(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5795743号
(P5795743)
(45)発行日　平成27年10月14日(2015.10.14)　(24)登録日　平成27年8月21日(2015.8.21)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 350C
G06F 17/30 340Z
G06F 17/30 170A

請求項の数 5　(全 8 頁)

(21)出願番号　特願2012-45250(P2012-45250)
(22)出願日　平成24年3月1日(2012.3.1)
(65)公開番号　特開2012-208924(P2012-208924A)
(43)公開日　平成24年10月25日(2012.10.25)
審査請求日　平成27年2月26日(2015.2.26)
(31)優先権主張番号　13/073,836
(32)優先日　平成23年3月28日(2011.3.28)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　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　９４３０４、パロ・アルト、コヨーテ・ヒル・ロード　３３３３
(74)代理人　100079049
弁理士　中島　淳
(74)代理人　100084995
弁理士　加藤　和詳

最終頁に続く

(54)【発明の名称】適応的重み付けを用いた様々な文書間類似度計算方法に基づいた文書比較方法および文書比較システム

(57)【特許請求の範囲】
【請求項１】
コンピュータ実行可能な文書比較方法であって、
コンピュータが、２つの文書に関する少なくとも２つの文書間類似値s_i（ａ，ｂ）を受信するステップであって、前記文書間類似値は、様々な文書間類似度計算方法ｉによって計算される、前記ステップと、
２つの文書ａ及びｂのそれぞれの個別の文書間類似度計算方法ｉの重み（α_i，β_i）を決定するステップと、
ここで、重みα_iは文書ａに対応し、重みβ_iは文書ｂに対応し、１つの文書間類似度計算方法の重みを決定するステップは、
前記重みを初期化するステップと、
前記重みを、ユーザからのフィードバックに基づいて更新するステップと、を含み、各文書の前記文書間類似度計算方法の前記重みを更新するステップは、ユーザからのフィードバックに基づいて、重みに学習アルゴリズムを適用するステップを含み、
前記２つの文書に関する前記個別の文書間類似度計算方法ｉの組み合わさった重みを計算するための重み組み合わせ関数ｆ（α_i，β_i）を決定するステップと、
ここで、前記重み組み合わせ関数は、前記文書ａ及びｂの間の比較の方向を定め、
コンピュータが、ｋ個の文書間類似度計算方法によって計算される前記文書間類似値と前記重み組み合わせ関数とに基づいて、

【数１】

$$S(a,b)=\sum_{i=1}^{k} f(\alpha_i,\beta_i)\cdot s_i(a,b)$$

として、組み合わさった類似値Ｓ（ａ，ｂ）を生成するステップとを含む、コンピュータ実行可能な文書比較方法。

【請求項２】

前記文書間類似度計算方法は、テキストに基づいた文書間類似度計算方法、視覚に基づいた文書間類似度計算方法、および、ソーシャルネットワークに基づいた文書間類似度計算方法のうちの１つ以上を含む、請求項１に記載の方法。

【請求項３】

各文書の前記文書間類似度計算方法の前記重みは、文書タイプ、文書構造、および、それらの文書の重みのうちの少なくとも１つに基づいて初期化される、請求項１に記載の方法。

【請求項４】

前記重み組み合わせ関数は、前記重みの、平均値、最小値、または、最大値を計算する実数値関数である、請求項１に記載の方法。

【請求項５】

文書間類似度レベルを推定するシステムであって、

文書間類似度レベルを推定するための、少なくとも１つのメモリに結合された少なくとも１つのハードウエア・プロセッサと、

２つの文書に関する少なくとも２つの文書間類似値ｓ$_i$（ａ，ｂ）を受信するように構成された受信機構であって、前記文書間類似値は、様々な文書間類似度計算方法ｉによって計算される、前記受信機構と、

