(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5979650号
(P5979650)
(45)発行日　平成28年8月24日(2016.8.24)　(24)登録日　平成28年8月5日(2016.8.5)

(51)Int.Cl.　F I
G06F 17/27 (2006.01)　G06F 17/27 670

請求項の数 20 (全 25 頁)

(21)出願番号　特願2014-152580(P2014-152580)
(22)出願日　平成26年7月28日(2014.7.28)
(65)公開番号　特開2016-31572(P2016-31572A)
(43)公開日　平成28年3月7日(2016.3.7)
審査請求日　平成28年1月5日(2016.1.5)

早期審査対象出願

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
INTERNATIONAL BUSINESS MACHINES CORPORATION
アメリカ合衆国10504 ニューヨーク州 アーモンク ニュー オーチャード ロード
New Orchard Road, Armonk, New York 10504, United States of America
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】用語を適切な粒度で分割する方法、並びに、用語を適切な粒度で分割するためのコンピュータ及びそのコンピュータ・プログラム

(57)【特許請求の範囲】
【請求項１】
用語を適切な粒度で分割する方法であって、コンピュータが抽出手段と分割手段とを備えており、前記方法は、
前記抽出手段が、コンテンツを格納したメモリ又は記憶装置から読み取り、構文解析により、粒度を規定する構成要素を前記コンテンツから抽出するステップを実行し、ここで、前記構成要素は、少なくとも１つの名詞又は記号を含む１又は複数の単語列であり、
前記分割手段が、前記用語がその一部に少なくとも１つの前記構成要素を含む場合に、前記用語を前記構成要素がある位置で分割し、当該分割した後の用語を当該分割した後の用語を入れるリストに格納するステップを実行し、
前記分割するステップが、
前記用語が当該用語の末尾から最長一致する前記構成要素を含む場合に、前記用語を前記末尾から最長一致する構成要素がある位置で分割するステップと、
前記用語から前記末尾から最長一致する前記構成要素を除いた後の用語が当該除いた後の用語の先頭から最長一致する前記構成要素を含む場合に、前記除いた後の用語を前記先頭から最長一致する構成要素がある位置で分割するステップと
を含む、
前記方法。
【請求項２】
前記用語を前記末尾から最長一致する前記構成要素がある位置で分割するステップが、

前記末尾から最長一致する前記構成要素を前記用語の主要語として保存するステップ
を含む、請求項１に記載の方法。
【請求項３】
前記除いた後の用語を前記先頭から最長一致する前記構成要素がある位置で分割するステップが、
前記先頭から最長一致する前記構成要素を前記用語の第１の修飾語として保存するステップ
をさらに含む、請求項１又は２に記載の方法。
【請求項４】
前記除いた後の用語を前記先頭から最長一致する前記構成要素がある位置で分割するステップが、
前記先頭から最長一致する前記構成要素以外の部分を第２の修飾語として保存するステップ
を含む、請求項３に記載の方法。
【請求項５】
前記構成要素を抽出するステップが、
前記コンテンツ中のテキストそれぞれに前記構文解析を適用して、文節を抽出するステップと、
前記抽出した文節のうちの名詞又は記号を含む文節から前記構成要素となりうる部分を抽出するステップと
を含む、請求項１～４のいずれか一項に記載の方法。
【請求項６】
前記構成要素を抽出するステップが、
前記コンテンツから、前記構成要素を抽出する対象のテキストを切り出すステップ
をさらに含み、
前記文節を抽出するステップが、前記切り出したテキストそれぞれに前記構文解析を適用して行われる、請求項５に記載の方法。
【請求項７】
前記構成要素を抽出するステップが、
前記切り出したテキストを事前定義した文字がある場所で分割するステップ
をさらに含み、
前記文節を抽出するステップが、前記分割したテキストそれぞれに前記構文解析を適用して行われる、請求項６に記載の方法。
【請求項８】
前記用語が用語リスト中の用語であり、
前記構成要素を抽出するステップが、
前記構成要素となりうる部分から前記用語リスト中にある同じ用語を削除し、当該削除した残りを前記構成要素とするステップ
をさらに含む、請求項５～７のいずれか一項に記載の方法。
【請求項９】
前記分割するステップが、
前記分割するステップに従い分割した分割回数と、予め設定された分割回数を規定する分割パラメータとを比較し、前記分割回数が前記分割パラメータよりも少ないことに応じて、前記用語を前記構成要素がある位置でさらに分割するステップ
を含む、請求項１～８のいずれか一項に記載の方法。
【請求項１０】
前記用語が用語リスト中の用語である、請求項１～７及び９のいずれか一項に記載の方法。
【請求項１１】
前記用語が、前記コンテンツ中の所定の長さよりも長い用語である、請求項１～７及び

１０のいずれか一項に記載の方法。
【請求項１２】
前記用語が名詞、記号又はそれらの組み合わせを含む単語列である、請求項１～１１のいずれか一項に記載の方法。
【請求項１３】
前記用語が複合名詞である、請求項１～１１のいずれか一項に記載の方法。
【請求項１４】
前記構成要素が、少なくとも１つの名詞又は記号を含む１又は複数の単語列である、請求項１～１３のいずれか一項に記載の方法。
【請求項１５】
用語を適切な粒度で分割するためのコンピュータであって、
コンテンツを格納するメモリ又は記憶装置と、
前記メモリ又は記憶装置からコンテンツを読み取り、構文解析により、粒度を規定する構成要素を前記コンテンツから抽出する抽出手段であって、前記構成要素は、少なくとも１つの名詞又は記号を含む１又は複数の単語列である、前記抽出手段と、
前記用語がその一部に少なくとも１つの前記構成要素を含む場合に、前記用語を前記構成要素がある位置で分割し、当該分割した後の用語を当該分割した後の用語を入れるリストに格納する分割手段と
を備えており、
前記分割手段が、
前記用語が当該用語の末尾から最長一致する前記構成要素を含む場合に、前記用語を前記末尾から最長一致する構成要素がある位置で分割すること、
前記用語から前記末尾から最長一致する前記構成要素を除いた後の用語が当該除いた後の用語の先頭から最長一致する前記構成要素を含む場合に、前記除いた後の用語を前記先頭から最長一致する構成要素がある位置で分割すること
を実行する、前記コンピュータ。
【請求項１６】
前記分割手段が、前記末尾から最長一致する前記構成要素を前記用語の主要語として保存することを実行する、請求項１５に記載のコンピュータ。
【請求項１７】
前記分割手段が、前記先頭から最長一致する前記構成要素を前記用語の第１の修飾語として保存することを実行する、請求項１５又は１６に記載のコンピュータ。
【請求項１８】
前記分割手段が、前記先頭から最長一致する前記構成要素以外の部分を第２の修飾語として保存することを実行する、請求項１７に記載のコンピュータ。
【請求項１９】
前記抽出手段が、
前記コンテンツ中のテキストそれぞれに前記構文解析を適用して、文節を抽出すること、
前記抽出した文節のうちの名詞又は記号を含む文節から前記構成要素となりうる部分を抽出すること
を実行する、請求項１５～１８のいずれか一項に記載のコンピュータ。
【請求項２０】
用語を適切な粒度で分割するためのコンピュータ・プログラムであって、コンピュータに、請求項１～１４のいずれか一項に記載の方法の各ステップを実行させる、前記コンピュータ・プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、用語の分割技法に関する。特には、本発明は、用語を適切な粒度で分割する

技法に関する。
【背景技術】
【０００２】
名詞や名詞相当の接辞が複数接続して、複数の単語（例えば、２～６語の単語）からなる複合名詞が限りなく作り出される。
【０００３】
システム開発において作成される用語集は、上記複合名詞を含む。しかしながら、複合名詞はその意味が一見して不明なことが多い。
【０００４】
特に、金融機関のシステム用の用語集は、例えば「定期預金毎月振替限度額」（漢字からなる複合名詞である）や「基準価額型金信解約合計金額」（漢字からなる複合名詞である）などの複合名詞を含みうる。
【０００５】
また、英語においても、語か複数連なって名詞句を形成する。例えば、「Beneficiary right seller's business security deposit」や「Financial instruments intermediary service」である。
【０００６】
複合名詞に形態素解析技術を使用して、複合名詞を分割する技術が知られている（例えば、下記非特許文献１～２を参照）。しかしながら、形態素解析技術では、形態素解析器が持つシステム辞書及び文法に基づいて複合名詞が分割されるために、必ずしも望ましい結果が得られていない。
【０００７】
形態素解析器が、例えば上記「定期預金毎月振替限度額」（漢字からなる複合名詞である）を形態素解析技術を使用して分割すると、「定期＋預金＋毎月＋振替＋限度額」（漢字からなる単語である）のように、単語ごとに細分化して分割してしまう。また、形態素解析器が、上記「基準価額型金信解約合計金額」（漢字からなる複合名詞である）を形態素解析技術を使用して分割すると、「基準＋価額＋型＋金＋信＋解約＋合計＋金額」（漢字からなる単語である）のように、本来一語である「金信」が漢字（すなわち、一つの単語であり、「金銭信託」（漢字である）の略語である）一文字に細分化して分割してしまう。
【０００８】
形態素解析器が、例えば上記「business security deposit」を形態素解析技術を使用して分割すると、「”business security”＋”deposit”」若しくは「”business”＋”security deposit”」に分割するのか、又は、分割せずに「”business security deposit”」のままにするのかを判定することは難しい。
【０００９】
下記特許文献１～１０は、文章の解析やキーワードの抽出を記載する。
【先行技術文献】
【特許文献】
【００１０】
【特許文献１】特開２００７－２５７３９０号公報
【特許文献２】特開平１０－２０７８９０号公報
【特許文献３】特開平７－８５１０１号公報
【特許文献４】特開２００７－２６４７１８号公報
【特許文献５】特開平８－３０５６９５号公報
【特許文献６】特開２００１－３２５２８４号公報
【特許文献７】特開２０１０－２０４８６６号公報
【特許文献８】特開２０１１－９６２４５号公報
【特許文献９】特開２００８－１４０３５９号公報
【特許文献１０】特開２０１２－２３４５１２号公報

