

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号
特開2012-118977
(P2012-118977A)

(43) 公開日 平成24年6月21日 (2012.6.21)

(51) Int.Cl.
G06F 17/30 (2006.01)

F I
G06F 17/30 350C
G06F 17/30 340B
G06F 17/30 220Z

テーマコード (参考)

審査請求 未請求 請求項の数 10 O L (全 11 頁)

(21) 出願番号	特願2011-250180 (P2011-250180)	(71) 出願人	502096543 パロ・アルト・リサーチ・センター・イン コーポレーテッド Palo Alto Research Center Incorporated アメリカ合衆国、カリフォルニア州 94 304、パロ・アルト、コヨーテ・ヒル・ ロード 3333
(22) 出願日	平成23年11月15日 (2011.11.15)	(74) 代理人	100079049 弁理士 中島 淳
(31) 優先権主張番号	12/955,799	(74) 代理人	100084995 弁理士 加藤 和詳
(32) 優先日	平成22年11月29日 (2010.11.29)	(74) 代理人	100093861 弁理士 大賀 真司
(33) 優先権主張国	米国 (US)		

最終頁に続く

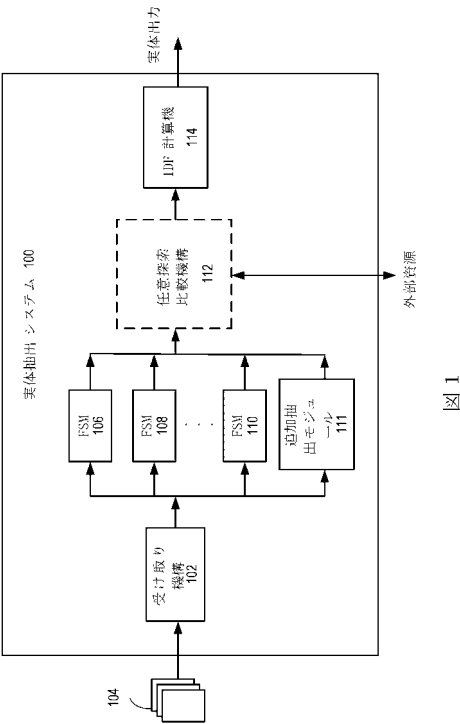
(54) 【発明の名称】 文書類似性計算の機械学習に基づく最適化およびカスタマイズのための方法およびシステム

(57) 【要約】 (修正有)

【課題】 文書類似性分析に関して、機械学習に基づき文書類似性計算を最適化してカスタマイズする。

【解決手段】 動作時に、システムは類似文書の収集をユーザに提示し、文書の類似性に関するフィードバックをユーザから収集し、文書類似性を計算するための一般的規則を生成し、ユーザにより提供されたフィードバックに基づきカスタマイズされた類似性計算を用いて文書をフィルタにかける。ユーザフィードバックは、収集内に間違っ

【選択図】 図 1



【特許請求の範囲】**【請求項 1】**

類似文書の収集をユーザにコンピュータにより提示することと、
前記文書の前記類似性に関するフィードバックを前記ユーザから収集することと、
文書類似性を計算するための一般的規則を前記コンピュータにより生成することと、
前記ユーザにより提供された前記フィードバックに基づきカスタマイズされた類似性計算を用いて文書をフィルタにかけることと、を含む、
文書類似性計算を最適化してカスタマイズするためのコンピュータで実現される方法。

【請求項 2】

前記ユーザフィードバックが、
前記収集内に間違っていて含まれている文書の指摘と、
前記収集内に含まれていない追加類似文書の指摘と、のうちの 1 つ以上を含む、
請求項 1 に記載の方法。

10

【請求項 3】

前記文書から多くの意味実体を抽出することと、
前記抽出された意味実体の逆文献頻度（IDF）値に基づき前記文書間の類似性測度を計算することと、
により前記文書類似性を計算することをさらに含む、請求項 1 に記載の方法。

【請求項 4】

文書類似性を計算するための前記一般的規則を生成することが、
前記収集されたユーザフィードバックに基づき各文書およびその関連文書から特徴を抽出することと、
前記抽出された特徴に基づき規則を生成するために機械学習技術を適用することと、を含む、請求項 1 に記載の方法。

