

(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2005-293582

(P2005-293582A)

(43) 公開日 平成17年10月20日(2005. 10. 20)

(51) Int.Cl.⁷

G06F 17/30

G06F 17/27

F I

G06F 17/30

G06F 17/30

G06F 17/27

330B

170A

E

テーマコード (参考)

5B075

5B091

審査請求 未請求 請求項の数 22 O L (全 16 頁)

(21) 出願番号 特願2005-95536 (P2005-95536)
 (22) 出願日 平成17年3月29日 (2005. 3. 29)
 (31) 優先権主張番号 04075975.5
 (32) 優先日 平成16年3月31日 (2004. 3. 31)
 (33) 優先権主張国 欧州特許庁 (EP)

(71) 出願人 593016732
 オセーテクノロジーズ ビービー
 オランダ国 5914 シーシー ヴェン
 ロ セイント ウルバヌスヴェーク 43
 番地
 (74) 代理人 100070150
 弁理士 伊東 忠彦
 (74) 代理人 100091214
 弁理士 大貫 進介
 (74) 代理人 100107766
 弁理士 伊東 忠重
 (72) 発明者 サミュエル イェー ドリーッセン
 オランダ国, 6544 ゼットヘー ナイ
 メーヘン オーバーデストラート 112

最終頁に続く

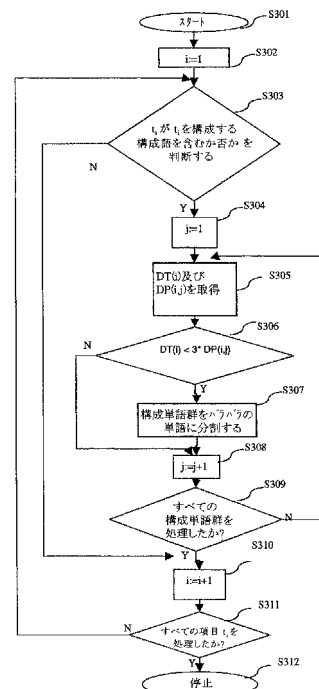
(54) 【発明の名称】 複合語を構成する単語を割り出す装置及びコンピュータ化された方法

(57) 【要約】

【課題】 複合語の構成単語を割り出す装置、コンピュータ・プログラム、及び、コンピュータ化された方法を提供すること。

【解決手段】 構成単語は、複合語を構成する。構成単語が分割判断基準を満たすとき、構成単語はバラバラに用いることができる。バラバラの構成単語は、検索において、文書の集合体から関連する文書を検索するのに用いられる。

【選択図】 図3



【特許請求の範囲】**【請求項 1】**

文書の集合体と、
前記文書の集合体の中で複合語を含む文書の数割り出す手段と、
前記文書の集合体の中で前記複合語を構成する構成単語を含む文書の数割り出す手段とを有する複合語の構成単語を割り出す装置であって、
前記複合語を含む文書の数と前記複合語を構成する構成単語を含む文書の数との比を求める手段と、
前記比が閾値より小さいときに前記複合語を前記構成単語に分割する手段と、を有することを特徴とする装置。

10

【請求項 2】

請求項 1 記載の装置であって、
前記複合語中の結合形態素を検出する手段を更に有する、ことを特徴とする装置。

【請求項 3】

請求項 1 記載の装置であって、
前記文書の集合体から求めた単語リストと、
前記単語リストを用いることによって前記構成単語を見つける手段とを更に有する、ことを特徴とする装置。

【請求項 4】

請求項 1 記載の装置であって、
前記文書の集合体から求めた単語リストと、
前記単語リストから前記複合語を含む文書の数求める手段とを更に有する、ことを特徴とする装置。

20

【請求項 5】

請求項 1 記載の装置であって、
前記文書の集合体から求めた単語リストと、
前記単語リストから前記複合語を構成する構成単語を含む文書の数求める手段とを更に有する、ことを特徴とする装置。

【請求項 6】

請求項 1 記載の装置であって、
前記閾値は前記文書の集合体に依存する、ことを特徴とする装置。

30

【請求項 7】

請求項 1 記載の装置であって、
前記閾値は 3 である、ことを特徴とする装置。

【請求項 8】

請求項 1 乃至 7 のいずれか一項記載の装置であって、
前記複合語及び前記構成単語は、文法上、名詞に属する、ことを特徴とする装置。

【請求項 9】

請求項 8 記載の装置であって、
前記単語リストは、各単語の単数形及び複数形を含む、ことを特徴とする装置。

40

【請求項 10】

情報検索システムであって、
単語を含む検索項目を入力する手段と、
前記単語を分割する手段とを有し、
前記 2 つの手段は、
請求項 1 乃至 9 のいずれか一項記載の装置を組み込んでおり、
構成単語を伝達し、
本システムは、更に、前記伝達された構成単語を検索項目として検索を実行する手段を有する、ことを特徴とするシステム。

