、それぞれのコンテンツ項目についてのスコアは、前記知識表現と該それぞれのコンテンツ項目の内容とに基づく、決定することと、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に関連付けられたスコアに基づいて、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目にラベルを割り当てることにより、前記機械学習分類器のための前記トレーニングデータを生成することと、を含む方法を実行するよう構成されている少なくとも１つのプロセッサを有する、システムを提供する。

【０００７】

いくつかの実施形態は、プロセッサ実行可能な命令を記憶している少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体であって、前記命令は、少なくとも１つのプロセッサにより実行されたときに、前記少なくとも１つのプロセッサに、機械学習分類

器のためのトレーニングデータを生成する方法を実行させ、前記方法は、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、コンテンツ項目の第１のセットを受け取ることであって、前記第１のセットは、ラベルを有しない１つ以上のコンテンツ項目を含み、前記ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する、受け取ることと、前記第１のセットのうちの１つ以上のそれぞれのコンテンツ項目についての１つ以上のスコアを決定することであって、それぞれのコンテンツ項目についてのスコアは、前記知識表現と該それぞれのコンテンツ項目の内容とに基づく、決定することと、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目に関連付けられたスコアに基づいて、前記第１のセットのうちの前記１つ以上のそれぞれのコンテンツ項目にラベルを割り当てることにより、前記機械学習分類器のための前記トレーニングデータを生成することと、を含む、少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体を提供する。

【０００８】

いくつかの実施形態は、機械学習分類器により、少なくとも１つのラベル付けされていないコンテンツ項目を分類する方法であって、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、トレーニングデータを受け取ることであって、前記トレーニングデータは、各コンテンツ項目を１つ以上のカテゴリに分類するラベルを有する１つ以上のラベル付けされたコンテンツ項目の第１のセットを含む、受け取ることと、前記知識表現から導出された１つ以上の属性に基づく、前記ラベル付けされたコンテンツ項目の少なくとも１つの特徴を用いて、前記機械学習分類器をトレーニングすることと、前記機械学習分類器により、前記知識表現から導出された前記１つ以上の属性に基づく、前記少なくとも１つのラベル付けされていないコンテンツ項目の少なくとも１つの特徴を用いて、前記少なくとも１つのラベル付けされていないコンテンツ項目を前記１つ以上のカテゴリに分類することと、を含む、方法を提供する。

【０００９】

いくつかの実施形態は、機械学習分類器により、少なくとも１つのラベル付けされていないコンテンツ項目を分類するためのシステムであって、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、トレーニングデータを受け取ることであって、前記トレーニングデータは、各コンテンツ項目を１つ以上のカテゴリに分類するラベルを有する１つ以上のラベル付けされたコンテンツ項目の第１のセットを含む、受け取ることと、前記知識表現から導出された１つ以上の属性に基づく、前記ラベル付けされたコンテンツ項目の少なくとも１つの特徴を用いて、前記機械学習分類器をトレーニングすることと、前記機械学習分類器により、前記知識表現から導出された前記１つ以上の属性に基づく、前記少なくとも１つのラベル付けされていないコンテンツ項目の少なくとも１つの特徴を用いて、前記少なくとも１つのラベル付けされていないコンテンツ項目を前記１つ以上のカテゴリに分類することと、を含む方法を実行するよう構成されている少なくとも１つのプロセッサを有する、システムを提供する。

【００１０】

いくつかの実施形態は、プロセッサ実行可能な命令を記憶している少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体であって、前記命令は、少なくとも１つのプロセッサにより実行されたときに、前記少なくとも１つのプロセッサに、機械学習分類器により、少なくとも１つのラベル付けされていないコンテンツ項目を分類する方法を実行させ、前記方法は、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、トレー

ニングデータを受け取ることであって、前記トレーニングデータは、各コンテンツ項目を１つ以上のカテゴリに分類するラベルを有する１つ以上のラベル付けされたコンテンツ項目の第１のセットを含む、受け取ることと、前記知識表現から導出された１つ以上の属性に基づく、前記ラベル付けされたコンテンツ項目の少なくとも１つの特徴を用いて、前記機械学習分類器をトレーニングすることと、前記機械学習分類器により、前記知識表現から導出された前記１つ以上の属性に基づく、前記少なくとも１つのラベル付けされていないコンテンツ項目の少なくとも１つの特徴を用いて、前記少なくとも１つのラベル付けされていないコンテンツ項目を前記１つ以上のカテゴリに分類することと、を含む、少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体を提供する。

【００１１】

いくつかの実施形態は、機械学習分類器に基づいて知識表現を変更する方法であって、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、検証データを受け取ることであって、前記検証データは、各コンテンツ項目を、前記関心オブジェクトと関連性があることが既知である第１のカテゴリと、前記関心オブジェクトと関連性がないことが既知である第２のカテゴリと、を含む１つ以上のカテゴリに分類するラベルを有する１つ以上のラベル付けされたコンテンツ項目の第１のセットを含む、受け取ることと、前記知識表現から導出された少なくとも１つの属性を、特徴として使用する機械学習分類器により、前記１つ以上のラベル付けされたコンテンツ項目の各々を、ａ）前記関心オブジェクトと関連性があるもの、及び、ｂ）前記関心オブジェクトと関連性がないもののうちの一方として、予測することと、前記第１のセットの各コンテンツ項目についての前記機械学習分類器による前記予測と、各コンテンツ項目のラベルと、の比較に基づいて、前記知識表現を変更することと、を含む、方法を提供する。

【００１２】

いくつかの実施形態は、機械学習分類器に基づいて知識表現を変更するためのシステムであって、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、検証データを受け取ることであって、前記検証データは、各コンテンツ項目を、前記関心オブジェクトと関連性があることが既知である第１のカテゴリと、前記関心オブジェクトと関連性がないことが既知である第２のカテゴリと、を含む１つ以上のカテゴリに分類するラベルを有する１つ以上のラベル付けされたコンテンツ項目の第１のセットを含む、受け取ることと、前記知識表現から導出された少なくとも１つの属性を、特徴として使用する機械学習分類器により、前記１つ以上のラベル付けされたコンテンツ項目の各々を、ａ）前記関心オブジェクトと関連性があるもの、及び、ｂ）前記関心オブジェクトと関連性がないもののうちの一方として、予測することと、前記第１のセットの各コンテンツ項目についての前記機械学習分類器による前記予測と、各コンテンツ項目のラベルと、の比較に基づいて、前記知識表現を変更することと、を含む方法を実行するよう構成されている少なくとも１つのプロセッサを有する、システムを提供する。

【００１３】

いくつかの実施形態は、プロセッサ実行可能な命令を記憶している少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体であって、前記命令は、少なくとも１つのプロセッサにより実行されたときに、前記少なくとも１つのプロセッサに、機械学習分類器に基づいて知識表現を変更する方法を実行させ、前記方法は、関心オブジェクトに基づく、非一時的なコンピュータ読み取り可能なデータ構造として符号化された知識表現を受け取ることであって、前記知識表現は、少なくとも１つの概念及び／又は２つ以上の概念間の関係を含む、受け取ることと、検証データを受け取ることであって、前記検証データは、各コンテンツ項目を、前記関心オブジェクトと関連性があることが既知である第１のカテゴリと、前記関心オブジェクトと関連性がないことが既知である第２のカテゴリと、

を含む１つ以上のカテゴリに分類するラベルを有する１つ以上のラベル付けされたコンテンツ項目の第１のセットを含む、受け取ることと、前記知識表現から導出された少なくとも１つの属性を、特徴として使用する機械学習分類器により、前記１つ以上のラベル付けされたコンテンツ項目の各々を、ａ）前記関心オブジェクトと関連性があるもの、及び、ｂ）前記関心オブジェクトと関連性がないもののうちの一方として、予測することと、前記第１のセットの各コンテンツ項目についての前記機械学習分類器による前記予測と、各コンテンツ項目のラベルと、の比較に基づいて、前記知識表現を変更することと、を含む、少なくとも１つの非一時的なコンピュータ読み取り可能な記憶媒体を提供する。

【００１４】

上記は、請求項により定められる本発明の非限定的な概要であり、この概要は、必ずしも、各請求項の主題を記載しているわけではなく、各請求項は、全ての実施形態ではないが１つ以上の実施形態に関連することを理解されたい。

【図面の簡単な説明】

【００１５】

添付の図面は、縮尺通りに描かれることを意図していない。実用的であるとき、類似の要素は、同一又は類似の参照符号により特定される。明確にするために、全てのコンポーネントが、全ての図面においてラベル付けされているとは限らない。

【図１】本発明のいくつかの実施形態を実施する際に使用される例示的なコンピューティングシステムを示すブロック図。

【図２】本発明のいくつかの実施形態に従った、教師あり学習技術のための例示的なワークフローを示すフローチャート。

【図３】本発明のいくつかの実施形態に従った、関心オブジェクトに基づいて知識表現を合成するための例示的なワークフローを示すフローチャート。

【図４】本発明のいくつかの実施形態に従った、例示的な関心オブジェクトの図。

【図５】本発明のいくつかの実施形態に従った、関心オブジェクトに基づいてトピック統一資源識別子を抽出するための例示的なワークフローの図。

【図６】いくつかの実施形態に従った、関心オブジェクトから抽出されたトピック統一資源識別子に基づいて知識表現を合成するための例示的なワークフローの図。

【図７】本発明のいくつかの実施形態に従った、コンテンツ項目をラベル付けするための例示的な方法を示すフローチャート。

【図８Ａ】本発明のいくつかの実施形態に従った、分類器モデルをトレーニングするための例示的なワークフローの図。

【図８Ｂ】本発明のいくつかの実施形態に従った、図８Ａの分類器モデルを使用してコンテンツ項目をラベル付けするための例示的なワークフローの図。

【図９】関心オブジェクトに基づく知識表現から導出された特徴に基づいてコンテンツ項目を分析するための例示的なワークフローの図。

【図１０Ａ】本発明のいくつかの実施形態に従った、関心オブジェクトから導出された知識表現からの特徴を使用して、分類器モデルをトレーニングするための例示的なワークフローの図。