２つの文書ａ及びｂのそれぞれの個別の文書間類似度計算方法ｉの重み（α$_i$，β$_i$）と、

前記２つの文書に関する前記個別の文書間類似度計算方法ｉの組み合わさった重みを計算するための重み組み合わせ関数ｆ（α$_i$，β$_i$）と、

を決定するように構成された決定機構であって、

ここで、重みα$_i$は文書ａに対応し、重みβ$_i$は文書ｂに対応し、１つの文書間類似度計算方法の重みの決定は、

前記重みを初期化することと、

前記重みを、ユーザからのフィードバックに基づいて更新することと、を含み、各文書の前記文書間類似度計算方法の前記重みを更新することは、ユーザからのフィードバックに基づいて、重みに学習アルゴリズムを適用することを含み、

前記重み組み合わせ関数は、前記文書ａ及びｂの間の比較の方向を定める、決定機構と、

ｋ個の文書間類似度計算方法によって計算される前記文書間類似値と前記重み組み合わせ関数とに基づいて、

【数２】

$$S(a,b)=\sum_{i=1}^{k} f(\alpha_i,\beta_i)\cdot s_i(a,b)$$

として、組み合わさった類似値Ｓ（ａ，ｂ）を生成するように構成された、値生成機構とを含む、文書間類似度レベルを推定するシステム。

【発明の詳細な説明】

【背景技術】

【０００１】

本開示は、一般的に文書類似度の分析に関する。より詳細には、本開示は、様々な文書間類似度計算方法に基づいた文書比較に関する。

【発明の概要】

【０００２】

本発明の一実施形態は、適応的重み付けを用いた様々な文書間類似度計算方法に基づいた文書比較システムを提示する。動作中に、システムは、２つの文書に関する少なくとも２つの文書間類似値を受信する。文書間類似値は、様々な文書間類似度計算方法によって計算される。次に、システムは、２つの文書それぞれの個別の文書間類似度計算方法の重みと、２つの文書に関する個別の文書間類似度計算方法の組み合わさった重みを計算するための重み組み合わせ関数とを決定する。次に、システムは、文書間類似値と重み組み合わせ関数とに基づいて組み合わさった類似値を生成する。

【０００３】

本実施形態における一変形形態では、文書間類似度計算方法は、１つまたは複数の、テキストに基づいた、視覚に基づいた、使用に基づいた、および、ソーシャルネットワークに基づいた文書間類似度計算方法を含む。

【０００４】

本実施形態における一変形形態では、文書間類似度計算方法の重みを決定している間、システムは、重みを初期化し、重みをユーザからのフィードバックに基づいて更新する。

【０００５】

さらなる一変形形態では、各文書の文書間類似度計算方法の重みは、文書タイプ、文書場所、文書構造、文書使用、および、それらの文書の重みのうちの少なくとも１つに基づいて初期化される。

