(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2016-532175
(P2016-532175A)
(43)公表日　平成28年10月13日(2016.10.13)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 17/30 (2006.01)	G06F 17/30 210A	

審査請求 有　予備審査請求 未請求　（全 25 頁）

(21)出願番号	特願2016-518124 (P2016-518124)
(86)(22)出願日	平成25年12月5日 (2013.12.5)
(85)翻訳文提出日	平成28年5月25日 (2016.5.25)
(86)国際出願番号	PCT/CN2013/088586
(87)国際公開番号	WO2015/043066
(87)国際公開日	平成27年4月2日 (2015.4.2)
(31)優先権主張番号	201310456381.X
(32)優先日	平成25年9月29日 (2013.9.29)
(33)優先権主張国	中国 (CN)

(71)出願人　515150025
ペキン　ユニバーシティ　ファウンダー　グループ　カンパニー，リミティド
中華人民共和国，ベイジン　１００８７１，ハイジャン　ディストリクト，チェンフー　ロード，ナンバー　２９８，ジョーングワンツゥン　ファウンダー　ビルディング，フィフス　フロア
(71)出願人　516020606
ファウンダー　アパビ　テクノロジー　リミティド
中華人民共和国，ベイジン　１０００８９　ハイジャン　ディストリクト，ジャーンホワ　ロード，ワーンヘユエン　シュウグワーン　ガーデン，ビルディング　５
最終頁に続く

(54)【発明の名称】キーワード拡張方法及びシステム並びに分類コーパス注釈方法及びシステム

(57)【要約】
キーワード拡張のための方法及びシステムである。初期キーワードにより検索を実行し、検索され且つ取得されたキーワードは次の検索の基礎となり、キーワード反復によりループ検索を実行する。２回連続して検索された単語のエラーが一定の範囲内にあるとき、検索されたキーワードは、初期キーワードの拡張キーワードとなる。このように、初期キーワードの多様な表現及び単語の多面的で黙示的な意味が取得され、且つ、初期キーワードは効果的且つ合理的に拡張され、コーパスが手動で構築されることを要する従来技術の問題を解決する。本方法は、容易であり、且つ、キーワード拡張のための高精度な方法である。また、これは、複数のコーパスの分類及び自動注釈のための方法及びシステムである。本方法は、各クラスについて１つ又は複数の初期主要キーワードを判定する。各クラスの拡張キーワードは、初期主要キーワード拡張により取得される。検索は、クラスに対応する拡張キーワードを使用して実行され、クラスコーパスは、クラスから選択され且つ注釈される。
【選択図】図１

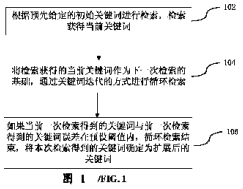

图 1　FIG. 1

102 Performing search according to predetermined initial keyword and obtaining current keyword
104 Obtained current keyword serving as basis for next search, and performing cyclical search by means of keyword iteration
106 If error between keyword obtained from one search and keyword obtained from previous search is within certain preset threshold, ending cyclical search and determining present obtained keyword to be expanded keyword

【特許請求の範囲】
【請求項１】
キーワード拡張方法であって、
所定の初期キーワードで検索して、現在のキーワードを取得するステップと、
検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行する、ステップと、を有し、
前記現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間のキーワードエラーが所定の閾値未満である場合は、前記ループ検索ステップを終了し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用することを特徴とする方法。
【請求項２】
請求項１に記載のキーワード拡張方法であって、現在のキーワードを取得する前記検索処理は、検索を通して取得された各単語の出現数をカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、ステップを有することを特徴とする方法。
【請求項３】
請求項１に記載のキーワード拡張方法であって、現在のキーワードを取得する前記検索処理は、検索を通して取得された単語の数及びこれらの出現数をカウントし、前記単語をこれらの出現数の降順にソートし、且つ、上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、ステップを有することを特徴とする方法。
【請求項４】
請求項２又は３に記載のキーワード拡張方法であって、検索を通して取得された単語を取得する前記方法は、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有する前記記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、ステップを有することを特徴とする方法。
【請求項５】
請求項４に記載のキーワード拡張方法であって、前記キーワード拡張方法は、単語分割の後にストップワードを削除するステップと、前記所定のキーワードと同時に現れる同時出現単語を取得するステップと、及び、これらの同時出現単語を検索を通して取得された単語として使用するステップと、を更に有することを特徴とする方法。
【請求項６】
請求項１乃至５のいずれか１項に記載のキーワード拡張方法であって、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間の前記キーワードエラーは、前記現在の検索で取得された前記キーワードの数と比較して、前記現在の検索と前記前の検索との間で異なるキーワードの数の割合であることを特徴とする方法。
【請求項７】
請求項６に記載のキーワード拡張方法であって、最初のｎ個のキーワードは、エラー評価用に前記現在の検索で取得されたキーワード及び前記前の検索を通して取得されたキーワードからそれぞれ取り出され、５≦ｎ≦１０であることを特徴とする方法。
【請求項８】
請求項１に記載のキーワード拡張方法であって、前記所定のエラー閾値は、２０％未満であることを特徴とする方法。
【請求項９】
請求項１に記載のキーワード拡張方法であって、前記現在の検索で取得されたキーワードが前記前の検索を通して取得されたキーワードと同じである場合は、前記現在の検索で取得された前記キーワードが拡張キーワードとして判定されることを特徴とする方法。
【請求項１０】
分類コーパスを注釈する方法であって、
各クラスについて１つ又は複数の初期主要キーワードを判定するステップと、

前記初期主要キーワードで、請求項１乃至９のいずれか１項に記載のキーワード拡張方法を使用して各クラスについて拡張キーワードを取得するステップと、
クラスに対応する前記拡張キーワードで検索して、分類コーパスを選択し、且つ、前記分類コーパスを注釈する、ステップと、
を有することを特徴とする方法。
【請求項１１】
キーワード拡張システムであって、
所定の初期キーワードで検索して、現在のキーワードを取得する取得ユニットと、
検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行するループ検索ユニットと、
前記現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間でキーワードエラーが所定の閾値未満であるか否かを判定する、判定ユニットであって、所定の閾値未満である場合は、前記ループ検索ユニットにループ検索処理を終了する指示を出し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用する、判定ユニットと、
を有することを特徴とするシステム。
【請求項１２】
請求項１１に記載のキーワード拡張システムであって、前記取得ユニットは、
記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、
検索を通して取得された各単語の出現数をそれぞれカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュールと、
を有することを特徴とするシステム。
【請求項１３】
請求項１１に記載のキーワード拡張システムであって、前記取得ユニットは、
記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、
検索を通して取得された前記単語の数及びこれらの出現数をカウントし、前記単語をこれらの出現数の降順にソートし、且つ、前記上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュール用の検索キーワード比較モジュールと、
を有することを特徴とするシステム。
【請求項１４】
請求項１２又は１３に記載のキーワード拡張システムであって、前記検索単語取得モジュールは、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、単語分割の後にストップワードを削除し、前記所定のキーワードと同時に現れる同時出現単語を取得し、及び、これらの同時出現単語を検索を通して取得された単語として使用することを特徴とするシステム。
【請求項１５】
請求項１１乃至１４のいずれか１項に記載のキーワード拡張システムであって、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間の前記キーワードエラーは、前記現在の検索で取得された前記キーワードの数と比較して、前記現在の検索と前記前の検索との間で異なるキーワードの数の割合であることを特徴とするシステム。
【請求項１６】
請求項１５に記載のキーワード拡張システムであって、最初のｎ個のキーワードは、エ

