(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2010-501096
(P2010-501096A)
(43)公表日　平成22年1月14日(2010.1.14)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	414A	5B075
		G06F	17/30	210D	
		G06F	17/30	419A	

審査請求 未請求　予備審査請求 未請求　（全 26 頁）

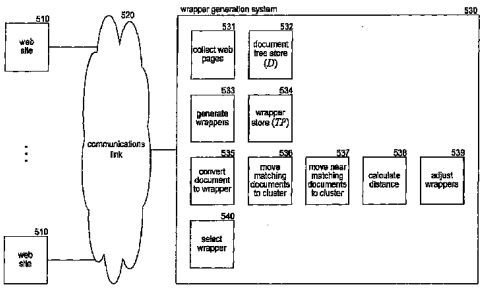
(21)出願番号　特願2009-524708 (P2009-524708)
(86)(22)出願日　平成19年8月16日 (2007.8.16)
(85)翻訳文提出日　平成21年2月16日 (2009.2.16)
(86)国際出願番号　PCT/US2007/018417
(87)国際公開番号　WO2008/021561
(87)国際公開日　平成20年2月21日 (2008.2.21)
(31)優先権主張番号　11/465,026
(32)優先日　平成18年8月16日 (2006.8.16)
(33)優先権主張国　米国 (US)

(71)出願人　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　チーロン　ウェン
アメリカ合衆国　９８０５２　ワシントン州　レッドモンド　ワン　マイクロソフト　ウェイ　マイクロソフト　コーポレーション　インターナショナル　パテンツ内

最終頁に続く

(54)【発明の名称】ラッパー生成およびテンプレート検出の協同最適化

(57)【要約】
テンプレート検出およびラッパー生成を協同して最適化することによって階層的に編成された文書のラッパーを生成する方法およびシステムを提供する。ラッパー生成システムは、文書ツリーのクラスタを識別することと、クラスタのラッパーツリーを生成することとによって、類似するテンプレートを有する文書のラッパーを生成する。ラッパーツリーは、クラスタのテンプレートと一致する文書のラッパーを定義する。ラッパー生成システムは、初期文書ツリーに基づいてクラスタのラッパーツリーを生成することによって、文書ツリーをクラスタ化する。ラッパー生成システムは、次に、任意の他の文書ツリーがクラスタのラッパーツリーと一致しまたはほぼ一致するかどうかを繰り返して判定し、そうである場合には、その文書ツリーをクラスタに追加し、適当にラッパーツリーを調整し、その結果、新たに追加された文書ツリーを含むすべての文書ツリーがラッパーツリーと一致するようにする。



【特許請求の範囲】
【請求項１】
階層的に編成された文書のラッパーを生成するコンピューティング装置での方法であって、各文書は、文書ツリーを有し、前記方法は、
文書ツリーのラッパーツリーを作成すること（５３５）と、
前記ラッパーツリーへのその距離がしきい値以内である文書ツリーを選択すること（５３７）と、
前記文書ツリーに基づいて前記ラッパーツリーを調整すること（５３９）と
を含み、前記ラッパーは、前記調整されたラッパーツリーに基づくことを特徴とする方法。
【請求項２】
前記ラッパーツリーの前記調整は、前記選択された文書ツリーのラッパーツリーおよび前記作成されたラッパーツリーをマージすることを含むことを特徴とする請求項１に記載の方法。
【請求項３】
文書ツリーのラッパーツリーの前記作成は、前記文書ツリー内のノードの連続するサブフォレストを組み合わせることを含むことを特徴とする請求項２に記載の方法。
【請求項４】
複数の文書ツリーについて前記選択および調整を実行することを含むことを特徴とする請求項１に記載の方法。
【請求項５】
前記ラッパーツリーは、前記ラッパーツリーがそれに関して作成される前記文書ツリーおよび前記ラッパーツリーがそれに関して調整される前記文書ツリーを含む文書ツリーのクラスタに関することを特徴とする請求項１に記載の方法。
【請求項６】
前記ラッパーツリーが生成された後に、文書ツリーのもう１つのクラスタについてもう１つのラッパーツリーを生成することを含むことを特徴とする請求項５に記載の方法。
【請求項７】
文書ツリーとラッパーツリーとの間の距離は、文書ノードと一致しないラッパーノードの個数およびラッパーノードと一致しない文書ノードの個数に基づくことを特徴とする請求項１に記載の方法。
【請求項８】
前記距離は、前記文書ツリーおよび前記ラッパーツリーの重さに基づいて正規化されることを特徴とする請求項７に記載の方法。
【請求項９】
複数のラッパーツリーが生成された時に、文書ツリーと前記ラッパーツリーとの間の距離に基づいて前記文書ツリーからデータを抽出するのに使用すべきラッパーを識別することを特徴とする請求項１に記載の方法。
【請求項１０】
前記しきい値は、適応式であることを特徴とする請求項１に記載の方法。
【請求項１１】
前記しきい値は、増やされたしきい値のゆえに選択される文書に基づく前記ラッパーツリーの前記調整が前記ラッパーツリーの有効性を下げるまで増やされることを特徴とする請求項１０に記載の方法。
【請求項１２】
階層的に編成された文書とラッパーツリーとの間の類似性を判定するコンピューティングシステムであって、前記文書は、文書ツリーを有し、前記システムは、
前記文書ツリーのノードを前記ラッパーツリーのノードに位置合せするコンポーネント（１１０１）と、
位置合せされないノードの個数からメトリックを生成するコンポーネント（１１０２～

１１０６）であって、前記メトリックは、前記文書ツリーと前記ラッパーツリーとの間の類似性を示す、コンポーネント（１１０２～１１０６）と
を含むことを特徴とするコンピューティングシステム。
【請求項１３】
前記メトリックは、前記ラッパーツリーのノードと位置合せされない前記文書ツリーのノードの個数および前記文書ツリーのノードと位置合せされない前記ラッパーツリーのノードの個数に基づくことを特徴とする請求項１２に記載のコンピューティングシステム。
【請求項１４】
位置合せされないノードの個数は、前記ツリーの重さに基づいて正規化されることを特徴とする請求項１３に記載のコンピューティングシステム。
【請求項１５】
前記メトリックは、次式
【数１】

$$\Psi(T_w,T_d)=\left(\frac{C_w(T_w,T_d)}{W(T_w)}+\frac{C_d(T_w,T_d)}{W(T_w)}\right)\Big/2$$

によって表されることを特徴とする請求項１２に記載のコンピューティングシステム。
【請求項１６】
ラッパーツリーごとに、
以前に選択されたことがない文書ツリーを選択すること（６０３）と、
前記選択された文書ツリーの前記ラッパーツリーを作成すること（６０４）と、
前記ラッパーからのその距離がしきい値未満である選択されていない文書ツリーが存在する時に、前記文書ツリーを選択し（６０６、６０７）、前記選択された文書ツリーに基づいて前記ラッパーツリーを調整することと
を含む方法によって、文書ツリーのラッパーツリーを生成するためにコンピューティングシステムを制御する命令を含むことを特徴とするコンピュータ可読媒体。
【請求項１７】
前記距離が０である時に、前記ラッパーツリーを調整せずに前記文書ツリーを選択することを特徴とする請求項１６に記載のコンピュータ可読媒体。
【請求項１８】
ラッパーツリーの前記選択された文書ツリーは、文書ツリーのクラスタを形成することを特徴とする請求項１７に記載のコンピュータ可読媒体。
【請求項１９】
前記距離は、前記文書ツリーと前記ラッパーツリーとの間の位置合せされないノードの個数に基づくことを特徴とする請求項１６に記載のコンピュータ可読媒体。
【請求項２０】
文書の文書ツリーからのその距離が最小である前記ラッパーツリーを識別することと、前記識別されたラッパーツリーを前記文書からデータを抽出するためのテンプレートとして使用することとによって、前記文書からデータを抽出することを含むことを特徴とする請求項１６に記載のコンピュータ可読媒体。
【発明の詳細な説明】
【背景技術】
【０００１】
ワールドワイドウェブ（「ウェブ」）は、ウェブページを介してアクセス可能な膨大な量の情報を提供する。ウェブページは、スタティックコンテンツまたはダイナミックコンテンツを含むことができる。スタティックコンテンツは、一般に、ウェブページの多数のアクセスにまたがって同一のままであることができる情報を指す。ダイナミックコンテンツは、一般に、ウェブデータベース内に格納され、検索要求に応答してウェブページに追加される情報を指す。ダイナミックコンテンツは、ディープウェブまたは隠しウェブと呼

ばれてきたものを表す。

【０００２】

多くの検索エンジンサービスは、ユーザがウェブのスタティックコンテンツを検索することを可能にする。ユーザが、検索用語を含む検索要求またはクエリをサブミットした後に、検索エンジンサービスは、これらの検索用語に関連する可能性があるウェブページを識別する。これらのウェブページが、検索結果である。関連するウェブページをすばやく識別するために、検索エンジンサービスは、ウェブページへのキーワードのマッピングを維持する場合がある。このマッピングは、各ウェブページのキーワードを識別するためにウェブを「クロール」することによって生成することができる。ウェブをクロールするために、検索エンジンサービスは、ルートウェブページを介してアクセス可能なすべてのウェブページを識別するためにルートウェブページのリストを使用する場合がある。特定のウェブページのキーワードは、見出しとしての単語の識別、ウェブページのメタデータ内で供給される単語、強調表示される単語など、さまざまな周知の情報検索技法を使用して識別することができる。