【非特許文献】
【００１１】
【非特許文献１】太田悟　等、「規則・用例融合型の日本語複合名詞構造解析法」、言語処理学会　第３回年次大会　発表論文集、313～316頁1997年3月、＜URL:http://www.anlp.jp/proceedings/annual_meeting/2003/pdf_dir/C6-2.pdf＞から入手可能
【非特許文献２】高橋允彦　等、「構造化規則を用いた日本語複合名詞解析」、言語処理学会　第９回年次大会　発表論文集、541～544頁、2003年3月、＜URL:http://www.anlp.jp/proceedings/annual_meeting/2003/pdf_dir/C6-2.pdf＞から入手可能
【非特許文献３】宮崎正弘　等、「構造化チャートパーザを用いた日本語複合名詞構造解析器」、言語処理学会、2008年、＜URL:http://www.languetech.co.jp/out/08nlp-miyazaki.pdf＞から入手可能
【発明の概要】
【発明が解決しようとする課題】
【００１２】
複合名詞は限りなく作り出される故に、当該複合名詞の全てを辞書に登録することは不可能である。その為に、複合名詞を分割することが試みられている。
【００１３】
しかしながら、上記した通り、作り出される複合名詞の数は膨大であるので、人が、用語集に登録される全ての用語が複合名詞であるかどうかをチェックし、複合名詞である場合に当該複合名詞を分割することは現実的でない。
【００１４】
また、人手によって複合名詞を分割する場合には、複合名詞の分割の仕方が作業者に依存してしまい、例えば何が主要語又は修飾語であるのかの判断が作業者の主観によって異なる。
【００１５】
また、主要語又は修飾語のリストは、事前に定義されていない場合がほとんどであり、また事前に定義されていたとしてもそのリストは不完全である。
【００１６】
また、例えばシステム開発において作成される用語集に登録される複合名詞の数は数百にものぼり、特に、巨大な業務システム開発において作成される用語集に登録される用語の数は１０００を越える場合がある。
【００１７】
さらに、特定のプロジェクトの用語辞書に含まれる用語を、単語ごと又は一文字までではなく、幾つかの単語のまとまりとして分割したいという要望がある。例えば上記「定期預金毎月振替限度額」（漢字からなる複合名詞である）を、「定期預金＋毎月＋振替限度額」（いずれも漢字からなる単語である）に分割したいという要望がある。
【００１８】
そこで、本発明は、特定のプロジェクトが独自に持つ粒度で用語を分割する技法を提供することを目的とする。
【００１９】
また、本発明は、作業者の主観に依存することなしに、用語を分割する技法を提供することを目的とする。
【００２０】
さらに、本発明は、主要語又は修飾語のリストを必要とすることなしに、用語を分割する技法を提供することを目的とする。
【課題を解決するための手段】
【００２１】
本発明は、用語を適切な粒度で分割する技法を提供する。当該技法は、用語を適切な粒度で分割する方法、並びに、用語を適切な粒度で分割するためのコンピュータ、そのコンピュータ・プログラム及びコンピュータ・プログラム製品を包含しうる。

【００２２】
本発明に従う第１の態様において、用語を適切な粒度で分割する方法は、コンピュータが、
（Ａ）構文解析により、コンテンツから粒度を規定する構成要素（element word）を抽出するステップと、
（Ｂ）上記用語がその一部に少なくとも１つの上記構成要素を含む場合に、上記用語を上記構成要素がある位置で分割するステップと
を実行することを含む。
【００２３】
本発明の一つの実施態様において、（Ａ）上記構成要素を抽出するステップが、
（Ａ－３）上記コンテンツ中のテキストそれぞれに上記構文解析を適用して、文節を抽出するステップと、
（Ａ－４）上記抽出した文節のうちの名詞又は記号を含む文節から上記構成要素となりうる部分を抽出するステップと
を含みうる。
【００２４】
本発明の一つの実施態様において、（Ａ）上記構成要素を抽出するステップが、
（Ａ－１）上記コンテンツから、上記構成要素を抽出する対象のテキストを切り出すステップ
をさらに含み、
上記文節を抽出するステップが、上記切り出したテキストそれぞれに上記構文解析を適用して行われうる。
【００２５】
本発明の一つの実施態様において、（Ａ）上記構成要素を抽出するステップが、
（Ａ－２）上記切り出したテキストを事前定義した文字がある場所で分割するステップ
をさらに含み、
上記文節を抽出するステップが、上記分割したテキストそれぞれに上記構文解析を適用して行われうる。
【００２６】
本発明の一つの実施態様において、上記用語が用語リスト中の用語であり、
（Ａ）上記構成要素を抽出するステップが、
（Ａ－５）上記構成要素となりうる部分のうちから上記用語リスト中にある用語を削除し、当該削除した残りを上記構成要素とするステップ
をさらに含みうる。
【００２７】
本発明の一つの実施態様において、（Ｂ）上記分割するステップが、
（Ｂ－１）上記用語が当該用語の末尾から最長一致する上記構成要素（第１の構成要素）を含む場合に、上記用語を上記末尾から最長一致する上記構成要素（第１の構成要素）がある位置で分割するステップ
を含みうる。
【００２８】
本発明の一つの実施態様において、（Ｂ－１）上記用語を上記末尾から最長一致する上記構成要素（第１の構成要素）がある位置で分割するステップが、
上記末尾から最長一致する上記構成要素（第１の構成要素）を上記用語の主要語として保存するステップ
を含みうる。
【００２９】
本発明の一つの実施態様において、（Ｂ）上記分割するステップが、
（Ｂ－２）上記用語から上記末尾から最長一致する上記構成要素（第１の構成要素）を除いた後の用語が当該除いた後の用語の先頭から最長一致する上記構成要素（第２の構成

要素）を含む場合に、上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割するステップ
を含みうる。
【００３０】
本発明の一つの実施態様において、（Ｂ－２）上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割するステップが、
上記先頭から最長一致する上記構成要素（第２の構成要素）を上記用語の第１の修飾語として保存するステップ
をさらに含みうる。
【００３１】
本発明の一つの実施態様において、（Ｂ－２）上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割するステップが、
上記先頭から最長一致する上記構成要素（第２の構成要素）以外の部分を第２の修飾語として保存するステップ
を含みうる。
【００３２】
本発明の一つの実施態様において、（Ｂ）上記分割するステップが、
予め設定された分割回数を規定する分割パラメータに従って、上記用語を上記構成要素がある位置で分割するステップ
を含みうる。
【００３３】
本発明に従う第２の態様において、用語を適切な粒度で分割するためのコンピュータは、
構文解析により、粒度を規定する構成要素をコンテンツから抽出する抽出手段と、
上記用語がその一部に少なくとも１つの上記構成要素を含む場合に、上記用語を上記構成要素がある位置で分割する分割手段と
を備えている。
【００３４】
本発明の一つの実施態様において、上記抽出手段が、上記コンテンツ中のテキストそれぞれに上記構文解析を適用して、文節を抽出し、上記抽出した文節のうちの名詞又は記号を含む文節から上記構成要素となりうる部分を抽出しうる。
【００３５】
本発明の一つの実施態様において、上記抽出手段が、さらに、上記コンテンツから、上記構成要素を抽出する対象のテキストを切り出し、当該切り出したテキストそれぞれに上記構文解析を適用して、上記文節を抽出しうる。
【００３６】
本発明の一つの実施態様において、上記抽出手段が、さらに、上記切り出したテキストを事前定義した文字がある場所で分割し、当該分割したテキストそれぞれに上記構文解析を適用して、上記文節を抽出しうる。
【００３７】
本発明の一つの実施態様において、上記用語が用語リスト中の用語であり、
上記抽出手段が、上記構成要素となりうる部分のうちから上記用語リスト中にある用語を削除し、当該削除した残りを上記構成要素としうる。
【００３８】
本発明の一つの実施態様において、上記分割手段が、上記用語が当該用語の末尾から最長一致する上記構成要素（第１の構成要素）を含む場合に、上記用語を上記末尾から最長一致する上記構成要素（第１の構成要素）がある位置で分割しうる。
【００３９】
本発明の一つの実施態様において、上記分割手段が、上記用語を上記末尾から最長一致する上記構成要素（第１の構成要素）がある位置で分割し、上記末尾から最長一致する上