20

【請求項 5】

前記各文書およびその関連文書の前記抽出された特徴が、
前記関連文書の類似性ランクと、
各文書および関連文書の文書重みと、
各文書および関連文書の実体出現の大きさと、
各文書および関連文書の実体出現平均と、
各文書および関連文書の間の共有された実体の個数と、
各文書および関連文書の間の前記共有された実体の平均実体重みと、
各文書および関連文書の間の前記共有された実体の最大実体重みと、
各文書および関連文書の間の前記共有された実体の最小実体重みと、
各文書および関連文書の間の前記共有された実体のタイプされた数、平均実体重み、最小実体重み、および最大実体重みと、
各文書および関連文書内の補足的（非共有）実体の個数と、
各文書および関連文書内の前記補足的実体の平均実体重みと、
各文書および関連文書内の前記補足的実体の最大実体重みと、
各文書および関連文書内の前記補足的実体の最小実体重みと、
各文書および関連文書内の前記補足的実体のタイプされた数、平均実体重み、最小実体重み、および最大実体重みと、のうちの 1 つ以上を含む、
請求項 4 に記載の方法。

30

40

【請求項 6】

管理された機械学習を用いて文書類似性を計算するための決定木を生成することをさらに含む、請求項 1 に記載の方法。

【請求項 7】

ユーザ用にカスタマイズされた類似性計算を用いて文書をフィルタにかけることが、
前記ユーザにより提供された前記フィードバックに基づき各文書およびその関連文書から特徴を抽出することと、

50

前記抽出された特徴に基づきフィルタリング規則を生成するために機械学習技術を適用することと、を含む、
請求項 1 に記載の方法。

【請求項 8】

類似文書の収集をユーザにコンピュータにより提示することと、
前記文書の前記類似性に関するフィードバックを前記ユーザから収集することと、
文書類似性を計算するための一般的規則を前記コンピュータにより生成することと、
前記ユーザにより提供された前記フィードバックに基づきカスタマイズされた類似性計算を用いて文書をフィルタにかけることと、を含む、
方法を、コンピュータにより実行されるとき、コンピュータに実行させる命令を保存する持続性コンピュータ可読記憶媒体。

10

【請求項 9】

前記ユーザフィードバックが、
前記収集内に間違っ含まれている文書の指摘と、
前記収集内に含まれていない追加類似文書の指摘と、のうちの 1 つ以上を含む、
請求項 8 に記載のコンピュータ可読記憶媒体。

【請求項 10】

前記方法が、
前記文書から多くの意味実体を抽出することと、
前記抽出された意味実体の逆文献頻度 (I D F) 値に基づき前記文書間の類似性測度を計算することと、により前記文書類似性を計算することをさらに含む、
請求項 8 に記載のコンピュータ可読記憶媒体。

20

【発明の詳細な説明】

【技術分野】

【0001】

本開示は、一般に、文書類似性分析に関する。より具体的には、本開示は、機械学習に基づき文書類似性計算を最適化してカスタマイズすることに関する。

【発明の概要】

【課題を解決するための手段】

【0002】

本発明の一実施形態は、文書類似性計算を最適化してカスタマイズするためのシステムを提供する。動作時に、システムは類似文書の収集をユーザに提示し、文書の類似性に関するフィードバックをユーザから収集し、文書類似性を計算するための一般的規則を生成し、ユーザにより提供されたフィードバックに基づきカスタマイズされた類似性計算を用いて文書をフィルタにかける。

30

【0003】

この実施形態に関する変形例では、ユーザフィードバックは、収集内に間違っ含まれている文書の指摘と、収集内に含まれていない追加類似文書の指摘と、のうちの 1 つ以上を含んでいる。

【0004】

この実施形態に関する変形例では、システムは、文書から多くの意味実体を抽出することと、抽出された意味実体の逆文献頻度 (I D F) 値に基づき文書間の類似性測度を計算することと、により文書類似性を計算する。

40

【0005】

この実施形態に関する変形例では、文書類似性を計算するための一般的規則を生成することは、収集されたユーザフィードバックに基づき各文書およびその関連文書から特徴を抽出することと、抽出された特徴に基づき規則を生成するために機械学習技術を適用することと、を含んでいる。