【０００３】

しかし、これらの検索エンジンサービスは、一般に、非クロール可能コンテンツとも考えられるダイナミックコンテンツの検索を提供しない。多くのウェブページは、構造化されたソース（たとえば、リレーショナルデータベース）から生成されたダイナミックページを含む。そのようなダイナミックコンテンツを含むウェブページが生成される時に、基礎になる構造化されたソースの構造化データは、構造化されない形または半構造化された形でウェブページ内においてエンコードされる。そのようなダイナミックコンテンツの検索に関連する１つの問題は、ウェブページから対応する構造化されたソースのスキーマを識別することが難しいことである。スキーマは、基礎になる構造化ソースに格納された情報または属性を定義する。この問題のゆえに、そのようなダイナミックコンテンツを有するウェブページの照会は、しばしば、不満足な結果をもたらす。

【０００４】

ウェブページのダイナミックコンテンツのスキーマを識別する試みが行われ、その結果、検索を容易にするためにより構造化されたフォーマットにコンテンツを変換できるようになってきた。構造化されたフォーマットでのウェブページからの情報およびその編成の抽出は、「ラッパー」と呼ばれるプログラムによって実行される。ウェブサイトのウェブページ用のラッパーを手作業で生成するのは、時間がかかる可能性がある。したがって、ダイナミックコンテンツを提供する数千個のウェブサイトの数百万個のウェブページのラッパーを手作業で生成するのは、非実用的である。

【０００５】

いくつかの自動ラッパー「誘導」システムまたは自動ラッパー生成システムが開発されてきた。ラッパー誘導は、ウェブページのダイナミックコンテンツのスキーマを学習することと、そのウェブページからデータを抽出し、抽出されたデータをスキーマによって識別される構造化フォーマットで格納するラッパーを生成することというプロセスである。これらの自動ラッパー誘導システムは、ラッパーの表現力に対して有効性をトレードオフするものである。有効性とは、ラッパー誘導プロセスには使用されないが、同一の「テンプレート」を共有するウェブページからコンテンツを抽出する際にラッパーがどれほど正確であるかを指す。ラッパー誘導システムは、ウェブページのトレーニングセットを使用してテンプレートに関してラッパーを生成する。次に、ラッパーは、同一のテンプレートを共有するウェブページからデータを抽出するのに使用される。表現力は、ラッパーのテンプレートによって識別されるラッパーによって処理できるウェブページの範囲を指す。ラッパーをより表現力のあるものにするために、ラッパー誘導システムは、一般に、ラッパーにワイルドカード（たとえば、「＊」）を導入し、その結果、より多くのウェブページがラッパーの範囲に含まれるようにする。しかし、一般に、ラッパーの表現力が高まるほど、その有効性が下がり、逆も同様である。

【０００６】

有効性と表現力との間の受け入れられるトレードオフを提供するために、通常のラッパー誘導システムは、トレーニングウェブページを、ウェブページ上のダイナミックコンテンツの編成を表すテンプレートに従ってクラスタ化する。したがって、類似する編成を用いる（たとえば、同一のテンプレートを有する）ウェブページは、一緒にクラスタ化される。これらのラッパー誘導システムは、あるクラスタ内のウェブページのラッパーを自動的に生成することができる。クラスタのウェブページが類似するので、そのようなラッパーは、表現力を高めながら許容できる有効性を達成するために、限られたワイルドカードを使用する可能性がある。

【先行技術文献】

【非特許文献】

【０００７】

【非特許文献１】Liu, B., Grossman, R., and Zhai, Y., “Mining Data Records in Web Pages,” Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 601-606

【発明の概要】

【発明が解決しようとする課題】

【０００８】

しかし、そのような通常のラッパー誘導システムによって生成されるラッパーの正確さは、大部分、同一のテンプレートを有するウェブページを正しくクラスタ化することの正確さに依存する。一部のラッパー誘導システムは、単純に、ウェブページのＵＲＬの間の類似性に基づいてウェブページをクラスタ化する。このクラスタ化の単純な手法は、ウェブサイトが、そのウェブサイトの同一のサブディレクトリ内の同一のテンプレートを使用するウェブページを格納する場合には適当である。その場合に、それらのＵＲＬは、サブディレクトリの位置を示すために同一のプレフィックスを有する。しかし、多くのウェブサイトは、ウェブページのＵＲＬを定義する時に、はるかにより複雑な手法を使用する。その結果、類似するＵＲＬを有するウェブページが、非常に異なるテンプレートを有する可能性があり、非常に異なるＵＲＬを有するウェブページが、非常に類似するテンプレートを有する場合がある。したがって、ウェブページの編成の類似性に基づいてウェブページを正確にクラスタ化することは、非常に難しい可能性があり、有効性と表現力との間の許容できないトレードオフを有するラッパーがもたらされる。

【課題を解決するための手段】

【０００９】

テンプレート検出とラッパー生成とを協同して最適化することによって階層的に編成された文書のラッパーを生成する方法およびシステムを提供する。ラッパー生成システムは、文書のテンプレートを検出すると同時に、文書のラッパーを動的に生成する。ラッパー生成システムは、ラベル付き葉ノードを有する文書ツリーとして文書を表す。ラッパー生成システムは、文書ツリーのクラスタを識別することと、クラスタのラッパーツリーを生成することとによって、類似するテンプレートを有する文書のラッパーを生成する。ラッパーツリーは、クラスタのテンプレートと一致する文書のラッパーを定義する。ラッパー生成システムは、初期文書ツリーに基づいてクラスタのラッパーツリーを生成することによって文書ツリーをクラスタ化する。次に、ラッパー生成システムは、他の文書ツリーがそのクラスタのラッパーツリーと一致しまたはほぼ一致するかどうかを繰り返して判定し、そうである場合には、その文書ツリーをクラスタに追加し、新たに追加された文書ツリーを含むすべての文書ツリーがラッパーツリーと一致するようにするために、適当にラッパーツリーを調整する。ラッパーツリーと一致するかほぼ一致する文書ツリーがもうない時には、そのクラスタのラッパーツリーの生成が完了する。次に、ラッパー生成システムは、このプロセスを繰り返して、文書ツリーの新しいクラスタを形成し、そのラッパーツリーを生成する。次に、ラッパー生成システムは、これらのラッパーツリーを使用して、ラッパーを定義する。新しい文書のデータを抽出するために、新しい文書の文書ツリーが生成され、最もよく一致するラッパーツリーが識別され、その識別されたラッパーツリー

のラッパーが、データを抽出するのに使用される。
【００１０】
この「課題を解決するための手段」は、下の「発明を実施するための形態」でさらに説明する概念の選択物を単純化された形で導入するために提供されたものである。この「課題を解決するための手段」は、請求される主題の主要な特徴または本質的特徴を識別することを意図されたものではなく、請求される主題の範囲を判定する際の助けとして使用されることを意図されたものでもない。
【図面の簡単な説明】
【００１１】
【図１】一実施形態のラッパー生成システムの動作の高水準概要を示すブロック図である。
【図２】文書ツリーのラッパーツリーへの変換を示す図である。
【図３】一実施形態でのラッパーツリーのマージを示す図である。
【図４】ラッパーツリーと文書ツリーとの位置合わせを示す図である。
【図５】一実施形態でのラッパー生成システムのコンポーネントを示すブロック図である。
【図６】一実施形態でのラッパー生成システムのラッパー生成コンポーネントの処理を示す流れ図である。
【図７】一実施形態でのラッパー生成システムのラッパーへの文書変換コンポーネントの処理を示す流れ図である。
【図８】一実施形態でのラッパー生成システムのクラスタへの一致する文書の移動コンポーネントの処理を示す流れ図である。
【図９】一実施形態でのラッパー生成システムのクラスタへのほぼ一致する文書の移動コンポーネントの処理を示す流れ図である。
【図１０】一実施形態でのラッパー生成システムのほぼ一致する文書のチェックコンポーネントの処理を示す流れ図である。
【図１１】一実施形態でのラッパー生成システムの距離計算コンポーネントの処理を示す流れ図である。
【図１２】一実施形態でのラッパー生成システムのラッパー調整コンポーネントの処理を示す流れ図である。
【図１３】一実施形態でのラッパー生成システムのラッパー位置合わせコンポーネントの処理を示す流れ図である。
【発明を実施するための形態】
【００１２】
テンプレート検出およびラッパー生成を協同して最適化することによって階層的に編成された文書のラッパーを生成する方法およびシステムを提供する。一実施形態で、ラッパー生成システムは、文書のテンプレートを検出すると同時に文書のラッパーを動的に生成する。ラッパー生成システムは、ウェブページなど、それぞれが文書ツリーと称するその階層のツリー構造によって表される、階層的に編成された文書のコレクションを与えられる。たとえば、ウェブページが、ドキュメントオブジェクトモデル（「ＤＯＭ」）に準拠する場合に、文書ツリーは、ＤＯＭ文書のタグに対応するノードを含む。ラッパー生成システムは、文書ツリーの葉ノードのラベルをも与えられる。これらのラベルは、基礎になる構造化データの識別子、フィールド、または属性に対応する。たとえば、自動車に関する情報を含むウェブページは、「メーカー」、「モデル」、「年式」、「色」、「価格」などとしてラベル付けされた葉ノードを有する可能性がある。ラッパー生成システムは、コレクションの文書ツリーを選択し、その文書ツリーに対応するラッパーツリーを生成することによって、類似するテンプレートを有する文書のラッパーを生成する。ラッパーツリーは、検出されたテンプレートと一致する文書のラッパーを定義する。ラッパー生成システムは、コレクションの任意の他の文書ツリーがラッパーツリーとほぼ一致するかどうかを判定する。そうである場合には、ラッパー生成システムは、文書ツリーが、ラッパー