【図１０Ｂ】本発明のいくつかの実施形態に従った、図１０Ａの分類器モデルを使用してコンテンツ項目をラベル付けするための例示的なワークフローの図。

【図１１】本発明のいくつかの実施形態に従った、コンテンツ項目に関連付けられた、属性に基づく特徴と、コンテンツ項目の既知の関連性と、分類器による予測と、を示す例示的な表の図。

【図１２】本発明のいくつかの実施形態に従った、コンテンツ項目に関連付けられた、属性に基づく特徴と、コンテンツ項目の既知の関連性と、第１の分類器及び第２の分類器による予測と、を示す例示的な表の図。

【図１３】本発明のいくつかの実施形態に従った、複数の分類器モデルによりなされた予測の結果を示す例示的な表の図。

【図１４Ａ】本発明のいくつかの実施形態に従った、分類器により予測されたラベルの精

度を示す例示的な表の図。
【図１４Ｂ】本発明のいくつかの実施形態に従った、分類器の精度のグラフィカル図。
【図１５Ａ】本発明のいくつかの実施形態に従った、変更された知識表現に関連付けられた属性の値と、分類器により予測されたラベルの精度と、を示す例示的な表の図。
【図１５Ｂ】本発明のいくつかの実施形態に従った、図１５Ａの分類器の精度のグラフィカル図。
【図１６】本発明のいくつかの実施形態に従った、分類器の結果に基づいて知識表現を変更するための例示的なフローチャートの図。
【発明を実施するための形態】
【００１６】
デジタル情報システムを介してアクセス可能な膨大な量のコンテンツは、情報検索のいくつかの課題を与えている。１つの難しい問題は、コンテンツの大きなセットにおけるどのような情報が、関連性がない情報でユーザを圧倒することなくこのような情報がユーザに提示され得るように、ユーザにとって関心があり得るかをどのように判別するかである。さらに、別の難しい問題は、関心があり得る情報を計算効率的にどのように識別するかである。
【００１７】
本開示に記載のシステム及び方法は、様々な形で実施可能である。図１は、本開示に記載の本発明のいくつかの実施形態を実施する際に使用される例示的なコンピューティングシステムを示すブロック図である。
【００１８】
コンピューティングデバイス１００は、１つ以上のプロセッサ（例えばマイクロプロセッサ）１１０及び１つ以上の非一時的な有形のコンピュータ読み取り可能な記憶媒体（例えばメモリ１２０）を含み得る。メモリ１２０は、非一時的な有形のコンピュータ読み取り可能な記憶媒体に、本開示に記載の機能のいずれかを実装するコンピュータ命令を記憶することができる。１つ以上のプロセッサ１１０は、メモリ１２０に接続され、そのようなコンピュータ命令を実行して機能を実現及び実行させることができる。コンピューティングデバイス１００はまた、ネットワーク入力／出力（Ｉ／Ｏ）インタフェース１３０を含み得、コンピューティングデバイス１００は、ネットワークＩ／Ｏインタフェース１３０を介して、他のコンピュータと（例えばネットワークを介して）通信することができる。いくつかの実施形態において、コンピューティングデバイスはまた、１つ以上のユーザＩ／Ｏインタフェース１４０を含み得、コンピューティングデバイスは、１つ以上のユーザＩ／Ｏインタフェース１４０を介して、ユーザに出力を提供し、ユーザから入力を受け取ることができる。ユーザＩ／Ｏインタフェース１４０は、キーボード、マウス、マイクロフォン、ディスプレイデバイス（例えばモニタ又はタッチスクリーン）、スピーカ、カメラ、及び／又は、様々な他のタイプのＩ／Ｏデバイス等のデバイスを含み得る。
【００１９】
本発明のいくつかの実施形態は、分類器をトレーニングするために、機械学習アルゴリズム、特に教師あり学習、を用いる。教師あり学習は、ラベル付けされたトレーニングデータから関数を推論するタスクを含む、機械学習の一分野である。このようなトレーニングデータは、トレーニング例のセットを含む。教師あり学習において、各トレーニング例は、オブジェクト（例えばコンテンツ項目）及び（例えば、コンテンツ項目を１つ以上のカテゴリに分類し得る）ラベルを含み得る。教師あり学習アルゴリズムは、トレーニングデータを分析し、推論される関数（例えば分類器モデル）をもたらすことができ、その後、これは、ラベル付けされていないコンテンツ項目についてのラベルを予測するために（又は、ラベル付けされていないコンテンツ項目を分類するために）使用され得る。
【００２０】
例えば、教師あり学習を用いてトレーニングされた機械学習分類器を使用して、ラベル付けされていないコンテンツ項目にラベルを割り当てることができる。しかしながら、分類器をトレーニングするためには、かなりの量のトレーニングデータが必要とされる。い

くつかの実施形態において、トレーニングデータは、各コンテンツ項目をカテゴリに分類するラベルを有するコンテンツ項目を含む。分類器をトレーニングするのに十分なトレーニングデータがなければ、機械学習アルゴリズムが、正確な結果をもたらすモデルを生成することは難しい。

【００２１】

図２は、本発明のいくつかの実施形態に従った、教師あり学習技術のための例示的なワークフローを示すフローチャートである。このプロセスは、未処理データ２１０で始まる。未処理データ２１０は、１つ以上のコンテンツ項目を含み得る。機械学習目的に有用であるために、未処理データ２１０は、未処理データ２１０における各コンテンツ項目にラベルを割り当てることにより、ラベルを有する特徴付けデータ２２０に変換される。ラベルは、実際上２値であってよい（例えば、「イエス」ラベル又は「ノー」ラベル、あるいは、「１」ラベル又は「０」ラベル）。しかしながら、３つ以上の可能なラベルが存在してもよい。ラベルは、任意の数の方法で、未処理データ２１０に割り当てられ得る。例えば、人間のオペレータが、手作業で、未処理データ２１０における各項目をレビューし、ラベルを割り当て得る。

【００２２】

ラベル付けされると、特徴付けデータ２２０は、トレーニングセット２３０と検証セット２４０とに分割され得る。特徴付けデータは、トレーニングセット２３０と検証セット２４０との間で任意の比率で分割され得る。次いで、トレーニングセット２３０は、モデル２５０を構築する（以下において、トレーニングすると呼ばれる）ための基礎として使用される。モデル２５０は、コンテンツ項目についての既知のラベルとそのコンテンツ項目の様々な特徴との間の相関から推論される関数として概念化され得る。モデル２５０の目的のうちの１つは、トレーニングセット２３０に割り当てられたラベル及びトレーニングセット２３０に関連付けられた特徴と整合するように、ラベル付けされていないコンテンツ項目の特徴に基づいて、ラベル付けされていないコンテンツ項目にラベルを割り当てる際に使用するためである。

【００２３】

モデル２５０がトレーニングされると、２６０において、モデル２５０は、検証セット２４０を使用して検証され得る。モデル２５０を検証する１つの方法は、モデル２５０を検証セット２４０に適用し、検証セット２４０における各コンテンツ項目についてのラベルを取得することである。検証セット２４０は、既知のラベルを有するコンテンツ項目を既に含んでいるので、次いで、既知のラベルが、モデル２５０により生成されたラベルと比較されて、ラベルを正しく予測することにおいてモデル２５０がどれだけ正確であるかが決定され得る。

【００２４】

次いで、検証されたモデル２５０が、ラベル付けされていないコンテンツ項目に適用され、ラベルを生成するために使用され得る。このようにして、大量のデータが、妥当な精度範囲内でモデルによりラベル付け及び／又は分類され得る。しかしながら、上述のように、一般に、モデルをトレーニングするために、大量のトレーニングデータが必要とされる。計算効率的であり、コンテンツ項目の手動ラベル付けを必要としない方法で、トレーニングデータを取得することが望ましいであろう。

【００２５】

図３は、本発明のいくつかの実施形態に従った、関心オブジェクトに基づいて知識表現を合成するための例示的なワークフローを示すフローチャートである。関心オブジェクト（「ＯＯＩ」（object of interest））３１０が分析され、ＯＯＩ３１０に含まれる用語及び／又は他の特性が使用されて、１つ以上のトピック統一資源識別子（トピックＵＲＩ）３２０が生成される。次いで、トピックＵＲＩ３２０が、いくつかの追加の概念及び／又は概念間の関係であって、これらのうちの一部はＯＯＩ３１０において見つけられない、いくつかの追加の概念及び／又は概念間の関係を含む知識表現３３０を合成するための基礎として使用される。

【００２６】
図４は、本発明のいくつかの実施形態に従った、例示的なＯＯＩの図である。図４における例示的なＯＯＩ３１０は、単一のウェブページであるが、ＯＯＩは、複数のウェブページをカバーする１つのウェブサイト又は複数のウェブサイトまで及ぶことがあり、関心項目を提示する１つ以上のコーパスから構成される任意の適切なタイプの非構造化データを含み得ることを理解されたい。いくつかの実施形態において、ＯＯＩ３１０は、トピック、ツイート、ウェブページ、ウェブサイト、文書、文書タイトル、メッセージ、広告、及び／又は検索クエリを含み得る。いくつかの実施形態において、ＯＯＩ３１０は、前述のオブジェクトの単一のインスタンス（例えば、単純なトピック、単一の文書等）であることもあるし、又は、前述のオブジェクトの集合（例えば、個人の関心と関連性があるトピックの集合、個人の関心と関連性がある文書のグループ等）を含むこともある。さらに、ＯＯＩ３１０は、特定のユーザと関連性がある様々なタイプの項目（例えば、ユーザのツイート及びその同じユーザにとって関心がある文書）を含み得、様々なユーザにわたる項目（例えば、特定の関心マーケット又は人口セグメントと関連性がある項目）まで及び得る。ＯＯＩ３１０は、オンラインオブジェクト及びオフラインオブジェクトの両方を含み得る。