【請求項 11】

50

文書の集合体の中で複合語を含む文書の数を割り出す工程と、

前記文書の集合体の中で前記複合語を構成する構成単語を含む文書の数を割り出す工程とを有する複合語の構成単語を割り出すコンピュータ化された方法であって、

前記複合語を含む文書の数と前記複合語を構成する構成単語を含む文書の数との比を求める工程と、

前記比が閾値より小さいときに前記複合語を前記構成単語に分割する工程と、を有することを特徴とする方法。

【請求項 12】

請求項 11 記載のコンピュータ化された方法であって、

前記複合語中の結合形態素を検出する工程を更に有する、ことを特徴とする方法。

10

【請求項 13】

請求項 11 記載のコンピュータ化された方法であって、

単語リストを用いることによって前記構成単語を見つける工程を更に有する、ことを特徴とする方法。

【請求項 14】

請求項 11 記載のコンピュータ化された方法であって、

単語リストから前記複合語を含む文書の数を求める工程を更に有する、ことを特徴とする方法。

【請求項 15】

請求項 11 記載のコンピュータ化された方法であって、

単語リストから前記複合語を構成する構成単語を含む文書の数を求める工程を更に有する、ことを特徴とする方法。

20

【請求項 16】

請求項 11 記載のコンピュータ化された方法であって、

オペレータによって検索を入力する工程を更に有する、ことを特徴とする方法。

【請求項 17】

請求項 11 記載のコンピュータ化された方法であって、

前記閾値を前記文書の集合体に応じて選択する工程を更に有する、ことを特徴とする方法。

【請求項 18】

請求項 11 記載のコンピュータ化された方法であって、

前記閾値は 3 である、ことを特徴とする方法。

30

【請求項 19】

請求項 11 乃至 18 のいずれか一項記載のコンピュータ化された方法であって、

前記複合語及び前記構成単語は、文法上、名詞に属する、ことを特徴とする方法。

【請求項 20】

請求項 19 記載のコンピュータ化された方法であって、

前記単語リストは、各単語の単数形及び複数形を含む、ことを特徴とする方法。

【請求項 21】

情報を検索するコンピュータ化された方法であって、

単語を含む検索項目を入力する工程と、

前記単語を請求項 11 乃至 20 のいずれか一項記載の方法に従って分割する工程と、

構成単語を伝達する工程と、

前記伝達された構成単語を検索項目として検索を実行する工程とを有する、ことを特徴とする方法。

40

【請求項 22】

コンピュータ上で実行されたときに請求項 11 記載のすべての工程を実行するコード手段を含むことを特徴とするコンピュータ・プログラム。

【発明の詳細な説明】

【技術分野】

50

【 0 0 0 1 】

本発明は、文書の集合体と、該文書の集合体から複合語を含む文書の数を割り出す手段と、上記文書の集合体から上記複合語を構成する単語を含む文書の数を割り出す手段とを有する複合語を構成する単語を割り出す装置に関する。また、本発明は、このような複合語を構成する単語を割り出す装置及びコンピュータ化された方法を組み込んだ情報検索システムにも関する。

【 背景技術 】

【 0 0 0 2 】

複合語は、ドイツ語、オランダ語、デンマーク語、ギリシア語、ノルウェー語、アイスランド語、及び、フィンランド語などの複数の言語において一般的である。複合語は自由に結合できるため、複合語により語彙サイズは莫大に増える。上記言語は、その言語のどの辞書にも載っていない非常に長い単語を含む可能性がある。このような単語の典型的な一例は、ドイツ語の複合語「Abschreibungs-möglichkeiten」である。この複合語は、結合形態素「s」によって2つの単語「Abschreibung」及び「möglichkeiten」を連結させることによって構成されている。以下の本発明の説明において、複合語をそれを構成する単語と（存在すれば）結合形態素とに分解することを複合語の細分化又は複合語を構成する単語をバラバラすると呼び、記号「+」によって隔てられた文字列によって表される。例えば、複合語「Abschreibungs-möglichkeiten」は、「Abschreibung+s+möglichkeiten」と表される。他の結合形態素としては、「-」や「es」などが用いられる。単語を復号化することは、上記言語において新しい単語を作り出す積極的な方法である。これは、機械翻訳、音声認識、テキスト分類、情報抽出、及び、情報検索などの複数のアプリケーション（CLEF2001：<http://www.ercim.org/publication/ws-proceedings/CLEF2/>）にチャレンジを与える。複合語を構成する単語を割り出すことは難しいことが判っている。本分野で既知の解決策が基本的には3つ存在する。

【 発明の開示 】

【 発明が解決しようとする課題 】

【 0 0 0 3 】

第一の解決策は、ある言語において存在する「すべての」複合語のリスト及び複合語の分割方法を保持しておくことである。ある言語におけるすべての複合語のリストを保持しておくことは不可能である点が欠点である。なぜなら、ある言語における複合語の量は際限がないからである。このため、この解決策の精度は低く、この方法は退屈である。この手法は一言語に適用される。

【 0 0 0 4 】

第二の解決策は、シンプルなものであって、規則に基づいて複合語を構成する単語を割り出すことである。この規則は、いつ分割するかを提示する統計データと組み合わせられる場合もある。ここでの問題は、この方法が、複合語を、存在しない単語、又は、複合語の意味とは関係のない単語の組み合わせ、に分割してしまうことが多いことである。これは一般的な複合語には有効であるが、強力ではない。この手法は、一言語に適用される。

【 0 0 0 5 】

第三の（最後の）解決策は、ディジタル辞書を用いる。辞書といくつかの規則とに基づいて、複合語を構成する単語の割り出しが実行される。この最後の解決策の一例は、米国特許公開第03/0097252号（複合語を構成する単語を割り出す方法）に開示されている。構成単語は、複合語の確率的区切り点群に基づいて割り出される。適切な語彙集から引かれたnグラムの分析に基づいて、複合語における区切り点に区切り点重みが割り当てられる。この方法は、複合語内の単語を割り出すものである。主たる欠点は、間違った単語分割が発生し得る点である。

【 0 0 0 6 】

従来技術の欠点を克服するために、本発明の目的の1つは、複合語を構成する単語を割

10

20

30

40

50

り出すことである。

【課題を解決するための手段】

【0007】

これは、ブリアンブル記載の装置であって、更に、複合語を含む文書の数と該複合語を構成する構成単語を含む文書の数との比を求める手段と、該比が閾値より小さいときに上記複合語を上記構成単語に分割する手段とを有する装置によって実現される。