を生成するのに使用された文書ツリーと同一のテンプレートを共有すると考える。ラッパー生成システムは、ほぼ一致する文書ツリーについてラッパーツリーを調整し、その結果、両方の文書ツリーがそのラッパーツリーと一致するようにする。文書ツリーが、たまたまラッパーツリーと正確に一致する場合には、ラッパー生成システムは、ラッパーツリーを調整する必要がない。次に、ラッパー生成システムは、コレクションの任意の他の文書ツリーが調整されたラッパーツリーとほぼ一致するかどうかの判定およびラッパーツリーの調整を、調整されたラッパーツリーとほぼ一致するコレクション内の文書ツリーがなくなるまで繰り返す。ラッパーツリーの生成および調整に使用される文書ツリーを、類似するテンプレートを有し、ラッパーツリーと一致する文書ツリーの「クラスタ」と称する。次に、ラッパー生成システムは、新しいクラスタを形成するための文書ツリーの選択、その文書ツリーのラッパーツリーの生成、およびクラスタに追加されるほぼ一致するツリーに関するラッパーツリーの動的調整というプロセスを繰り返す。この形で、ラッパー生成システムは、動的に生成されるラッパーツリーに基づいてテンプレートを検出し、動的に識別されたテンプレートに基づいてラッパーツリーを生成する。次に、ラッパー生成システムは、ラッパーツリーを使用してラッパーを定義する。

【００１３】

一実施形態で、ラッパー生成システムは、距離メトリックを使用して、文書ツリーがラッパーツリーと一致し、またはほぼ一致するかどうかを判定する。ラッパーツリーは、類似するテンプレートを有する文書ツリーを定義するラッパーノードの階層であり、テンプレートを有する文書ツリーからデータを抽出するのに使用されるラッパーを表す。ラッパーツリーのラッパーノードは、ラッパーのノードと一致する文書ツリーの対応する文書ノードを定義するタグ、ラベル、またはワイルドカードを含むことができる。文書ツリーとラッパーツリーとの間の距離を判定するために、ラッパー生成システムは、文書ノードをラッパーノードに位置合せする。次に、ラッパー生成システムは、ラッパーノードと位置合せされないツリーノードの個数と、ツリーノードと位置合せされないラッパーノードの個数とをカウントする。ラッパー生成システムは、これらの位置合せされないノードの個数に基づいて距離メトリックを計算する。ラッパー生成は、距離メトリックを正規化し、その結果、同一の距離しきい値を、そのサイズに関わりなく文書ツリーとラッパーツリーとの間のほぼ一致を定義するのに使用できるようにすることもできる。たとえば、５つの位置合せされないノードを有する文書ツリーは、１００個のノードを有するラッパーツリーとほぼ一致するが、１０個のノードのみを有するラッパーツリーとほぼ一致はしないものとすることができる。

【００１４】

一実施形態で、ラッパー生成システムは、固定しきい値または適応しきい値のいずれかを使用して、文書ツリーがラッパーツリーとほぼ一致するかどうかを判定することができる。固定しきい値を使用する時には、ラッパー生成システムは、ラッパー生成プロセス全体を通じて、文書ツリーがラッパーツリーとほぼ一致するかどうかを判定するのに同一のしきい値を使用する。適応しきい値を使用する時には、ラッパー生成システムは、当初に、文書ツリーをクラスタ化する時に小さいしきい値を使用することができる。クラスタのラッパーツリーとほぼ一致する文書ツリーがもうない場合に、ラッパー生成システムは、しきい値を増やし、その後、増やされたしきい値を使用し、ラッパーツリーを適当に調整することによって、ラッパーツリーとほぼ一致する文書ツリーをクラスタに追加することができる。その後、ラッパー生成システムは、増やされたしきい値に基づいて調整されたラッパーツリーが、以前のしきい値を用いるラッパーツリーよりよく動作するかどうかをテストする。ラッパー生成システムは、新たに調整されたラッパーツリーに基づくラッパーを使用してクラスタの文書ツリーからデータを抽出することによって、性能をテストすることができる。この性能が、大幅によりよいものではない場合には、ラッパー生成システムは、増やされたしきい値のゆえに追加された文書ツリーを除外するように最終的なクラスタをセットすることと、最終的なラッパーツリーに、増やされたしきい値のゆえに調整される前のラッパーツリーをセットすることとによって、増やされたしきい値の影響を

ロールバックする。しかし、調整されたラッパーツリーの性能の方がよい場合には、ラッパー生成システムは、もう一度しきい値を増やし、ほぼ一致する文書ツリーをクラスタに追加する。ラッパー生成システムは、増やされたしきい値に基づいて調整されたラッパーツリーが以前のしきい値を用いるラッパーツリーよりよくまたは大幅によりよく動作しなくなるまで、このプロセスを継続する。

【００１５】

図１は、一実施形態のラッパー生成システムの動作の高水準概要を示すブロック図である。このラッパー生成システムは、ウェブページなどの文書のトレーニングセットを与えられる１０１。このラッパー生成システムは、次に、ウェブページを解析して１０２、文書ツリーを生成し、その葉ノードのラベルを入力する。このラッパー生成システムは、テンプレートを協同して検出し１０３、動的に調整されるラッパーツリーに基づいて文書ツリーをクラスタ化することによってラッパーを生成する１０３。次に、このラッパー生成システムは、ラッパーツリーに基づいて定義されたラッパーを出力する１０４。次に、このラッパー生成システムは、ラッパーを使用してウェブページからデータを抽出することができる。このラッパー生成システムは、新しいウェブページを受け取る１０５時に、新しいウェブページを解析して１０６、文書ツリーを生成する。次に、このラッパー生成システムは、距離メトリックに基づいて、生成された文書ツリーに最も近いラッパーツリーを選択し１０７、選択されたラッパーツリーに対応するラッパーを使用してツリーからデータを抽出する１０８。一実施形態で、文書は、ＸＭＬフォーマットで表すことができる。

【００１６】

ラッパー生成システムは、「括弧」ノードを含めることができる各ラッパーノードに割り当てられた記号を有する修正ＤＯＭツリー（ｍｏｄｉｆｉｅｄ　ＤＯＭ　ｔｒｅｅ）としてラッパーツリーを表す。ラッパー生成システムは、ラッパーノードの記号を使用して、文書ツリーとラッパーツリーとの間の距離を計算する時に、文書ノードおよびラッパーノードが位置合せされているかどうかを判定する。ラッパー生成システムは、ラッパーノードσの記号Ｓ（σ）が、１、整数Ｎ（Ｎ≧２）、またはワイルドカードすなわち？、＋、および＊のうちの１つになるように定義する。１の記号は、ラッパーノードが１つのツリーノードだけと一致することができることを意味する。Ｎの記号は、ラッパーノードがＮ個の連続するツリーノードだけと一致することができることを意味する。？の記号は、ラッパーノードが０個または１個のツリーノードだけと一致することができることを意味する。＋の記号は、ラッパーノードが連続するＮ個のツリーノードと一致することができ、Ｎ≧１であることを意味する。＊の記号は、ラッパーノードが０個または連続するＮ個のツリーノードと一致することができ、Ｎ≧１であることを意味する。？または＊の記号を有するラッパーノードは、ツリーノードと一致しない可能性があるので、「ソフト」ノードと考えられる。他のすべてのラッパーノードは、「ハード」ノードと考えられる。ラッパーツリーには、「ブロック」ノードと称する特殊なノードを含めることができる。括弧ノードは、タグを有しておらず、括弧の対のように振る舞い、したがって、葉ノードになることはできない。ラッパーツリーの他のすべてのノードは、「タグ」ノードと呼ばれる。

【００１７】

一実施形態で、ラッパー生成システムは、文書ツリーとラッパーツリーとの間の距離メトリックを、文書ツリーおよびラッパーツリーの重さに基づいて正規化する。ラッパー生成システムは、ツリーノードσのツリーノード重さＷ（σ）を、σをルートとするサブツリーのノード数と等しくなるように定義する。そのルートノードがτであるツリーT_dの文書ツリー重さＷ（T_d）は、Ｗ（τ）である。ラッパー生成システムは、ラッパーノードσのラッパーノード重さＷ（σ）を、そのラッパーノードがソフトノードである時には０になり、そのラッパーノードがハードノードである時にはその子ノードの重さの合計になり、そのラッパーノードがハード葉タグノードである。時には１になり、そのラッパーノードがハード非葉タグノードである時には１にその子ノードの重さの合計を加えたもの