【０００６】

さらなる一変形形態では、各文書の文書間類似度計算方法の重みは、ユーザからのフィードバックに基づいて重みに学習アルゴリズムを適用することによって更新される。

【０００７】

本実施形態における一変形形態では、重み組み合わせ関数は、重みの、平均値、最小値、または、最大値を計算する実数値関数である。

【図面の簡単な説明】

【０００８】

【図１】本発明の一実施形態による類似度組み合わせシステムを示す図である。

【図２】本発明の一実施形態による文書比較プロセスを示すフローチャートである。

【図３】本発明の一実施形態による一文書の文書間類似度計算方法の重みを決定するプロセスを示すフローチャートである。

【図４】本発明の一実施形態による文書比較用の例示的なコンピュータシステムである。
【発明を実施するための形態】
【０００９】
本発明の実施形態は、様々な文書間類似度計算方法に基づいた文書比較問題を解決する。動作中に、システムは、２つの文書に関する少なくとも２つの文書間類似値であって様々な文書間類似度計算方法によって計算される文書間類似値を受信する。次に、システムは、２つの文書それぞれの個別の文書間類似度計算方法の重みと、２つの文書に関する個別の文書間類似度計算方法の組み合わさった重みを計算するための重み組み合わせ関数とを決定する。システムは、続いて、文書間類似値と重み組み合わせ関数とに基づいて組み合わさった類似値を生成する。
【００１０】
メッセージ間または会話間の類似度から導き出すために、計算方法の中には、メッセージ中の、意味のある語または「エンティティ」の出現を比較するものもある。他の方法は、文書使用、または、文書が移動するときに行われる演算の順序を検出することによって、文書類似度を推定する。画像ファイルおよびプレゼンテーションスライドが比較されるとき、多くの場合、視覚類似度が発見される。幅の広い人気を得たソーシャルネットワークに関しては、ソーシャルネットワークに基づいた方法は、文書間のソーシャル接続に基づいて文書類似度を決定する。
【００１１】
本発明の実施形態は、様々な文書タイプ、ユーザの行動様式、および、ユーザの好みを説明しながら、これらの独立した文書間類似度計算方法を効果的に組み合わせるための方法を提示する。例えば、文書ａと文書ｂとの類似度を計算するとき、ｋ類似度結果s_i（ａ，ｂ）（ここで、ｉ＝１・・・ｋ）を取得する多数のｋ計算方法が展開される。ｋ結果を組み合わせるために、重みα_iが文書ａの文書間類似度計算方法ｉに割り当てられ、もう１つの重みβ_iが文書ｂの文書間類似度計算方法ｉに割り当てられる。次に、文書間類似度計算方法ｉの組み合わさった重みを計算する重み組み合わせ関数ｆ（α_i，β_i）が決定される。組み合わさった類似値Ｓ（ａ，ｂ）を、
【数１】

$$S(a,b)=\sum_{i=1}^{k} f(\alpha_i,\beta_i)\cdot s_i(a,b) \qquad (1)$$

のように計算できる。
【００１２】
図１は、本発明の一実施形態にかかる類似度組み合わせシステムを示す図である。類似度組み合わせシステム１００は、組み合わさった類似度計算器１０２と、多数の入力１０４‐１１０と、出力１１２とを含む。動作中に、組み合わさった類似度計算器１０２は、類似度結果s_i（ａ，ｂ）の入力１０４と、重みα_iの入力１０６と、重みβ_iの入力１０８と、重み組み合わせ関数ｆ（α_i，β_i）の入力１１０とを受信する。組み合わさった類似度計算器１０２は、次に、入力に基づいて組み合わさった類似値Ｓ（ａ，ｂ）の出力１１２を計算する。
【００１３】
文書間類似度計算方法ｉの組み合わさった重みを計算するための重み組み合わせ関数ｆ（α_i，β_i）は、重みのパラメータの、平均値、最小値、または、最大値を計算する実数値関数といった、様々な形態をとっていてもよい。例えば、ｆ（α_i，β_i）は、α_iとβ_iとの線形結合、
【数２】

$$f(\alpha_i,\beta_i)=x\cdot\alpha_i+y\cdot\beta_i \qquad (2)$$

ここでｘ＋ｙ＝１として規定されてもよい。ｘ＝ｙであれば、ｆ（α_i，β_i）は、α_iとβ_iとの平均値に等しい。さらに、文書ａと文書ｂとの類似度が対称的ではない、すなわちs_i（ａ，ｂ）≠s_i（ｂ，ａ）ならば、関数ｆ（α_i，β_i）は、設定ｘ≠ｙによってこのことを説明することができる。具体的には、ｘ＞ｙの場合、組み合わさった重みｆ（α_i，β_i）は、文書ａから文書ｂへの比較方向を強調する。

【００１４】

図２は、本発明の一実施形態による文書比較プロセスを示すフローチャートである。動作中に、システムは、様々な文書間類似度計算方法によって計算される多数の類似値を受信する（演算２０２）。次に、システムは、各文書間類似度計算方法の重みを決定する（演算２０４）。次に、システムは、重み組み合わせ関数を決定する（演算２０６）。続いて、システムは、組み合わさった類似度を生成する（演算２０８）。

【００１５】

文書間類似度計算方法の重みを決定するとき、文書比較システムは、重み値を初期化して、重み値をユーザからのフィードバックに基づいて更新する必要がある。図３は、本発明の一実施形態にかかる文書の文書間類似度計算方法の重みを決定するプロセスを示すフローチャートである。動作中に、システムは、文書タイプ、文書場所、および、文書構造に基づいて、重み値を初期化し（演算３０２）、重みをユーザからのフィードバックに基づいて更新する（演算３０４）。