記構成要素（第１の構成要素）を上記用語の主要語として保存しうる。

【００４０】

本発明の一つの実施態様において、上記分割手段が、上記用語から上記末尾から最長一致する上記構成要素（第１の構成要素）を除いた後の用語が当該除いた後の用語の先頭から最長一致する上記構成要素（第２の構成要素）を含む場合に、上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割しうる。

【００４１】

本発明の一つの実施態様において、上記分割手段が、上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割し、上記先頭から最長一致する上記構成要素（第２の構成要素）を上記用語の第１の修飾語として保存しうる。

【００４２】

本発明の一つの実施態様において、上記分割手段が、上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割し、上記先頭から最長一致する上記構成要素（第２の構成要素）以外の部分を第２の修飾語として保存しうる。

【００４３】

本発明の一つの実施態様において、上記分割手段が、予め設定された分割回数を規定する分割パラメータに従って、上記用語を上記構成要素がある位置で分割しうる。

【００４４】

また、本発明に従う第３の態様において、コンピュータ・プログラム及びコンピュータ・プログラム製品は、上記コンピュータに、本発明に従う第１の態様に記載の用語を適切な粒度で分割する方法の各ステップを実行させる。

【００４５】

本発明の実施態様に従うコンピュータ・プログラムはそれぞれ、一つ又は複数のフレキシブル・ディスク、ＭＯ、ＣＤ－ＲＯＭ、ＤＶＤ、ＢＤ、ハードディスク装置、ＵＳＢに接続可能なメモリ媒体、ＲＯＭ、ＭＲＡＭ、ＲＡＭ等の任意のコンピュータ読み取り可能な記録媒体に格納することができる。当該コンピュータ・プログラムは、記録媒体への格納のために、通信回線で接続する他のデータ処理システム、例えばコンピュータからダウンロードしたり、又は他の記録媒体から複製したりすることができる。また、本発明の実施態様に従うコンピュータ・プログラムは、圧縮し、又は複数に分割して、単一又は複数の記録媒体に格納することもできる。また、様々な形態で、本発明の実施態様に従うコンピュータ・プログラム製品を提供することも勿論可能であることにも留意されたい。本発明の実施態様に従うコンピュータ・プログラム製品は、例えば、上記コンピュータ・プログラムを記録した記憶媒体、又は、上記コンピュータ・プログラムを伝送する伝送媒体を包含しうる。

【００４６】

本発明の上記概要は、本発明の必要な特徴の全てを列挙したものではなく、これらの構成要素のコンビネーション又はサブコンビネーションもまた、本発明となりうることに留意すべきである。

【００４７】

本発明の実施態様において使用されるコンピュータの各ハードウェア構成要素を、複数のマシンと組み合わせ、それらに機能を配分し実施する等の種々の変更は当業者によって容易に想定され得ることは勿論である。それらの変更は、当然に本発明の思想に包含される概念である。ただし、これらの構成要素は例示であり、そのすべての構成要素が本発明の必須構成要素となるわけではない。

【００４８】

また、本発明は、ハードウェア、ソフトウェア、又は、ハードウェア及びソフトウェアの組み合わせとして実現可能である。ハードウェアとソフトウェアとの組み合わせによる実行において、上記コンピュータ・プログラムをインストールされたコンピュータにおける当該プログラムの実行が典型的な例として挙げられる。かかる場合、当該コンピュータ・プログラムが当該コンピュータのメモリにロードされて実行されることにより、当該コ

ンピュータ・プログラムは、当該コンピュータを制御し、本発明にかかる処理を実行させる。当該コンピュータ・プログラムは、任意の言語、コード、又は、表記によって表現可能な命令群から構成されうる。そのような命令群は、当該コンピュータが特定の機能を直接的に、又は、１．他の言語、コード若しくは表記への変換及び、２．他の媒体への複製、のいずれか一方若しくは双方が行われた後に、実行することを可能にするものである。

【発明の効果】

【００４９】

本発明の実施態様に従うと、コンテンツ（例えば、マニュアル、業務手順書）から抽出される構成要素に従う粒度で、用語（例えば、用語辞書中の用語）を分割することが可能になる。コンテンツ又はコンテンツの技術分野が異なれば当該コンテンツから抽出される構成要素は異なる。従って、本発明の実施態様に従い、コンテンツ又はコンテンツの技術分野に従って粒度が動的に変わることから、コンテンツ又はコンテンツの技術分野に適した粒度で用語が分割されることが可能になる。

【００５０】

また、本発明の実施態様に従うと、作業者の主観に依存することなしに、コンテンツから抽出される構成要素に従う粒度で、用語が分割される。

【００５１】

さらに、本発明の実施態様に従うと、主要語又は修飾語のリストを必要とすることなしに、コンテンツから抽出される構成要素に従う粒度で、用語が分割される。

【図面の簡単な説明】

【００５２】

【図１】本発明の実施態様において使用されうるコンピュータの一例を示した図である。

【図２Ａ】本発明の実施態様に従い、用語リスト中の用語（英語である）を、コンテンツから抽出された粒度を規定する構成要素がある位置で分割する例を示す。

【図２Ｂ】本発明の実施態様に従い、用語リスト中の用語（漢字を含む）を、コンテンツから抽出された粒度を規定する構成要素がある位置で分割する例を示す。

【図３Ａ】本発明の実施態様に従い、粒度を規定する構成要素をコンテンツから抽出する処理の為のフローチャートを示す。

【図３Ｂ】本発明の実施態様に従い、用語を構成要素がある位置で分割する処理の為のフローチャートを示す。

【図４】図１に従うハードウェア構成を好ましくは備えており、図３Ａ及び図３Ｂそれぞれに示すフローチャートに従って本発明の実施態様を実施するコンピュータの機能ブロック図の一例を示す図である。

【発明を実施するための形態】

【００５３】

本発明の実施形態を、以下に図面に従って説明する。以下の図を通して、特に断らない限り、同一の符号は同一の対象を指す。本発明の実施形態は、本発明の好適な態様を説明するためのものであり、本発明の範囲をここで示すものに限定する意図はないことを理解されたい。

【００５４】

図１は、本発明の実施態様において使用されうるコンピュータの一例を示した図である。

【００５５】

本発明の実施態様に従うコンピュータは、１又は複数のコンピュータから構成されうる。

【００５６】

図１は、本発明の実施態様において使用されうるコンピュータを実現するためのハードウェア構成の一例を示した図である。

【００５７】

コンピュータ（１０１）は例えば、コンピュータ（例えば、デスクトップ・コンピュー

タ、ノートブック・コンピュータ、ウルトラブック・コンピュータ、サーバ・コンピュータ）でありうる。

【００５８】

コンピュータ（１０１）は、ＣＰＵ（１０２）とメイン・メモリ（１０３）とを備えており、これらはバス（１０４）に接続されている。ＣＰＵ（１０２）は好ましくは、３２ビット又は６４ビットのアーキテクチャに基づくものである。当該ＣＰＵ（１０２）は例えば、インテル社のＣｏｒｅ（商標　ｉ）シリーズ、Ｃｏｒｅ（商標）　２シリーズ、Ａｔｏｍ（商標）シリーズ、Ｘｅｏｎ（登録商標）シリーズ、Ｐｅｎｔｉｕｍ（登録商標）シリーズ若しくはＣｅｌｅｒｏｎ（登録商標）シリーズ、ＡＭＤ（Advanced Micro Devices）社のＡシリーズ、Ｐｈｅｎｏｍ（商標）シリーズ、Ａｔｈｌｏｎ（商標）シリーズ、Ｔｕｒｉｏｎ（商標）シリーズ若しくはＳｅｍｐｒｏｎ（商標）、又は、インターナショナル・ビジネス・マシーンズ・コーポレーションのＰｏｗｅｒ（商標）シリーズでありうる。

【００５９】

バス（１０４）には、ディスプレイ・コントローラ（１０５）を介して、ディスプレイ（１０６）、例えば液晶ディスプレイ（ＬＣＤ）が接続されうる。また、液晶ディスプレイ（ＬＣＤ）は例えば、タッチパネル・ディスプレイ又はフローティング・タッチ・ディスプレイであってもよい。ディスプレイ（１０６）は、コンピュータ（１０１）上で動作中のソフトウェア、例えば本発明の実施態様に従うコンピュータ・プログラムが稼働することによって表示される情報（例えば、用語リスト中の用語、コンテンツ、構成要素、又は分割された用語）を、適当なグラフィック・インタフェースで表示するために使用されうる。