になるように定義する。ラッパーツリー重さは、ラッパーツリーT_wのW（T_w）は、T_wが文書ツリーT_dによって生成される時にはW（T_d）であり、ラッパーツリーT_wがラッパーツリー

【００１８】

【数１】

$$T_{w_1}$$

【００１９】

および

【００２０】

【数２】

$$T_{w_2}$$

【００２１】

の組合せによって生成される時には

【００２２】

【数３】

$$max\left(W\left(T_{w_1}\right),W\left(T_{w_2}\right)\right)$$

【００２３】

である。

【００２４】

ラッパー生成システムは、ラッパーツリーを生成するのに使用される文書ツリーの個数として「ラッパーレベル」を定義する。レベル１ラッパーツリーは、文書ツリーをラッパーツリーに変換することによって生成されるラッパーツリーである。文書ツリーのラッパーツリーへの変換は、次の式によって表される。

$$T_d \rightarrow T_w$$

ここで、T_dは、文書ｄの文書ツリーを表し、T_wは、ラッパーｗのラッパーツリーを表す。定義により、レベル１ラッパーの重さは、それがそこから生成された文書ツリーの重さである。文書ツリーをラッパーツリーに変換する時に、ラッパー生成システムは、反復パターン組合せアルゴリズムを実行して、T_wをT_dより簡潔にする。一実施形態で、ラッパー生成システムは、参照によって本明細書に組み込まれている非特許文献１に記載のアルゴリズムに類似するアルゴリズムを使用する。このアルゴリズムが、文書ノードの同一の連続するサブツリーを検出する場合に、このアルゴリズムは、それらをラッパーツリーの１つのラッパーノードにマージし、その記号に同一のサブツリーの個数をセットする。このアルゴリズムが、文書ノードの同一の連続するサブフォレストを識別する場合に、このアルゴリズムは、ラッパーツリー内のブロックノードの下の１つのサブフォレストとしてそれらをマージし、その記号に同一のサブフォレストの個数をセットする。このアルゴリズムは、文書ツリーをラッパーツリーに変換する時に、葉ノードのラベルを検討する。図２に、文書ツリーのラッパーツリーへの変換を示す。文書ツリー２１０が、ラッパーツリー２２０に変換される。文書ツリーのルートノードＡは、ＢノードおよびＣノードの反復サブフォレストを含むので、ラッパー生成システムは、括弧ノードＸをラッパーツリーに追加し、その記号に２をセットする。各Ｂノードは、連続するノードD_1を含む（添字は、ノードのラベルを表す）ので、ラッパー生成システムは、これらのノードをラッパーツリー内で組み合わせ、その記号に２をセットする。

【００２５】

ラッパー生成システムは、２つの低水準ラッパーツリーの位置合わせに基づいて高水準ラッパーツリーを生成する。ラッパー生成システムは、トップダウンの順序でレイヤごとに２つのラッパーツリー

【００２６】

【数４】
T_{w_1}

【００２７】
および
【００２８】
【数５】
T_{w_2}

【００２９】
を位置合せする。ルートノードから同一の深さにあるノードは、同一レイヤに含まれ、ブロックノードは、１つのレイヤに含まれるとは考えられない。ラッパー生成システムは、同一レイヤ内のノードだけを位置合せする。ラッパー生成システムは、ラッパーノード
【００３０】
【数６】
σ_{w_1}

および
【００３１】
【数７】
σ_{w_2}

が次の条件を満足する時に、これらのノードが一致すると考える。
【００３２】
【数８】
σ_{w_1}

【００３３】
および
【００３４】
【数９】
σ_{w_2}

【００３５】
が、両方とも非葉ノードであるか、両方とも葉ノードであり、
【００３６】
【数１０】
$$T(\sigma_{w_1}) = T(\sigma_{w_2}),$$

【００３７】
であり、
【００３８】
【数１１】
σ_{w_1}

【００３９】
および
【００４０】

【数１２】
σ_{w_2}

【００４１】
が両方とも葉ノードである場合に
【００４２】
【数１３】
$L(\sigma_{w_1}) = L(\sigma_{w_2})$

【００４３】
である。
ここで、Ｔ（σ）は、ノードσのタグを表し、Ｌ（σ）は、ノードσのラベルを表す。各レイヤで、ラッパー生成システムは、位置合わせ関数
【００４４】
【数１６】
$A(F_{w_1}, F_{w_2})$

【００４５】
を呼び出すことによって表されるように、サブツリーのアレイ
【００４６】
【数１４】
F_{w_1}

【００４７】
と
【００４８】
【数１５】
F_{w_2}

【００４９】
との間のシーケンス位置合わせを実行する。ラッパー生成システムは、最小コスト位置合わせを入手するために動的計画法を使用する。
【００５０】
【数１７】
F_{w_1}

【００５１】
および
【００５２】
【数１８】
F_{w_2}

【００５３】
内のすべての不一致のルートノードは、その重みをコストとして
【００５４】
【数１９】
$A(F_{w_1}, F_{w_2})$

【００５５】
に寄与する。非葉ノードである一致するノードの対

【００５６】
【数２０】
σ_{w_2}
【００５７】
および
【００５８】
【数２１】
σ_{w_2}

【００５９】
について、ラッパー生成システムは、
【００６０】
【数２２】
$A\left(childF_{w_1}, childF_{w_2}\right)$

【００６１】
を再帰的に呼び出し、ここで、
【００６２】
【数２３】
$childF_{w_1}$

【００６３】
および
【００６４】
【数２４】
$childF_{w_2}$

【００６５】
は、
【００６６】
【数２５】
σ_{w_2}

【００６７】
および
【００６８】
【数２６】
σ_{w_2}
【００６９】
の子ノードをルートとするサブツリーからなるサブフォレストである。ラッパー生成システムは、関数
【００７０】
【数２７】
$A\left(childF_{w_1}, childF_{w_2}\right)$

【００７１】
を呼び出すことによって計算されたコストを
【００７２】

【数２８】

$$A\left(F_{w_1}, F_{w_2}\right)$$

【００７３】

のコストに加算する。ラッパー生成システムは、ラッパーノードをトップダウンの再帰的な形で位置合せするので、２つのラッパー内のノードの両方がルートノードであるか、その親ノードが互いに位置合せされている場合に限って、それらのノードの位置合わせを試みる。

【００７４】

図３は、一実施形態でのラッパーツリーのマージを示す図である。ラッパーツリー３１０および３２０がマージされて、ラッパーツリー３３０を形成する。影付きのノードＤおよびＧは、他方のラッパーツリー内に一致するノードを有しない。その結果、マージされたラッパーツリーの対応するラッパーノードは、ソフトノードである。ラッパー生成システムは、当初に、これらのラッパーツリーのルートノードを渡して、Ａ（Ａ，Ａ）として位置合わせ関数を呼び出す。この関数は、第２レイヤのラッパーノードを渡してＡ（Ｂ（Ｃ^{3}ＤＥ＊）$^{?}$，ＢＣ^{3}Ｅ）としてそれ自体を再帰的に呼び出す。ワイルドカード？はソフトなので、この関数は、Ａ（ＢＣ^{3}ＤＥ＊，ＢＣ^{3}Ｅ）およびＡ（Ｂ，ＢＣ^{3}Ｅ）としてそれ自体を再帰的に呼び出す。前者が、より低いコスト（すなわち、よりよい一致）をもたらすので、この関数は、前者を位置合わせとして選択する。関数Ａ（Ｂ（Ｃ^{3}ＤＥ＊）$^{?}$，ＢＣ^{3}Ｅ）の動的計画法プロセス中に、関数Ａ（Ｆ^{2}，ＦＧ$^{+}$）が、Ａ（ＢＣ^{3}ＤＥ＊，ＢＣ^{3}Ｅ）とＡ（Ｂ，ＢＣ^{3}Ｅ）との両方によって再帰的に呼び出されて、この２つの解のコストを計算する。位置合わせが、２つのラッパーツリーの間の最適解を入手した後に、ラッパー生成システムは、次の記号生成関数Ｆを使用して新しいラッパーツリーを構築する。

Ｆ（１，ＮＵＬＬ） ＝ ？	Ｆ（？，Ｎ） ＝ ＊			
Ｆ（？，ＮＵＬＬ） ＝ ？	Ｆ（？，＋） ＝ ＊			
Ｆ（ｎ，ＮＵＬＬ） ＝ ＊	Ｆ（１，＊） ＝ ＊			
Ｆ（＋，ＮＵＬＬ） ＝ ＊	Ｆ（Ｎ，＊） ＝ ＊			
Ｆ（＊，ＮＵＬＬ） ＝ ＊	Ｆ（？，＊） ＝ ＊			
Ｆ（１，１） ＝ １	Ｆ（＋，＊） ＝ ＊	（１）		
Ｆ（Ｎ，Ｎ） ＝ Ｎ	Ｆ（１，Ｎ） ＝ ＋			
Ｆ（＋，＋） ＝ ＋	Ｆ（Ｎ，＋） ＝ ＋			
Ｆ（？，？） ＝ ？	Ｆ（１，＋） ＝ ＋			
Ｆ（＊，＊） ＝ ＊	Ｆ（N_1，N_2） ＝ ＋			
Ｆ（１，？） ＝ ？				