【0006】

さらなる変形例では、各文書およびその関連文書の抽出された特徴は、関連文書の類

50

似性ランクと、各文書および関連文書の文書重みと、各文書および関連文書の実体出現の大きさと、各文書および関連文書の実体出現平均と、各文書および関連文書の間の共有された実体の個数と、各文書および関連文書の間の共有された実体の平均実体重みと、各文書および関連文書の間の共有された実体の最大実体重みと、各文書および関連文書の間の共有された実体の最小実体重みと、各文書および関連文書の間の共有された実体のタイプされた数、平均実体重み、最小実体重み、および最大実体重みと、各文書および関連文書内の補足的（非共有）実体の個数と、各文書および関連文書内の補足的実体の平均実体重みと、各文書および関連文書内の補足的実体の最大実体重みと、各文書および関連文書内の補足的実体の最小実体重みと、各文書および関連文書内の補足的実体のタイプされた数、平均実体重み、最小実体重み、および最大実体重みと、のうちの1つ以上を含んでいる。

10

【0007】

この実施形態に関する変形例では、システムは管理された機械学習を用いて文書類似性を計算するための決定木を生成する。

【0008】

この実施形態に関する変形例では、ユーザ用にカスタマイズされた類似性計算を用いて文書をフィルタにかけることは、ユーザにより提供されたフィードバックに基づき各文書およびその関連文書から特徴を抽出することと、抽出された特徴に基づきフィルタリング規則を生成するために機械学習技術を適用することと、を含んでいる。

【図面の簡単な説明】

20

【0009】

【図1】本発明の実施形態の実体抽出システムを説明する図である。

【図2】本発明の実施形態の、文書類似性計算の最適化およびカスタマイズのプロセスを説明するフローチャートである。

【図3】本発明の実施形態の、機械学習に基づき文書類似性を計算するプロセスを説明するフローチャートである。

【図4】本発明の実施形態の、類似文書から抽出された例示的特徴集合を説明する図である。

【図5】本発明の一実施形態の、文書類似性計算を最適化してカスタマイズするための例示的コンピュータシステムを示す図である。

30

【0010】

図面では、同様の参照番号は、同じ図面要素を示す図である。

【発明を実施するための形態】

【0011】

本発明の実施形態は文書類似性計算を最適化してカスタマイズするための解決法を提供する。本発明の一実施形態では、文書類似性計算システムは、文書の類似性に関するフィードバックを収集するために、類似文書の収集をユーザに提示する。ユーザにより提供されたフィードバックに基づき、システムは将来の類似文書を特定するための一般的規則を生成する。また、システムはユーザからのフィードバックに基づきカスタマイズされた類似性計算を用いて文書をフィルタにかけることができる。

40

【0012】

文書の間の従来の類似性計算は、通常、単語の出現を数えて比較することにより関連文書の文字列を照合することを利用する。例えば、地方の天気を話題にする電子メールメッセージはすべて雨、雪、または風のような単語を含んでいる可能性がある。したがって、文字列を比較することにより、2つのメッセージ間の類似性を推定できる。しかしながら、このような方法は非効率的である場合があり、多くの間違った結果を生じる可能性がある。例えば、決まり文句の文章を含む文書では、2つの文書間で決まり文句の共出現が多い可能性があるが、2つの文書間の類似性は実際には低い可能性がある。この問題を克服するために、各単語の出現を数える代わりに、文書間の類似性を導き出すために「実体」として定義される重要な単語の出現を比較することを利用する実体抽出方法を提案する。

50

【0013】

このような実体抽出プロセスを図1に示す。実体抽出システム100は、受け取り機構102と、多くの有限状態機械(FSM)106~110と、任意探索比較機構112と、逆文献頻度(IDF)計算機114と、を含んでいる。動作時に、受け取り機構102は実体抽出のための入力文書104を受け取る。その後、受け取られた文書の文字列はFSM106~110を含む多くのFSMに送られる。これらのFSMは、事前に定義された異なる群に属する意味実体を認識するように異なって設計されている。意味実体は単語、単語の組み合わせ、または人名、会社名、日時、住所、業界固有の用語、電子メールアドレス、ユニフォーム・リソース・ロケータ(URL)、電話番号のような固有の意味を有する配列である可能性がある。事前に定義された群に属さない追加的意味実体を追加抽出モジュール111により抽出できる。