【００６０】

バス（１０４）には任意的に、例えばＳＡＴＡ又はＩＤＥコントローラ（１０７）を介して、記憶装置（１０８）、例えばハードディスク又はソリッド・ステート・ドライブに接続されうる。

【００６１】

バス（１０４）には任意的に、例えばＳＡＴＡ又はＩＤＥコントローラ（１０７）を介して、記憶装置（１０８）、ドライブ（１０９）、例えばＣＤ、ＤＶＤ又はＢＤドライブが接続されうる。

【００６２】

バス（１０４）には、周辺装置コントローラ（１１０）を介して、例えばキーボード・マウス・コントローラ又はＵＳＢバスを介して、任意的に、キーボード（１１１）及びマウス（１１２）が接続されうる。

【００６３】

記憶装置（１０８）には、オペレーティング・システム、Ｗｉｎｄｏｗｓ（登録商標）ＯＳ、ＵＮＩＸ（登録商標）、Ｌｉｎｕｘ（登録商標）（例えば、ＲｅｄＨａｔ（登録商標）、Ｄｅｂｉａｎ（登録商標））、ＭａｃＯＳ（登録商標）、及びＪ２ＥＥなどのＪａｖａ（登録商標）処理環境、Ｊａｖａ（登録商標）アプリケーション、Ｊａｖａ（登録商標）仮想マシン（ＶＭ）、Ｊａｖａ（登録商標）実行時（ＪＩＴ）コンパイラを提供するプログラム、本発明の実施態様に従うコンピュータ・プログラム、及びその他のプログラム、並びにデータ（例えば、用語リスト、コンテンツ）が、メイン・メモリ（１０３）にロード可能なように記憶されうる。

【００６４】

記憶装置（１０８）は、コンピュータ（１０１）内に内蔵されていてもよく、当該コンピュータ（１０１）がアクセス可能なようにケーブル（例えば、ＵＳＢケーブル又はＬＡＮケーブル）を介して接続されていてもよく、又は、当該コンピュータ（１０１）がアクセス可能なように有線又は無線ネットワークを介して接続されていてもよい。

【００６５】

ドライブ（１０９）は、必要に応じて、例えばＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ又はＢＤ

－ＲＯＭからプログラム、例えばオペレーティング・システム又はアプリケーションを記憶装置（１０８）にインストールするために使用されうる。

【００６６】

通信インタフェース（１１４）は、例えばイーサネット（登録商標）・プロトコルに従う。通信インタフェース（１１４）は、通信コントローラ（１１３）を介してバス（１０４）に接続され、コンピュータ（１０１）を通信回線（１１５）に有線又は無線接続する役割を担い、コンピュータ（１０１）のオペレーティング・システムの通信機能のＴＣＰ／ＩＰ通信プロトコルに対して、ネットワーク・インタフェース層を提供する。通信回線は例えば、有線ＬＡＮ接続規格に基づく有線ＬＡＮ環境、又は無線ＬＡＮ接続規格に基づく無線ＬＡＮ環境、例えばＩＥＥＥ８０２．１１ａ／ｂ／ｇ／ｎなどのＷｉ－Ｆｉ無線ＬＡＮ環境、若しくは携帯電話網環境（例えば、３Ｇ、又は４Ｇ（ＬＴＥを含む）環境）でありうる。

【００６７】

コンピュータ（１０１）は、通信回線（１１５）を介して例えば他の装置（例えば、コンピュータ又はネットワーク・アタッチト・ストレージ）からのデータを受信し、記憶装置（１０８）上に格納しうる。

【００６８】

図２Ａ及び図２Ｂはそれぞれ、本発明の実施態様に従い、粒度を規定する構成要素をコンテンツから抽出し、そして用語リスト中の用語を、当該抽出された構成要素がある位置で分割する例を示す。

【００６９】

図２Ａは、コンテンツ及び用語リストが英語である場合の例を示す。

【００７０】

ユーザは、用語を分割する為の粒度を規定する構成要素を抽出する為のコンテンツ（２０１）及び、分割対象の用語を含む用語リスト（２０２）を用意する。コンテンツ（２０１）及び用語リスト（２０２）それぞれの内容は、図２Ａに記載の通りである。コンテンツ（２０１）はコンピュータ分野の業務手順書であり、及び用語リスト（２０２）もまたコンピュータ分野の用語リストであるとする。

【００７１】

コンピュータ（１０１）は、大別して、粒度を規定する構成要素をコンテンツから抽出する工程、及び、用語リスト中の用語を、当該抽出された構成要素がある位置で分割する工程を含む。

【００７２】

（粒度を規定する構成要素をコンテンツから抽出する工程）

【００７３】

コンピュータ（１０１）は、コンテンツ（２０１）及び用語リスト（２０２）を入力として受け取り、例えばメモリ（１０３）又は記憶装置（１０８）に格納する。

【００７４】

コンピュータ（１０１）は、コンテンツ（２０１）から、構成要素を抽出する対象のテキストを切り出す。例えば、コンピュータ（１０１）は、コンテンツ（２０１）から、例えば変更履歴やコメント又は注釈を削除して、例えば本文のテキストを切り出しうる。

【００７５】

コンピュータ（１０１）は、上記切り出したテキストを、事前定義した文字がある場所（すなわち、事前定義した文字がある場所の前後）で分割する。事前定義した文字は例えば、広義の句読点でありうる。広義の句読点は例えば、狭義の句読点（句点、読点）、疑問符、感嘆符、省略符、括弧（例えば、丸括弧、鉤括弧、角括弧、波括弧、亀甲括弧、山括弧、若しくは、隅付き括弧）、又は、その他文章に使う様々な記号を含みうる。コンピュータ（１０１）は例えば、コンテンツ（２０１）中の「- AAA」、「- BBB」及び「- ZUR」の記号 - の前後でテキストを分割する。

【００７６】

コンピュータ（１０１）は、上記事前定義した文字で分割したテキストそれぞれに、当業者に知られている任意の構文解析技術を適用して、文節を抽出する。

【００７７】

コンピュータ（１０１）は、上記抽出した文節のうちの名詞又は記号を含む文節から構成要素となりうる部分（すなわち、構成要素の候補）（２０３）を抽出する。名詞は、所謂文法上の名詞に分類される文字でありうる。記号は、自然言語処理において、辞書中に存在しない単語である未知語や省略語を含みうる。構成要素は、少なくとも１つの名詞又は記号を含む１又は複数の単語列でありうる。構成要素の候補（２０３）の内容は、図２Ａに記載の通りである。構成要素の候補（２０３）中、「AAA」、「BBB」、「PPP」、「QQQ」及び「RRR」はいずれも、名詞又は記号を含む文節である。また、構成要素の候補（２０３）中、「ZUR」は固有名詞であり、「EOF」は「End-Of-File」の省略形である。

【００７８】

コンピュータ（１０１）は、構成要素の候補（２０３）が、用語リスト（２０２）中にある用語を含むかを判断する。当該用語は、名詞、記号又はそれらの組み合わせを含む単語列でありうる。また、当該用語は例えば、複合名詞でありうる。コンピュータ（１０１）は、構成要素の候補（２０３）が用語リスト（２０２）中にある用語「ZUR EOF mark」を含む為に、構成要素の候補（２０３）から構成要素「ZUR EOF mark」を削除し、当該構成要素「ZUR EOF mark」を削除した残りを、粒度を規定する構成要素（２０４）とする。

【００７９】

（用語リスト中の用語を、当該抽出された構成要素がある位置で分割する工程）

【００８０】

コンピュータ（１０１）は、用語リスト（２０２）から一つの用語を取り出し、当該取り出した用語が、当該用語の末尾から最長一致する構成要素（２０４）を含むかを判断する。コンピュータ（１０１）は、当該取り出した用語が、当該用語の末尾から最長一致する構成要素（２０４）を含むことに応じて、当該用語をその末尾から最長一致する構成要素がある位置で分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語を、分割した後の用語を入れるリストＬ（２０５）に格納する。

【００８１】

コンピュータ（１０１）は、用語リスト（２０２）中の「Beneficiary right seller's business security deposit」を取り出し、当該「Beneficiary right seller's business security deposit」が、その末尾から最長一致する構成要素（２０４）を含むか判断する。コンピュータ（１０１）は、当該「Beneficiary right seller's business security deposit」が、その末尾から最長一致する構成要素「business security deposit」を含んでいることに応じて、「Beneficiary right seller's business security deposit」をその末尾から最長一致する構成要素「business security deposit」がある位置（すなわち、「business security deposit」の直前）で分割し、すなわち、「Beneficiary right seller's」と「business security deposit」とに分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語「business security deposit」を、上記リストＬ（２０５）に格納する。従って、「Beneficiary right seller's business security deposit」は、コンテンツ（２０１）中の構成要素「business security deposit」の粒度に従って分割されている。