ここで、ＮＵＬＬは、ラッパーノードの不一致を表す。たとえば、Ｆ（１，ＮＵＬＬ）は、その記号が１であるラッパーノードの不一致を示す。

【００７５】

ラッパー生成システムは、次を除いて、ラッパーツリーが位置合せされる方法に似た形で文書ツリーおよびラッパーツリーを位置合せする。ラッパーツリーは、１対１の形で位置合せされるが、その記号が＋または＊であるラッパーノード（タグノードのみ）を、複数の文書ノードと位置合わせすることができる。また、２つのラッパーノードを位置合せできるかどうかを判定する時には、ラッパー生成システムは、ノードのラベルを考慮に入れない。ラッパーツリーT_wと文書ツリーT_dとの間の位置合わせについて、ラッパー生成システムは、不一致の文書ノードによって寄与される総コストを示すためにC_d（T_w，T_d）を使用する。C_dは、ラッパーノードと一致しない文書ツリーT_dのノードのカウントを表し、C_wは、文書ノードと一致しないラッパーツリーT_wのノードを表す。

【００７６】

【数２９】
$$T_{w1} + T_{w2} \rightarrow T_{w3}$$

【００７７】
の場合には、すべての文書ツリーＴ$_d$について、
【００７８】
【数３０】
$$T_{w3}$$

【００７９】
とＴ$_d$との間の位置合わせは、少なくとも
【００８０】
【数３１】
$$T_{w1}$$

【００８１】
とＴ$_d$との間ならびに
【００８２】
【数３２】
$$T_{w2}$$

【００８３】
とＴ$_d$との間と同数の一致する対を作る。というのは、
【００８４】
【数３３】
$$T_{w3}$$

が、少なくとも
【００８５】
【数３４】
$$T_{w1}$$

【００８６】
および
【００８７】
【数３５】
$$T_{w2}$$

【００８８】
と同数のワイルドカードを有するからである。また、
【００８９】
【数３６】
$$T_{w3}$$

【００９０】
に現れるが
【００９１】
【数３７】
$$T_{w1}$$

【００９２】
と
【００９３】
【数３８】
T_{w2}

【００９４】
との両方には存在しないラッパーノードは、ソフトノードであり、したがって、コストには寄与しない。したがって、すべての文書ツリーＴ$_{d}$について、次の条件が満足される。
【００９５】
【数３９】

$$C_w\left(T_{w_3},T_d\right)\le C_w\left(T_{w_i},T_d\right)$$
$$C_d\left(T_{w_3},T_d\right)\le C_d\left(T_{w_i},T_d\right)\quad(i=1,2) \qquad (2)$$

【００９６】
ラッパー生成システムは、次の式によってラッパーツリーＴ$_{w}$と文書ツリーＴ$_{d}$との間の距離を定義する。
【００９７】
【数４０】

$$\Psi\left(T_w,T_d\right)=\left(\frac{C_w\left(T_w,T_d\right)}{W\left(T_w\right)}+\frac{C_d\left(T_w,T_d\right)}{W\left(T_w\right)}\right)\Big/2 \qquad (3)$$

【００９８】
この式は、すべてのラッパーツリーＴ$_{w}$および文書ツリーＴ$_{d}$について、０≦Ψ（Ｔ$_{w}$，Ｔ$_{d}$）≦１であるという特性を有する。また、すべての文書ツリーＴ$_{d}$について、次の条件が満足される。
【００９９】
【数４１】

$$\Psi\left(T_{w_3},T_d\right)\le\Psi\left(T_{w_i},T_d\right)\quad(i=1,2) \qquad (4)$$

【０１００】
図４は、ラッパーツリーと文書ツリーとの位置合わせを示す図である。ラッパーノードおよび文書ノードは、破線によって示されているように位置合せされる。
【０１０１】
一実施形態で、ラッパー生成システムは、文書ツリーがラッパーツリーとほぼ一致するかどうかを判定する時に、適応しきい値を使用する。ラッパー生成システムは、当初に、小さいしきい値から開始する。文書ツリーが現在のしきい値内にない時には、ラッパー生成システムは、小さい量だけしきい値を増やす。ラッパー生成システムは、増やされたしきい値がラッパーツリーによって表されるラッパーの性能を大きくは高めなくなるまで、クラスタ化を繰り返す。ラッパー生成システムは、対応するクラスタ化された文書ツリーをテストすることによって、増やされたしきい値の下で生成されたラッパーツリーＴ$_{w}$’の性能を評価する。ラッパー生成システムは、次に、前のラッパーツリーおよび現在のラッパーツリーの精度ｐ、リコールｒ、およびＦ１　ｆを計算する。ラッパー生成システムは、次の式によって、大きい改善を表すことができる。
【０１０２】

【数４２】

$$f_{T_i} > f_{T_n}$$
$$r_{T_i} - r_{T_n} > \lambda \qquad (5)$$

【０１０３】

ここで、λは、しきい値増加が生成されるラッパーのリコールの大きい改善につながることを保証するのに使用される小さい値（たとえば、０．００５）である。ラッパー生成システムは、初期しきい値と１との間になるように定義される停止値に達する時に、しきい値の増加を終了することもできる。

【０１０４】

図５は、一実施形態でのラッパー生成システムのコンポーネントを示すブロック図である。ラッパー生成システム５３０は、通信リンク５２０を介してウェブサイト５１０に接続される。ラッパー生成システムは、ウェブページ収集コンポーネント５３１および文書ツリーストア５３２を含む。ウェブページ収集コンポーネントは、さまざまなウェブサイトをクロールして、ラッパー生成用のトレーニングデータとしてウェブページを収集する。ウェブページ収集コンポーネントは、各ウェブページのドキュメントオブジェクトモデル（「ＤＯＭ」）表現を生成し、その表現を文書ツリーとして文書ツリーストア内に格納する。ラッパー生成システムは、ユーザが文書ツリーの葉ノードに手作業でラベルを付けるためのユーザインターフェースをも提供することができる。ラッパー生成システムは、これらのラベルを文書ツリーストア内に格納する。したがって、ラッパー生成用のトレーニングデータは、葉ノードのラベルと一緒に文書ツリーを含む。ラッパー生成システムは、ラッパー生成コンポーネント５３３およびラッパーストア５３４をも含む。ラッパー生成コンポーネントは、類似するテンプレートを共有するウェブページをクラスタ化し、ツリーストアのトレーニングデータのラッパーツリーを生成し、そのラッパーツリーをラッパーストアに格納する。ラッパー生成コンポーネントは、類似するテンプレートを共有するウェブページを動的にクラスタ化し、クラスタ化中にラッパーツリーを動的に調整する。したがって、クラスタのウェブページは、それらがクラスタのラッパーツリーとどれほどよく一致するかに基づいて選択され、ラッパーツリーは、クラスタの選択されたウェブページに基づいて調整される。ラッパー生成コンポーネントは、ラッパーへの文書変換コンポーネント５３５、クラスタへの一致する文書の移動コンポーネント５３６、クラスタへのほぼ一致する文書の移動コンポーネント５３７、距離計算コンポーネント５３８、およびラッパー調整コンポーネント５３９を呼び出す。ラッパーへの文書変換コンポーネントは、トレーニングデータの文書ツリーをラッパーツリーに変換する。クラスタへの一致する文書の移動コンポーネントは、クラスタの現在のラッパーツリーと一致するトレーニングデータの文書ツリーを識別し、これらの識別された文書ツリーをクラスタに移動する。クラスタへのほぼ一致する文書の移動コンポーネントは、クラスタの現在のラッパーツリーとほぼ一致するトレーニングデータの文書ツリーを識別し、これらの識別された文書ツリーをクラスタに移動し、これらの識別された文書ツリーに基づいてクラスタのラッパーツリーを調整する。距離計算コンポーネントは、文書ツリーとラッパーツリーとの間の距離を計算して、文書ツリーがどれほどよく一致するかを判定する。ラッパー調整コンポーネントは、新しい文書ツリーがクラスタに追加される時のクラスタのラッパーツリーの動的調整をもたらすために、２つのラッパーツリーをマージする。ラッパー生成コンポーネントがウェブページをクラスタ化した後に、ラッパー生成コンポーネントは、各クラスタの動的に生成されたラッパーツリーをラッパーストアに格納する。ラッパー生成システムは、ラッパー選択コンポーネント５４０をも含む。ラッパー選択コンポーネントは、ウェブページを受け取り、そのウェブページの文書ツリーを生成し、そのツリーに最も近いラッパーストアのラッパーツリーを識別し、識別されたラッパーツリーのラッパーを使用してウェブページからデータを抽出する。

【０１０５】

ラッパー生成システムを実施できるコンピューティング装置は、中央処理装置、メモリ、入力装置（たとえば、キーボードおよびポインティング装置）、出力装置（たとえば、ディスプレイ装置）、およびストレージ装置（たとえば、ディスクドライブ）を含むことができる。物理メモリおよびストレージ装置は、ラッパー生成システムを実施する命令およびデータ構造を含むことができるコンピュータ可読媒体である。さらに、このデータ構造および命令を、格納するか、通信リンク上の信号などのデータ伝送媒体を介して伝送することができる。インターネット、ローカルエリアネットワーク、広域ネットワーク、ポイントツーポイントダイヤルアップネットワーク、セル電話網など、さまざまな通信リンクを使用して、システムのコンポーネントを接続することができる。