【００２７】
理解されるように、ＯＯＩ３１０内に含まれる内容の範囲を変えることにより、それから抽出されるトピックＵＲＩ３２０の量及び深さは変わることになり、それから生じる合成された知識表現３３０も変わることになる。したがって、フォーカスの程度と知識表現３３０内に含まれる概念及び／又は関係の量との間のバランスをとることが可能である。例えば、より広範なデータを含むＯＯＩ３１０は、より大きなドメインを表し、トピックＵＲＩ３２０を生成し、データを合成するためのより大きな範囲を提供する。対照的に、少しの単語からなる句等、トピックをＯＯＩ３１０として使用するとき、そのトピックは、データを合成するための唯一のレンズとなるであろう。

【００２８】
図５は、本発明のいくつかの実施形態に従った、関心オブジェクトに基づいてトピックＵＲＩを抽出するための例示的なワークフローの図である。明確にするために、図５に示されているＯＯＩ３１０は、図４に示されているのと同じウェブページである。ＯＯＩ３１０から生成されたトピックＵＲＩ３２０が、直線的形態及び階層的形態の両方で示されている。確認できるように、用語「skin（皮膚）」が識別され、conditions（状態）、symptoms（症状）、及びclogs（詰まらせる）というサブトピックが、skinに関連するものとして示されている。トピックＵＲＩ３２０の階層的表現及び直線的表現は、同じ関係を表現する２つの異なるやり方であることに留意されたい。

【００２９】
図６は、いくつかの実施形態に従った、関心オブジェクト３１０から生成されたトピックＵＲＩ３２０に基づいて知識表現３３０を合成するための例示的なワークフローの図である。確認できるように、知識表現３３０は、追加の概念、すなわち、「organ（器官）」、「brain（脳）」、「cancer（癌）」、及び「redness（発赤）」を含み、これらは、トピックＵＲＩ３２０における他の概念との関係を有するものとして識別される。これらの追加の概念及び概念間の関係は、例えば、関心オブジェクトの外部にあるソースから取得され得る。例えば、シソーラス、辞書、他のコンテンツ項目、又は様々なトートロジーを用いて、追加の概念及び関係を識別することができる。上述のように、知識表現３３０は、例えば、ユーザにとって関心があることが既知である文書の集合といった、広範囲の非構造化データを含む様々なタイプの関心オブジェクトの共通部分（product）であり得ることを理解されたい。

【００３０】
本開示において特定の例が提供されているが、知識表現３３０を合成する多くの方法が存在することに留意されたい。さらなる例は、例えば、その全体が参照により組み込まれる米国特許第９３７８２０３号及び米国特許出願第１３／３４５６３７号において見出す

ことができる。いくつかの実施形態において、原子的知識表現モデル（ＡＫＲＭ：atomic knowledge representation model）は、１つ以上の要素的（elemental）データ構造及び１つ以上の知識処理ルールを含み得る。いくつかの実施形態において、ルールは、１つ以上の複合的知識表現（ＫＲ）を分解（分析）して要素的データ構造を生成するために、システムにより使用され得る。例えば、システムは、１つ以上のコンピュータプロセッサ及び１つ以上のコンピュータメモリハードウェアコンポーネントを含み得、メモリは、１つ以上のコンピュータプロセッサにより実行されたときに１つ以上のコンピュータプロセッサに１つ以上の複合的ＫＲの分析においてルールを使用させてＡＫＲＭの要素的データ構造を生成させるコンピュータ実行可能な命令で符号化され得る。メモリはまた、ルールを使用して要素的データ構造から新しい複合的ＫＲを合成するように１つ以上のコンピュータプロセッサをプログラムする命令で符号化され得る。いくつかの実施形態において、コンピュータメモリは、実行されたときに１つ以上のプロセッサに本開示に記載の機能のいずれかを実行させるコンピュータ実行可能な命令で符号化された１つ以上の非一時的な有形のコンピュータ読み取り可能な記憶媒体として実現され得る。

【００３１】

合成された知識表現３３０を使用して、他のコンテンツ項目の関連性を評価することができる。例えば、図６からの例を用いると、他のコンテンツ項目が、合成された知識表現において見つけられる概念のうちの１つ以上の存在について分析され得る。ＯＯＩ３１０に対するコンテンツ項目の潜在的な関連性の指標として使用され得るスコアが、そのような分析に基づいて生成され得る。さらに、重みが、合成された知識表現３３０における異なる概念に割り当てられ得る。そのような重みは、知識表現３３０における特定の概念が、知識表現における別の概念と比較して、関連性と相関がある可能性がより高いことを示すために使用され得る。コンテンツ項目についてのスコアは、例えば、概念の重み付け合計をとることにより決定され得る。いくつかの実施形態において、コンテンツ項目についてのスコアは、知識表現３３０における１つ以上の概念とコンテンツ項目の内容との共通部分（intersection）に基づいて決定され得る。

【００３２】

図７は、本発明のいくつかの実施形態に従った、スコアに従ってコンテンツ項目をラベリングするための例示的な方法を示すフローチャートである。７１０において、１つ以上のスコア付けされたコンテンツ項目が得られる。いくつかの実施形態において、コンテンツ項目は、関心オブジェクト３１０に基づく知識表現３３０に基づいてスコア付けされる。７２０において、各コンテンツ項目が、予め定められたスコア閾値と比較される。コンテンツ項目についてのスコアが閾値よりも大きい場合、７３０において、コンテンツ項目は、「推奨する（recommend）」とラベル付けされる。コンテンツ項目についてのスコアが、スコア閾値以下である場合、コンテンツ項目は、「推奨しない（do not recommend）」とラベル付けされる。「推奨する」ラベルを有するコンテンツ項目及び「推奨しない」ラベルを有するコンテンツ項目は、ラベル付けされたコンテンツ項目のセット７５０を形成する。いくつかの実施形態において、コンテンツ項目は、スコアによりランク付けされ得、ラベルは、純粋にスコアに基づいてではなく、セット内でのコンテンツ項目のランキングに基づいて割り当てられ得る。例えば、７１０において受け取られた１０個のスコア付けされたコンテンツ項目について、７２０において、コンテンツ項目が、合計１０個のコンテンツ項目のうちの上位３個にランク付けされたかどうかが判定され得る。

【００３３】

図７は、コンテンツ項目についての２つの可能なラベルを示しているが、本発明は、２値ラベル（例えば「イエス」又は「ノー」）ではなく、３つ以上のラベルの使用、及び多層型ラベル（例えば多数のラベルオプション）も企図している。例えば、ラベルは、「推奨する」を含み得、次の低い層が「薦める（suggested）」であり、その後に「推奨しない」が続く。層は、任意の適切な数のラベル及び／又はカテゴリへと拡張され得る。さらに、ラベルは、階層的である必要はなく、連想（associative）カテゴリ（例えば、スポーツ、政治、テクノロジ等）を確立してもよい。

【００３４】
上述のように、ラベル付けされたコンテンツ項目のセット７５０は、関連付けられたラベルを有する１つ以上のコンテンツ項目を含む。図８Ａは、本発明のいくつかの実施形態に従った、分類モデル８２０をトレーニングするための例示的なワークフローの図である。ラベル付けされたコンテンツ項目のセット８１０は、機械学習分類器モデル８２０を構築するために、トレーニングデータ（例えば、本開示に記載のトレーニングセット２３０）として使用され得る。いくつかの実施形態において、ラベル付けされたコンテンツ項目のセット８１０は、図７に記載のプロセスにより生成される。
【００３５】
いくつかの実施形態において、ラベル付けされたコンテンツ項目８１０に基づいて分類モデル８２０をトレーニングすることは、各ラベル付けされたコンテンツ項目について、当該コンテンツ項目の様々な特徴と当該コンテンツ項目に関連付けられたラベルとを比較することを含む。いくつかの実施形態において、これらの特徴は、標準的な特徴を含み得、その例は、タイトル、長さ、作者、単語頻度、及び／又は逆文書頻度を含む。いくつかの実施形態において、これらの特徴は、知識表現の１つ以上の属性を含み得る。分類モデル８２０は、特定のコンテンツ項目の特徴に基づいて、どのラベルが特定のコンテンツ項目に割り当てられるべきであるかを決定するために使用され得る推論される関数を含む。
【００３６】
図８Ｂは、本発明のいくつかの実施形態に従った、図８Ａの分類モデル８２０を使用してコンテンツ項目をラベル付けするための例示的なワークフローの図である。分類モデル８２０が、トレーニングされ、任意的に、追加のラベル付けされた検証データを用いて検証されると、分類モデル８２０を使用して、１つ以上のラベル付けされていないコンテンツ項目のセット８３０にラベルを割り当てることができる。そのような分類は、分類モデル８２０をトレーニングするために使用されたのと同じ特徴に基づく。そのような特徴（例えば、タイトル、長さ、作者、単語頻度、及び／又は逆文書頻度）は、人間のオペレータにより必要な最小限の分析で又は分析なしで、コンテンツ項目から容易に確認され得ることを理解されたい。したがって、分類モデル８２０を使用してラベル付けされていないコンテンツ項目８３０をラベル付けすることは、比較的迅速且つ効率的に実行され得る。分類モデル８２０をラベル付けされていないコンテンツ項目８３０に適用することは、ラベル付けされたコンテンツ項目の第２のセット８２０をもたらす。次いで、好適なラベル（例えば「推奨する」又は「薦める」）が割り当てられたラベル付けされたコンテンツ項目８５０が、関心がある可能性が高いコンテンツとして、ユーザ又はユーザグループに送られ得る。
【００３７】
前述のプロセスは、教師なし学習手法に類似していると考えられ得る。なぜならば、分類器は、有用な分類をもたらすためにいかなる教師（すなわち、トレーニングデータ又はラベル付けされていないコンテンツ）も必要としないからである。教師なし学習は、ラベル付けされていないデータから隠れ構造を記述するための関数を推論する機械学習タスクである。知識表現３３０とラベル付けされていないコンテンツ項目８３０との間の共通部分を識別することにより、ラベル付けされていないコンテンツ項目８３０の有用な表現／特徴（すなわち、関心オブジェクト３１０に固有の意味的表現を含む隠れ構造）が生成される。さらに、図７のプロセスは、（以下でさらに詳細に説明される、）知識表現の属性から導出された特徴の周りでコンテンツ項目をクラスタ化するための技術に拡張され得る。
【００３８】
図９は、関心オブジェクトに基づく知識表現から導出された属性を使用して、機械学習分類器により、ラベル付けされていないコンテンツ項目を分類するための例示的なワークフローの図である。図３～図６を参照して上述したように、関心オブジェクト３１０を使用して、少なくとも１つの概念、及び／又は、２つ以上の概念間の関係を含む知識表現３３０を合成することができる。いくつかの実施形態において、知識表現は、非一時的なコ