【0008】

この装置は、複合語を含む文書の数と該複合語を構成する構成単語を含む文書の数との比が閾値より小さいときには、その複合語は見つかった構成単語に正確に分割されるという観測に基づくものである。

10

【0009】

別の実施形態において、本発明に係る装置は、上記複合語中の結合形態素を検出する手段を更に有する。複合語中の結合形態素を検出することは、構成単語を見つけるのに必須である。

【0010】

別の実施形態において、本発明に係る装置は、上記文書の集合体から求めた単語リストと、該単語リストを用いることによって上記構成単語を見つける手段とを更に有する。複合語を構成する構成単語の検出は、上記文書の集合体のすべての単語を含む単語リストが推測的に用意されるときに、より効果的である。

【0011】

20

次の実施形態において、本発明に係る装置は、上記文書の集合体から求めた単語リストと、該単語リストから上記複合語を含む文書の数を求める手段とを更に有する。複合語を含む文書の数と導くことができる単語リストがあると効率的である。

【0012】

次の実施形態において、本発明に係る装置は、上記文書の集合体から求めた単語リストと、該単語リストから上記複合語を構成する構成単語を含む文書の数とを求める手段とを更に有する。複合語を構成する構成単語を含む文書の数と導くことができる単語リストがあると効率的である。

【0013】

次の実施形態において、本発明に係る装置は、情報検索システムに内蔵される。ユーザは、検索項目を入力し、この項目が複合語である場合、構成単語に分割されて、分割された構成単語について検索が実行される。このように、オペレータは、複合語を構成する構成単語であって分割判断基準を満たす構成単語を含むすべての文書を見つける検索を創成することができる。

30

【0014】

次の実施形態において、本発明に係る装置は、上記閾値が上記文書の集合体に応じて選択される。これにより、用いられた文書の集合体に応じた閾値を選択することが可能となる。これは、複合語の分割を最適化する。

【0015】

別の実施形態において、本発明に係る装置は、上記閾値が3である。これは、オランダ語文書の集合体にとって最適な閾値であることを示している。

40

【0016】

別の実施形態において、本発明に係る装置は、上記複合語及び上記構成単語が、文法上、名詞に属する。

【0017】

別の実施形態において、本発明に係る装置は、上記単語リストが、各単語の単数形及び複数形を含む。これにより、複合語を構成するすべての構成単語が見つかる。

【発明を実施するための最良の形態】

【0018】

以下、本発明を下記の本発明の例示的实施形態を参照して説明し、図面を参照すること

50

により例示する。これら実施形態は、本発明を例示するためのものであって、本発明を限定するものと解釈されるべきではない。

【 0 0 1 9 】

以下、添付図面を参照して本発明を詳細に説明する。

【実施例】

【 0 0 2 0 】

図 1 は、本発明に係る情報検索システムの一実施形態を示している。ここで、情報検索システムとは、例えば、文書管理システムや、ウェブ検索システムなどである。

【 0 0 2 1 】

文書管理システムは、オペレータ・コンソール (1 0 1) と、検索エンジン・フロント
10 エンド (1 0 4) と、検索エンジン (1 0 5) と、文書データベース (1 0 6) とを有す
る。オペレータ・コンソールは、本発明に係る本システムにユーザ・インターフェイスを
提供する。オペレータ・コンソールには、ディスプレイ (1 0 2) と、例えばキーボード
(1 0 3) であるデータ入力手段とが備えられている。検索エンジン・フロントエンド (1 0 4) は、オペレータ・コンソールから受信されたすべてのデータを検索エンジンによ
って実行される検索へ変換する。検索エンジン (1 0 5) は、この検索をデータベース上
20 で実行し、検索結果が提出者へ返されるように取り計らう。データベース (1 0 6) は、
検索対象の文書を含む。本システムの様々な変形例が考えられる。第一の変形例として、
図 1 のすべての構成要素が 1 つのパーソナル・コンピュータ内に内蔵されてもよい。別の
変形例として、検索エンジン (1 0 5) 、検索エンジン・フロントエンド (1 0 4) 、及
20 び文書データベース (1 0 6) が 1 つのサーバ内に実現され、ワークステーションとして
実現される数多くのオペレータ・コンソール (1 0 1) がクライアントとしてサーバにア
クセスし、ウェブ・ブラウザをサーバへのアクセス手段として用いる、ようにしてもよい
。

【 0 0 2 2 】

図 2 は、本発明に係る検索エンジン (1 0 5) 及び文書データベース (1 0 6) の一実
施形態を示している。文書の集合体は、報告書、新聞記事、及び、ウェブ・ページなどの
あらゆる形の文字情報を含み得る。実際の検索プロセスを開始する前に、ある文書の集合
30 体 (2 0 1) のすべての単語が例えば意味ネットワーク (2 0 2) によって収集され、単
語リスト (2 0 3) として順に並べられる。自動翻訳又は情報検索にとって最も良い結果
は、文書の集合体に存在する単語のみを考慮することによって得られる。これは、文書の
集合体からすべての単語を収集し、これらの単語を順番に単語リストに並べることによ
って行われる。

【 0 0 2 3 】

検索プロセスは、検索エンジン (1 0 5) が検索エンジン・フロントエンド (1 0 4)
から検索を受信したときに開始される。T 個の項目を含む検索の一例は、 t_1 且つ t_2 且
つ $\dots t_T$ である。ここで、項目 t_i は、用いられた項目群の中の一項目である (t_i
($t_1 \dots t_T$)) 。一項目は、任意の単語でよい。項目間には、AND (且つ) 、
OR (又は) 、NOT などの任意の論理演算子を用いることができる。当業者には明らか
40 かなように、検索は T 個の項目から成る。ここで、T は少なくとも 1 である。当業者には明
らかなように、本発明の一実施形態として、検索内で 1 以上の項目が処理されてもよい。