【０１０６】

ラッパー生成システムの実施形態は、パーソナルコンピュータ、サーバコンピュータ、マルチプロセッサシステム、マイクロプロセッサベースのシステム、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータ、上記のシステムまたは装置のいずれかを含む分散コンピューティング環境などを含むさまざまなオペレーティング環境で実施し、使用することができる。ユーザコンピューティング装置は、セル電話機、携帯情報端末、スマートホン、パーソナルコンピュータ、プログラマブル消費者エレクトロニクス、ディジタルカメラなどを含むことができる。

【０１０７】

ラッパー生成システムは、１つまたは複数のコンピュータまたは他の装置によって実行される、プログラムモジュールなどのコンピュータ実行可能命令の全体的な文脈で説明することができる。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か特定の抽象データ型を実施する、ルーチン、プログラム、オブジェクト、コンポーネント、データ構造などを含む。通常、プログラムモジュールの機能性は、さまざまな実施形態で望み通りに組み合わせるか分散させることができる。ラッパー生成システムによって使用される文書には、ウェブページ、ＸＭＬベースの文書、ＨＴＭＬベースの文書など、すべての階層的に編成された文書を含めることができる。

【０１０８】

図６は、一実施形態でのラッパー生成システムのラッパー生成コンポーネントの処理を示す流れ図である。このコンポーネントは、類似するテンプレートを有するウェブページを動的にクラスタ化し、ウェブページの各クラスタのラッパーツリーを動的に生成する。このコンポーネントは、当初に、トレーニングコレクションＤ内にトレーニングデータのすべての文書ツリーを有する状態で開始する。ブロック６０１～６０９では、このコンポーネントは、ウェブページのクラスタを識別し、クラスタのラッパーツリーを動的に調整しながらループする。判断ブロック６０１で、トレーニングコレクションが空の場合には、このコンポーネントは完了し、そうでない場合には、このコンポーネントはブロック６０２で継続する。ブロック６０２では、このコンポーネントは、新しいクラスタＴＰを作成する。ブロック６０３で、このコンポーネントは、コレクションからの文書ツリーT_dをクラスタの最初の文書ツリーとして選択する。一実施形態で、このコンポーネントは、文書ツリーをランダムに選択する。ブロック６０４で、このコンポーネントは、ラッパーへの文書変換コンポーネントを呼び出して、選択された文書ツリーT_dを新しいクラスタＴＰの初期ラッパーツリーT_wに変換する。ブロック６０５で、このコンポーネントは、コレクションＤからの選択された文書ツリーT_dを新しいクラスタＴＰに移動する。ブロック６０６で、このコンポーネントは、クラスタへの一致する文書の移動コンポーネントを呼び出して、初期ラッパーツリーと一致する文書ツリーをコレクションＤから新しいクラスタＴＰに移動する。ブロック６０７で、このコンポーネントは、クラスタへのほぼ一致する文書の移動コンポーネントを呼び出して、調整されたラッパーツリーとほぼ一致する文書ツリーをコレクションＤから新しいクラスタＴＰに移動し、移動された文書に基づいてラッパーツリーを調整する。ブロック６０８で、このコンポーネントは、クラスタＴＰをクラスタのコレクションＲに追加する。ブロック６０９で、このコンポーネントは、ラッパーツリーT_wをラッパーツリーのコレクションＷに追加する。その後、このコンポ

ーネントは、ブロック６０１にループして、クラスタの識別を継続する。

【０１０９】

図７は、一実施形態でのラッパー生成システムのラッパーへの文書変換コンポーネントの処理を示す流れ図である。このコンポーネントは、文書ツリーのルートノードを渡され、ラッパーツリーを生成するためにこのコンポーネント自体を再帰的に呼び出す。このコンポーネントは、再帰的であるものとして図示されているが、当業者は、このコンポーネントを、その代わりに非再帰的な形で実施できることを了解するであろう。判断ブロック７０１で、渡されたノードが葉ノードである場合には、このコンポーネントはリターンし、そうでない場合には、このコンポーネントはブロック７０２で継続する。ブロック７０２～７０４では、このコンポーネントは、渡されたノードの各子ノードを選択し、このコンポーネントを再帰的に呼び出しながらループする。ブロック７０２で、このコンポーネントは、渡されたノードの次の子ノードを選択する。判断ブロック７０３で、すべての子ノードを既に選択し終えている場合に、このコンポーネントは、ブロック７０５で継続し、そうでない場合には、このコンポーネントは、ブロック７０４で継続する。ブロック７０４で、このコンポーネントは、選択されたノードを渡して、ラッパーへの文書変換コンポーネントを再帰的に呼び出し、その後、ブロック７０２にループして、次の子ノードを選択する。ブロック７０５で、このコンポーネントは、渡されたノードの連続するサブツリーを組み合わせる。ブロック７０６で、このコンポーネントは、渡されたノードの連続するサブフォレストを組み合わせる。その後、このコンポーネントは、リターンする。

【０１１０】

図８は、一実施形態でのラッパー生成システムのクラスタへの一致する文書の移動コンポーネントの処理を示す流れ図である。このコンポーネントは、ラッパーツリーT_wと一致するコレクションＤのすべての文書ツリーT_dをクラスタＴＰに移動する。ブロック８０１で、このコンポーネントは、コレクションＤの次の文書ツリーT_dを選択する。判断ブロック８０２で、すべての文書ツリーが既に選択済みである場合には、このコンポーネントは、リターンし、そうでない場合には、このコンポーネントは、ブロック８０３で継続する。ブロック８０３では、このコンポーネントは、距離計算コンポーネントを呼び出して、選択された文書ツリーT_dとラッパーツリーT_wとの間の距離を計算する。判断ブロック８０４で、距離が０である場合には、選択された文書ツリーは、ラッパーツリーと一致し、このコンポーネントは、ブロック８０５で継続し、そうでない場合には、このコンポーネントは、ブロック８０１にループして、次の文書ツリーを選択する。ブロック８０５で、このコンポーネントは、選択された文書ツリーT_dをクラスタＴＰに移動し、その後、ブロック８０１にループして、次の文書ツリーを選択する。

【０１１１】

図９は、一実施形態でのラッパー生成システムのクラスタへのほぼ一致する文書の移動コンポーネントの処理を示す流れ図である。このコンポーネントは、ラッパーツリーT_wとほぼ一致するコレクションＤの文書ツリーをクラスタＴＰに移動し、ラッパーツリーを動的に調整する。このコンポーネントは、コレクション内のどの文書も、調整されたラッパーツリーとほぼ一致しなくなるまでこの処理を繰り返す。ブロック９０１～９０６では、このコンポーネントは、ほぼ一致する文書ツリーをクラスタに移動しながらループする。ブロック９０１で、このコンポーネントは、文書ツリーT_dがラッパーツリーT_wとほぼ一致するかどうかを判定するために、ほぼ一致する文書のチェックコンポーネントを呼び出す。判断ブロック９０２で、ほぼ一致する文書ツリーが見つかった場合には、このコンポーネントは、ブロック９０３で継続し、そうでない場合には、このコンポーネントは、ブロック９０７で継続する。ブロック９０３で、このコンポーネントは、ラッパーへの文書の変換コンポーネントを呼び出して、ほぼ一致する文書ツリーをラッパーツリーT_w’に変換する。ブロック９０４で、このコンポーネントは、ラッパー調整コンポーネントを呼び出して、クラスタに追加される文書ツリーから生成されたラッパーツリーT_w’に基づいて、ラッパーツリーT_wを調整する。ブロック９０５で、このコンポーネントは、文書ツリーT_dをコレクションＤからクラスタＴＰに移動する。ブロック９０６で、この

コンポーネントは、クラスタへの一致する文書の移動コンポーネントを呼び出して、調整されたラッパーツリーと一致するコレクションのすべての文書ツリーをクラスタに移動する。その後、このコンポーネントは、ブロック９０１にループして、調整されたラッパーツリーとほぼ一致するさらなる文書ツリーをチェックする。

【０１１２】

ブロック９０７～９１１で、このコンポーネントは、クラスタへの文書ツリーの移動に関する適応しきい値を実施する。このコンポーネントが固定しきい値を使用した場合には、このコンポーネントは、ブロック９０７で継続するのではなく、リターンする。ブロック９０７では、このコンポーネントは、現在のしきい値が、前のしきい値を使用して生成されたラッパーツリーより改善されたラッパーツリーをもたらしたかどうかを判定する。判断ブロック９０８で、改善が大きい場合には、このコンポーネントは、ブロック９０９でしきい値を増やし、ブロック９０１にループして、増やされたしきい値に基づいて文書ツリーをクラスタに移動する。しかし、改善が大きくはない場合には、このコンポーネントは、ブロック９１０で継続する。判断ブロック９１０では、改善がある場合に、このコンポーネントは、リターンし、そうでない場合には、このコンポーネントは、ブロック９１１で継続する。ブロック９１１で、このコンポーネントは、改善をもたらさなかった現在のしきい値に関する文書ツリー移動の影響をロールバックし、その後、リターンする。