【００１６】

各文書は、様々な計算方法に関する文書間類似度方法の重みのベクトルを有している。このベクトルの初期値を、文書タイプ、文書場所、および、文書構造から導き出すことができる。重みベクトルをまた、それらの文書の重み、それらの文書の場所、または、文書使用の周波数に基づいて初期化できる。

【００１７】

しかし、初めに割り当てられた重み値は、正確ではない場合があり、その場合は使用中に調整される必要がある。極めて重要なのは、ユーザが、類似度組み合わせ結果を順位付けし、これらの結果の不一致を指摘することによって、フィードバックを提示することである。ユーザからのフィードバックに基づいて、重みを更新でき、各類似度計算方法に関して絶えず精密化できる。このために、類似値を入力と捉え、重みを出力と捉え、ユーザからのフィードバックをグラウンドトルースと捉える、機械学習アルゴリズムを適用できる。一実施形態では、システムは、Ａｄａｂｏｏｓｔといったブースティングアルゴリズムを適用して、重みを確認および更新する。ＡｄａＢｏｏｓｔは、不一致の結果として選び抜かれたこれらの実現値を支持して、重みを適応できるように微調整する。従って、より精密化された重みは、不一致を正すことができる。

【００１８】

図４は、本発明の一実施形態による文書比較のための例示的な一コンピュータシステムを示す。一実施形態では、コンピュータ通信システム４００は、プロセッサ４０２と、メモリ４０４と、記憶装置４０６とを含む。記憶装置４０６は、文書比較アプリケーション４０８、および、アプリケーション４１０、４１２といった他のアプリケーションを記憶する。動作中に、文書比較アプリケーション４０８を、記憶装置４０６からメモリ４０４に読み込み、次に、プロセッサ４０２によって実行する。プログラムを実行している間、プロセッサ４０２は、前述の関数を行う。コンピュータ通信システム４００は、任意の表示装置４１４と、キーボード４１６と、ポインティングデバイス４１８とに連結している。

【００１９】

本詳細な説明において説明したデータ構造およびコードは、通常、コンピュータ読取り可能記憶媒体に格納されている。記憶媒体は、コンピュータシステムによって用いられるコードおよび／またはデータを格納できる任意の装置または媒体であってもよい。コンピュータ読取り可能記憶媒体は、揮発性メモリ、不揮発性メモリ、ディスクドライブなどの磁気記憶装置および光学記憶装置、磁気テープ、ＣＤ（コンパクトディスク）、ＤＶＤ（

デジタル多用途ディスクまたはデジタルビデオディスク）、または、現在知られているか、または、今後開発されるコンピュータ読取り可能媒体を格納できる他の媒体を含むが、これらに限定されるものではない。

【００２０】

本詳細な説明部分において説明した方法およびプロセスを、上述したようなコンピュータ読取り可能記憶媒体に格納されうるコードおよび／またはデータとして実施できる。コンピュータシステムが、コンピュータ読取り可能記憶媒体に格納されたコードおよび／またはデータを読み取り、実行するとき、コンピュータシステムは、データ構造およびコードとして実施され、コンピュータ読取り可能記憶媒体内に格納された方法およびプロセスを行う。

【００２１】

さらに、本発明において説明した方法およびプロセスを、ハードウェアモジュールまたはハードウェア機器に含むことができる。これらのモジュールまたは機器は、特定用途向けＩＣ（ＡＳＩＣ）チップ、書替え可能ゲートアレイ（ＦＰＧＡ）、特定のソフトウェアモジュールまたは特定の時間においてコードの一部を実行する専用プロセッサまたは共用プロセッサ、および／または、現在知られているまたは今後開発される他のプログラマブルロジックデバイスを含んでいてもよいが、それらに制限されるものではない。ハードウェアモジュールまたはハードウェア機器を起動するとき、それらは、それらに含まれる方法およびプロセスを行う。

【００２２】

様々な複数の実施形態に関する上述の説明は、図解および説明するためにのみ示されたものである。これらの説明は、完全であること、あるいは、開示した形態に本発明を限定することを意図したものではない。従って、多くの変更形態および変形形態が、当業者には明らかであろう。さらに、上述の開示は、本発明を限定することを意図したものではない。