10

【0014】

無意味な単語が意味実体としてFSM106~110により誤って認識されることを回避するために、受け取られた文書の文字列から特定された実体のうちの一定の種類は、外部資源を探索して比較するために任意探索比較機構112に送られる。その後、実体候補がそれらのIDF値を計算するIDF計算機114に送られる。実体候補の有意性を測定するためにIDF値を使用できる。低いIDF値は、その実体候補が言語資料全体にわたって広く使用されており、したがって、決まり文句、統計的外れ値、または間違っただけである可能性が高いことを示す場合が多い。対照的に、高いIDF値は、このような実体候補が真に意味のある、または有意な意味実体であり、文書から抽出されるのにふさわしいことを示している。最後に、所定の範囲内のIDF値を有する実体候補は抽出されるが、この範囲外のIDF値を有する実体候補は無視される。

20

【0015】

次に、有意な実体であると考えられる抽出された意味実体は文書間の類似性計算に使用され得る。2つの文書が多く重複する有意な実体を有しているとき、システムはこれらの2つの文書が似ている高い可能性を有している、したがって、高い類似性値を有していると判断できる。文書の中の有意な実体の出現を数えることに加えて、また、文書類似性を計算するときには、発生論的実体重みを考慮に入れる。異なる群に属する実体には異なる重みを割り当てる。例えば、人名の群に属する実体には、住所の群に属する実体とは異なる重みを割り当てる。異なる実体群の重要性および言語資料の状況に応じて、重みは適宜に調整できる。例えば、人事労働者にとっては人々の名前は専門用語よりも重みを持つが、技術者にとっては、その反対のことが当てはまる。

30

【0016】

文書間の類似性を判断するために多くの異なる測度を計算され得る。例えば、第1の測度は、2つの文書間で重複する実体のIDF値の加重和と、各文書内の実体のIDF値の加重和の比率を計算する。第1の測度と同様の他の測度は、各文書内の重み付けされたIDF値を別々に合計する代わりに、2つの文書の和集合内の実体の重み付けされたIDF値を使用する。その後、文書は、それらの文書の実体出現に基づく所与の文書に関する類似性に基づいた順序に並べられる。2つの文書は、それらの2つの文書の実体出現に基づく類似性レベルの間の差が所定の閾値よりも小さいとき、同じレベルの類似性を有している。

40

【0017】

本発明の実施形態は文書類似性計算の機械学習に基づく最適化およびカスタマイズのためのシステムを提供する。このシステムは、文書内の意味実体を抽出し、複数のユーザを通じた違いに対処するためにそれらの文書の類似性を計算するとき、さまざまなユーザ選択およびユーザ構成を考慮する。

【0018】

本発明の実施形態では、システムは、元の文書と、候補文書の言語資料との間の類似性を、これらの文書から抽出された意味実体に基づいて計算する。結果としてもたらされた類似文書の収集は、誤検出、すなわち、収集内に間違っただけの文書、および検出

50

漏れ、すなわち、収集内に含まれていない追加類似文書、を含む可能性があることが分かった。文書類似性の今後の決定を改良し、ユーザを通じて類似性計算をカスタマイズするために提案される方法は、最適化およびカスタマイズの2つの段階で構成されている。

【0019】

第1段階の最適化の目的は、ユーザフィードバックを組み込むことにより広範囲の類似性計算を促進することである。第1段階では、システムは元の文書に関連する類似文書の収集をシステムユーザに提示し、文書の類似性に関するフィードバックをシステムユーザから収集する。ユーザは収集内に間違っ

10

【0020】

て含まれている文書、および収集内に含まれない元の候補からの追加類似文書を指摘してもよい。ユーザのフィードバックは、管理された学習用の訓練データとして機械学習サブシステムに提供される。機械学習サブシステムはユーザから収集されたフィードバックに基づき文書類似性を計算するための一般的規則の集合を生成する。機械学習サブシステムが生成した一般的規則は、既存の類似性計算の枠組みに組み込む前にシステム設計者が見直すことができる。生成された規則は、文書類似性計算に適用したときのそれらの誤検出率および正検出率により評価できる。

20

【0021】

第2のカスタマイズ段階は、各ユーザに対する類似文書を見つけるための個々の調整を提供することを目指す。この段階は類似性計算を改良するためにユーザが絶えずフィードバックを与えてもよい反復プロセスである。この段階は個々のユーザのフィードバックを採取することと、管理された機械学習アルゴリズムをユーザフィードバックに適用することと、を含んでいる。機械学習アルゴリズムで生成された分類規則は、各ユーザに対する類似文書をフィルタにかけるのに使用できる。ユーザは誤検出率、正検出率、または誤検出対正検出率に基づき規則を選択してもよい。