【０１１３】

図１０は、一実施形態でのラッパー生成システムのほぼ一致する文書のチェックコンポーネントの処理を示す流れ図である。このコンポーネントは、ほぼ一致である文書ツリーが見つかるか、一致する文書ツリーがないと判定されるまで、コレクションＤの文書ツリーとラッパーツリーT_wとの間の距離をチェックしながらループする。ブロック１００１で、このコンポーネントは、コレクションＤの次の文書ツリーT_dを選択する。判断ブロック１００２で、コレクションの文書ツリーのすべてが既に選択済みである場合には、ほぼ一致である文書ツリーはなく、このコンポーネントは、ほぼ一致がないことを示してリターンし、そうでない場合には、このコンポーネントは、ブロック１００３で継続する。ブロック１００３で、このコンポーネントは、ラッパーツリーT_wおよび文書ツリーT_dを渡して距離計算コンポーネントを呼び出して、文書ツリーとラッパーツリーとの間の距離を計算する。判断ブロック１００４で、距離がしきい値未満である場合に、このコンポーネントは、ほぼ一致を示して文書ツリーを返し、そうでない場合には、このコンポーネントは、ブロック１００１にループして、コレクションの次の文書ツリーを選択する。代替案では、このコンポーネントは、最も近いほぼ一致するが同一ではない文書ツリー、最も遠いほぼ一致する文書ツリー、またはランダムに選択されたほぼ一致する文書ツリーを返すことができる。

【０１１４】

図１１は、一実施形態でのラッパー生成システムの距離計算コンポーネントの処理を示す流れ図である。このコンポーネントは、ラッパーツリーT_wおよび文書ツリーT_dを渡され、そのラッパーツリーとその文書ツリーとの間の距離を計算する。ブロック１１０１で、このコンポーネントは、文書ツリーをラッパーツリーに位置合せする。ブロック１１０２で、このコンポーネントは、ラッパーツリーの位置合せされないノードの個数C_wをカウントする。ブロック１１０３で、このコンポーネントは、ラッパーツリーの重さW（T_w）を計算する。ブロック１１０４で、このコンポーネントは、文書ツリーの位置合せされないノードの個数C_dをカウントする。ブロック１１０５で、このコンポーネントは、文書ツリーの重さW（T_d）を計算する。ブロック１１０６で、このコンポーネントは、式５を使用して距離を計算する。その後、このコンポーネントは、リターンする。

【０１１５】

図１２は、一実施形態でのラッパー生成システムのラッパー調整コンポーネントの処理を示す流れ図である。このコンポーネントは、ラッパーツリーの対を渡され、これらを単一のラッパーツリーにマージする。ブロック１２０１で、このコンポーネントは、ラッパー位置合わせコンポーネントを呼び出して、ラッパーツリーのノードを位置合せする。ラ

ッパー位置合わせコンポーネントは、ラッパーツリーのサブツリーのすべての可能な位置合わせをテストする再帰コンポーネントである。このコンポーネントは、動的計画法技法を使用して、以前にテストされた解の再テストを防ぐことができる。ブロック１２０２で、このコンポーネントは、ラッパーツリーのルートノードの次の解を選択する。判断ブロック１２０３で、すべての解が既に選択済みである場合には、このコンポーネントは、ブロック１２０５で継続し、そうでない場合には、このコンポーネントは、ブロック１２０４で継続する。ブロック１２０４で、このコンポーネントは、解のコストを集計し、その後、ブロック１２０２にループして、次の解を選択する。ブロック１２０５で、このコンポーネントは、最小のコストを有する解を選択し、その後、リターンする。

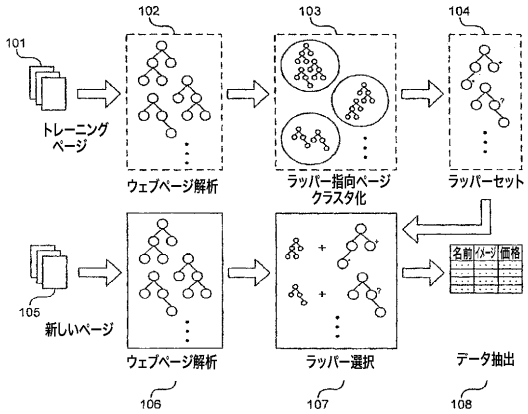
【０１１６】

図１３は、一実施形態でのラッパー生成システムのラッパー位置合わせコンポーネントの処理を示す流れ図である。このコンポーネントは、すべての可能な解のコストを判定するために再帰的に呼び出される。判断ブロック１３０１で、両方のラッパーツリーのレイヤがまだある場合には、このコンポーネントは、ブロック１３０２で継続し、そうでない場合には、このコンポーネントは、リターンする。ブロック１３０２～１３０５では、このコンポーネントは、新しい解を選択し、ラッパー位置合わせコンポーネントを再帰的に呼び出しながらループする。ブロック１３０２で、このコンポーネントは、現在のレイヤの次の解を選択する。判断ブロック１３０３で、すべての解が既に選択済みである場合には、このコンポーネントは、リターンし、そうでない場合には、このコンポーネントは、ブロック１３０４で継続する。ブロック１３０４で、このコンポーネントは、ラッパーツリーの次のレイヤのノードを位置合せするために、ラッパー位置合わせコンポーネントを再帰的に呼び出す。ブロック１３０５で、このコンポーネントは、レイヤの選択された解のコストをセットし、その後、ブロック１３０２にループして、次の解を選択する。

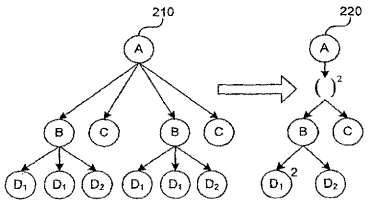
【０１１７】

本主題を、構造的特徴および／または方法論的動作に固有の言葉で説明してきたが、添付の特許請求の範囲で定義される本主題が、必ずしも上で説明した特定の特徴または動作に限定されないことを理解されたい。そうではなく、上で説明した特定の特徴および動作は、特許請求の範囲を実施する例の形として開示されたものである。したがって、本発明は、添付の特許請求の範囲によるものを除いて限定されない。