ンピュータ読み取り可能なデータ構造として符号化される。

【００３９】

図９に示されているように、ラベル付けされたコンテンツ項目の集合が、機械学習アルゴリズムのためのトレーニングデータとして使用され得る。知識表現３３０を、機械学習分類モデルのための特徴のソースとして使用することが可能であり得る。知識表現から導出された属性に基づくコンテンツ項目特徴の例は、知識表現とコンテンツ項目との間の概念の共通部分の総数、知識表現とコンテンツ項目との間のより狭い概念の共通部分の数、又は、共通概念の重みの合計を含み得る。当業者は、これらの属性の多くの組み合わせ及び重み付け合計も、機械学習分類モデルのための特徴として使用されてよいことを理解するであろう。さらなる例は、以下においてより詳細に説明される。

【００４０】

図１０Ａは、本発明のいくつかの実施形態に従った、関心オブジェクトから導出された知識表現からの特徴を使用して分類器モデルをトレーニングするための例示的なワークフローの図である。トレーニングデータは、１つ以上のラベル付けされたコンテンツ項目の第１のセット１０１０を含む。いくつかの実施形態において、各ラベルは、コンテンツ項目を１つ以上のカテゴリに分類する。いくつかの実施形態において、１つ以上のラベル付けされたコンテンツ項目の第２のセットを含む検証データのセットも存在し得る。

【００４１】

機械学習分類器１０２０は、知識表現から導出された１つ以上の属性に基づく特徴及びラベル付けされたコンテンツ項目１０１０からトレーニングされる。分類モデル１０２０をトレーニングすることは、知識表現から導出された属性のうちの１つ以上と、トレーニングデータセットにおける、コンテンツ項目に関連付けられたラベルと、に基づく、トレーニングデータセットにおけるコンテンツ項目の少なくとも１つの特徴を使用して、１つ以上の関数を推論することを含み得る。

【００４２】

図１０Ｂは、本発明のいくつかの実施形態に従った、図１０Ａの分類器モデルを使用してラベル付けされていないコンテンツ項目を分類するための例示的なワークフローの図である。少なくとも１つのラベル付けされていないコンテンツ項目のセット１０３０が、分類モデル１０２０により分析され、分類モデル１０２０は、少なくとも１つのラベル付けされていないコンテンツ項目１０３０を、少なくとも１つのラベル付けされたコンテンツ項目の第２のセット１０５０へと分類する。いくつかの実施形態において、この分類することは、知識表現から導出された１つ以上の属性に基づく少なくとも１つの特徴と、ラベル付けされていないコンテンツ項目の内容と、を用いる。

【００４３】

図１１は、本発明のいくつかの実施形態に従った、コンテンツ項目に関連付けられた属性と、コンテンツ項目の既知の関連性と、分類器による予測と、を示す例示的な表の図である。図１１において、ラベル付けされたコンテンツ項目からなるトレーニングセットは、コンテンツ項目１～５０を含む。この例において、ラベルは、コンテンツ項目を、関心オブジェクトと関連性があるものとして、又は、関心オブジェクトと関連性がないものとして、分類する（この表において、「イエス」の場合はＹで示され、「ノー」の場合はＮで示されている）。さらに、コンテンツ項目５１～１００で示されている、ラベル付けされたコンテンツ項目の第２のセットを含む検証データセットが存在する。コンテンツ項目１０１～ｎは、ラベル（したがって、ＯＯＩとの関連性）が知られていないラベル付けされていないコンテンツ項目である。

【００４４】

図１１は、知識表現の第１の属性及び第２の属性（「属性１」及び「属性２」としてそれぞれ示されている）に基づく、トレーニングデータ項目及び検証データ項目の特徴についての対応する値をさらに示している。２つの属性に基づく特徴についての値が図１１に示されているが、いくつかの実施形態は、分類器モデルをトレーニングするために１つの属性に基づく特徴を使用することができる。いくつかの実施形態はまた、分類器モデルを

トレーニングするために複数の属性に基づく２つ以上の特徴を使用することができる。図１１の例におけるラベルは、２値（すなわち、選択肢はＹ又はＮだけである）であるが、これは、必ずしもそうである必要はなく、本開示に記載のシステム及び方法を使用して、コンテンツ項目を３つ以上のグループに分類することもできる。

【００４５】

図１１の表における各コンテンツ項目についての属性１及び属性２に基づく特徴についての値は、これらの属性と当該コンテンツ項目の内容との共通部分に基づき得る。属性は、合成された知識表現から導出され、例えば、知識表現とコンテンツ項目との間で共通する概念の総数、知識表現とコンテンツ項目との間で共通するより広い概念の数、知識表現とコンテンツ項目との間で共通するより狭い概念の数、及び、知識表現とコンテンツ項目との間で共通する概念の重みの合計を含み得る。

【００４６】

図４に示されている例示的なＯＯＩ３１０及び図６に示されている合成された知識表現３３０を例として用いると、知識表現３３０は、異なる幅を伴う概念の階層を含むことが分かる。例えば、「organ」という概念は、知識表現３３０における「brain」及び「skin」という概念よりも広い。同様に、「redness」という概念は、「symptoms」という概念よりも狭い。知識表現３３０におけるこれらの概念の各々はまた、関連付けられた重みを有し得る。

【００４７】

この例を続けると、図１１における属性１は、「より広い概念の数」であり、属性２は、「より狭い概念の数」である。コンテンツ項目１が、知識表現３３０と比較され得、且つ／又は、知識表現３３０との共通部分について調べられ得、コンテンツ項目１が、７という属性１のスコアについて、単語「organ」の７つのインスタンスを含み、３０という属性２のスコアについて、単語「redness」の３０個のインスタンスを含む、と判別され得る。同様に、コンテンツ項目２は、１５という属性１のスコアについて、単語「brain」の５つのインスタンス及び単語「skin」の１０個のインスタンスを含み得る。コンテンツ項目２はまた、１５という属性２のスコアについて、単語「cancer」の１５個のインスタンスを含み、単語「redness」のインスタンスを含み得ない。属性１及び属性２は、知識表現３３０の特性であるが、図１１の表におけるスコアは、知識表現３３０から導出された属性と共通するコンテンツ項目の特徴に基づくことを、当業者は理解するであろう。

【００４８】

「既知の関連性」ラベルが、コンテンツ項目１及び２（より一般的にはコンテンツ項目１～５０）について既に知られているので、分類器モデルは、特徴が基づく属性１及び属性２の値と結果のラベルとの間の検出された関係に基づいて構築され得る。合成された知識表現３３０から導出された属性が、機械学習分類器をトレーニングするための特徴として使用されるとき、機械学習分類器のトレーニングのためのコンテンツ項目のいわゆる「標準的な特徴」（例えば、タイトル、長さ、作者等）の使用と比較して、より少ないトレーニングデータ項目が、予め規定された精度レベルを伴う分類器モデルを構築するために必要とされ得ることを理解されたい。

【００４９】

合成手段を利用するとともに、そのもたらされた属性を特徴として利用することは、ＯＯＩ３１０又はラベル付けされていないコンテンツ項目のいずれかにおけるデータの表現があまりにも限られていて「標準的な特徴」のみを使用して有用な分類をもたらすことができない場合のデータ希薄の問題を回避するのを助けることができる。例えば、ＯＯＩ３１０の１つのツイート又はラベル付けされていないコンテンツ項目１０３０のタイトルは、分類器がコンテンツ項目を正確に分類するのに十分な情報を単に提供することができない（場合によっては、いくつかのラベル付けされていないコンテンツ項目１０３０におけるタイトルの文字どおりの表現は、互いから十分に区別され得ない）。その代わりに、知識表現をソースとする、特徴が基づく属性を利用することにより、従来の機械学習手法で適用されていた特徴に依拠することの欠点を回避することができるより多くのデータが、

システムに提供される。

【００５０】

機械学習分類器が、トレーニングデータを使用してトレーニングされると、分類器モデルが、検証データ項目（例えば、図１１におけるコンテンツ項目５１～１００）を用いて検証され得る。確認できるように、属性１及び属性２に基づく特徴についての値が、各検証データ項目について決定され、属性１及び属性２に基づく特徴についての値に基づく分類器による予測が、図１１の最も右側の列に示されている。コンテンツ項目５１及び１００に割り当てられたラベルは正しいが、コンテンツ項目９９について分類器により予測されたラベルは正しくない。いくつかの実施形態において、分類器モデルについての最低限の精度レベルが必要とされ得る。分類器モデルについての正しい予測の割合が、許容可能な閾値未満である場合、システムは、追加のトレーニングデータを使用して、機械学習分類器をさらにトレーニングし得る。例えば、いくつかの実施形態において、トレーニングデータセットのみに基づく分類器が、必要な精度レベルを達成しなかった場合には、機械学習分類器は、トレーニングデータセット及び検証データセットの両方を使用してトレーニングされ得る。

【００５１】

機械学習分類器がトレーニングされると、機械学習分類器を使用して、ラベルを有しないコンテンツ項目を分類することができる。図１１に示されているように、コンテンツ項目１０１及び１０２は、「既知の関連性」列において「？」を含み、機械学習分類器は、「Ｎ」ラベル及び「Ｙ」ラベルをそれぞれ予測している。

【００５２】

いくつかの実施形態において、２つ以上の機械学習分類器がトレーニングされ得る。例えば、図１２を参照すると、第１の分類器（「分類器１」）は、属性１及び属性２に基づく、トレーニングデータコンテンツ項目の特徴を使用して、トレーニングされ得る。第２の分類器（「分類器２」）は、これらの属性の異なるサブセットに基づく特徴を使用して、トレーニングされ得る。例えば、第２の分類器は、属性１に基づく特徴のみを使用して、トレーニングされ得る。確認できるように、第１の分類器及び第２の分類器により予測されるラベルは、特定の属性又は属性のセットと、特定のコンテンツ項目についてこれに関連付けられたラベルと、の間の相関に応じて、同じコンテンツ項目でも異なり得る。