30

【0022】

図2は、本発明の実施形態の、文書類似性計算の最適化およびカスタマイズのプロセスを説明するフローチャートを示している。動作時に、システムは類似文書の収集をユーザに提示する(動作202)。その後、システムは文書の類似性に関するフィードバックをユーザから収集する(動作204)。一実施形態では、ユーザフィードバックは収集内に間違っ

40

【0023】

て含まれている文書の指摘、および／または収集内に含まれていない追加類似文書の指摘を含んでいる。その後、システムは収集されたユーザフィードバックに基づき類似文書の計算を最適化するために一般的規則を生成する(動作206)。各ユーザからのフィードバックは各ユーザに対する類似文書のフィルタリングをカスタマイズするために使用してもよい(動作208)。また、システムはユーザに対する状況情報に基づき類似文書を必要に応じて見つけることができる(動作210)。

50

類規則はシステム性能を最適化するように配置でき、他方、各ユーザのフィードバックから推定された分類規則はユーザに対してカスタマイズされた類似性計算を促進する。他の実施形態では、機械学習アルゴリズムのための文書特徴のユーザ入力用のユーザインタフェースが提供される。

【0024】

図3は、本発明の実施形態の、機械学習に基づき文書類似性を計算するプロセスを説明するフローチャートを示している。動作時に、システムは類似文書の収集内に間違っ

て含まれている文書の指摘、および／または収集内に含まれていない追加類似文書の指摘を含むユーザフィードバックを収集する（動作302）とともに、元の文書と関連文書とから特徴を抽出する（動作304）。一実施形態では、システムは抽出された特徴に機械学習を適用して（動作306）、文書類似性を計算するための一般的規則を生成する（動作308）。また、各ユーザからのフィードバックは、各ユーザに対する文書類似性を計算するためのカスタマイズされた規則を生成する（動作310）ために使用できる。その後、システムは文書を類似性に基づく順序に並べる（動作312）。

【0025】

一実施形態では、システムは文書類似性の計算を最適化するための一般的規則を生成するために管理された機械学習を適用する。管理された学習は、訓練例の集合で構成された管理された訓練データから分類規則を推定するタスクである。類似文書の発見を向上させるために、システムは収集内に間違っ

て含まれている文書、および／または収集内に含まれていない追加類似文書を指摘するユーザフィードバックを収集する。ユーザフィードバックは管理された機械学習のための訓練データを提供して、管理された機械学習アルゴリズムがユーザフィードバックを分析して分類規則の集合を推定してもよい。推定された分類規則は将来の文書の類似性を予測する際に使用できる。

【0026】

分類規則を推定するには、入力訓練データから特定の属性または特徴が抽出され、その結果、抽出された属性または特徴が分類結果に関連するようになっている必要がある。本発明の実施形態では、元の文書と関連文書とから特徴の4つの群を抽出する。第1の特徴は文書の類似性の広範囲のランクである。類似文書は意味実体出現類似性に基づいて計算され、類似性ランクの順序でユーザに提示される。特徴の第2の群は、2つの文書間の共有された意味実体を含んでいる。

【0027】

図4に示す実施例では、意味実体抽出を実行した後に、システムは元の文書402に対する実体集合400と、関連文書412に対する実体集合410と、を決定する。実体集合400と実体集合410の間の交わりが共有された実体集合420を形成する。同様に、元の文書402と関連文書414または416との間で、他の共有された実体集合を決定できる。特徴のこの群は、共有された実体集合内の共有された実体の個数および重みに基づく。

- ・ Shared Count : 2つの文書間で共有された実体の個数、
- ・ Shared Average : 2つの文書間で共有された実体に対する平均実体重み、
- ・ Shared Max : 2つの文書間で共有された実体に対する最大実体重み、
- ・ Shared Min : 2つの文書間で共有された実体に対する最小実体重み、
- ・ タイプされた共有された実体値 : 人、会社、および場所のような異なる種類の実体であり、上述の特徴は下記の4個の異なる種類で区別でき、
- ・ Shared Type X Count
- ・ Shared Type X Average
- ・ Shared Type X Max
- ・ Shared Type X Min

ここで、Xは{人、組織、話題、Capitalized Sequence、略語、URL、電子メールアドレス、電話番号、住所、場所、日時、署名、…}のうちの1つである。