ラー評価用に前記現在の検索で取得されたキーワード及び前記前の検索を通して取得された前記キーワードからそれぞれ取り出され、５≦ｎ≦１０であることを特徴とするシステム。
【請求項１７】
請求項１１乃至１６のいずれか１項に記載のキーワード拡張システムであって、前記所定のエラー閾値は、２０％未満であることを特徴とするシステム。
【請求項１８】
請求項１１乃至１７に記載のキーワード拡張システムであって、前記現在の検索で取得されたキーワードが前記前の検索を通して取得されたキーワードと同じである場合は、前記現在の検索で取得された前記キーワードが拡張キーワードとして判定されることを特徴とするシステム。
【請求項１９】
分類コーパス注釈システムであって、
各クラスについて１つ又は複数の初期主要キーワードを判定するキーワード判定ユニットと、
前記初期主要キーワードで、請求項１１乃至１８のいずれか１項に記載の前記キーワード拡張システムを使用して各クラスの拡張キーワードを取得する、キーワード拡張ユニットと、
クラスに対応する前記拡張キーワードで検索して、分類コーパスを選択し、且つ、前記分類コーパスを注釈する、注釈ユニットと、
を有することを特徴とするシステム。
【請求項２０】
コンピュータによって実行されたときにキーワード拡張方法を実行するコンピュータ実行可能命令が保存された１つ又は複数のコンピュータ可読媒体であって、前記方法は、
所定の初期キーワードで検索して、現在のキーワードを取得するステップと、
検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行するステップであって、
前記現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間でキーワードエラーが所定の閾値未満である場合は、前記ループ検索処理を終了し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用する、ステップと、
を有することを特徴とする方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、キーワード拡張の方法及び分類コーパスを自動的に注釈する方法に関し、電子デジタルデータ処理の分野に関する。
【背景技術】
【０００２】
一般に、キーワードは、何らかの関連のある用語を総合して表し得る単語である。キーワードにより包含される事柄の包括性を改善するために、各キーワードは、一般に、いくつかの関連のある意味に対応する。キーワードベースの検索のヒット率を高めるためには、キーワードに対応する関連のある単語を取得するべく、特定の初期キーワードの拡張を実行するのが一般的であり、これは同時検索で使用される。キーワード拡張方法は、従来技術で提供され、最初にキーワード、用語及び識別コードを含むデータベースを構築するステップと、次に各キーワードを少なくとも１つの用語に対応させるステップと、関連のあるキーワードを識別コードに対応させるステップと、ユーザにより入力されたキーワードに従って、データベース中にあるキーワードに対応する識別コードを判定するステップと、識別コードに従って、識別コードに対応する関連のあるキーワードを抽出するステップと、関連のあるキーワードに従って、各キーワードに対応する用語を問い合わせるステ

ップと、を有する。この構成は、事前構築されたシソーラスに基づく自動キーワード拡張による検索方法を提供する。不十分に構築されたシソーラスは、キーワード拡張の正確性に深刻な影響を与える可能性がある。また、シソーラスの構築は、多くの人為的な経験を必要とし、ある程度主観的であることにより、分類の正確性に影響を及ぼす。

【０００３】

コーパス注釈は、主としてコーパスの分類特徴情報を記録することに関し、コーパスの表面的な分析の主要部分である。コーパス注釈は、情報検索、機械翻訳、主題事項分析及びテキスト処理といったような多くの分野で適用される。コーパス注釈の正確性は、テキスト分析やテキスト処理の正確性に直接影響を与える。

【０００４】

教師付きテキスト分類には、例えば、ＳＶＭ（サポートベクターマシン）を使用するテキスト分類があり、分類システムが決定された後に、注釈されたコーパスは、分類モデルを訓練するために分類システムの分類ごとに用意される。分類コーパス注釈は、一般に人為的に実行される。即ち、コーパス注釈を担当する人は、彼又は彼女の知識に従ってどのクラスにコーパス要素が属するのかを判定する。しかし、膨大な量のコーパス要素が注釈されるためには、人為的なコーパス分類は、以下のような問題をもつ。（１）高い人為的コスト、（２）長時間の人工的注釈、（３）人工的注釈における主観的影響、即ち、同じコーパス要素について、異なる人々により異なるクラスに分類される可能性がある、（４）膨大な量のコーパス要素の場合、大量の注釈のためにエラーが生じる傾向にある。

【０００５】

ＢＰニューラルネットワークに基づくコーパス注釈システムは、従来技術に開示されており、コーパスメモリ、注釈コーパスバッファリングメモリ、コーパス注釈結果コンパレータ及びＢＰニューラルネットワーク処理ユニットを有する。注釈中、ＢＰニューラルネットワーク処理ユニットは、コーパスメモリ中の注釈すべきコーパスを注釈し、且つ、その注釈結果を注釈コーパスバッファリングメモリに保存する。コーパス注釈結果コンパレータは、バッファリングメモリ中の結果を比較する。上記の技術的解決において、ＢＰニューラルネットワーク処理ユニットは、少なくとも２つの分類プロセッサを有する。注釈結果の処理において、少なくとも２つの分類プロセッサが注釈されるべきコーパスの注釈結果に対する幾つかの比較係数を特定の基準で満たすときのみ、注釈されるべきコーパスに注釈が実行され、且つ、コーパスメモリに保存される。この解決は、ＢＰニューラルネットワークアルゴリズムに基づくものである。このアルゴリズムは、複雑且つ計算量が多く、低い収束率で、且つ、膨大な量のコーパス要素を処理するときに時間が消費される。更に、少なくとも２つの分類プロセッサは、分類処理に必要であるので、多くのメモリが占有される。一方で、ニューラルネットワークを訓練するために、幾つかの大規模に注釈された複数のコーパスは、事前に準備されなければないが、しかしこれはコストがかかる。

【発明の概要】

【発明が解決しようとする課題】

【０００６】

本発明で解決されるべき技術的問題は、従来技術におけるキーワード拡張が、より強い主観性を有し、シソーラスを構築するために多くの仕事量を必要とし、且つ、キーワード拡張が低い精度をもつということである。客観的、単純且つ容易で、正確なキーワード拡張の解決が提供される。

【０００７】

本発明で解決されるべき別の問題は、従来技術で採用されたコーパス注釈方法が、ＢＰニューラルネットワークアルゴリズムに基づき、複雑且つ計算量が多く、低い収束率で、且つ、多くのメモリを占有するということである。一方、コーパス注釈において、幾つかの大規模に注釈された複数のコーパスは、分類プロセッサを訓練するために、事前に手動で準備されなければならないが、しかし、注釈された複数のコーパスを準備することは、コストがかかる。分類コーパスを自動的に注釈するための機械補助による方法を提供する