【００５３】

図１３は、本発明のいくつかの実施形態に従った、複数の分類器モデル（すなわち、分類器モデルの集合）によりなされた予測の結果を示す例示的な表の図である。図示されているように、ラベル付けされていないコンテンツ項目１０１～１０５についての予測されたラベルが、異なる属性に基づく特徴の異なるセット又はサブセットを使用してトレーニングされた第１の分類器、第２の分類器、及び第３の分類器について示されている。いくつかの実施形態において、全体ラベルが、異なる分類器からの予測の重み付け合計に従って生成され得る。図１３に示されている例において、各分類器は、２値（イエス又はノーというラベルが割り当てられることを意味する）であり、各分類器は、等しく重み付けされる。したがって、「イエス」という予測には、１という値が与えられ得、「ノー」という予測には、０という値が与えられ得る。次いで、平均スコアが、全体ラベルについて、コンテンツ項目ごとにとられ得る。図１３に示されている例において、０．５を超える平均スコアを有するコンテンツ項目には、「イエス」というラベルが与えられているのに対し、０．５を下回るスコアを有するコンテンツ項目には、「ノー」というラベルが与えられている。いくつかの実施形態において、コンテンツ項目にラベルを割り当てることは、コンテンツ項目を複数の可能なカテゴリのうちの１つ（例えば、「ＯＯＩと関連性がある」又は「ＯＯＩと関連性がない」）に分類することに類似している。

【００５４】

特定の分類器に割り当てられる閾値及び重みは、所望の結果を達成するように選択され得ることに留意されたい。２値分類器を有する図１３に示されている等しく重み付けされるシステムは単に例である。いくつかの実施形態において、分類器は、等しく重み付けさ

れず、分類器は、任意的に、３つ以上の可能なラベル（例えば「推奨する」、「薦める」、及び「推奨しない」）を生成してもよい。別の例示的な分類は、「複製である（duplicate）」又は「複製ではない（not duplicate）」であり、ここでは、分類器は、推奨する関心オブジェクトにあまりに類似しているもの（例えば、記事が、関心オブジェクトと同じテキストを含むが、異なる公開ソースによりリリースされている場合）として、コンテンツ項目を分類するようにトレーニングされ得る。

【００５５】

上述のように、いくつかの実施形態において、知識表現は、関心オブジェクトの内容に基づいて合成される。いくつかの実施形態において、この合成することは、少なくとも１つの概念及び／又は２つ以上の概念間の関係を生成することであって、概念及び／又は関係は、関心オブジェクトにおいて列挙されていない、生成することを含む。いくつかの実施形態において、合成された知識表現は、新しい複合的知識表現を合成するために生成ルールを要素的データ構造に適用し得る。知識表現はまた、少なくとも１つの概念に関連付けられた重みを含み得る。さらに、知識表現は、特定のユーザ又は特定のユーザグループの特性を考慮し得る。例えば、特定のユーザがテニスに関心があることが既知である場合、このことが、例えばテニスに関連する概念により大きな重みを割り当てることにより、又は、テニスに関連する追加の概念を生成し、追加の概念を知識表現に付加することにより、知識表現において考慮され得る。したがって、特定のユーザ又はユーザグループについてカスタマイズされた知識表現からの属性を使用してトレーニングされた機械学習分類器は、特定のユーザ又はユーザグループに固有ではない知識表現からの属性を使用してトレーニングされた分類器と比べて、同じコンテンツ項目について異なる分類をもたらし得る。

【００５６】

いくつかの実施形態において、機械学習分類モデルをトレーニングするために使用されるトレーニングデータは、図７を参照して上述したプロセスに従って生成される。

【００５７】

上述のように、教師あり学習アルゴリズムにより推論される関数は、トレーニングデータコンテンツ項目の様々な特徴とトレーニングデータコンテンツ項目に割り当てられたラベルとの間の観測又は検出された相関に基づき得る。コンテンツ項目の特徴の一部は、結果のラベルにほとんど又は全く関係しないものであり得るのに対し、他の特徴は、結果のラベルとの非常に密接な相関を有し得ることが可能である。例えば、コンテンツ項目に含まれる、知識表現におけるより狭い概念の数に関する属性に基づく特徴は、より広い概念の数に関する属性に基づく特徴よりもラベルとの強い相関を有し得る。

【００５８】

トレーニングデータオブジェクトに割り当てられたラベルとのより高い相関を有する傾向がある、トレーニングデータコンテンツ項目の特徴及び／又は知識表現の属性を識別することが望ましいであろう。そのような特徴又は属性をより重視して機械学習分類器をトレーニングすることは、分類器のための、又は、潜在的には、許容可能な精度範囲へと分類器をトレーニングするために必要とされるトレーニングデータ項目量の低減のための、より優れた全体精度をもたらし得る。さらに、そのような特徴及び／又は属性の識別はまた、各トレーニングデータ項目について評価されるべき属性又は特徴の数を低減することができる。このことは、機械学習アルゴリズムの計算複雑性、及びそれに関連する処理コストを低減することができる。

【００５９】

図１４Ａは、本発明のいくつかの実施形態に従った、分類器により予測されたラベルの精度を示す例示的な表の図である。確認できるように、機械学習分類器は、検証コンテンツ項目１、２、４、及び６を正しく分類しているのに対し、検証コンテンツ項目３、５、７、及び８は、誤って分類されている。したがって、分類器は、これらの８つの検証コンテンツ項目について、５０％という正解率を有する。

【００６０】

図１４Ｂは、図１４Ａに示されている結果の精度のグラフィカル図である。確認できるように、コンテンツ項目７及び８は、関連性があるが、機械学習分類器により「見逃され」、図示されている円の外側にある。コンテンツ項目１、２、及び４は正しく分類されている。コンテンツ項目３及び５は、関連性がなく、関連性があるものとして誤って分類されている。コンテンツ項目６は、関連性がなく、正しく分類されている。

【００６１】

分類器の精度を評価する１つの指標は、適合率という概念であり、これは、正しくラベル付けされた関連性がある項目の数を、分類器により関連性があるものとしてラベル付けされた項目の総数で除算した比と定義される。図１４Ｂの例において、分類器の適合率は３／５になる（なぜならば、コンテンツ項目１、２、及び４は正しく推奨されているのに対し、コンテンツ項目３及び５は誤って推奨されているからである）。

【００６２】

分類器の精度を評価する別の指標は、再現率という概念であり、これは、正しくラベル付けされた関連性がある項目の数を、関連性がある項目の総数で除算した比と定義される。図１４Ｂの例において、分類器の再現率は、３／５になる（なぜならば、コンテンツ項目１、２、及び４は推奨されているのに対し、コンテンツ項目３及び５はそうでないからである）。

【００６３】

再現率のフォーカスは、関連性がない項目も含まれているとしても全ての関連性がある項目を識別することであるのに対し、適合率のフォーカスは、関連性がない項目を除外しながら関連性がある項目のみを識別することであるので、適合率及び再現率は、しばしば、分類器にとっての相反する目標である。言い換えれば、フォルスネガティブを避けることが、フォルスポジティブを避けることよりも優先されるときには、再現率が適合率よりも優先され得、フォルスポジティブを避けることが、フォルスネガティブを避けることよりも優先されるときには、適合率が再現率よりも優先され得る。

【００６４】

いくつかの実施形態において、機械学習分類器によりなされた予測が、各検証コンテンツデータ項目に割り当てられたラベルと比較され得る。その後、知識表現が、この比較に基づいて変更され得る。

【００６５】

例えば、属性２に基づく特徴についてのより高いスコアが、「Ｙ」ラベルと相関し、属性１に基づく特徴についての値が、相対的にそれほど相関しないことが、図１４Ａにおいて観測され得る。この傾向を踏まえて、ＯＯＩについての知識表現が、そのような相関を反映するように変更され得る。例えば、図１４Ａの例における属性２に基づく特徴は、各検証コンテンツ項目と共通する、知識表現におけるより狭い概念の数である場合、知識表現（例えば図６）は、追加のより狭い概念を含むように変更され得る。例えば、「sunburn（日焼け）」という概念が、「conditions」階層に付加され、「itching（かゆみ）」という概念が、「symptoms」階層に付加され得る。別の例として、知識表現における特定の概念（すなわち、より狭い概念）に関連付けられた重みが増加され得る、且つ／又は、特定の概念（すなわち、より広い概念）に関連付けられた重みが減少され得る。

【００６６】

その後、変更された知識表現を使用して、各属性に基づく、コンテンツ項目の特徴についての値を生成することができる。図１５Ａは、検証コンテンツ項目１～８について、知識表現における属性に基づく各特徴についての更新された値を反映している。確認できるように、属性２（この例では、より狭い概念の数に対応する）に基づく特徴についての値は、図１４Ａにおける値と比べて大きくなっている。図１５Ａにおいて、属性１についての値は変わっていないが、これは、必ずしもそうであるとは限らず、本開示において示されている値は単に例である。

【００６７】

変更された知識表現を使用して属性１及び属性２に基づく特徴についての値を使用する

ことで、機械学習分類器は、その後、検証コンテンツ項目についてのラベルを新たに予測することができる。上述のように、属性は、知識表現と１つ以上のラベル付けされたコンテンツ項目との間で共通する概念の総数、知識表現と１つ以上のラベル付けされたコンテンツ項目との間で共通するより広い概念の数、知識表現と１つ以上のラベル付けされたコンテンツ項目との間で共通するより狭い概念の数、及び、知識表現と１つ以上のラベル付けされたコンテンツ項目との間で共通する概念の重みの合計、のうちの少なくとも１つを含み得る。図１５Ａから分かるように、特定のコンテンツ項目についての予測された関連性は、変更された知識表現を用いると異なる。