【 0 0 2 4 】

分割複合判断モジュール (2 0 4) は、一検索内のすべての項目について、その項目が
該検索内で別々にされるべき複数の構成単語を含むか否かを判断する。

【 0 0 2 5 】

分割複合判断モジュール (2 0 4) の結果は、複合語を構成する単語を別体として含む
修正された検索である。文書検索・ランク付けモジュール (2 0 5) は、この修正された
検索と単語リストを用いて、参照文書を取得して参照文書ごとに関連性スコアを計算し、
見つかった参照文書をランク付けして、このランク付けされた参照文書をそれらの関連性
50 スコアと共に検索エンジン・フロントエンド (1 0 3) へ伝達する。図 3 は、本発明に係

る分割複合判断モジュール(204)の一実施形態を示している。分割複合判断モジュール(204)は、一検索の項目 t_i ごとに、その項目が複合語であるか否か、及び、それらを構成する単語をバラバラにすべきか否か、を判断する。スタート位置301からスタートすると、まずステップ302において、項目カウンタ i が値1にセットされる。ステップ303において、項目 t_i が項目 t_i を構成する単語を含むか否かが判断される。これら構成単語は、単語リスト(204)から導かれる。同じ複合語を構成する単語の組み合わせは複数存在し得る。各組み合わせを構成単語群と呼ぶ。単語リストは、複合語を構成する単語の組み合わせを見つけるのに用いられる。構成単語が見つかったとき(ステップ303のY)、ステップ304において構成単語群カウンタ j が値1に初期化される。次のステップ305は、複合語 t_i を含む文書の数 $DT(i)$ と10
呼ぶ。ステップ305は、構成単語群 j を含む文書の数 $DP(i, j)$ を求める。 $DT(i)$ 及び $DP(i, j)$ の関数結果は、後に実例を挙げて説明するように、単語リストを用いて見つけることができる。ステップ306において、分割判断基準 $DT(i) < 3 * DP(i, j)$ が、各 j 番目の構成単語群について判定される。 $DT(i) < 3 * DP(i, j)$ を満たす場合(Y)、次いで、ステップ307において、複合語 t_i を構成する単語群 j がバラバラの単語に分離される。ステップ306において $DT(i) < 3 * DP(i, j)$ を満たさなかった場合(N)、処理はS308へ進む。分割判断基準はすべての項目 t_i について判定され、最終的に修正された検索が得られる。構成単語群カウンタの値 j は、ステップ308において1つインクリメントされ、ステップ309においてすべての構成単語群が処理されたか否かが判定される。処理されていない構成単語群もある20
場合(N)、次の構成単語群についてこの方法がステップ305において繰り返される。すべての構成単語群が処理されたとき(Y)、ステップ310において、項目カウンタ値 i が1つインクリメントされる。ステップ303において複合語を構成する単語が見つからずにステップ303が反復されるとき(N)も同様である。次のステップ311では、すべての項目 t_i が処理されたか否かが判定される。すべての項目 t_i が処理されていないとき(N)、本方法はステップ303へ戻る。すべての項目 t_i が処理済みのとき(Y)、本方法はステップ312において停止する。

【0026】

当業者には明らかなように、本発明の一実施形態として、検索内の1以上の項目を処理し、ステップ306の分割判断基準に合致した構成単語である項目を分離してもよい。既存の複数の単語を組み合わせることによって複合語が創成される。単語を組み合わせる際、接合部分に結合形態素文字を挿入することができる。このような結合形態素文字を挿入できる場合についてのシンプルな規則は存在しないため、すべての2つの単語の間についてチェックしなければならない。ドイツ語の複合語では、「s」及び「es」が結合形態素(linking morpheme)として用いられる。オランダ語では、結合形態素として「s」が可能である。30

【0027】

図4は、項目 t_i を構成する単語を割り出す方法(ステップ303)の一実施形態のフロー図を示している。図4の実施形態は、結合形態素として「s」のみを考えているため、オランダ語にとって典型的な例と言える。当業者には明らかなように、他の複合言語について、他の結合形態素及び2以上の結合形態素をチェックし、それに応じて対処することが可能である。本方法はステップ401からスタートする。ステップ402において、項目 t_i が文字列 w_1^L へコピーされる。ここで、 L は項目 t_i の総文字数を示す。さらに、単語文字スタート位置 s が1に、単語レベル・カウンタの値が1に、フラグが「偽」に、単語レベル上の潜在的単語の最後の文字位置 $c(level = 0)$ が値0に、単語文字終了位置 $c(level = 0)$ が値2に、それぞれ初期化される。ステップ403において、文字位置 $c(level)$ が文字列 w_s^L の最後から2番目の文字に到達していないとき、すなわち $L - c(level) \geq 2$ のとき(Y)、ステップ404において、単語リストを参照して、文字列 $w_s^{c(level)}$ が単語であるか判定する。 $w_s^{c(level)}$ が単語でないとき(S404のN)、ステップ408において、単語文字終了50