ことが望ましい。
【課題を解決するための手段】
【０００８】
上記の技術的問題を解決するために、本発明は、以下の技術的解決を提供する。
【０００９】
キーワード拡張方法は、所定の初期キーワードで検索して、現在のキーワードを取得するステップと、検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行する、ステップと、を有し、前記現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間のキーワードエラーが所定の閾値未満である場合は、前記ループ検索ステップを終了し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用する。
【００１０】
任意の選択で、現在のキーワードを取得する前記検索処理は、検索を通して取得された各単語の前記出現数をカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、ステップを有する。
【００１１】
任意の選択で、現在のキーワードを取得する前記検索処理は、検索を通して取得された単語の数及びこれらの出現数をカウントし、前記単語をこれらの出現数の降順にソートし、且つ、上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、ステップを有する。
【００１２】
任意の選択で、検索を通して取得された単語を取得する前記方法は、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有する前記記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された前記単語として使用する、ステップを有する。
【００１３】
任意の選択で、単語分割の後にストップワードを削除するステップと、前記所定のキーワードと同時に現れる同時出現単語を取得するステップと、及び、これらの同時出現単語を検索を通して取得された前記単語として使用するステップと、を更に有する。
【００１４】
任意の選択で、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間の前記キーワードエラーは、前記現在の検索で取得された前記キーワードの数と比較して、前記現在の検索と前記前の検索との間で異なるキーワードの数の割合である。
【００１５】
任意の選択で、最初のｎ個のキーワードは、エラー評価用に前記現在の検索で取得されたキーワード及び前記前の検索を通して取得されたキーワードからそれぞれ取り出され、５≦ｎ≦１０である。
【００１６】
任意の選択で、前記所定のエラー閾値は、２０％未満である。
【００１７】
任意の選択で、前記現在の検索で取得されたキーワードが前記前の検索を通して取得されたキーワードと同じである場合は、前記現在の検索で取得された前記キーワードが拡張キーワードとして判定される。
【００１８】
本明細書中に記載のキーワード拡張方法を使用して分類コーパスを注釈する方法は、各クラスについて１つ又は複数の初期主要キーワードを判定するステップと、前記初期主要キーワードで、キーワード拡張方法を使用して各クラスについて拡張キーワードを取得するステップと、クラスに対応する前記拡張キーワードで検索して、分類コーパスを選択し、且つ、前記分類コーパスを注釈する、ステップと、を有する。

【００１９】
キーワード拡張システムは、所定の初期キーワードで検索して、現在のキーワードを取得する取得ユニットと、検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行するループ検索ユニットと、前記現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間でキーワードエラーが所定の閾値未満であるか否かを判定する、判定ユニットであって、所定の閾値未満である場合は、前記ループ検索ユニットにループ検索処理を終了する指示を出し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用する、判定ユニットと、を有する。
【００２０】
任意の選択で、取得ユニットは、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、検索を通して取得された各単語の出現数をそれぞれカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュールと、を有する。
【００２１】
任意の選択で、取得ユニットは、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、検索を通して取得された前記単語の数及びこれらの出現数をカウントし、前記単語をこれらの出現数の降順にソートし、且つ、前記上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュール用の検索キーワード比較モジュールと、を有する。
【００２２】
任意の選択で、キーワード拡張システムにおいて、前記検索単語取得モジュールは、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、単語分割の後にストップワードを削除し、前記所定のキーワードと同時に現れる同時出現単語を取得し、及び、これらの同時出現単語を検索を通して取得された前記単語として使用する。
【００２３】
任意の選択で、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間の前記キーワードエラーは、前記現在の検索で取得された前記キーワードの数と比較して、前記現在の検索と前記前の検索との間で異なるキーワードの数の割合である。
【００２４】
任意の選択で、最初のｎ個のキーワードは、エラー評価用に前記現在の検索で取得されたキーワード及び前記前の検索を通して取得された前記キーワードからそれぞれ取り出され、５≦ｎ≦１０である。
【００２５】
任意の選択で、キーワード拡張システムにおいて、前記所定のエラー閾値は、２０％未満である。
【００２６】
任意の選択で、キーワード拡張システムにおいて、前記現在の検索で取得されたキーワードが前記前の検索を通して取得されたキーワードと同じである場合は、前記現在の検索で取得された前記キーワードが拡張キーワードとして判定される。
【００２７】
本キーワード拡張システムを使用して分類コーパス注釈システムは、各クラスについて１つ又は複数の初期主要キーワードを判定するキーワード判定ユニットと、前記初期主要キーワードで、上記に記載の前記キーワード拡張システムを使用して各クラスの拡張キー

ワードを取得する、キーワード拡張ユニットと、クラスに対応する前記拡張キーワードで検索して、分類コーパスを選択し、且つ、前記分類コーパスを注釈する、注釈ユニットと、を有する。

【発明の効果】

【００２８】

本開示の上記の技術的解決は、従来技術よりも１つ又はそれ以上の利点を持つ。

【００２９】

（１）本開示のキーワード拡張方法の一実施形態では、所定の初期キーワードで検索するステップを通して、次の検索の基礎として使用されるキーワードを取得するために、キーワード反復を通してループ検索を実行し、現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間のキーワードエラーが一定の範囲内である場合は、現在の検索で取得されたキーワードを初期キーワードの拡張キーワードとして使用し、この方法は初期キーワードの複数の表現及び複数の意味を得る可能性があり、効果的で意味のある初期キーワードの拡張を実現し、且つ、従来技術におけるシソーラスの手動構築の問題を解決し得る。本キーワード拡張方法は、容易な実施と高い精度において有利である。

【００３０】

（２）本キーワード拡張方法では、検索を通して取得された各単語の数の出現数をカウントするステップを通して、所定の閾値よりも大きな閾値を有する単語を検索を通して取得されたキーワードとして獲得する。または、検索を通して取得された単語の数及びそれらの出現数をカウントし、これらの出現数の降順に単語をソートし、且つ、上位の割合を占める単語を検索を通して取得されたキーワードとして獲得する。取得されたキーワードは、統計的有意性をもち、そのキーワードのあらゆる意味と関連するこれらの単語を容易に見つける。

【００３１】

（３）本キーワード拡張方法では、高い関連性を有する記事を取得するために、単語は記事リポジトリ中の検索を通して取得され、単語分割を実行し、ストップワードを削除し、且つ、同時出現単語を取得する。様々なフィルタリングステップの後、不要な単語は削除され、且つ、効果的な単語が取得されるかもしれない。

【００３２】

（４）本キーワード拡張方法では、現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間のキーワードエラーが一定の範囲内であるときに、検索処理は終了し、且つ、拡張されたキーワードが取得される。キーワード反復及び収束を通して理想的なキーワードが取得されるので、処理速度が速くなり且つ作業効率が向上する。

【００３３】

（５）本キーワード拡張方法では、現在の検索で取得されたキーワードが前の検索を通して取得されたキーワードと同じであるときに、現在の検索で取得されたキーワードは拡張キーワードとして判定され、且つ、拡張キーワードの正確性が向上する。

【００３４】

（６）本発明は、分類コーパス注釈方法を提供し、分類コーパスを取得するために検索処理において拡張されたキーワードが使用され、分類コーパス注釈の実効性及び正確性を向上し得る。この分類コーパス注釈方法は、この技術分野におけるＢＰニューラルネットワークアルゴリズムに基づき分類コーパス注釈方法の問題を効果的に回避し得る。即ち、このアルゴリズムは、複雑且つ計算量が多く、低い収束率で、且つ、膨大な量のコーパス要素を処理するときに時間が消費される。さらに、分類処理のために少なくとも２つの分類処理が要求されるので、多くのメモリが占有される。一方で、ニューラルネットワークを訓練するために、幾つかの大規模に注釈された複数のコーパスが事前に準備されなければならず、故にコストがかかる。