【００６８】

図１５Ｂは、機械学習分類器とともに、変更された知識表現を用いた、分類器の結果のグラフィカル図を提供している。確認できるように、コンテンツ項目１、２、４、及び７は、ここでは正しく推奨されているのに対し、コンテンツ項目８は、分類器により「見逃されている」ままである。コンテンツ項目３は、誤って推奨されているままであるのに対し、コンテンツ項目５は、関連性がないものとして、ここでは正しくラベル付けされている。コンテンツ項目６は、関連性がないものとして正しくラベル付けされているままである。したがって、この例における、変更された知識表現を用いた分類器の適合率は、ここでは４／５である（４つのコンテンツ項目が正しく推奨されており、合計５つのコンテンツ項目が推奨されている）。変更された知識表現を用いた分類器の再現率は、ここでは４／５である（４つのコンテンツ項目が正しく推奨されており、１つの関連性があるコンテンツ項目が抜かされている）。したがって、機械学習分類器の適合率及び再現率の両方が、この例において向上している。いくつかの実施形態は、知識表現を変更することにより、適合率及び再現率の１つのみを向上させ得ることに留意されたい。

【００６９】

いくつかの実施形態において、知識表現は、目標適合率閾値又は目標再現率閾値が達成されるまで、反復的に変更され得る。例えば、目標適合率が９／１０（すなわち９０％）である場合、図１５Ａ及び図１５Ｂにおける結果は、図１５Ａ及び図１５Ｂが４／５（すなわち８０％）という適合率を示していることを踏まえて、コンテンツ項目８を関連性があるものとしてラベル付けする又はコンテンツ項目３を関連性がないものとしてラベル付けするために、知識表現に対するさらなる変更を必要とするであろう。

【００７０】

図１６は、本発明のいくつかの実施形態に従った、分類器の結果に基づいて知識表現を変更するための例示的なフローチャートの図である。１６１０において、機械学習分類器は、ラベル付けされたコンテンツ項目（例えば検証コンテンツ項目）のセットの関連性を予測する。１６２０において、機械学習分類器による予測が、検証データ項目の既知のラベル（この例では関連性に対応する）と比較される。１６３０において、目標が機械学習分類器により達成されたかどうかが判定される。いくつかの実施形態において、目標は、適合率目標である。いくつかの実施形態において、目標は、再現率目標である。いくつかの実施形態において、目標は、適合率目標と再現率目標との組み合わせである。

【００７１】

目標が達成されていない場合、１６４０において、知識表現が変更される。１６５０において、検証コンテンツ項目についての予測されたラベルが除去される。次いで、このプロセスは１６１０に戻り、１６１０において、検証コンテンツ項目の関連性が予測される。

【００７２】

１６３０において、目標結果が達成されている場合、このプロセスは終了し、機械学習分類器及び知識表現を使用して、ラベル付けされていないコンテンツ項目の関連性を予測することができる。

【００７３】

したがって、本開示に記載の方法、システム、及びコンピュータ読み取り可能な媒体のいくつかの実施形態は、関心オブジェクトと関連性があるコンテンツを識別するために機

械学習分類器をトレーニングするためのトレーニングデータを生成すること、機械学習分類器をトレーニングするために、関心オブジェクトに基づく合成された知識表現からの特徴の属性を使用すること、及び、トレーニングされた機械学習分類器に関連付けられた観測された適合率値及び／又は再現率値に基づいて、合成された知識表現を変更すること、のうちの１つ以上において有用であり得る。

【００７４】

本開示はまた、新たな関係及び／又は概念が知識表現内で形成され得るように、前述のフィードバックループを適用して、知識表現を変更することを企図している。例えば、知識表現内における特定の概念又は関係の発生が、機械学習分類器の向上した性能をもたらす場合、本方法は、その向上した性能を認識し、この概念又は関係を含むように知識表現を変更することができる。特定の概念又は関係を有する又は有しないコンテンツ部分を考慮することにより、知識表現の要素構成に影響を及ぼすように、コンテンツ内のパターンが識別され得る。

【００７５】

また、本開示に記載のタイプの機械学習分類器を使用して、新たな概念を形成するために知識表現内に既に含まれている２つの独立した概念が組み合わせられ得るかどうか（又は代替的に何らかの関係タイプにより結び付けられ得るかどうか）を、例えば、２つの独立している結び付けられていない概念（例えば、「fire（火災）」及び「hose（ホース）」）が、関係（例えば、「fire」は「hose」に関連する（”fire” is-related-to ”hose”））で結び付けられるべきであること、又は、新たな概念（例えば、「firehose（火災ホース）」）を形成するために組み合わせられるべきであることを識別することにより、判定することもできる。

【００７６】

様々な発明のコンセプトが、１つ以上の方法として具現化されてもよく、そのうちの複数の例が、本開示において提供されていることに留意されたい。方法の一部として実行される動作は、任意の適切な形で順序付けられてよい。したがって、例示的な実施形態において逐次的な動作として示されていても、いくつかの動作を同時に実行することを含み得る、示されているのとは異なる順序で動作が実行される実施形態が構成されることもあり、その逆も同様である。

【００７７】

本発明の上述の実施形態は、多数の方法のうちの任意の方法で実装され得る。例えば、実施形態は、ハードウェア、ソフトウェア、又はこれらの組み合わせを使用して実装され得る。ソフトウェアで実装される場合、ソフトウェアコードは、単一のコンピュータにおいて提供されるか又は複数のコンピュータ間で分散されるかにかかわらず、任意の適切なプロセッサ又はプロセッサの集合（例えば、１つのマイクロプロセッサ又は複数のマイクロプロセッサ）上で実行され得る記憶されるプログラム命令として具現化され得る。

【００７８】

コンピュータは、ラックマウント型コンピュータ、デスクトップコンピュータ、ラップトップコンピュータ、又はタブレットコンピュータ等の多数の形態のうちの任意の形態で具現化され得ることを理解されたい。さらに、コンピュータは、一般にコンピュータとはみなされないが適切な処理能力を有する、パーソナルデジタルアシスタント（ＰＤＡ）、スマートフォン、タブレット、リーダ、又は任意の他の適切なポータブル電子デバイス若しくは固定型電子デバイスを含むデバイスで具現化され得る。

【００７９】

また、コンピュータは、１つ以上の入力デバイス及び出力デバイスを有することができる。これらのデバイスは、とりわけ、ユーザインタフェースを提示するために使用され得る。ユーザインタフェースを提供するために使用され得る出力デバイスの例は、出力の視覚的提示のためのプリンタ又はディスプレイスクリーン、及び、出力の可聴提示のためのスピーカ又は他の音発生デバイスを含む。ユーザインタフェース用に使用され得る入力デバイスの例は、キーボード、マイクロフォン、及び、マウス、タッチパッド、デジタルタ

ブレット等のポインティングデバイスを含む。

【００８０】

そのようなコンピュータは、ローカルエリアネットワーク（ＬＡＮ）、又は、企業ネットワーク、インテリジェントネットワーク（ＩＮ）、インターネット等のワイドエリアネットワーク（ＷＡＮ）等のネットワークを含む、任意の適切な形態の１つ以上のネットワークにより、相互接続され得る。そのようなネットワークは、任意の適切な技術に基づいてよく、任意の適切なプロトコルに従って動作することができ、無線ネットワーク、有線ネットワーク、及び／又は光ファイバネットワークを含んでよい。

【００８１】

本開示において概説された様々な方法又はプロセスは、様々なオペレーティングシステム又はプラットフォームのうちの任意の１つを利用する１つ以上のプロセッサ上で実行可能なソフトウェアとしてコード化され得る。さらに、そのようなソフトウェアは、多数の適切なプログラミング言語及び／又はプログラミングツール若しくはスクリプトツールを使用して記述されてよく、仮想マシン又は適切なフレームワーク上で実行される実行可能な機械語コード又は中間コードとしてコンパイルされてもよい。

【００８２】

この点において、様々な発明のコンセプトは、少なくとも１つの非一時的な有形のコンピュータ読み取り可能な記憶媒体（例えば、コンピュータメモリ、１つ以上のフロッピー（登録商標）ディスク、コンパクトディスク、光ディスク、磁気テープ、フラッシュメモリ、フィールドプログラマブルゲートアレイにおける回路構成、又は他の半導体デバイス等）の製品であって、１つ以上のコンピュータ又は他のプロセッサ上で実行されたときに、本発明の様々なプロセスの実施形態を実施する１つ以上のプログラムで符号化された、少なくとも１つの非一時的な有形のコンピュータ読み取り可能な記憶媒体の製品として具現化され得る。少なくとも１つの非一時的なコンピュータ読み取り可能な媒体は、そこに記憶されている１つ以上のプログラムが、上述の本発明の様々な態様を実施するために任意の適切なコンピュータリソースにロードされ得るように、可搬型であってよい。

【００８３】

「プログラム」又は「ソフトウェア」という用語は、上述の実施形態の様々な態様を実施するようにコンピュータ又は他のプロセッサをプログラムするために使用され得る任意のタイプのコンピュータコード又はコンピュータ実行可能な命令のセットを指すよう、本開示において一般的な意味で使用される。さらに、一態様に従うと、実行されたときに本発明の方法を実行する１つ以上のコンピュータプログラムは、単一のコンピュータ又はプロセッサ上に存在する必要はなく、本発明の様々な態様を実施するために異なるコンピュータ又はプロセッサ間においてモジュール形式で分散されてもよいことを理解されたい。

【００８４】

コンピュータ実行可能な命令は、１つ以上のコンピュータ又は他のデバイスにより実行される、プログラムモジュール等の多くの形態であってよい。一般に、プログラムモジュールは、特定のタスクを実行する又は特定の抽象データ型を実装するルーチン、プログラム、アイテム、コンポーネント、データ構造等を含む。通常、プログラムモジュールの機能は、様々な実施形態において、必要に応じて、組み合わせられることもあるし、分散されることもある。