位置の値 $c(\text{level})$ が 1 つ増やされる。が単語のとき (S 4 0 4 の Y)、ステップ 4 0 5 において、その単語は実際の単語レベル上に単語ストアとして記憶される。例えば、最初に見つかった (よって、そのレベルはレベル 1 である) 単語 $w_{s, c(\text{level})}$ は、ストア $\text{Word}(\text{level} = 1)$ として記憶される。その後、このレベル値は次の単語レベルへ更新され、新しいレベル $c(\text{level})$ 値の上での文字位置の値が $c(\text{level} - 1)$ へ初期化される。次のステップ 4 0 6 において、後に詳述する 8 つの状況の 1 つである状況の値が求められ、ステップ 4 0 7 へ進む。状況値及びフラグ値に依存するステップ 4 0 7 においては、後に詳述するように、単語文字開始位置 s 及び単語文字終了位置 $c(\text{level})$ が新しい文字位置へ適合され、場合によっては 1 つの単語が見つかり、記憶される、などのいくつかのアクションが実行される。ステップ 4 0 7 の後、本方法はステップ 4 0 3 へ進む。ステップ 4 0 3 において、文字列 $w_{s, L}$ の最後から 2 番目の文字に達したとき、すなわち $L - c(\text{level}) < 2$ となったとき (N)、本方法はステップ 4 0 9 へ進む。ステップ 4 0 9 は、フラグステータスが「真」のとき、肯定 (Y) となり、ステップ 4 0 7 へ進む。これは、例えば、単語 $w_{L, c(\text{level}) + 1}$ を分割しようとする試みが終了し、本処理が続けて単語 $w_{L, c(\text{level}) + 2}$ を分割しようとする場合などである。その後、フラグが「偽」となり、ステップ 4 0 9 が否定 (N) となると、ステップ 4 1 0 へ進む。ステップ 4 1 0 では、単語レベルが 1 つ減らされる。ステップ 4 1 1 において、すべての単語レベルが他の単語を見つけたと判定されたか否かがチェックされる。単語レベル > 0 のとき (S 4 1 1 の Y)、本処理は、まずステップ 4 1 2 において単語文字終了位置の値 $c(\text{level})$ を 1 つ増やし、単語文字開始位置の値 s を $c(\text{level}) + 1$ で更新することによって単語を別の単語へ広げ、ステップ 4 0 3 へ進む。ステップ 4 1 1 におけるチェックにおいて、すべての単語レベルが別の単語へ単語拡張されたと判定されたとき (N)、ステップ 4 1 3 において、 t_i を構成する単語が見つかったか否かをチェックされ、その後、本方法はステップ 4 1 3 で終了する。

【0028】

図 5 は、表 1 (S 4 0 6) 及び表 2 (S 4 0 7) の内部の一実施形態を詳細に示している。表 1 (S 4 0 6) において、4 つの判断基準が状況値を判断する。これら判断基準は、フラグステータス (真 / 偽)、残りの部分 $w_{L, c(\text{level}) + 1}$ は単語である (はい / いいえ)、文字位置 $w_{c(\text{level}) + 1, c(\text{level}) + 1}$ は「s」である (はい / いいえ)、及び、残りの部分 $w_{L, c(\text{level}) + 2}$ は単語である (はい / いいえ)、である。表 1 (S 4 0 6) の第 5 列には、割り当てられた状況値が示されている。最終行の最後のセルに示されているように、フラグ値が真のときには状況値の更新はない。その後、表 2 (S 4 0 7) に更新後の最終的な状況が入力される。表 2 (S 4 0 7) において、2 つの判断基準：状況値 ($\{1, 2, \dots, 8\}$) 及びフラグステータス (はい / いいえ) が必要なアクションを判断する。表 2 (S 4 0 7) の判断基準の「フラグ」は、表 1 (S 4 0 6) の判断基準の「フラグ」と同じものである。図 5 の表 2 (S 4 0 7) のインジケータ $\text{CONSTITUTENT}()$ は、複合語を構成すると割り出された単語群 $\{\text{word}(h)_{h=1}^{h=\text{level}}\}$ が構成単語群リストにセットで追加されることを示している。一例を挙げると、表 2 (S 4 0 7) に状況 = 1 及びフラグ = 偽が入力されると、フラグが (後で $w_{L, c(\text{level}) + 2}$ を分割できるように) 真にセットされ、単語 $w_{L, c(\text{level}) + 1}$ がストア単語 (level) 内にコピーされ、構成単語がセットで構成単語群リストに追加され、その後、単語 $w_{L, c(\text{level}) + 2}$ がストア単語 (level) 内にコピーされ、構成単語がセットで構成単語群リストへ追加される。その後、文字位置値 $c(\text{level}) + 2$ 及び $c(\text{level}) + 3$ が後で (最初に単語 $w_{L, c(\text{level}) + 1}$ の分割を試みた後で単語 $w_{L, c(\text{level}) + 2}$ の 2 つの単語への分割を試みるのに) 用いるためにそれぞれヘルパー定数「sf」及び「cf」として記憶され、文字位置値 s 及び $c(\text{level})$ がそれぞれ $c(\text{level}) + 1$ 及び $c(\text{level}) + 2$ で更新され、本方法は表を離れ、ステップ 4 0 3 へ進み、単語 $w_{L, c(\text{level}) + 1}$ の 2 つの単語への分離を試みる。

【0029】

(具体例)

本システムの仕組みを説明するために、複合語内で分割判断基準を満たす構成単語を判断する具体例をいくつか紹介する。

【 0 0 3 0 】

(具体例 1)

第一の具体例において、検索はオランダ語の項目 t_i : 「basketballkampioenschappen」を含む。ここで、 $i = 1$ である。英語訳は「basketball championships (バスケットボール選手権)」である。分割複合判断モジュール (202) は、文書の集合体から引き出したすべての名詞を含む単語リストを用いて複合語を構成するすべての構成名詞を見つけ出す。一例として図6の単語リストが用いられる。この単語リストの第1列は、文書の集合体において利用可能なすべての名詞をアルファベット順に含む。第2列は、各名詞について、文書の集合体の中でその名詞を含むすべての参照文書を含む。第3列は、文書の集合体の中でその名詞を含む文書の数を含む。単語リストの一実施形態として、最初の2列だけを含むものとしてもよい。第3列は、第2列から求めることができる。当業者には明らかなように、単語リストは、第3列などの追加的情報で拡張することも、或いは、動詞などの他の単語種類も含み、名詞は名詞インジケータで示すようにすることも可能である。当業者には図6から明らかなように、複合語 t_i (第1列) を含む文書数 (第3列) は、この単語リストから求めることができる。