【００８２】

同様に、コンピュータ（１０１）は、用語リスト（２０２）中の「Financial instruments intermediary service」を取り出し、当該「Financial instruments intermediary service」が、その末尾から最長一致する構成要素（２０４）を含むか判断する。コンピュータ（１０１）は、当該「Financial instruments intermediary service」が、その末尾から最長一致する構成要素「intermediary service」を含んでいることに応じて、「Financial instruments intermediary service」をその末尾から最長一致する構成要素「intermediary service」がある位置（すなわち、「intermediary service」の直前）で分割し

、すなわち、「Financial instruments」と「intermediary service」とに分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語「intermediary service」を、上記リストＬ（２０５）に格納する。従って、「Financial instruments intermediary service」は、コンテンツ（２０１）中の構成要素「intermediary service」の粒度に従って分割されている。

【００８３】

同様に、コンピュータ（１０１）は、用語リスト（２０２）中の「ZUR EOF mark」を取り出し、当該「ZUR EOF mark」が、その末尾から最長一致する構成要素（２０４）を含むか判断する。コンピュータ（１０１）は、当該「ZUR EOF mark」が、その末尾から最長一致する構成要素「mark」を含んでいることに応じて、「ZUR EOF mark」をその末尾から最長一致する構成要素「mark」がある位置（すなわち、「mark」の直前）で分割し、すなわち、「ZUR EOF」と「mark」とに分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語「mark」を、上記リストＬ（２０５）に格納する。従って、「ZUR EOF mark」は、コンテンツ（２０１）中の構成要素「mark」の粒度に従って分割されている。

【００８４】

次に、コンピュータ（１０１）は、上記取り出した用語から上記末尾から最長一致する構成要素を分割した用語を除いた後の用語が、当該除いた後の用語の先頭から最長一致する構成要素（２０４）を含むかを判断する。コンピュータ（１０１）は、上記取り出した用語から上記末尾から最長一致する構成要素を分割した用語を除いた後の用語が当該除いた後の用語の先頭から最長一致する構成要素を含むことに応じて、当該除いた後の用語をその先頭から最長一致する構成要素がある位置で分割する。そして、コンピュータ（１０１）は、上記先頭から最長一致する構成要素を分割した用語を、上記リストＬ（２０５）に格納する。

【００８５】

コンピュータ（１０１）は、用語リスト（２０２）中の「Beneficiary right seller's business security deposit」中の「business security deposit」を除いた用語「Beneficiary right seller's」がその先頭から最長一致する構成要素（２０４）を含むか判断する。コンピュータ（１０１）は、当該「business security deposit」がその先頭から最長一致する構成要素（２０４）を含まないことに応じて、当該「business security deposit」を、上記リストＬ（２０５）に格納して、分割処理を終了する。

【００８６】

同様に、コンピュータ（１０１）は、用語リスト（２０２）中の「Financial instruments intermediary service」中の「intermediary service」を除いた用語「Financial instruments」がその先頭から最長一致する構成要素（２０４）を含むか判断する。コンピュータ（１０１）は、当該「Financial instruments」がその先頭から最長一致する構成要素（２０４）を含まないことに応じて、当該「Financial instruments」をそのまま、上記リストＬ（２０５）に格納して、分割処理を終了する。

【００８７】

同様に、コンピュータ（１０１）は、用語リスト（２０２）中の「ZUR EOF mark」中の「mark」を除いた用語「ZUR EOF」がその先頭から最長一致する構成要素（２０４）を含むか判断する。当該除いた用語「ZUR EOF」はその先頭から最長一致する構成要素「ZUR EOF」を含んでいる。しかしながら、両者は同一であるから分割出来ない。従って、コンピュータ（１０１）は、当該「ZUR EOF」をそのまま、上記リストＬ（２０５）に格納して、分割処理を終了する。

【００８８】

従って、上記分割処理後のリストＬ（２０５）は、「Beneficiary right seller's」、「business security deposit」、「Financial instruments」、「intermediary service」、「ZUR EOF」、及び「mark」を含む。

【００８９】

また、コンピュータ（１０１）は、上記分割処理後のリストＬ（２０５）を、図２Ａに示すように、末尾から最長一致する最初の構成要素を主要語として、及び、先頭から最長一致する最初の構成要素を修飾語１として、並びに、次に、末尾から最長一致する構成要素を修飾語２（ある場合）として、表示装置（１０６）上に表示しうる。

【００９０】

代替的には、コンピュータ（１０１）は、上記分割処理後のリストＬ（２０５）を、下記のように、分割した箇所を示す記号、例えば、　|　を入れて、表示装置（１０６）上に表示しうる。例えば、以下の通りである。

Beneficiary right seller's | business security deposit
Financial instruments | intermediary service
ZUR EOF | mark

【００９１】

用語リスト（２０２）中の用語「Beneficiary right seller's business security deposit」、「Financial instruments intermediary service」、及び「ZUR EOF mark」それぞれは、従来技術に従う形態素解析器に従うと、単語は通常空白によってわかち書きされている為に、単語ごとに分割される。一方、本願発明の態様に従うと、上記分割処理後のリストＬ（２０５）に示すように、コンテンツ（２０１）から抽出された構成要素（２０４）の粒度に従い用語が分割される。

【００９２】

図２Ｂは、コンテンツ及び用語リストが日本語（漢字を含む）である場合の例を示す。

【００９３】

ユーザは、用語を分割する為の粒度を規定する構成要素を抽出する為のコンテンツ（２１１）及び、分割対象の用語を含む用語リスト（２１２）を用意する。コンテンツ（２１１）及び用語リスト（２１２）それぞれの内容は、図２Ｂに記載の通りである。コンテンツ（２１１）は金融分野の業務手順書であり、及び用語リスト（２１２）もまた金融分野の用語リストであるとする。

【００９４】

コンピュータ（１０１）は、大別して、粒度を規定する構成要素をコンテンツから抽出する工程、及び、用語リスト中の用語を、当該抽出された構成要素がある位置で分割する工程を含む。

【００９５】

（粒度を規定する構成要素をコンテンツから抽出する工程）

【００９６】

コンピュータ（１０１）は、コンテンツ（２１１）及び用語リスト（２１２）を入力として受け取り、例えばメモリ（１０３）又は記憶装置（１０８）に格納する。

【００９７】

コンピュータ（１０１）は、コンテンツ（２１１）から、構成要素を抽出する対象のテキストを切り出す。当該テキストの切り出しとは、上記において述べた通りである。

【００９８】

コンピュータ（１０１）は、上記切り出したテキストを、事前定義した文字がある場所（すなわち、事前定義した文字がある場所の前後）で分割する。事前定義した文字とは、上記において述べた通りである。コンピュータ（１０１）は例えば、コンテンツ（２１１）中の「（金信）」の括弧書き　（　の前後、及び括弧書き　）の前後でテキストを分割する。