10

20

30

40

50

【0028】

特徴の第3の群は元の文書内にだけ存在する実体に関連する。

- ・SourceCompCount：元の文書内の共有されていない実体の個数、
- ・SourceCompAverage：元の文書内の共有されていない実体の平均重み、
- ・SourceCompMax：元の文書内の共有されていない実体の最大重み、
- ・SourceCompMin：元の文書内の共有されていない実体の最小重み、
- ・タイプされた元の補足的実体値：下記の4個の異なる種類で区別される元の補足的実体個数、平均値、最大値、および最小値であり、
- ・SourceTypeXCount
- ・SourceTypeXAverage
- ・SourceTypeXMax
- ・SourceTypeXMin

ここで、Xは{人、組織、話題、CapitalizedSequence、略語、URL、電子メールアドレス、電話番号、住所、場所、日時、署名、…}のうちの1つであり、

- ・SourceDocumentWeight：文書内の実体の個数および重みにより計算される元の文書の重み、
- ・SourceOccurrenceMagnitude：元の文書内の最大実体重み、
- ・SourceOccurrenceAverage：元の文書内の平均実体重み。

【0029】

特徴の第4の群は関連文書内にだけ存在するそれらの実体を含み、

- ・RelatedCompCount：潜在的に関連する文書内の共有されていない実体の個数、
- ・RelatedCompAverage：潜在的に関連する文書内の共有されていない実体の平均重み、
- ・RelatedCompMax：潜在的に関連する文書内の共有されていない実体の最大重み、
- ・RelatedCompMin：潜在的に関連する文書内の共有されていない実体の最小重み、
- ・タイプされた関連する補足的実体値：関連文書内の補足的実体のタイプされた数、平均値、最大値、最小値、
- ・RelatedTypeXCount
- ・RelatedTypeXAverage
- ・RelatedTypeXMax
- ・RelatedTypeXMin

ここで、Xは{人、組織、話題、CapitalizedSequence、略語、URL、電子メールアドレス、電話番号、住所、場所、日時、署名、…}のうちの1つであり、

- ・RelatedDocumentWeight：文書内の実体の個数および重みにより計算される潜在的に関連する文書の重み、
 - ・RelatedOccurrenceMagnitude：潜在的に関連する文書内の最大実体重み、
 - ・RelatedOccurrenceAverage：潜在的に関連する文書内の平均実体重み、
- を含む。

【0030】

上述の特徴はユーザのフィードバックに基づいて類似文書の計算を最適化するための一般的規則を生成するために使用できる。各ユーザに対する類似文書の発見のカスタマイズはユーザのフィードバックだけを用いて実現可能である。また、ユーザの場所、電子メー

10

20

30

40

50

ルからの社会的状況、時間情報、およびユーザのタスクのようなユーザ状況情報は、計算をさらにカスタマイズするために利用できる。

【0031】

図5は、本発明の一実施形態の、文書類似性を推定するための例示的コンピュータシステムを示している。一実施形態では、コンピュータ・コミュニケーション・システム500がプロセッサ502、メモリ504、および記憶デバイス506を含んでいる。記憶デバイス506は文書類似性推定アプリケーション508、ならびにアプリケーション510および512などの他のアプリケーションを保存する。動作時に、文書類似性推定アプリケーション508は記憶デバイス506からメモリ504へロードされ、その後、プロセッサ502により実行される。プログラムを実行している間、プロセッサ502は上述の機能を実行する。コンピュータ・コミュニケーション・システム500は任意の表示部514、キーボード516、およびポインティングデバイス518と接続される。