【００３５】

本発明のより簡単で且つ明確な理解のために、本発明の詳細な説明が以下の図面を参照するとともに与えられる。

【図面の簡単な説明】
【００３６】
【図１】本発明の一実施形態によるキーワード拡張方法のフローチャートである。
【図２】本発明の一実施形態による分類コーパス注釈方法のフローチャートである。
【図３】本発明の一実施形態によるキーワード拡張システムの構造図である。
【図４】本発明の一実施形態による分類コーパス注釈システムの構造図である。
【発明を実施するための形態】
【００３７】
実施形態１
本実施形態は、キーワード拡張方法を提供し、図１に示すように、本方法は以下のステップを有する。
【００３８】
ステップ１０２：　所定の初期キーワードで検索して、現在のキーワードを取得する。本実施形態において、検索は、記事レポジトリ中で初期キーワードにより実行され、高い関連性を有する記事を取得する。次に、単語分割は、高い関連性を有するこれらの記事で実行され、且つ、単語分割の結果は、検索を通して取得された単語として使用される。各単語の出現数はカウントされ、且つ、所定の閾値である５０よりも大きな出現数を有する単語は、検索（記事レポジトリのサイズ及びキーワードのポピュラリティに従って特定される）を通して取得されたキーワードとして使用される。この方法で取得されたキーワードは、統計的有意性をもち、そのキーワードのあらゆる意味と関連するこれらの単語を容易に見つける。
【００３９】
ステップ１０４：　検索を通して取得された現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行する。この検索処理は、ステップ１０２の特定の処理と類似する。このステップにおいて、検索は、前の検索で取得されたキーワードをこの検索処理で使用されるキーワードとして使用して実行される。検索を通して取得されたキーワードは、次の検索処理で使用されるキーワードとして順番に使用される。このように、検索は、キーワード反復を通して実行される。
【００４０】
ステップ１０６：各検索の後、現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間のキーワードエラーが所定の閾値よりも少ない場合は、ループ検索処理を終了し、且つ、現在の検索で取得されたキーワードを拡張キーワードとして使用する。例えば、現在の検索で取得されたキーワードは、前の検索で取得されたこれらのキーワードと比較され、同一である場合に、現在の検索で取得されたキーワードは、拡張キーワードとして使用される。このように、拡張キーワードの正確性は、向上し得る。
【００４１】
上記の実施形態による本キーワード拡張方法において、所定の初期キーワードで検索して、次の検索の基礎として使用されるキーワードを取得する間に、キーワード反復を通してループ検索を実行し、現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間のキーワードエラーが一定の範囲内にある場合は、現在の検索で取得されたキーワードを拡張キーワードとして使用する。本方法は、初期キーワードの複数の表現及び複数の意味を得る可能性があり、効果的で意味のある初期キーワードの拡張を実現し、且つ、従来技術におけるシソーラスの手動構築の問題を解決し得る。本キーワード拡張方法は、容易な実施と高い正確性において有利である。
【００４２】
別の代替実施形態として、現在の検索で取得されたキーワードは、前の検索で取得されたこれらのキーワードと比較してもよく、全キーワードに対する異なるキーワードの割合が所定の閾値、例えば、２０％よりも小さい場合に、現在の検索で取得されたキーワードは、拡張キーワードとして判定される。

【００４３】
実施形態２
（１）所定の初期キーワードで検索して、現在のキーワードを取得する。
（２）検索を通して取得された現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行する。
【００４４】
ステップ（１）及び（２）の検索処理において、本検索方法は、以下の通りである。
所定のキーワードで記事レポジトリ中を検索して、高い関連性を有する記事を取得する。次に、高い関連性を有するこれらの記事について単語分割を実行する。単語分割の後にストップワードを削除する。所定のキーワードと同時に現れる同時出現単語を取得し、及び、これらの同時出現単語を検索を通して取得された単語として使用する。ここで、同時出現単語は、スライディングウィンドウ法を使用して取得されてもよい。
【００４５】
上記の実施形態において、単語は、単語分割を通して取得され、ストップワードを削除し、且つ、同時出現単語を取得する。様々なフィルタリングステップの後、不要な単語は削除され、効果的な単語が取得されうる。
【００４６】
検索を通して取得された単語の数及びこれらの出現数をカウントし、単語をこれらの出現数の降順にソートし、且つ、上位の割合、例えば５０％（ここで、割合は必要に応じて指定されてもよい）を占める単語を検索を通して取得された現在のキーワードとして獲得する。例えば、１００単語が検索を通して取得された場合、頻繁に現れる単語の上位２０％は、検索を通して取得されたキーワードとして獲得される。
【００４７】
ここで、別の代替実施形態として、出現数は、事前に正規化される。正規化の方法は、検索を通して取得された様々な単語について、これらの出現数の合計を計算し、単語毎に、この単語の出現数／合計の値を正規化した値として使用する。正規化した値を降順にソートし、且つ、上位の値をもつ単語の割合を検索を通して取得されたキーワードとして獲得する。
【００４８】
この処理において、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間のエラーは、現在の検索で取得されたキーワードの数と比較して、現在の検索と前の検索との間で異なるキーワードの数の割合として定義される。このエラーが１０％より小さいときは、この検索処理は終了し、且つ、現在の検索で取得されたキーワードは拡張キーワードとして使用される。
【００４９】
別の代替実施形態として、このエラーは、例えば、最初の５個又は１０個のキーワードといったように、最初のｎ個のキーワードから計算されてもよい。エラーが２０％より小さいときは、処理が終了し、且つ、拡張キーワードが取得される。
【００５０】
現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間のエラーが一定の範囲内であるときに、検索処理は終了し、且つ、拡張キーワードが取得される。キーワード反復及び収束を通して理想的なキーワードが取得されるので、処理速度は向上し、且つ、作業効率が改善される。
【００５１】
実施形態３
図３は、本発明の実施形態によるキーワード拡張システムの構造図である。
図３に示すように、キーワード拡張システムは、
（１）　所定の初期キーワードで検索して、現在のキーワードを取得する取得ユニット３１を有する。本実施形態において、取得ユニットは、記事リポジトリ中の所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事に

ついて単語分割を実行し、且つ、単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、検索を通して取得された各単語の出現数をそれぞれカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュールと、を有する。

【００５２】

代替実施形態として、取得ユニットは、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、検索を通して取得された単語の数及びこれらの出現数をカウントし、これらの出現数の降順に単語をソートし、且つ、上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュール用の検索キーワード比較モジュールと、を有する。

【００５３】

（２）検索を通して取得された現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行されるループ検索ユニット３２。

【００５４】

上記に記載の検索処理は、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、単語分割の結果を検索を通して取得された単語として使用する。本キーワード拡張システムにおいて、ストップワードは、単語分割の後にも削除され、且つ、所定のキーワードと同時に現れる同時出現単語が取得され、且つ、検索を通して取得された単語として使用される。検索単語取得モジュール又は検索キーワード比較モジュールは、検索を通して取得された単語について統計的に実行して、検索を通して取得されたキーワードを取得する。

【００５５】

（３） 現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間のキーワードエラーが所定の閾値（例えば、１０％）よりも少ないか否かを判定する判定ユニット３３は、エラーが１０％未満である場合、ループ検索ユニットにループ検索処理を終了する指示を出し、且つ、現在の検索で取得されたキーワードを拡張キーワードとして使用する。現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間のエラーは、現在の検索で取得されたキーワードの数と比較して、現在の検索と前の検索との間で異なるキーワードの数の割合として定義される。代替実施形態として、エラー評価は、最初のｎ個のキーワードを使用して実行されてもよく、例えば、５≦ｎ≦１０である。