10

【 0 0 3 1 】

20

この具体例では、最初のメインステップ (S303) において、図6の単語リスト (205) を用いることによって、「basketballkampioenschappen」を構成する群と呼ばれるすべての構成単語可能性を判断する。オランダ語の複合語「basketballkampioenschappen」について、分割複合判断モジュール (202) は、8つの構成名詞群を見つける。図7では、見つかった8つすべての構成単語群が第2列に示されている。第一の見つかった構成単語群 ($j = 1$) : 「basketball」+「kampioenschappen」を図7の第2行の第2セルに示す。ステップ305において、複合語 t_i : 「basketballkampioenschappen」を含む文書数 $DT(i)$ が求められる。複合語「basketballkampioenschappen」は、意味ネットワーク・モジュール (206) によって、文書の集合体 (207) 内のどの文書にも見つからない。図6の単語リストの第1列において、複合語「basketballkampioenschappen」は利用可能でないため、 $DT(i) = 0$ である。ステップ305における次のアクションは、すべての構成単語群 j を含む文書数 $DP(i, j)$ を求めることである。構成単語を別々に1つの文書内に含む文書数 $DP(i, j = 1)$ は、図6に示した単語リストから導くことができる。この単語リストは、名詞「basketball」は文書 [1] ~ [8] において用いられており、名詞「kampioenschappen」は文書 [1] ~ [6] 及び [9] において用いられている。この単語リストから、これら2つの名詞は文書 [1] 及び [6] において双方とも用いられていることが導かれるため、 $DP(i, j = 1) = 6$ である。これと同じことを残りの7つの構成単語群について実行すると、 $DP(i, j) = 0$ が導かれる。図7の第4列に $DP(i, j)$ 値が示されている。ステップ306において、分割判断基準 $DT(i) < 3 * DP(i, j)$ が構成単語群 j ごとに判定される。この判定の結果は、図7の第5列に示されている。第一の ($j = 1$) 構成単語群のみがこの基準を満たす。この例では、構成単語「basketball」+「kampioenschappen」のみが分割基準を満たし、ステップ307において別々の「basketball」及び「kampioenschappen」に分割される。この例でのステップ307の結果は、図7の第6列に示されている。

30

40

【 0 0 3 2 】

分割複合判断モジュールを使用せずに検索項目「basketballkampioenschappen」を検索エンジン (105) に適用すると、修正された検索が得られず

50

、文書の集合体から関連文書が一切見つけられない。しかし、分割複合判断モジュールを検索項目「basketballkampioenschappen」に適用すると、修正された検索は、例えば、

(「basketball」AND「kampioenschappen」)

となる。ここで、例えば、

(「basketball」AND「kampioenschappen」)OR「basketballkampioenschappen」

も妥当である。

【0033】

この修正された検索により、場合によっては図6の単語リストに助けられて、関連文書 [1]、[2]、[3]、[4]、[5]、及び、[6]が見つかる。

【0034】

(具体例2)

ここでは、オランダ語の例を挙げて、複合語 t_i の構成単語がステップ303においてどのように見つけ出されるのかを説明する。この単語の英語訳は「course of life (人生航路)」である。

【0035】

【数1】

複合語	= w_1^{10} =	L E V E N S L O O P	20
文字位置	=	1 2 3 4 5 6 7 8 9 10	

構成単語を見つけている間に複数のイベントを識別することができる。これらイベントを表1に示す。ここで、単語リストは「levens」を定義していないことに注意。英語訳は「life (人生)」である。

【0036】

【表 1】

表1

イベント	説明
1	ステップ(402):初期化
2	ステップ(407):名詞「leven」が見つかる⇒単語(level=1):=「leven」
3	ステップ(408): 「sloop」は名詞であり、「s」から始まり、「loop」は名詞である⇒状況1 ステップ(409): 単語(2):=「sloop」。構成単語群をリスト(1)=leven+sloopへ追加。 英語:「life」+「demolition」 単語(2):=「loop」。構成単語群をリスト(2)=leven+loopへ追加。 英語:「life」+「stream」
3.1	ステップ(404):「sloop」の分割を試みる
3.2	ステップ(405):「sloop」は分割できない
3.3	ステップ(404):「loop」の分割を試みる
3.4	ステップ(405):「loop」は分割できない
4	ステップ(414):レベル「levens」に戻る。「levens」を別の単語へ拡張する。
5	ステップ(405):「levens」は別の単語へ拡張できない
6	ステップ(415):終了

10

20

表1のイベントは、(名詞)レベル、名詞開始位置s、及び、2つのレベルc(level=1)及びc(level=2)についての単語文字終了位置c(level)、のイベントステータスと共に、表2に入れられる。

【0037】

30

【表 2】

表 2

イベント	1	2	3	3.1	3.2	3.3	3.4	4	5	6
レベル	1	1	2	2	2	2	2	1	1	0
s=	1	1	6	6	6	7	7	1	1	1
c(1)=	2	5	5	5	5	5	5	6	9	9
c(2)=	--	--	5	7	9	8	9	9	9	9

40

(具体例3)

別の例を挙げる。

【0038】

【数 2】

複合語 = W_1^{17} = A L P E N V E R K E E R S P L A N
 文字位置 = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

構成単語を見つけている間に表 3 に示す複数のイベントを識別することができる。ここで、単語リストは名詞「verkeersplan」（英語：「traffic plan」（交通計画）」を含まない。