【００９９】

コンピュータ（１０１）は、上記事前定義した文字で分割したテキストそれぞれに、当業者に知られている任意の構文解析技術を適用して、文節を抽出する。

【０１００】

コンピュータ（１０１）は、上記抽出した文節のうちの名詞又は記号を含む文節から構成要素となりうる部分（すなわち、構成要素の候補）（２１３）を抽出する。構成要素の

候補（２１３）の内容は、図２Ｂに記載の通りである。
【０１０１】
コンピュータ（１０１）は、構成要素の候補（２１３）が、用語リスト（２１２）中にある用語を含むかを判断する。コンピュータ（１０１）は、構成要素の候補（２１３）が用語リスト（２１２）中にある用語「金信期日後収益金税額」、「延滞元金額」及び「補正計算元本額」を含む為に、構成要素の候補（２１３）から構成要素「金信期日後収益金税額」、「延滞元金額」及び「補正計算元本額」を削除し、これらの構成要素を削除した残りを、粒度を規定する構成要素（２１４）とする。
【０１０２】
（用語リスト中の用語を、当該抽出された構成要素がある位置で分割する工程）
【０１０３】
コンピュータ（１０１）は、用語リスト（２１２）から一つの用語を取り出し、当該取り出した用語が、当該用語の末尾から最長一致する構成要素（２１４）を含むかを判断する。コンピュータ（１０１）は、当該取り出した用語が、当該用語の末尾から最長一致する構成要素（２１４）を含むことに応じて、当該用語をその末尾から最長一致する構成要素がある位置で分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語を、分割した後の用語を入れるリストＬ（２１５）に格納する。
【０１０４】
コンピュータ（１０１）は、用語リスト（２１２）中の「金信期日後収益金税額」を取り出し、当該「金信期日後収益金税額」が、その末尾から最長一致する構成要素（２１４）を含むか判断する。コンピュータ（１０１）は、当該「金信期日後収益金税額」が、その末尾から最長一致する構成要素「税額」を含んでいることに応じて、「金信期日後収益金税額」をその末尾から最長一致する構成要素「税額」がある位置（すなわち、「税額」の直前）で分割し、すなわち、「金信期日後収益金」と「税額」とに分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語「税額」を、上記リストＬ（２１５）に格納する。従って、「金信期日後収益金税額」は、コンテンツ（２１１）中の構成要素「税額」の粒度に従って分割されている。
【０１０５】
同様に、コンピュータ（１０１）は、用語リスト（２１２）中の「延滞元金額」を取り出し、当該「延滞元金額」が、その末尾から最長一致する構成要素（２１４）を含むか判断する。コンピュータ（１０１）は、当該「延滞元金額」が、その末尾から最長一致する構成要素「元金額」を含んでいることに応じて、「延滞元金額」をその末尾から最長一致する構成要素「元金額」がある位置（すなわち、「元金額」の直前）で分割し、すなわち、「延滞」と「元金額」とに分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語「元金額」を、上記リストＬ（２１５）に格納する。従って、「延滞元金額」は、コンテンツ（２１１）中の構成要素「元金額」の粒度に従って分割されている。
【０１０６】
同様に、コンピュータ（１０１）は、用語リスト（２１２）中の「補正計算元本額」を取り出し、当該「補正計算元本額」が、その末尾から最長一致する構成要素（２１４）を含むか判断する。コンピュータ（１０１）は、当該「補正計算元本額」が、その末尾から最長一致する構成要素「元本額」を含んでいることに応じて、「補正計算元本額」をその末尾から最長一致する構成要素「元本額」がある位置（すなわち、「元本額」の直前）で分割し、すなわち、「補正計算」と「元本額」とに分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語「元本額」を、上記リストＬ（２１５）に格納する。従って、「補正計算元本額」は、コンテンツ（２１１）中の構成要素「元本額」の粒度に従って分割されている。
【０１０７】
次に、コンピュータ（１０１）は、上記取り出した用語から上記末尾から最長一致する

構成要素を分割した用語を除いた後の用語が、当該除いた後の用語の先頭から最長一致する構成要素（２１４）を含むかを判断する。コンピュータ（１０１）は、上記取り出した用語から上記末尾から最長一致する構成要素を分割した用語を除いた後の用語が当該除いた後の用語の先頭から最長一致する上記構成要素を含むことに応じて、当該除いた後の用語をその先頭から最長一致する構成要素がある位置で分割する。そして、コンピュータ（１０１）は、上記先頭から最長一致する構成要素を分割した用語を、上記リストＬ（２１５）に格納する。

【０１０８】

コンピュータ（１０１）は、用語リスト（２１２）中の「金信期日後収益金税額」中の「税額」を除いた用語「金信期日後収益金」の先頭から最長一致する構成要素（２１４）を含むか判断する。コンピュータ（１０１）は、当該「金信期日後収益金」がその先頭から最長一致する構成要素「金信」を含んでいることに応じて、「金信期日後収益金」をその先頭から最長一致する構成要素「金信」がある位置（すなわち、「金信」の直後）で分割し、すなわち、「金信」と「期日後収益金」とに分割する。そして、コンピュータ（１０１）は、上記先頭から最長一致する構成要素を分割した用語「金信」を、上記リストＬ（２１５）に格納する。従って、「金信期日後収益金」は、コンテンツ（２１１）中の構成要素「金信」の粒度に従って分割されている。

【０１０９】

同様に、コンピュータ（１０１）は、用語リスト（２１２）中の「延滞元金額」中の「元金額」を除いた用語「延滞」の先頭から最長一致する構成要素（２１４）を含むか判断する。コンピュータ（１０１）は、当該「延滞」がその先頭から最長一致する構成要素（２１４）を含まないことに応じて、当該「延滞」をそのまま、上記リストＬ（２１５）に格納して、分割処理を終了する。

【０１１０】

同様に、コンピュータ（１０１）は、用語リスト（２１２）中の「補正計算元本額」中の「元本額」を除いた用語「補正計算」の先頭から最長一致する構成要素（２１４）を含むか判断する。コンピュータ（１０１）は、当該「補正計算」がその先頭から最長一致する構成要素「補正計算」を含むが、両者は同一である為に分割できない。そこで、コンピュータ（１０１）は、当該「補正計算」をそのまま、上記リストＬ（２１５）に格納して、分割処理を終了する。

【０１１１】

次に、コンピュータ（１０１）は、上記「金信期日後収益金税額」のうち、残った用語「期日後収益金」が、当該用語の末尾から最長一致する構成要素（２１４）を含むかを判断する。当該残った用語「期日後収益金」はその先頭から最長一致する構成要素「期日後収益金」を含んでいる。しかしながら、両者は同一であるから分割出来ない。従って、コンピュータ（１０１）は、当該「期日後収益金」をそのまま、上記リストＬ（２１５）に格納して、分割処理を終了する。

【０１１２】

従って、上記分割処理後のリストＬ（２１５）は、「金信」、「期日後収益金」、「税額」、「延滞」、「元金額」、「補正計算」、及び「元本額」を含む。

【０１１３】

また、コンピュータ（１０１）は、上記分割処理後のリストＬ（２１５）を、図２Ｂに示すように、末尾から最長一致する最初の構成要素を主要語として、及び、先頭から最長一致する最初の構成要素を修飾語１として、並びに、次に、末尾から最長一致する構成要素を修飾語２（ある場合）として、表示装置（１０６）上に表示しうる。

【０１１４】

代替的には、コンピュータ（１０１）は、上記分割処理後のリストＬ（２１５）を、下記のように、分割した箇所を示す記号、例えば、 ｜ を入れて、表示装置（１０６）上に表示しうる。例えば、以下の通りである。

金信 ｜ 期日後収益金 ｜税額

延滞 | 元金額
補正計算| 元本額

【０１１５】

用語リスト（２１２）中の用語「金信期日後収益金税額」、「延滞元金額」、及び「補正計算元本額」それぞれは、従来技術に従う形態素解析器に従うと、例えば「金　信　期　日　後　収益金　税額」、「延滞　元金　額」、及び「補正　計算　元本　額」というように、従来技術に従う形態素解析器が持つシステム辞書及び文法に基づいて複合名詞が分割される。従って、必ずしも、望ましい箇所で用語が分割されていない。一方、本願発明の態様に従うと、上記分割処理後のリストＬ（２１５）に示すように、コンテンツ（２１１）から抽出された構成要素（２１４）の粒度に従い用語が分割される。

【０１１６】

図２Ａ及び図２Ｂそれぞれに示す本願発明の実施態様に従い、コンテンツから抽出された構成要素の粒度に従い用語を分割することによって得られた語は、下記（１）～（３）それぞれに示す場面において利用されうる。

【０１１７】

（１）主要語又は例えば、修飾語１若しくは修飾語２で、分割された語をソートする。当該語をソートをすることによって、ユーザは、例えば、統一すべき類似の又は同義の語を発見したり、追加すべき語を洗い出したり、又は短い主要語のみからなる語を発見したりすることが用意になる。

【０１１８】

（２）主要語の抽象度が高い場合、ユーザは、当該主要語を複数の具体的な用語に置き換えることが容易になる。抽象度が高い主要語は例えば、「実行金額」である。

【０１１９】

（３）ユーザは、主要語が非常に長い用語を発見することが容易になる。このことは、ユーザが、当該用語が仕様書に説明がない用語だと判定することを容易にしたり、又は、関係者間で共通の理解を持つ為に、当該用語の説明の為の記述を仕様書に追加することをしたりすることを可能にする。例えば、上記分割の結果、コンテンツ（２１１）から抽出された構成要素（２１４）中に該当する構成要素がなく、「分割可能上限金額」及び「利息算出対象元金」が主要語として得られたとする。このような場合には、コンピュータ（１０１）は、当該長い主要語を持つ用語「分配可能上限金額」及び「利息算出対象元金」として、抽出する。

【０１２０】

図３Ａ及び図３Ｂは、本発明の実施態様に従い、粒度を規定する構成要素をコンテンツから抽出し、そして用語リスト中の用語を、当該抽出された構成要素がある位置で分割する処理の為のフローチャートを示す。

【０１２１】

図３Ａは、本発明の実施態様に従い、粒度を規定する構成要素をコンテンツから抽出する処理の為のフローチャートを示す。

【０１２２】

ステップ３０１において、コンピュータ（１０１）は、コンテンツから上記構成要素を抽出する処理を開始する。

【０１２３】

ステップ３０２において、コンピュータ（１０１）は、粒度を規定する構成要素をそこから抽出する為のコンテンツを、コンテンツを記録した記録媒体（３３１）から読み取る。コンテンツは例えば、索引を作成することが必要な文書、例えばビジネス文書（例えば、仕様書、業務手順書、ビジネスツール定義書）でありうるがこれらに限定されるものでない。コンテンツが例えば、仕様書である場合には、例えばデータ項目やビジネス・プロセスが説明とともに記載されている。コンテンツが属する技術分野や適用分野が異なると、それから抽出される構成要素も異なってくる。すなわち、コンテンツが変われば、用語を分割する為の粒度も変わってくる。