【００５６】

代替実施形態として、検索精度を向上するために、現在の検索で取得されたキーワードが前の検索を通して取得されたこれらのキーワードと同じである場合にのみ、現在の検索で取得されたキーワードは、拡張キーワードとして判定される。

【００５７】

実施形態４

個別の適用例は、以下により与えられる。

検索は、初期キーワード“コップ”で実行される。記事レポジトリ（５００個の記事）は、単語“コップ”で検索され、且つ、一連のキーワード“水”、“ケトル”、“ティーカップ”、“水ディスペンサ”、“飲み物”は、上述の検索方法及びキーワードを取得する方法で取得される。

検索は、上記で取得された一連の単語で再び実行され、一連のキーワード“水”、“ティーカップ”、“ケトル”、“サーモスボトル”、“バケット”が取得される。

エラー４０％は、上記２つの検索結果の比較を通して判定される。故に、検索は、キーワードとして上記の検索結果で更に実行され、“水”、“ティーカップ”、“カップ”、“水グラス”、“ケトル”という結果が取得される。

エラー４０％は、この検索結果と前の検索結果との比較を通して判定され、閾値２０％を満たさず、そして、上記キーワードで検索処理をし続け、検索結果“水”、“ティーカップ”、“カップ”、“水グラス”、“ケトル”を取得する。

２０％未満のエラーは、この検索結果と前の検索結果の比較を通して判定され、閾値の基準を満たすので、検索処理は終了する。現在の検索の結果である“水”、“ティーカップ”、“カップ”、“水グラス”、“ケトル”は、キーワード“カップ”の拡張の後にキーワードとして使用される。

【００５８】

実施形態５

本実施形態は、キーワード拡張方法を使用した分類コーパス注釈の方法を提供し、図２のフローチャートに示すように、以下のステップを有する。

ステップ２０２：各クラスについて１つ又は複数の初期主要キーワードを判定する。

ステップ２０４：初期主要キーワードで、上記に記載のキーワード拡張方法を使用して各クラスの拡張キーワードを取得する。

ステップ２０６：クラスに対応する拡張キーワードで検索して、分類コーパスを選択し、且つ、分類コーパスを注釈する。

【００５９】

実施形態６

図４は、本発明の一実施形態による分類コーパス注釈システムの構造図である。

図４に示すように、キーワード拡張システムを使用した分類コーパス注釈のシステムは、各クラスについて１つ又は複数の初期主要キーワードを判定するキーワード判定ユニット４１と、初期主要キーワードで、キーワード拡張システムを使用して各クラスの拡張キーワードを取得する、キーワード拡張ユニット４２であって、所定の初期主要キーワードで検索して、現在のキーワードを取得する取得サブユニットと、検索を通して取得された現在のキーワードを次の検索の基礎として使用し、且つ、キーワード反復を通してループ検索を実行する、ループ検索サブユニットと、現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間でキーワードエラーが所定の閾値未満であるか否かを判定する、判定サブユニットであって、所定の閾値未満である場合は、ループ検索ユニットにループ検索処理を終了する指示を出し、且つ、現在の検索で取得されたキーワードを拡張キーワードとして使用する、判定サブユニットと、を有する、キーワード拡張ユニット４２と、クラスに対応する拡張キーワードで検索して、分類コーパスを選択し、且つ、分類コーパスを注釈する、注釈ユニット４３と、を有する。

【００６０】

実施形態７

キーワード拡張方法を使用した分類コーパス注釈方法は、１つの適用例を参照して説明される。

Ｓ１：　各クラスについて１つ又は複数の初期主要キーワードを判定する。

分類システムにおいて、３つのクラス｛軍事、経済、スポーツ｝が与えられる。各クラスについて１つ又は複数の初期主要キーワードが手動で判定される。“軍事”を例にとると、キーワード｛戦争、難民、死傷者｝は、初期主要キーワードとして判定される。全てのテキストレポジトリは、新聞及び定期刊行物データベースから選択された記事で構築されている。

Ｓ２：　初期主要キーワードの拡張を通して各クラスの拡張キーワードを取得する。

ステップＳ２において、各クラスの拡張キーワードは、反復的検索を通して取得され、以下のステップを有する。

Ｓ２１：　クラスの初期主要キーワードで、検索を通してこのクラスの拡張キーワードの候補を取得する。

Ｓ２１０：　“軍事”クラスの初期主要キーワード｛戦争、難民、死傷者｝を選択する。

Ｓ２１１：　初期主要キーワード｛戦争、難民、死傷者｝で検索をし、且つ、これらの

関連性に従って、最初の１０００個の記事を取得する。
他の実施形態では、記事の数はｎ個であり、ここで、ｎ≧２、ｎは整数である。ｎの値は、３０≦ｎ≦２０００の範囲である。ｎの値は、５０、１００、５００、７００、１２００、１７００、２０００及び他の異なる値を選択してもよく、且つ、ユーザの要望及びクラス特性に従って選択されてもよい。
Ｓ２１２：クラス“軍事”の１０００個の記事について単語分割を実行し、且つ、ストップワードを削除する。
【００６１】
本実施形態において、ＮＬＰＩＲトークナイザは、ｎ個の記事及びストップワードについて単語分割を実行されるために使用される。ストップワードは、単語分割の後にストップワード辞書を使用して除去されてもよい。使用されるＮＬＰＩＲトークナイザは、中国語の単語分割、ＰＯＳタギング、言語要素識別、ユーザ辞書、マイクロブログ単語分割、新しい単語マイニング及びキーワード抽出といった機能を有し、且つ、ＧＢＫ、ＵＴＦ８、ＢＩＧ５のエンコード形式をサポートする。このトークナイザは、完全な機能、高速な処理速度及び高い信頼性をもつ。
【００６２】
別の実施形態において、ＣＪＫトークナイザ又はＩＫトークナイザは、ｎ個の記事について単語分割を実行し、且つ、ストップワードを削除するために使用されてもよい。ストップワードは、単語分割の後にストップワード辞書を使用して除去されてもよい。中国語のテキストレポジトリのために、中国語テキスト文書の処理専用で、高速な処理速度を有し、安定性及び信頼性があるＣＪＫトークナイザが使用されてもよい。また、ＩＫトークナイザも適している。ストップワードは、単語分割の後にストップワード辞書を使用して除去されてもよく、又は、ＩＫトークナイザのストップワード辞書の設定を通して除去されてもよい。前方向且つ後方向における完全な分割及び前方向且つ後方向における最大一致分割は、辞書ベースの分割に基づき実現されてもよい。このトークナイザは、辞書のストレージを最適化し、少ないメモリを消費し、高速な処理速度及び高い信頼性を有する。
【００６３】
Ｓ２１３： スライディングウィンドウ法を使用して、キーワードの周囲にサイズ７のスライディングウィンドウを有する単語を拡張キーワードの候補として取得する。主要キーワードの前の３つの単語及び主要キーワードの後の３つの単語及び３つの単語自体は、拡張キーワードの候補として使用される。主要キーワードの前又は後が３つの単語よりも少ない場合、主要キーワードの前又は後の全ての単語は、選択される。
【００６４】
別の実施形態において、主要キーワードの前の６つのキーワード及び主要キーワード自体は、拡張キーワードの候補として使用されてもよい。または、主要キーワードの前の４つの単語、主要キーワードの後の２つの単語及び主要キーワード自体は、拡張キーワードの候補として使用されてもよい。または、主要キーワードの前の２つの単語、主要キーワードの後の４つの単語及び主要キーワード自体は拡張キーワードの候補として使用されてもよい。拡張キーワードの前又は後に十分な単語が存在しない場合、主要キーワードの前又は後の全ての単語が選択されてもよい。
【００６５】
別の実施形態において、スライディングウィンドウはサイズＳを有し、ここで、Ｓ≧２であり、Ｓは、整数である。サイズＳのスライディングウィンドウは、３≦Ｓ≦１０の値を有する。スライディングウィンドウの値は、４、５、６、８、９、１０及び他の異なる値から選択されてもよく、又は、ユーザの要望に従って選択されてもよい。
【００６６】
本発明の分類コーパスを自動的に注釈する方法において、キーワードは、スライディングウィンドウ法を使用して取得される。本方法は、ウィンドウサイズの制限を通して許容できる単語の最大数を制御してもよい。本アルゴリズムは、単純であり、高速な処理速度及び高い正確性を有する。