10

【図1】

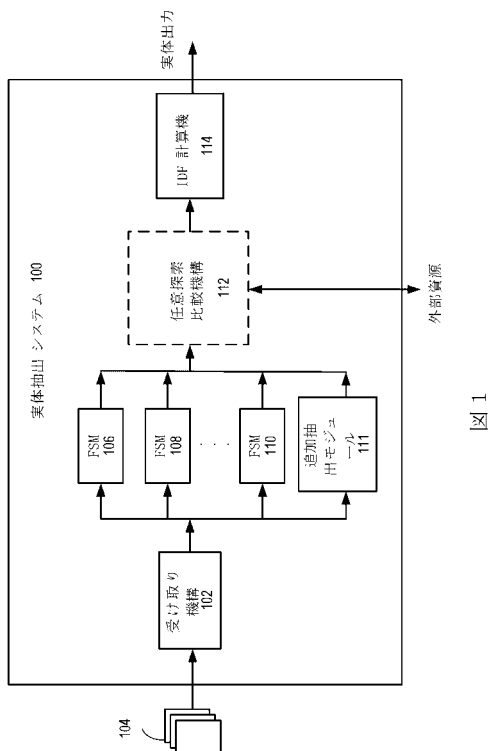


図 1

【図2】

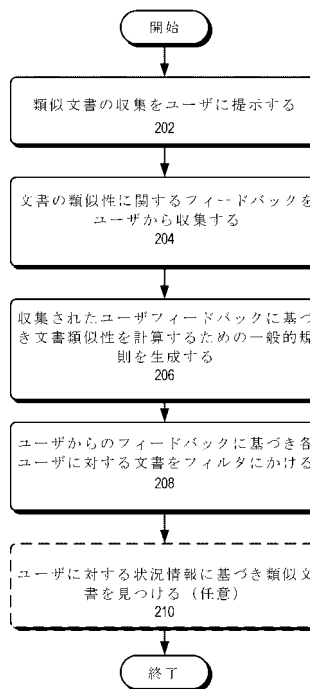


図 2

【図 3】

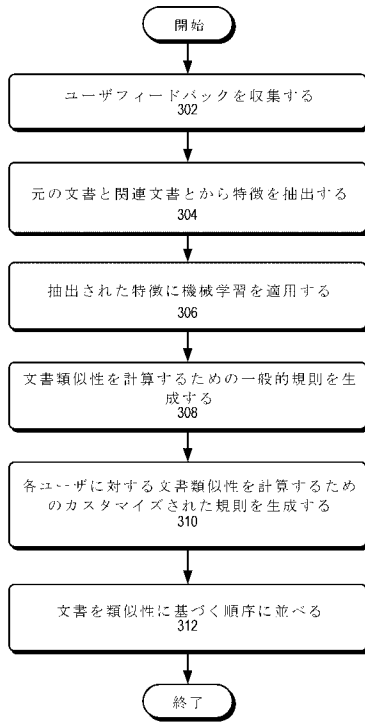


図 3

【図 4】

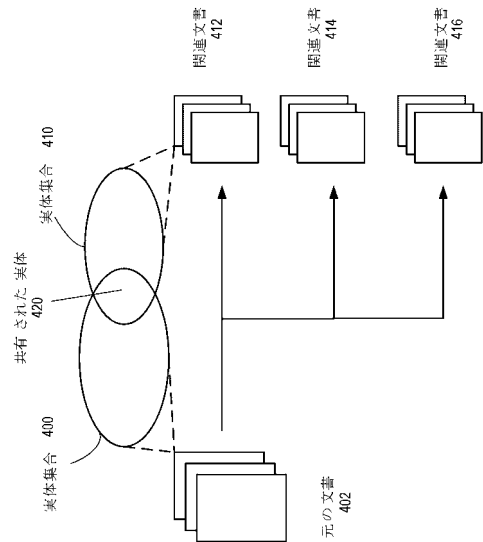


図 4

【図 5】

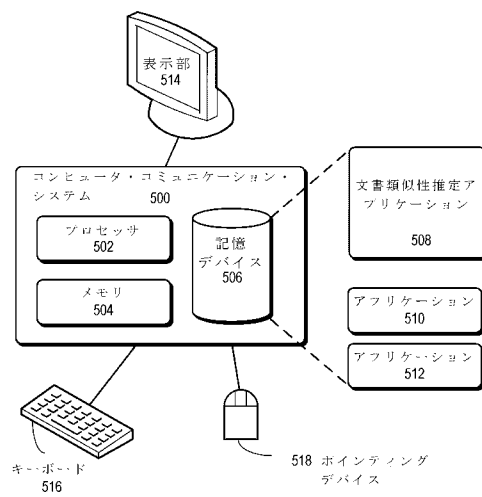


図 5

フロントページの続き

(74)代理人 100129218

弁理士 百本 宏之

(72)発明者 オリヴァー・ブルディスカ

アメリカ合衆国 カリフォルニア州 9 4 3 0 4 マウンテン・ビュー オルテガ・アヴェニュー
5 6 5 ナンバー 1 1