【０１２４】
引き続き、ステップ３０２において、コンピュータ（１０１）は、任意的に、当該読み取ったコンテンツから、上記構成要素を抽出する対象のテキストを切り出しうる。例えば、コンピュータ（１０１）は、コンテンツ（３３１）から、例えば変更履歴やコメント又は注釈を削除して、例えば本文のテキストを切り出しうる。
【０１２５】
ステップ３０３において、コンピュータ（１０１）は、ステップ３０２で切り出したテキストを、事前定義した文字がある場所（すなわち、事前定義した文字がある場所の前後）で分割する。事前定義した文字は例えば、広義の句読点でありうる。広義の句読点は例えば、狭義の句読点（句点、読点）、疑問符、感嘆符、省略符、括弧（例えば、丸括弧、鉤括弧、角括弧、波括弧、亀甲括弧、山括弧、若しくは、隅付き括弧）、又は、その他文章に使う様々な記号を含みうる。例えば、テキストが、「Confirm the item number (check whether the order is valid) in this process.」（英語である）である場合には、コンピュータ（１０１）は、「Confirm the item number| (| check whether the order is valid |) | in this process」（|　は、分割する位置を示す）と、丸括弧の前後で分割する。同様に、例えば、テキストが、「品物番号を確認（オーダーと同じか）する」（日本語である）である場合には、コンピュータ（１０１）は、「品物番号を確認|（|オーダーと同じか|）|する」（日本語である）（|　は、分割する位置を示す）と、丸括弧の前後で分割する。
【０１２６】
ステップ３０４において、コンピュータ（１０１）は、ステップ３０３で分割したテキストそれぞれに、当業者に知られている任意の構文解析技術を適用して、文節を抽出する。例えば、テキストが、「In the calculation / of business security deposit / , PPP / , QQQ / and RRR / are used.」（英語である）である場合には、コンピュータ（１０１）は、「In the calculation / of business security deposit / , PPP / , QQQ / and RRR / are used.」（/　は、文節を抽出する区切りを示す）として文節を抽出する。同様に、例えば、テキストが、「中途解約受取金額は預金期間から決まる」（日本語である）である場合には、コンピュータ（１０１）は、「中途解約受取金額は／預金期間から／決まる」（日本語である）（／　は、文節を抽出する区切りを示す）として文節を抽出する。
【０１２７】
ステップ３０５において、コンピュータ（１０１）は、ステップ３０４で抽出した文節のうち名詞又は記号を含む文節から構成要素となりうる部分を抽出する。コンピュータ（１０１）は、当該抽出した部分を、構成要素の候補としてリストしうる。名詞は、所謂文法上の名詞に分類される文字でありうる。記号は、自然言語処理において、辞書中に存在しない単語である未知語や省略語を含みうる。構成要素は、少なくとも１つの名詞又は記号を含む１又は複数の単語列でありうる。構成要素となりうる部分を抽出することは例えば、英語の場合、冠詞を除くことを含む。また、構成要素となりうる部分を抽出することは、当該抽出された部分について、複数形を単数形に変換したり、大文字を小文字に変形したり、旧字体を新字体に変形したりするような変形処理をすることを含みうる。
【０１２８】
ステップ３０６において、コンピュータ（１０１）は、任意的に、ステップ３０５で抽出した部分に、当該部分の直前の文字列を補完するかどうかを判断する。例えば、文字列が専門用語の場合には、任意の構文解析技術を使用すると、一つの単語が複数に分割される場合がある。そこで、ステップ３０５で抽出した部分に、当該部分の直前の文字列を補完することによって、コンピュータ（１０１）は、複数に分割された一つの用語を、本来の一つの用語になるように直前の文字列で補完しうる。例えば、上記抽出した部分が「File:Open / Menu」（英語である）（/　は、区切りを示す）である場合、「File:Open」はスペースがなく且つ記号「:」が挿入されているために、当該「File:Open」は未知語として検出され、及び、「Menu」は名詞として別々に検出される。従って、上記抽出した部分

「File:Open / Menu」は、「Menu」の直前に「File:Open」を補って、「File:Open Menu」と本来の一つのまとまりのある用語にする。例えば、上記抽出した部分が、「採／番」（漢字である）（／　は、区切りを示す）である場合には、「番」の直前に「採」を補って、「採番」（漢字である；「採番」とは、データ管理のために、それぞれのデータに固有の番号を与えることを意味する）と本来の一つの用語にする。コンピュータ（１０１）は、上記直前の文字列を補完することに応じて、処理をステップ３０７に進める。一方、コンピュータ（１０１）は、上記直前の文字列を補完しないことに応じて、処理をステップ３０８に進める。

【０１２９】

ステップ３０７において、コンピュータ（１０１）は、ステップ３０５で抽出した部分に、当該部分の直前の文字列を補完する。そして、コンピュータ（１０１）は、当該補完した文字列で、構成要素の候補の上記リストを更新しうる。引き続き、コンピュータ（１０１）は、ステップ３０６に戻り、さらに補完する必要があるかどうか判断しうる。

【０１３０】

ステップ３０８において、コンピュータ（１０１）は、用語を例えばメモリ（１０３）中に読み取る。用語は、名詞、記号（未知語や省略語を含む）又はそれらの組み合わせを含む単語列でありうる。当該用語の読み取りは、例えば、用語を格納した用語リストを、当該用語リストを記録した記録媒体（３３２）から読み取ることによって行われうる。また、代替的には、コンピュータ（１０１）は、ユーザによって指定された用語を入力として読み取りうる。さらに、代替的には、コンピュータ（１０１）は、コンテンツ（３３１）中の所定の長さよりも長い（例えば、構成要素の平均文字長よりも長い、又は例えば、１０文字よりも長い）用語を、分割対象の用語として読み取りうる。引き続き、コンピュータ（１０１）は、当該用語リスト中に、ステップ３０５で抽出した、構成要素となりうる部分と同じ用語があるかを判断する。用語リスト中に上記構成要素となりうる部分と同じ用語がある場合には、当該構成要素となりうる部分は、用語リスト中の用語を分割する為の構成要素となり得ないからである。コンピュータ（１０１）は、用語リスト中に上記構成要素となりうる部分と同じ用語がある場合には、処理をステップ３０９に進める。一方、コンピュータ（１０１）は、用語リスト中に上記構成要素となりうる部分と同じ用語がない場合には、処理をステップ３１０に進める。

【０１３１】

ステップ３０９において、コンピュータ（１０１）は、用語リスト中に上記構成要素となりうる部分と同じ用語がある場合には、当該同じ用語を、上記構成要素の候補の上記リストから削除する。用語リストは例えば、データ項目やビジネス・プロセスが記載されている。用語リストは例えば、Kwds＝｛k_1, k_2, k_3, ・・・, k_n｝で表されうる。

【０１３２】

ステップ３１０において、コンピュータ（１０１）は、上記構成要素の候補の上記リストを、コンテンツから粒度を規定する構成要素として、例えば構成要素を格納する記録媒体（３３３）に格納する。

【０１３３】

ステップ３１１において、コンピュータ（１０１）は、コンテンツから上記構成要素を抽出する処理を終了する。

【０１３４】

図３Ｂは、本発明の実施態様に従い、用語を構成要素がある位置で分割する処理の為のフローチャートを示す。

【０１３５】

ステップ３２１において、コンピュータ（１０１）は、用語を、ステップ３１０で作成した構成要素がある位置で分割する処理を開始する。

【０１３６】

ステップ３２２において、コンピュータ（１０１）は、用語を例えば用語リスト（３３２）から一つ取り出す。そして、コンピュータ（１０１）は、当該取り出した一つの用語

が、当該用語の末尾から最長一致する構成要素（記憶媒体（３３３）に格納されている）を含むかを判断する。コンピュータ（１０１）は、当該取り出した用語が、当該用語の末尾から最長一致する構成要素（２０４）を含むことに応じて、当該用語をその末尾から最長一致する構成要素がある位置で分割する。そして、コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語を、分割した後の用語を入れるリストＬ（３３４）に格納する。コンピュータ（１０１）は、上記末尾から最長一致する構成要素を分割した用語を、主要語としてリストＬ（３３４）に格納しうる。

【０１３７】

ステップ３２３において、コンピュータ（１０１）は、ステップ３２２での上記分割ができたことに応じて、用語の分割回数ｄを１つ増加し（ｄ＋＋）、処理をステップ３２４に進める。一方、コンピュータ（１０１）は、ステップ３２２での上記分割ができなかったことに応じて、処理をステップ３２８に進める。

【０１３８】

ステップ３２４において、コンピュータ（１０１）は、分割回数ｄと用語の分割回数を規定する分割パラメータＱとを比較する。コンピュータ（１０１）は、分割回数ｄが分割パラメータＱよりも少ないことに応じて、さらに分割処理をする為に、処理をステップ３２５に進める。一方、コンピュータ（１０１）は、分割回数ｄが分割パラメータＱ以上であることに応じて、これ以上分割処理をしない為に、処理をステップ３２８に進める。