【0039】

10

【表 3】

表3

イベント	説明
1	ステップ(402):初期化
2	ステップ(407):名詞「alpen」が見つかる⇒単語(1):=「alpen」
3	ステップ(409): 「verkeersplan」は名詞でなく、「s」から始まっておらず、「erkeersplan」は名詞でない⇒状況8
4	ステップ(403):「verkeersplan」の分割を試みる
5	ステップ(407):名詞「verkeer」が見つかる⇒単語(2):=「verkeer」
6	ステップ(408): 「splan」は名詞でなく、「s」から始まり、「plan」は名詞である⇒状況5 ステップ(409): 単語(3):=「plan」 構成単語群をリスト(1)=alpen+verkeer+planへ追加。 (英語:「alps」+「traffic」+「plan」)
6.1	ステップ(404):「plan」の分割を試みる
6.2	ステップ(405):「plan」は分割できない
6.3	ステップ(404):「splan」の分割を試みる
6.4	ステップ(405):「splan」は分割できない
7	ステップ(414):レベル「alpen」+「verkeers」に戻る。「verkeers」の別の単語への拡張を試みる。
8	ステップ(405):「verkeersplan」の他の分割方法が見つからなかった
9	ステップ(414):レベル「alpenv」に戻る。「alpenv」の別の単語への拡張を試みる。
10	ステップ(405):「alpenverkeerplan」はこれ以上分割できない
11	ステップ(415):終了

20

30

40

表 4 では、名詞レベル、名詞開始位置 s 、及び、3つのレベル c ($level = 1$)、 c ($level = 2$)、及び、 c ($level = 3$) についての単語文字終了位置 c ($level$)、というイベントステータスが与えられる。

【0040】

50

【表 4】

表 4

イベント	1	2	3	4	5	6	6.1	6.2	6.3	6.4	7	8	9	10	11
レベル	1	1	2	2	2	3	3	3	3	3	2	2	1	1	0
=															
s=	1	1	6	6	6	13	14	14	13	13	6	6	1	1	1
c(1)=	2	5	5	5	5	5	5	5	5	5	5	5	6	16	16
c(2)=	-	--	5	7	12	12	12	12	12	12	13	16	16	16	16
c(3)=	-	--	--	--	--	12	15	16	14	16	16	16	16	16	16

10

以上、本発明について説明したが、本発明は多くの方法で変化させ得ることは明らかである。このような変形例は、本発明の意図及び範囲を逸脱するものと考えられるべきではなく、このような変更のすべては、当業者には明らかであり、請求項の範囲内に含まれることが意図されている。