Ｓ２２２：　最初の１０個の拡張キーワードの候補を新しい主要キーワードとして選択する。

【００６７】

別の実施形態において、最初のｍ個の拡張キーワードの候補は、新しい主要キーワードとして選択されてもよく、ここで、ｍ≧２、ｍは、整数であり、ｍの値は、５≦ｍ≦３０の範囲で、ｍの値は、５、７、１３、１７、２５、２７、３０及び他の異なる値から選択されてもよく、且つ、ユーザの要望及びクラス特性に従って選択されてもよい。

【００６８】

Ｓ２２３：　新しい主要キーワードが変化せず且つ特定のキーワードの組に収束するまで、ステップＳ２１１に戻り、且つ、新しい主要キーワードで検索する。

クラス”軍事”の初期主要キーワードの拡張を通して取得された１０個のキーワードは、初期主要キーワードに基づき反復的方法で取得された拡張キーワード｛難民、イラク、戦争、アフリカ、家、強制される、アフガニスタン、ヨルダン、戦闘、再定住｝である。

Ｓ２３：　キーワードの組をチェックし、且つ、クラス特性に適合しないキーワードを削除して、このクラスの拡張キーワードを取得する。

ユーザが軍事問題を研究していると仮定すると、クラス”軍事”の特性に適合しないキーワード｛家、再定住｝は削除されてもよい。

キーワードの組のチェックを通して、幾つかのクラス特性に適合しないキーワードは、削除されてもよく、取得された拡張キーワードは、より正確になる。

Ｓ３：　クラスに対応する拡張キーワードで検索して、分類コーパスを選択し、且つ、注釈を実行するには、以下のステップを有する。

Ｓ３１：　全テキストレポジトリ中を拡張キーワード｛難民、イラク、戦争、アフリカ、強制される、アフガニスタン、ヨルダン、戦闘｝で検索し、且つ、関連性の降順でソートする。

Ｓ３２：　最初の１０００個の記事をチェックし、且つ、分類コーパスを選択し、且つ、それを”軍事”として注釈する。

【００６９】

他の実施形態において、最初のＫ個の記事は、チェック用に選択されてもよく、ここで、Ｋ≧１０であり、Ｋは、整数で、Ｋの値は、１００≦ｍ≦２０００の範囲であり、Ｋの値は、１５００、１７００、２０００及び他の異なる値から選択されてもよく、且つ、クラス特性に従って選択されてもよい。

最初のＫ個の記事のチェックにおいて、幾つかのクラス特性に適合しない記事は、削除されてもよく、クラス特性に適合するその他残りの記事をこのクラスのコーパスとして注釈する。

本発明の分類コーパスを自動的に注釈する方法において、各検索で取得される記事の数を限定することを通して、処理すべき記事の数を減らし、処理速度が向上され得る。一方で、低い関連性を有する記事は、削除されてもよく、取得された新しい主要キーワードはより正確になる。

【００７０】

本発明の分類コーパスを自動的に注釈する方法において、各検索は、全テキストでマッチングが実行される全文検索であり、結果として高い再現率となり、且つ、注釈されたコーパスはより正確になる。

本発明の分類コーパスを自動的に注釈する方法において、拡張キーワードの検索を通して取得されたコーパスをチェックすることを通して、幾つかのクラス特性に合致しない記事を削除し、且つ、このクラスのコーパスとして残りの記事を注釈し、コーパスの注釈が

より正確になる。
【００７１】
実施形態８
本実施形態は、分類コーパスを注釈する方法の別の特定の実施形態を提供する。
ステップ１： 分類システムにおいて、３つのクラス｛軍事、経済、スポーツ｝が与えられる。各クラスについて、１つ又は複数の初期主要キーワードを手動で判定する。“軍事”を例にとると、キーワード｛戦争、難民、死傷者｝は、初期主要キーワードとして判定される。全てのテキストレポジトリは、新聞及び定期刊行物データベースから選択された記事で構築される。
ステップ２： クラス”軍事”のために、初期主要キーワード｛戦争、難民、死傷者｝での全文検索を通して最初の１０００個の記事を取得する。
ステップ３： 取得された１０００個の記事について単語分割を実行し、且つ、ストップワードを削除する。
ステップ４： スライディングウィンドウ法を使用して、サイズ６のスライディングウィンドウ中のキーワードの周囲にあるキーワードを取得する。
ステップ５： キーワードの出現数をカウントし、且つ、これらの出現数の降順にキーワードをソートする。
ステップ６： ステップ５で取得されたキーワードから、最初の１０個のキーワードを新しい主要キーワードとして選択する。
ステップ７： 最初の１０個のキーワードに変化が生じなくなるまで、ステップ２からステップ６を繰り返す。即ち、最初の１０個のキーワードが特定のキーワードの組に収束する。取得された１０個のキーワードは、初期主要キーワードに基づき反復的方法で取得された拡張キーワード｛難民、イラク、戦争、アフリカ、家、強制される、アフガニスタン、ヨルダン、死傷者、再定住｝である。
ステップ８： 拡張キーワードを手動でチェックして、クラス特性に適合しないキーワード｛家、再定住｝を削除する。
ステップ９： 全テキストレポジトリ中をこのクラスに対応する拡張キーワード｛難民、イラク、戦争、アフリカ、強制される、アフガニスタン、ヨルダン、戦闘｝で検索して、最初の１０００個の記事を取得し、このクラスのコーパスの候補を形成する。
ステップ１０： これらの１０００個の記事を手動でチェックして、このクラスのコーパスを選択する。
ステップ１１： 全クラスについて、ステップ２からステップ１０を繰り返して、分類システムにおいて各クラスの注釈コーパスを取得する。
【００７２】
明らかに、上記の実施形態は、明確な説明のために与えられた例にすぎず、本発明を限定するものではない。当業者によって、上記の説明に基づき他の変更及び変形がなされてもよく、本明細書中に網羅的に記載され且つ記載できるものではない。派生したこれらの明らかな変更又は変形は、本発明の保護の範囲内にある。
【００７３】
本発明は、コンピュータによって実行されたときにキーワード拡張方法を実行するコンピュータ実行可能命令が保存された１つ又は複数のコンピュータ可読媒体を更に提供し、本方法は、所定の初期キーワードで検索して、現在のキーワードを取得するステップと、検索を通して取得された現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行する、ステップと、現在の検索で取得されたキーワードと前の検索で取得されたこれらのキーワードとの間でキーワードエラーが所定の閾値未満である場合は、ループ検索処理を終了し、且つ、現在の検索で取得されたキーワードを拡張キーワードとして使用する、ステップと、を有する。
【００７４】
本発明は、コンピュータによって実行されたときに上述の分類コーパスを注釈する方法を実行するコンピュータ実行可能命令が保存された１つ又は複数のコンピュータ可読媒体