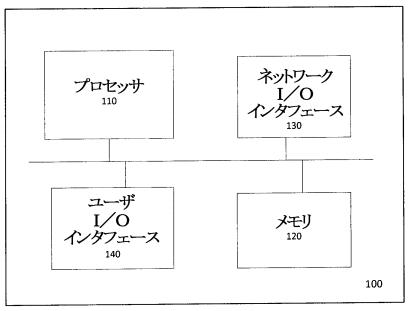
【００８５】

また、データ構造は、任意の適切な形態の非一時的な有形のコンピュータ読み取り可能な記憶媒体の製品に記憶され得る。説明を簡単にするために、データ構造は、データ構造内の位置を通じて関連するフィールドを有するように示されることがある。そのような関係は、フィールド間の関係を伝達する、非一時的な有形のコンピュータ読み取り可能な媒体内の位置を用いてフィールド用の記憶域を割り当てることによっても同様に実現され得る。しかしながら、ポインタ、タグ、又は、データ要素間の関係を確立する他のメカニズムの使用を介することを含め、任意の適切なメカニズムが、データ構造のフィールド内の情報間の関係を確立するために使用され得る。

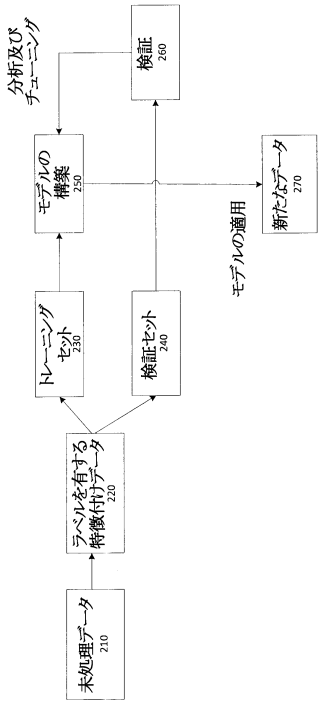
【００８６】
本開示において定義及び使用される全ての定義は、辞書の定義、参照により組み込まれる文書中の定義、及び／又は定義された用語の通常の意味を支配するものと理解されたい。
【００８７】
本開示において使用される不定冠詞「a」及び「an」は、そうでないことが明確に示されていない限り、「少なくとも１つ」を意味するものと理解されたい。
【００８８】
本開示において使用される場合、１つ以上の要素のリストに関する「少なくとも１つ」という句は、要素のリスト内の要素のうちの任意の１つ以上の要素から選択される少なくとも１つの要素を意味するものと理解されるべきであるが、必ずしも、要素のリスト内に具体的に列挙されるありとあらゆる要素のうちの少なくとも１つを含むわけではなく、要素のリスト内の要素の任意の組み合わせを排除するわけではない。この定義はまた、「少なくとも１つ」という句が指す要素のリスト内で具体的に特定される要素以外の要素が、具体的に特定される要素に関連するか又は関連しないかにかかわらず、任意的に存在し得ることを許容する。したがって、非限定的な例として、「Ａ及びＢのうちの少なくとも１つ」（若しくは「Ａ又はＢのうちの少なくとも１つ」、又は「Ａ及び／又はＢのうちの少なくとも１つ」）は、一実施形態においては、Ｂが存在しない（任意的にＢ以外の要素を含む）少なくとも１つの（任意的に２つ以上を含む）Ａを指すことができ、別の実施形態においては、Ａが存在しない（任意的にＡ以外の要素を含む）少なくとも１つの（任意的に２つ以上を含む）Ｂを指すことができ、さらに別の実施形態においては、少なくとも１つの（任意的に２つ以上を含む）Ａ及び少なくとも１つの（任意的に２つ以上を含む）Ｂ（任意的に他の要素を含む）を指すことができる、等である。
【００８９】
本開示において使用される「及び／又は」という句は、そのように結合された要素の「いずれか又は両方」、すなわち、いくつかの場合においては結合的に存在し、他の場合においては選言的に存在する要素を意味するものと理解されたい。「及び／又は」を用いて列挙される複数の要素は、同じように解釈されるべきである、すなわち、そのように結合された要素のうちの「１つ以上」であると解釈されるべきである。「及び／又は」の句により具体的に特定される要素以外の他の要素が、具体的に特定される要素に関連するか又は関連しないかにかかわらず、任意的に存在し得る。したがって、非限定的な例として、「有する、備える、含む」等のオープンエンドの言葉とともに使用されるときの「Ａ及び／又はＢ」との言及は、一実施形態においては、Ａのみ（任意的にＢ以外の要素を含む）を指すことができ、別の実施形態においては、Ｂのみ（任意的にＡ以外の要素を含む）を指すことができ、さらに別の実施形態においては、Ａ及びＢの両方（任意的に他の要素を含む）を指すことができる、等である。
【００９０】
本開示において使用される場合、「又は」は、上記で定義されたような「及び／又は」と同じ意味を有するものと理解されるべきである。例えば、リスト内の項目を分離している場合、「又は」又は「及び／又は」は、包括的であるとして、すなわち、複数の要素又は要素のリストのうちの少なくとも１つ（２つ以上も含む）を含み、任意的にリストにない追加の項目を含むとして、解釈されるべきである。
【００９１】
本開示において使用される表現及び用語は、説明を目的としており、限定とみなされるべきではない。「含む」、「備える」、「有する」、及びこれらの派生語の使用は、その後に列挙される項目及び追加の項目を包含することを意味する。
【００９２】
本発明のいくつかの実施形態を詳細に説明したが、様々な変形及び改良が、当業者には容易に思いつくであろう。そのような変更及び改良は、本発明の主旨及び範囲内にあることが意図されている。したがって、前述の説明は、例示にすぎず、限定することを意図し

ているものではない。

【図１】



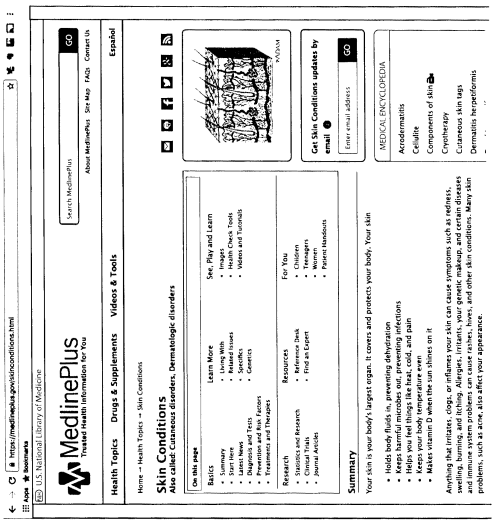
【図２】



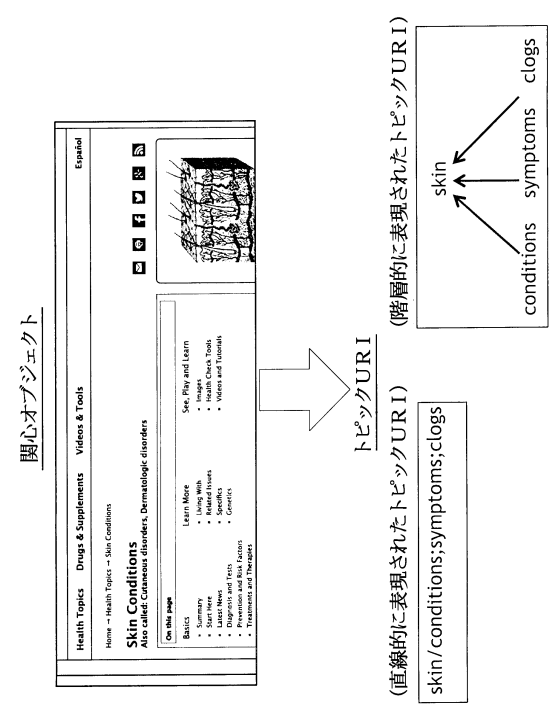
【図３】


関心オブジェクト
310
トピックURI
320
知識表現
330

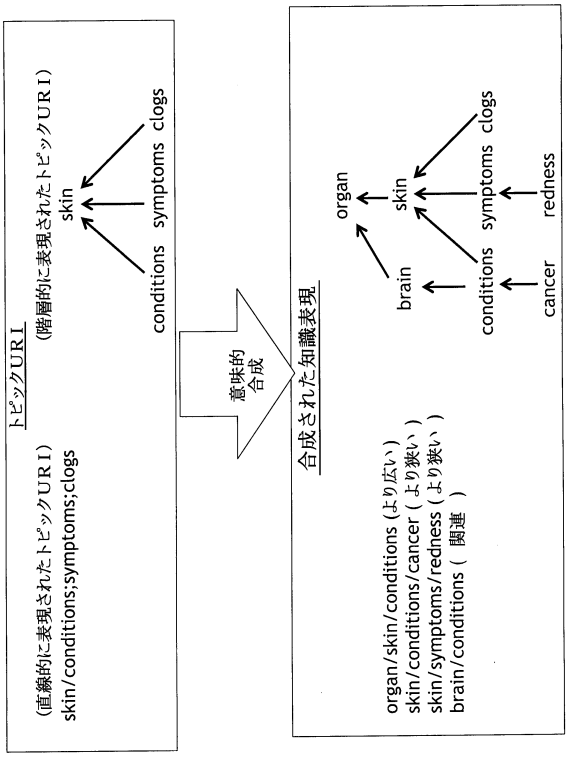
【図４】


例示的な関心オブジェクト

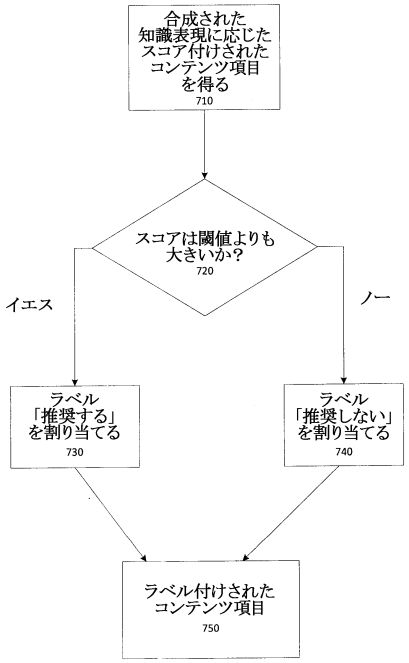
【図５】


関心オブジェクト
トピックURI
(直線的に表現されたトピックURI)
skin/conditions;symptoms;clogs
(階層的に表現されたトピックURI)
skin
conditions
symptoms
clogs