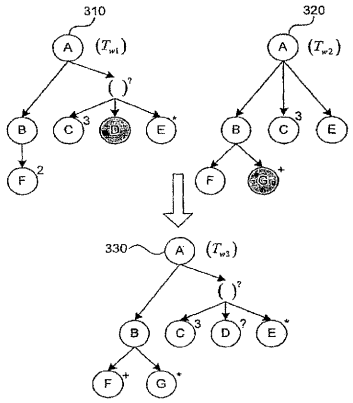
【図１】



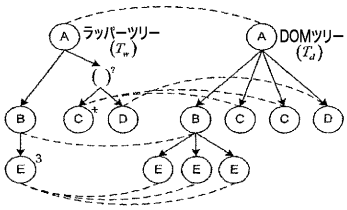
【図２】



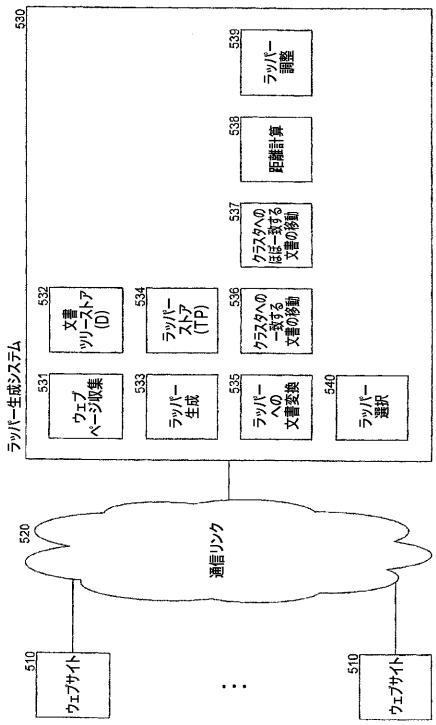
【図３】



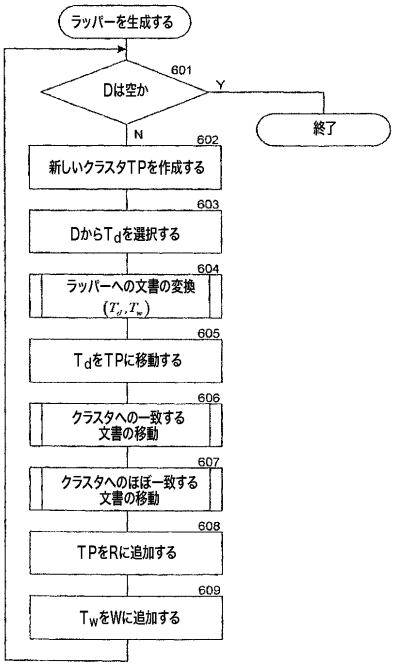
【図４】



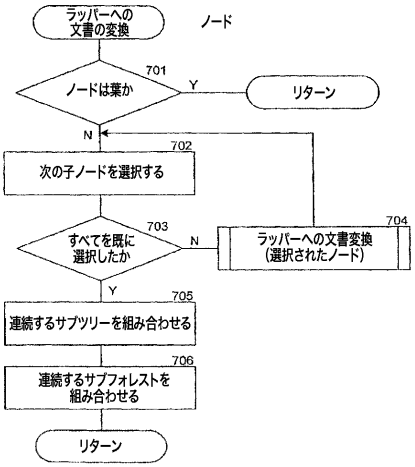
【図５】



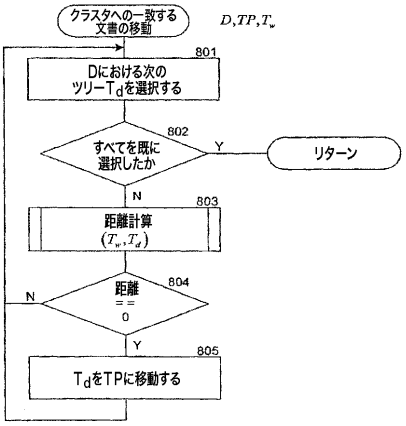
【図６】



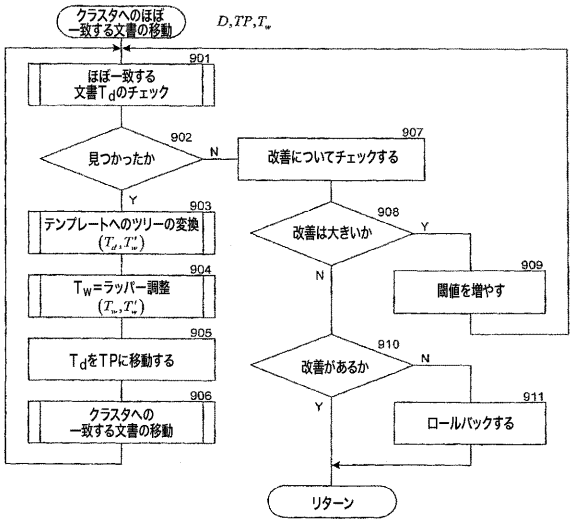
【図７】


ラッパーへの
文書の変換
ノード
701
ノードは葉か
Y
リターン
N
702
次の子ノードを選択する
703
すべてを既に
選択したか
N
704
ラッパーへの文書変換
(選択されたノード)
Y
705
連続するサブツリーを組み合わせる
706
連続するサブフォレストを
組み合わせる
リターン

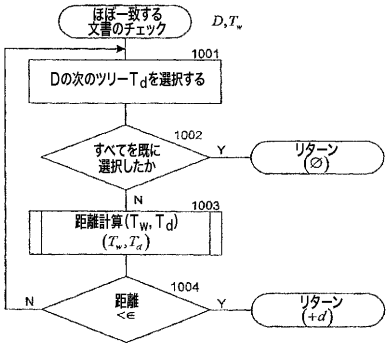
【図８】


クラスタへの一致する
文書の移動
D,TP,Tw
801
Dにおける次の
ツリーTdを選択する
802
すべてを既に
選択したか
Y
リターン
N
803
距離計算
(Tw,Td)
804
距離
==
0
N
Y
805
TdをTPに移動する

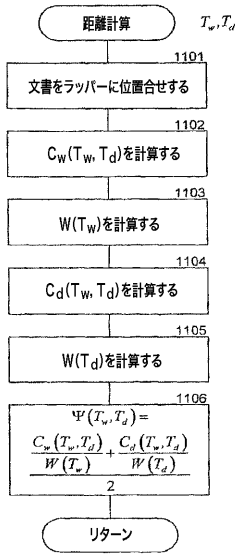
【図９】


クラスタへのほぼ
一致する文書の移動
D,TP,Tw
901
ほぼ一致する
文書Tdのチェック
902
見つかったか
N
907
改善についてチェックする
Y
903
テンプレートへのツリーの変換
(Td,Tw')
908
改善は大きいか
Y
909
閾値を増やす
N
904
Tw=ラッパー調整
(Tw,Tw')
905
TdをTPに移動する
910
改善があるか
N
906
クラスタへの
一致する文書の移動
Y
911
ロールバックする
リターン

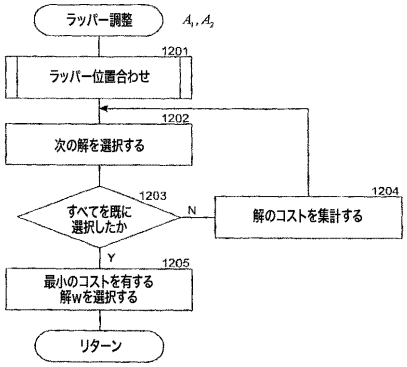
【図１０】


ほぼ一致する
文書のチェック
D,Tw
1001
Dの次のツリーTdを選択する
1002
すべてを既に
選択したか
Y
リターン
(∅)
N
1003
距離計算(Tw,Td)
(Tw,Td)
1004
距離
<∈
N
Y
リターン
(+d')

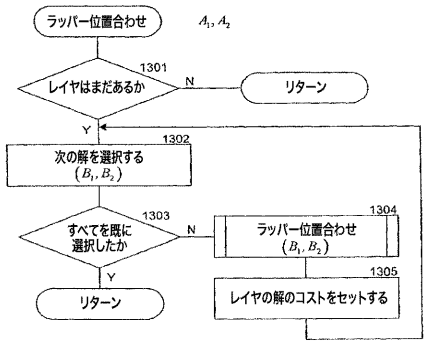
【図１１】


距離計算
T_w, T_d
1101
文書をラッパーに位置合せする
1102
$C_W(T_W, T_d)$を計算する
1103
$W(T_W)$を計算する
1104
$C_d(T_W, T_d)$を計算する
1105
$W(T_d)$を計算する
1106
$\Psi(T_w, T_d) = \frac{\frac{C_w(T_w, T_d)}{W(T_w)} + \frac{C_d(T_w, T_d)}{W(T_d)}}{2}$
リターン

【図１２】


ラッパー調整
A_1, A_2
1201
ラッパー位置合わせ
1202
次の解を選択する
1203
すべてを既に選択したか
N
1204
解のコストを集計する
Y
1205
最小のコストを有する解Wを選択する
リターン

【図１３】


ラッパー位置合わせ
A_1, A_2
1301
レイヤはまだあるか
N
リターン
Y
1302
次の解を選択する
(B_1, B_2)
1303
すべてを既に選択したか
N
1304
ラッパー位置合わせ
(B_1, B_2)
Y
1305
レイヤの解のコストをセットする
リターン

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2007/018417

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/00(2006.01)i, G06F 17/24(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)
IPC8 G06F 17/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean Utility models and applications for Utility models since 1975
Japanese Utility models and applications for Utility models since 1975

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
ekipass "wrapper, tree, web-page, cluster, parsing, query, extract, hierarchy, document, node, metadata"

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	WO 2005/072072 A2 (KIM, SUN KWON) 11 AUG. 2005 See abstract; figures 2,6~8,11. claims 1~6,24,29,35.	1-11,16-20 12-15
Y A	WO 2006/036376 A1 (GOOGLE INC.) 06 APR. 2006 See abstract; figures 7~9b. claims 1,5~8,18,19,24,26,27,30~32,42.	1-11,16-20 12-15
A	US 2004/0068697 A1 (GEORGES HARIK et al.) 08 APR. 2004 See abstract; figures 22~25. claims 1,21,41,61.	1-20
A	KR 1020030069639 A (LEE, UI BUM) 27 AUG. 2003 See abstract; figure 3. claims 1,2.	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 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
01 FEBRUARY 2008 (01.02.2008)	**01 FEBRUARY 2008 (01.02.2008)**

Name and mailing address of the ISA/KR
Korean Intellectual Property Office
920 Dunsan-dong, Seo-gu, Daejeon 302-701,
Republic of Korea
Facsimile No. 82-42-472-7140

Authorized officer
KIM, Jung Jin
Telephone No. 82-42-481-5962

Form PCT/ISA/210 (second sheet) (April 2007)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2007/018417

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO2005072072A2	11.08.2005	CN1906615A	31.01.2007
		KR2005077681A	03.08.2005
		US2007110047A1	17.05.2007
		WO2005072072A3	06.10.2005
WO2006036376A1	06.04.2006	AU2005290154A1	06.04.2006
		CA2581713A1	06.04.2006
		EP01800226A1	27.06.2007
		KR1020070058685A	08.06.2007
		US2006074907AA	06.04.2006
US20040068697A1	08.04.2004	US2007208772AA	06.09.2007
		US7231393BA	12.06.2007
KR1020030069639A	27.08.2003	None	

Form PCT/ISA/210 (patent family annex) (April 2007)

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,NA,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IS,IT,LT,LU,LV,MC,MT,NL,PL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BH,BR,BW,BY,BZ,CA,CH,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,IS,JP,KE,KG,KM,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,PG,PH,PL,PT,RO,RS,RU,SC,SD,SE,SG,SK,SL,SM,SV,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,ZA,ZM,ZW

(72)発明者 ミン ワン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内
(72)発明者 ルイホワ ソン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内
(72)発明者 ウェイ－イン マ
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内
(72)発明者 シュイ ゼン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内

Ｆターム(参考) 5B075 ND35 NK43 NR06