【図面の簡単な説明】

【0041】

20

【図1】本発明に係る装置の一実施形態のブロック図である。

【図2】本発明に係る検索エンジン及び文書データベースの一実施形態のブロック図である。

【図3】分割複合判断モジュールの一実施形態のフロー図である。

【図4】複合語を構成する単語を割り出す方法の一実施形態のフロー図である。

【図5】表1及び2の一実施形態である。

【図6】単語リストの一例の一実施形態である。

【図7】複合語「basketballkampioenschappen」についての方法中間結果の一実施形態を示す図である。

【符号の説明】

30

【0042】

101 オペレータ・コンソール

102 ディスプレイ

103 キーボード

104 検索エンジン・フロントエンド

105 検索エンジン

106 文書データベース

201 文書収集

202 意味ネットワーク

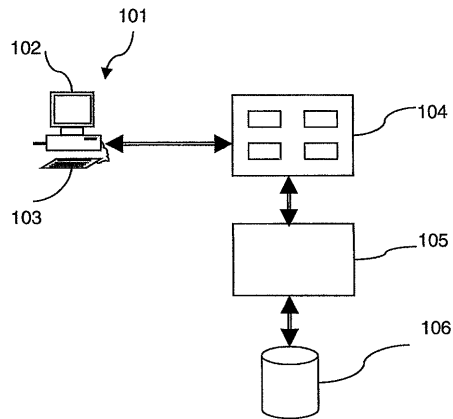
203 単語リスト

204 分割複合判断モジュール

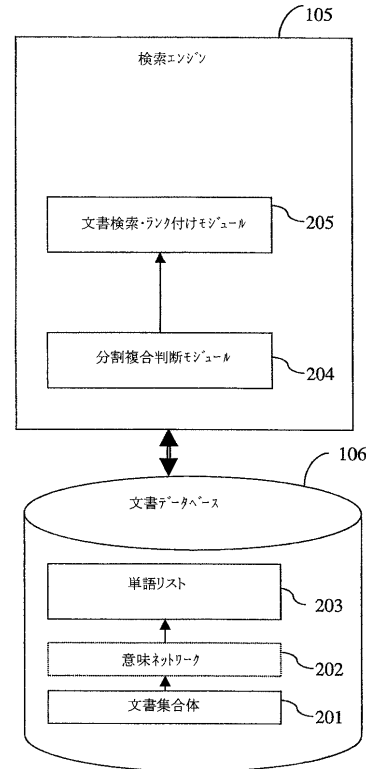
205 文書検索・ランク付けモジュール

40

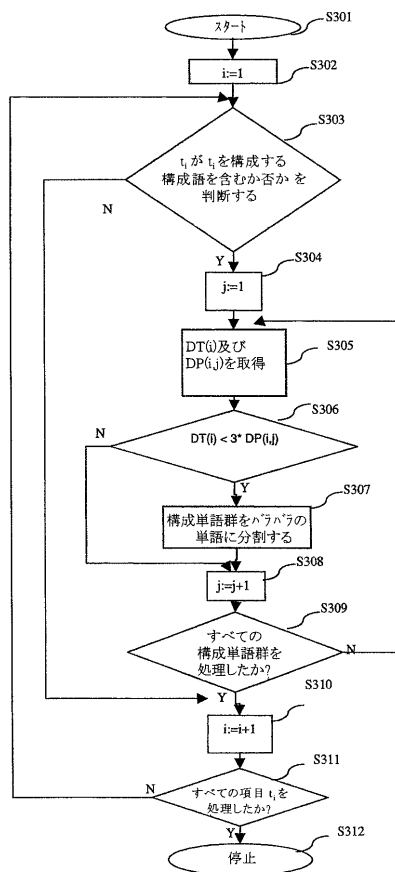
【図 1】



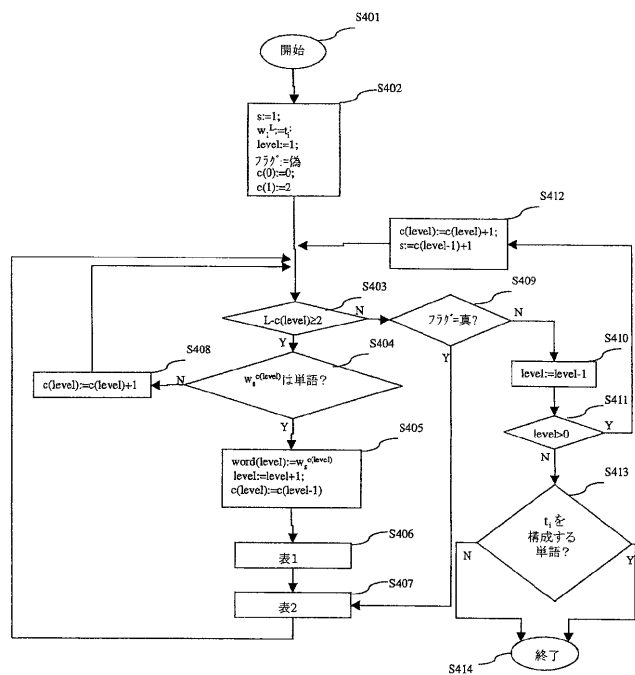
【図 2】



【図 3】



【図 4】



【図 5】

表 1:

判断基準				判断結果
フラグステータス	$W_{c(level)+1}^L$ は単語である	$W_{c(level)+1}^{c(level)+1}$ ='s'	$W_{c(level)+2}^L$ は単語である	状況 :=
偽	はい	はい	はい	1
偽	はい	はい	いいえ	2
偽	はい	いいえ	はい	3
偽	はい	いいえ	いいえ	4
偽	いいえ	はい	はい	5
偽	いいえ	はい	いいえ	6
偽	いいえ	いいえ	はい	7
偽	いいえ	いいえ	いいえ	8
真	任意	任意	任意	—

表 2:

状況 :=	フラグ=偽	フラグ=真
1	フラグ := 真 ; CONSTITUENT(Word(level):= $W_{c(level)+1}^L$); CONSTITUENT(Word(level):= $W_{c(level)+2}^L$); sf:=c(level)+2; cf:=c(level)+3; s:=c(level)+1; c(level):=c(level)+2;	フラグ := 偽 ; c(level) := cf; s := sf
2	フラグ := 真 ; CONSTITUENT(Word(level):= $W_{c(level)+1}^L$); sf:=c(level)+2; s:=c(level)+1; c(level):=c(level)+2;	フラグ := 偽 ; c(level) := sf+1
3	CONSTITUENT(Word(level):= $W_{c(level)+1}^L$); s:=c(level)+1; c(level):=c(level)+2;	
4	CONSTITUENT(Word(level):= $W_{c(level)+1}^L$); s:=c(level)+1; c(level):=c(level)+2;	
5	フラグ := 真 ; CONSTITUENT(Word(level):= $W_{c(level)+2}^L$); s:=c(level)+2; c(level):=c(level)+3;	フラグ := 偽 ; c(level) := c(level) + 2
6	フラグ := 真 ; c(level) := c(level) + 2	フラグ := 偽 ; c(level) := c(level) + 3
7	c(level) := c(level) + 2	
8	c(level) := c(level) + 2	

CONSTITUENT() は:

構成単語群 $\{Word(h)_{h=1}^{h=level}\}$ を
構成単語群リストへ追加する

【図 6】

単語(アルファベット順)	単語が見つかった文書[参照文書]	参照された文書の数
Basketbal	[1],[2],[3],[4],[5],[6],[7],[8]	8
Kam	[10]	1
Kampioen	[11]	1
Kampioenschap	[12],[13]	2
Kampioenschappen	[1],[2],[3],[4],[5],[6],[9]	7
Oen	[14]	1
Pen	[15]	1
Pi	[16],[17]	2
Pioen	[18]	1
Schap	[19]	1
Schappen	[20]	1

【図 7】

構成単語群	複合語 「Basketbalkampioenschappen」を 構成する単語	DT(i)	DP(i,1)	DT(i)<3*DP(i,1)	別々の単語
1	basketbal + kampioenschappen	0	6	はい	basketbal, kampioensch appen
2	basketbal + kam + pi + oen + schappen	0	0	いいえ	
3	basketbal + kam + pi + oen + schap + pen	0	0	いいえ	
4	basketbal + kam + pioen + schappen	0	0	いいえ	
5	basketbal + kam + pioen + schap + pen	0	0	いいえ	
6	basketbal + kampioen + schappen	0	0	いいえ	
7	basketbal + kampioen + schap + pen	0	0	いいえ	
8	basketbal + kampioenschap + pen	0	0	いいえ	

フロントページの続き

(72)発明者 パフェル エム イレイン

オランダ国, 5 9 1 2 イェーカー フェンロー マティルデ・ウィバウトストラート 6 6

F ターム(参考) 5B075 ND03 NK02 NS10

5B091 AA15 AB11 CA02