を更に提供する。

【００７５】

当業者は、本出願の実施形態は、方法、システム、又はコンピュータプログラムのプロダクトとして提供することができることを理解すべきである。従って、本出願は、全体的にハードウェアの実施形態、全体的にソフトウェアの実施形態、又はソフトウェアとハードウェアを組み合わせた実施形態の形態を使用することができる。更には、本出願は、コンピュータによって実行可能なプログラミングコードを有する１つ又は複数の記憶媒体（限定を伴うことなしに、ディスクメモリ、ＣＤ－ＲＯＭ、光メモリなどを含む）上において実行されるコンピュータプログラムプロダクトの形態を使用することもできる。

【００７６】

本出願は、本発明の実施形態による方法、機器（システム）、及びコンピュータプログラムプロダクトのフローチャート及び／又はブロックダイアグラムを参照して記述されている。フローチャート及び／又はブロックダイアグラム中のそれぞれのフロー及び／又はブロックのみならず、フローチャート及び／又はブロックダイアグラム中のフロー及び／又はブロックの組合せは、コンピュータプログラム命令を通じて実現可能であることを理解されたい。このようなコンピュータプログラム命令は、フローチャート中の１つ又は複数のフロー及び／又はブロックダイアグラムの１つのブロック又は複数のブロック内において規定されている機能を実現する装置が、コンピュータ又はプログラム可能なデータ処理機器の任意のその他のプロセッサによって実行される命令によって生成されるように、機械を生成するべく、汎用コンピュータ、特殊目的コンピュータ、組込み型プロセッサ、又はプログラム可能なデータ処理機器の任意のその他のプロセッサに提供されることができる。

【００７７】

また、このようなコンピュータプログラム命令は、コンピュータの可読メモリ内において保存されたコマンドがコマンド装置のプロダクトを生成するように、特定のスタイルにおける動作にコンピュータ又はその他のプログラム可能なデータ処理機器を導きうるコンピュータの可読メモリ内に保存可能であり、このような命令装置は、フローチャート中の１つ又は複数のフロー及び／又はブロックダイアグラムの１つ又は複数のブロック内に規定されている機能を実現することができる。

【００７８】

また、このようなコンピュータプログラム命令は、コンピュータ又はその他のプログラム機器によって実行される命令が、フローチャート中の１つ又は複数のフロー及び／又はブロックダイアグラムの１つのブロック又は複数のブロック内において規定されている機能を実現するように、コンピュータ又はその他のプログラム可能な機器上において一連の動作ステップを実行してコンピュータによって実現されるプロセスを生成するように、コンピュータ又はその他のプログラム可能なデータ処理機器上に読み込むこともできる。

【００７９】

以上、本願の好適な実施形態について説明したが、当業者であれば、基本的な創造的概念を理解すれば、これらの実施形態の更なる変更及び変形を実施することができる。従って、添付の請求項は、好適な実施形態と、本願の範囲内のすべての変更及び変形と、を包含するべく解釈されることを意図している。

【図１】

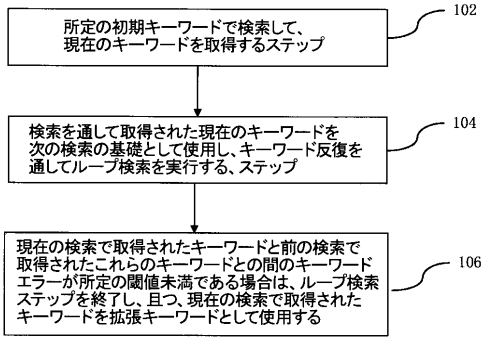

図1

【図２】

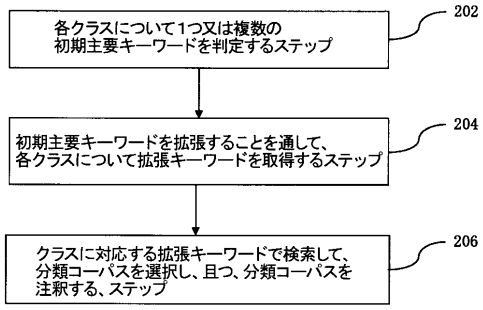

図2

【図３】

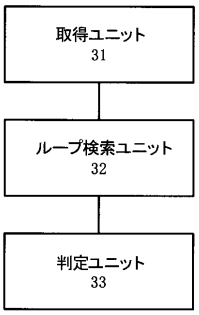

図3

【図４】

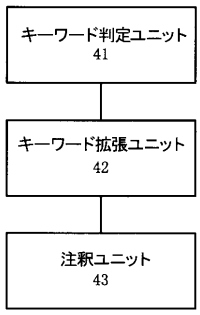

図4

【手続補正書】
【提出日】平成28年5月25日(2016.5.25)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
キーワード拡張方法であって、
所定の初期キーワードで検索して、現在のキーワードを取得するステップと、
検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行する、ステップと、を有し、
前記現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間のキーワードエラーが所定の閾値未満である場合は、前記ループ検索ステップを終了し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用することを特徴とする方法。
【請求項２】
請求項１に記載のキーワード拡張方法であって、現在のキーワードを取得する前記検索処理は、検索を通して取得された各単語の出現数をカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、ステップを有することを特徴とする方法。
【請求項３】
請求項１に記載のキーワード拡張方法であって、現在のキーワードを取得する前記検索処理は、検索を通して取得された単語の数及びこれらの出現数をカウントし、前記単語を

これらの出現数の降順にソートし、且つ、上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、ステップを有することを特徴とする方法。
【請求項４】
請求項２又は３に記載のキーワード拡張方法であって、検索を通して取得された単語を取得する前記方法は、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有する前記記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された前記単語として使用する、ステップを有することを特徴とする方法。
【請求項５】
請求項４に記載のキーワード拡張方法であって、前記キーワード拡張方法は、単語分割の後にストップワードを削除するステップと、前記所定のキーワードと同時に現れる同時出現単語を取得するステップと、及び、これらの同時出現単語を検索を通して取得された単語として使用するステップと、を更に有することを特徴とする方法。
【請求項６】
請求項１乃至５のいずれか１項に記載のキーワード拡張方法であって、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間の前記キーワードエラーは、前記現在の検索で取得された前記キーワードの数と比較して、前記現在の検索と前記前の検索との間で異なるキーワードの数の割合であること、
及び／又は、
前記所定のエラー閾値は、２０％未満であること、
及び／又は、
前記現在の検索で取得されたキーワードが前記前の検索を通して取得されたキーワードと同じである場合は、前記現在の検索で取得された前記キーワードが拡張キーワードとして判定されること、
を特徴とする方法。
【請求項７】
請求項６に記載のキーワード拡張方法であって、最初のｎ個のキーワードは、エラー評価用に前記現在の検索で取得されたキーワード及び前記前の検索を通して取得されたキーワードからそれぞれ取り出され、５≦ｎ≦１０であることを特徴とする方法。
【請求項８】
分類コーパスを注釈する方法であって、
各クラスについて１つ又は複数の初期主要キーワードを判定するステップと、
前記初期主要キーワードで、請求項１乃至７のいずれか１項に記載のキーワード拡張方法を使用して各クラスについて拡張キーワードを取得するステップと、
クラスに対応する前記拡張キーワードで検索して、分類コーパスを選択し、且つ、前記分類コーパスを注釈する、ステップと、
を有することを特徴とする方法。
【請求項９】
キーワード拡張システムであって、
所定の初期キーワードで検索して、現在のキーワードを取得する取得ユニットと、
検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行するループ検索ユニットと、
前記現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間でキーワードエラーが所定の閾値未満であるか否かを判定する、判定ユニットであって、所定の閾値未満である場合は、前記ループ検索ユニットにループ検索処理を終了する指示を出し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用する、判定ユニットと、
を有することを特徴とするシステム。
【請求項１０】
請求項９に記載のキーワード拡張システムであって、前記取得ユニットは、