【図１】

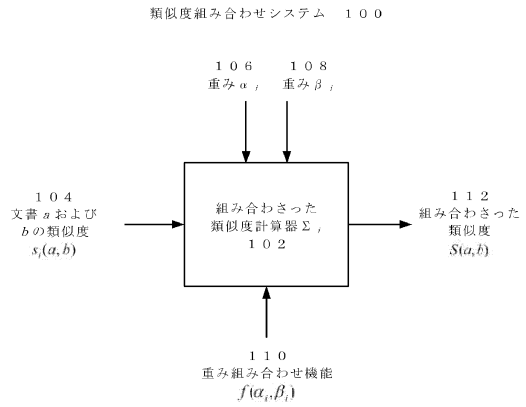

図１

【図２】

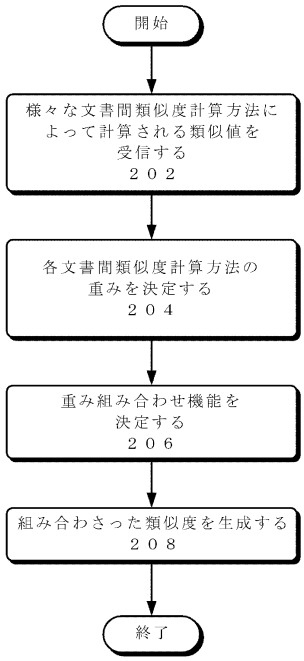

図２

【図３】

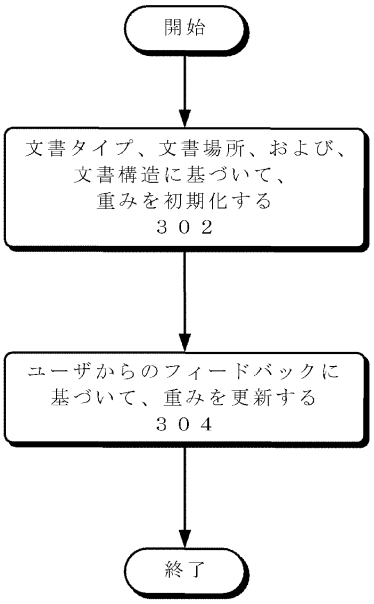
開始
文書タイプ、文書場所、および、
文書構造に基づいて、
重みを初期化する
３０２
ユーザからのフィードバックに
基づいて、重みを更新する
３０４
終了

図３

【図４】

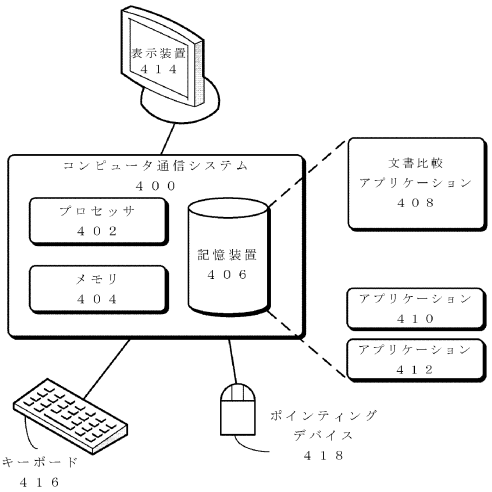
表示装置
４１４
コンピュータ通信システム
４００
プロセッサ
４０２
メモリ
４０４
記憶装置
４０６
文書比較
アプリケーション
４０８
アプリケーション
４１０
アプリケーション
４１２
キーボード
４１６
ポインティング
デバイス
４１８

図４

フロントページの続き

(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　９４３０４　マウンテン・ヴュー　オルテガ・アヴェニュー　５６５　ナンバー１１
(72)発明者　ペトロ・イザレフ
アメリカ合衆国　カリフォルニア州　９４３０９　パロ・アルト　シェリダン・アヴェニュー　４１０　アパートメント　３４０

審査官　田中　秀樹

(56)参考文献　特開２０００－２００２８５（ＪＰ，Ａ）
特開平０９－２１２５０９（ＪＰ，Ａ）
特開２００１－１５５０２７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０