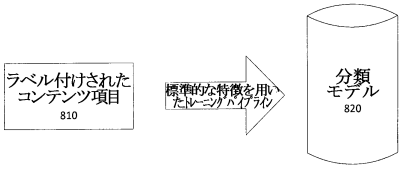
【図６】


トピックURI
(直線的に表現されたトピックURI)
skin/conditions;symptoms;clogs
(階層的に表現されたトピックURI)
skin
conditions
symptoms
clogs
意味的合成
合成された知識表現
organ/skin/conditions (より広い)
skin/conditions/cancer (より狭い)
skin/symptoms/redness (より狭い)
brain/conditions (関連)
organ
brain
skin
conditions
symptoms
clogs
cancer
redness

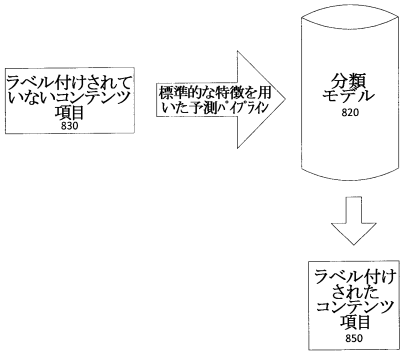
【図７】



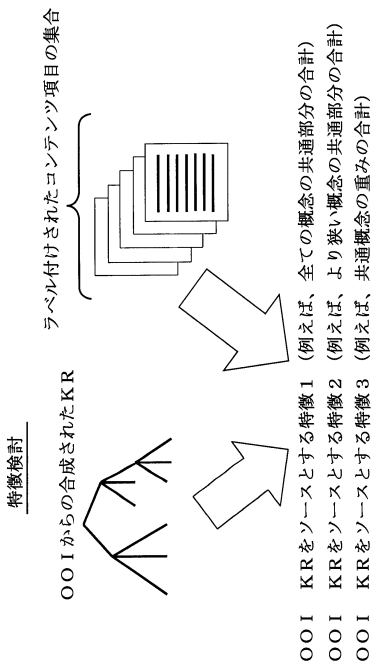
【図８Ａ】



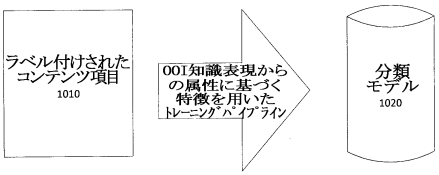
【図８Ｂ】



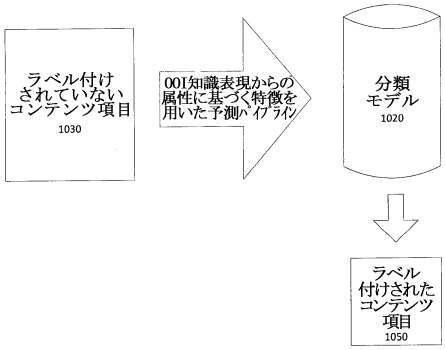
【図９】



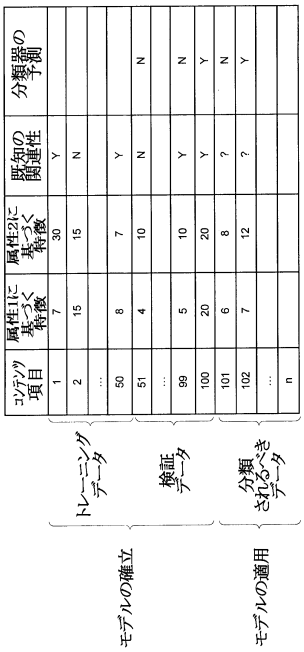
【図１０Ａ】



【図１０Ｂ】



【図１１】

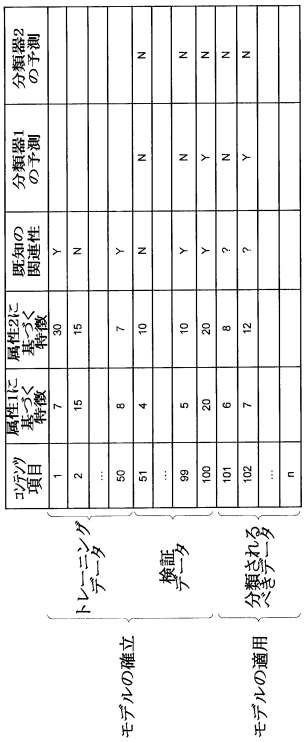


コンテンツ項目	属性1に基づく特徴	属性2に基づく特徴	既知の関連性	分類器の予測
1	7	30	Y	
2	15	15	N	
…	…	…		
50	8	7	Y	N
51	4	10	N	
…	…	…		
99	5	10	Y	N
100	20	20	Y	Y
101	6	8	?	N
102	7	12	?	Y
…	…	…		
n				

トレーニングデータ　検証データ　分類されるべきデータ

モデルの確立　モデルの適用

【図１２】



コンテンツ項目	属性1に基づく特徴	属性2に基づく特徴	既知の関連性	分類器1の予測	分類器2の予測
1	7	30	Y		
2	15	15	N		
…	…	…			
50	8	7	Y		N
51	4	10	N	N	
…	…	…			
99	5	10	Y	N	N
100	20	20	Y	Y	N
101	6	8	?	N	N
102	7	12	?	Y	N
…	…	…			
n					

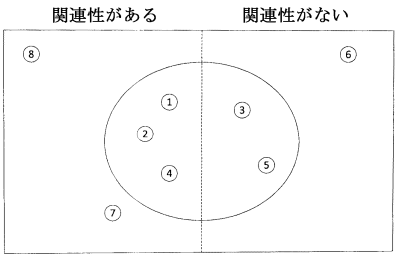
トレーニングデータ　検証データ　分類されるべきデータ

モデルの確立　モデルの適用

【図１３】

コンテンツ項目	分類器1の予測	分類器2の予測	分類器3の予測	全体	推奨する?
101	イエス	イエス	イエス	1	イエス
102	イエス	ノー	イエス	0.67	イエス
103	ノー	ノー	イエス	0.33	ノー
104	イエス	ノー	ノー	0.33	ノー
105	ノー	ノー	ノー	0	ノー

【図１４Ｂ】



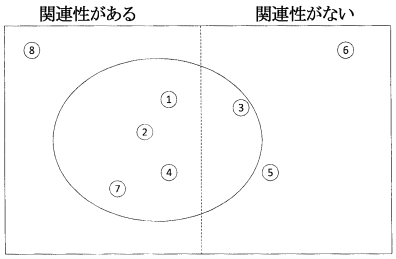
【図１４Ａ】

コンテンツ項目	属性1に基づく特徴	属性2に基づく特徴	…	既知の関連性	予測された関連性	分類器は正しいか?
1	10	10		Y	Y	Y
2	9	2		Y	Y	Y
3	8	3		N	Y	N
4	7	10		Y	Y	Y
5	6	6		N	Y	N
6	5	5		N	N	Y
7	4	10		Y	N	N
8	3	10		Y	N	N

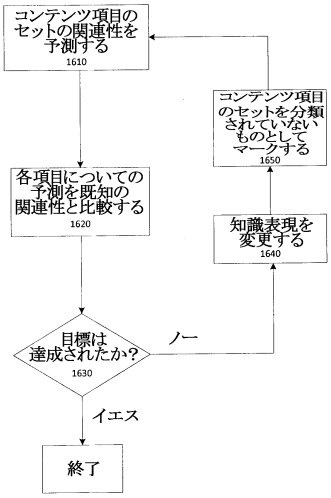
【図１５Ａ】

コンテンツ項目	属性1に基づく特徴	属性2に基づく特徴	…	既知の関連性	予測された関連性	分類器は正しいか?
1	10	15		Y	Y	Y
2	9	10		Y	Y	Y
3	8	3		N	Y	N
4	7	16		Y	Y	Y
5	6	6		N	N	Y
6	5	5		N	N	Y
7	4	18		Y	Y	Y
8	3	17		Y	N	N

【図１５Ｂ】


関連性がある
関連性がない
8
6
1
3
2
4
5
7

【図１６】


コンテンツ項目のセットの関連性を予測する
1610
各項目についての予測を既知の関連性と比較する
1620
目標は達成されたか?
1630
ノー
イエス
終了
知識表現を変更する
1640
コンテンツ項目のセットを分類されていないものとしてマークする
1650

フロントページの続き

(72)発明者 ウィルソン，マシュー ウィットニー
カナダ国 オンタリオ州 エヌ２エイチ ５エム９，キッチナー，ルイーザ ストリート ２７２
(72)発明者 イリヤス，イハブ
カナダ国 オンタリオ州 エヌ２ティー １アール４，ウォータールー，ウィルトシャー プレイス ２９４
(72)発明者 スウィーニー，ピーター ジェイ.
カナダ国 オンタリオ州 エヌ２アール １ビー８，キッチナー，キルバーニー コート ５６

審査官 久保 光宏

(56)参考文献 米国特許第８９５４３５８（ＵＳ，Ｂ１）
特開２００８－２７６５７１（ＪＰ，Ａ）
特開２００７－２６１１５（ＪＰ，Ａ）
山本剛士（外２名），「少数の正例とラベルなし事例による文書の自動分類」，電子情報通信学会論文誌，日本，一般社団法人 電子情報通信学会，２０１２年 ９月 １日，Vol.J95-D，No.9，第１７９４～１８０１頁，ISSN: 1880-4535
吉田真吾（外６名），「Software Design plusシリーズ ＡＷＳエキスパート養成読本」，日本，株式会社技術評論社，２０１６年 ４月 １日，初版，第５９頁，ISBN: 978-4-7741-7992-6

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ３／００－９９／００
Ｇ０６Ｆ１６／００－１６／９５８
ＣＳＤＢ（日本国特許庁）
ＩＥＥＥＸｐｌｏｒｅ（ＩＥＥＥ）