記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、
検索を通して取得された各単語の出現数をそれぞれカウントし、且つ、所定の閾値よりも大きな出現数を有する単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュールと、を有すること、
又は、
前記取得ユニットは、
記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、且つ、前記単語分割の結果を検索を通して取得された単語として使用する、検索単語取得モジュールと、
検索を通して取得された前記単語の数及びこれらの出現数をカウントし、前記単語をこれらの出現数の降順にソートし、且つ、前記上位の割合を占める単語を検索を通して取得された現在のキーワードとして獲得する、検索キーワード取得モジュール用の検索キーワード比較モジュールと、を有すること、
を特徴とするシステム。
【請求項１１】
請求項１０に記載のキーワード拡張システムであって、前記検索単語取得モジュールは、記事レポジトリ内を所定のキーワードで検索して、高い関連性を有する記事を取得し、高い関連性を有するこれらの記事について単語分割を実行し、単語分割の後にストップワードを削除し、前記所定のキーワードと同時に現れる同時出現単語を取得し、及び、これらの同時出現単語を検索を通して取得された単語として使用することを特徴とするシステム。
【請求項１２】
請求項９乃至１１のいずれか１項に記載のキーワード拡張システムであって、現在の検索を通して取得されたキーワードと前の検索で取得されたキーワードとの間の前記キーワードエラーは、前記現在の検索で取得された前記キーワードの数と比較して、前記現在の検索と前記前の検索との間で異なるキーワードの数の割合であること、
及び／又は、
前記所定のエラー閾値は、２０％未満であること、
及び／又は、
前記現在の検索で取得されたキーワードが前記前の検索を通して取得されたキーワードと同じである場合は、前記現在の検索で取得された前記キーワードが拡張キーワードとして判定されること、
を特徴とするシステム。
【請求項１３】
請求項１２に記載のキーワード拡張システムであって、最初のｎ個のキーワードは、エラー評価用に前記現在の検索で取得されたキーワード及び前記前の検索を通して取得された前記キーワードからそれぞれ取り出され、５≦ｎ≦１０であることを特徴とするシステム。
【請求項１４】
分類コーパス注釈システムであって、
各クラスについて１つ又は複数の初期主要キーワードを判定するキーワード判定ユニットと、
前記初期主要キーワードで、請求項９乃至１３のいずれか１項に記載の前記キーワード拡張システムを使用して各クラスの拡張キーワードを取得する、キーワード拡張ユニットと、
クラスに対応する前記拡張キーワードで検索して、分類コーパスを選択し、且つ、前記分類コーパスを注釈する、注釈ユニットと、
を有することを特徴とするシステム。

【請求項１５】
コンピュータによって実行されたときにキーワード拡張方法を実行するコンピュータ実行可能命令が保存された１つ又は複数のコンピュータ可読媒体であって、前記方法は、
所定の初期キーワードで検索して、現在のキーワードを取得するステップと、
検索を通して取得された前記現在のキーワードを次の検索の基礎として使用し、キーワード反復を通してループ検索を実行するステップであって、
前記現在の検索で取得されたキーワードと前の検索で取得されたキーワードとの間でキーワードエラーが所定の閾値未満である場合は、前記ループ検索処理を終了し、且つ、前記現在の検索で取得された前記キーワードを拡張キーワードとして使用する、ステップと、
を有することを特徴とする方法。

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/088586

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

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)

IPC: G06F 17/-

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

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

CNPAT; CNKI ;EPODOC;WPI; keyword, key word, search, expand, cycle, iterative, error, threshold, judg+, classified, classification corpus, label, selection

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 1341899 A (IBM CORP.) 27 March 2002 (27.03.2002) see claim 1, figure 1	1-20
A	CN 102682119 A (CUI, Zhiming et al.) 19 September 2012 (19.09.2012) see the whole document	1-20

☐ 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

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 May 2014	11 June 2014
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LI, Xinyu Telephone No. (86-10) 62089924

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/088586

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 1341899 A	27 March 2002	CN 1145899 C	14 April 2004
CN 102682119 A	19 September 2012	CN 102682119 B	05 March 2014

国际检索报告

国际申请号
PCT/CN2013/088586

A. 主题的分类

G06F 17/30 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F 17/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT; CNKI ;EPODOC;WPI; 关键词,关键字,检索,搜索,扩展,循环,迭代,误差,阈值,判断,分类语料,标注,选择,keyword, key word, search, expand, cycle, iterative, error, threshold, judg+, classified, label, selection

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 1341899 A (国际商业机器公司) 2002年 3月 27日 (2002-03-27) 参见权利要求1、图1	1-20
A	CN 102682119 A (崔志明 等) 2012年 9月 19日 (2012-09-19) 参见全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2014年 5月 21日	2014年 6月 11日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451	受权官员 李昕宇 电话号码 (86-10)62089924

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/088586

检索报告引用的专利文件	公布日（年/月/日）	同族专利	公布日（年/月/日）
CN 1341899 A	2002年 3月 27日	CN 1145899 C	2004年 4月 14日
CN 102682119 A	2012年 9月 19日	CN 102682119 B	2014年 3月 05日

表 PCT/ISA/210（同族专利附件）（2009年7月）

フロントページの続き

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

(71)出願人 510117919
ペキン ユニバーシティ
中華人民共和国, 北京 １００８７１, ハイディアン ディストリクト, イヘユアン ロード ナンバー５
(74)代理人 100099759
弁理士 青木 篤
(74)代理人 100092624
弁理士 鶴田 準一
(74)代理人 100114018
弁理士 南山 知広
(74)代理人 100160716
弁理士 遠藤 力
(74)代理人 100180806
弁理士 三浦 剛
(72)発明者 イエ マオ
中華人民共和国, ベイジン １０００８９ ハイジャン ディストリクト, ジャーンホワ ロード, ワーンヘユエン シュウグワーン ガーデン, ビルディング ５
(72)発明者 ターン ジー
中華人民共和国, ベイジン １０００８９ ハイジャン ディストリクト, ジャーンホワ ロード, ワーンヘユエン シュウグワーン ガーデン, ビルディング ５
(72)発明者 シュイ ジエンボー
中華人民共和国, ベイジン １０００８９ ハイジャン ディストリクト, ジャーンホワ ロード, ワーンヘユエン シュウグワーン ガーデン, ビルディング ５
(72)発明者 レイ チャオ
中華人民共和国, ベイジン １０００８９ ハイジャン ディストリクト, ジャーンホワ ロード, ワーンヘユエン シュウグワーン ガーデン, ビルディング ５
(72)発明者 ジン リーフオン
中華人民共和国, ベイジン １０００８９ ハイジャン ディストリクト, ジャーンホワ ロード, ワーンヘユエン シュウグワーン ガーデン, ビルディング ５