【０１３９】

ステップ３２５において、コンピュータ（１０１）は、ステップ３２２で取り出した用語から上記末尾から最長一致する構成要素を分割した用語を除いた後の用語が、当該除いた後の用語の先頭から最長一致する構成要素を含むかを判断する。コンピュータ（１０１）は、上記取り出した用語から上記末尾から最長一致する構成要素を分割した用語を除いた後の用語が当該除いた後の用語の先頭から最長一致する構成要素を含むことに応じて、当該除いた後の用語をその先頭から最長一致する構成要素がある位置で分割する。そして、コンピュータ（１０１）は、上記先頭から最長一致する構成要素を分割した用語を、分割した後の用語を入れるリストＬ（３３４）に格納する。コンピュータ（１０１）は、上記先頭から最長一致する構成要素を分割した用語を、修飾語１としてリストＬ（３３４）に格納しうる。

【０１４０】

ステップ３２６において、コンピュータ（１０１）は、ステップ３２５での上記分割ができたことに応じて、用語の分割回数ｄを１つ増加し（ｄ＋＋）、処理をステップ３２７に進める。一方、コンピュータ（１０１）は、ステップ３２５での上記分割ができなかったことに応じて、処理をステップ３２８に進める。

【０１４１】

ステップ３２７において、コンピュータ（１０１）は、分割回数ｄと用語の分割回数を規定する分割パラメータＱとを比較する。コンピュータ（１０１）は、分割回数ｄが分割パラメータＱよりも少ないことに応じて、さらに分割処理をする為に、処理をステップ３２２に戻す。コンピュータ（１０１）は、引き続き、末尾から最長一致する構成要素があるか、そして、先頭か最長一致する構成要素があるかを繰り返して行いうる。一方、コンピュータ（１０１）は、分割回数ｄが分割パラメータＱ以上であることに応じて、これ以上分割処理をしない為に、処理をステップ３２８に進める。

【０１４２】

ステップ３２８において、コンピュータ（１０１）は、分割した語と、分割した語の残りがある場合には当該残りとを分割後の語を格納する用語リストＬ（３３４）に格納する。

【０１４３】

ステップ３２９において、コンピュータ（１０１）は、任意的に、用語リストＬの内容（上記分割した語と、分割した語の残りがある場合には当該残りを分割後の語）を、例えば表示装置（１０６）上に表示しうる。

【０１４４】
ステップ３３０において、コンピュータ（１０１）は、用語を、上記構成要素がある位置で分割する処理を終了する。
【０１４５】
図４は、図１に従うハードウェア構成を好ましくは備えており、図３Ａ及び図３Ｂそれぞれに示すフローチャートに従って本発明の実施態様を実施するコンピュータの機能ブロック図の一例を示す図である。
【０１４６】
コンピュータ（４０１）は、図１Ａに示すコンピュータ（１０１）に示されている構成、例えばＣＰＵ（１０２）、メイン・メモリ（１０３）、記憶装置（１０８）、及びディスク（１０８）を備えている。
【０１４７】
コンピュータ（４０１）は、抽出手段（４１１）、分割手段（４１２）、及び表示手段（４１３）を備えている。
【０１４８】
抽出手段（４１１）は、構文解析により、粒度を規定する構成要素をコンテンツから抽出する。
【０１４９】
また、抽出手段（４１１）は、上記コンテンツ中のテキストそれぞれに上記構文解析を適用して、文節を抽出し、上記抽出した文節のうちの名詞又は記号を含む文節から上記構成要素となりうる部分を抽出する。
【０１５０】
また、抽出手段（４１１）は、上記コンテンツから、上記構成要素を抽出する対象のテキストを切り出し、当該切り出したテキストそれぞれに上記構文解析を適用して、上記文節を抽出する。
【０１５１】
また、抽出手段（４１１）は、上記切り出したテキストを事前定義した文字がある場所で分割し、当該分割したテキストそれぞれに上記構文解析を適用して、上記文節を抽出する。
【０１５２】
また、抽出手段（４１１）は、上記用語が用語リスト中の用語である場合に、上記構成要素となりうる部分のうちから上記用語リスト中にある用語を削除し、当該削除した残りを上記構成要素とする。
【０１５３】
抽出手段（４１１）は、図３Ａに記載の各ステップを実行しうる。
【０１５４】
分割手段（４１２）は、上記用語がその一部に少なくとも１つの上記構成要素を含む場合に、上記用語を上記構成要素がある位置で分割する。
【０１５５】
また、分割手段（４１２）は、上記用語が当該用語の末尾から最長一致する上記構成要素（第１の構成要素）を含む場合に、上記用語を上記末尾から最長一致する上記構成要素（第１の構成要素）がある位置で分割する。
【０１５６】
また、分割手段（４１２）は、上記用語を上記末尾から最長一致する上記構成要素（第１の構成要素）がある位置で分割し、上記末尾から最長一致する上記構成要素（第１の構成要素）を上記用語の主要語として保存する。
【０１５７】
また、分割手段（４１２）は、上記用語から上記末尾から最長一致する上記構成要素（第１の構成要素）を除いた後の用語が当該除いた後の用語の先頭から最長一致する上記構成要素（第２の構成要素）を含む場合に、上記除いた後の用語を上記先頭から最長一致す

る上記構成要素（第２の構成要素）がある位置で分割する。

【０１５８】

また、分割手段（４１２）は、上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割し、上記先頭から最長一致する上記構成要素（第２の構成要素）を上記用語の第１の修飾語として保存する。

【０１５９】

また、分割手段（４１２）は、上記除いた後の用語を上記先頭から最長一致する上記構成要素（第２の構成要素）がある位置で分割し、上記先頭から最長一致する上記構成要素（第２の構成要素）以外の部分を第２の修飾語として保存する。

【０１６０】

また、分割手段（４１２）は、予め設定された分割回数を規定する分割パラメータに従って、上記用語を上記構成要素がある位置で分割する。

【０１６１】

分割手段（４１２）は、図３Ｂに記載のステップ３２２～３２８を実行しうる。

【０１６２】

表示手段（４１３）は、用語リストＬの内容を、例えば表示装置（１０６）上に表示する。

【０１６３】

表示手段（４１３）は、図３Ｂに記載のステップ３２９を実行しうる。

【０１６４】

本発明の実施態様に従うと、上記した通り、コンテンツから抽出される構成要素に従う粒度で、用語を分割することが可能である。従って、コンテンツが属する技術分野や適用分野が異なると、それから抽出される構成要素も異なってくる為に、用語を分割する為の構成要素の粒度も異なってくる。このようにして分割された語は、下記（１）及び（２）それぞれに示す場面において利用されうる。

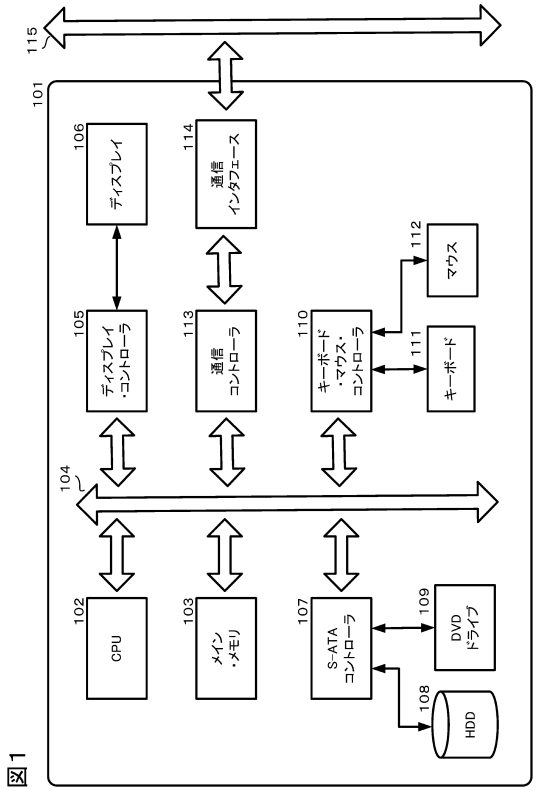
【０１６５】

（１）例えば、システムの大規模改修やシステム統合により新システムに移行するという場面では、例えば以前のシステムの設計で用いられていた用語を見直す必要がある。例えば、新システムで使わない用語を削除したり、新システムで新規に検討しなくてはならない用語（例えば、新しいビジネス・プロセス関連の用語）を洗い出したり、曖昧に使用されていた用語を見直したり、同じ意味で異なる用語が異なるシステム間で使用されている場合に、当該異なる用語を統一したりする必要がある。このような場合に、本発明の実施態様に従い、コンテンツから抽出した構成要素で用語を分割することによって当該用語が適切な粒度で分割される為に当該分割された結果に基づいて上記用語の見直しが可能となる点で、本発明の実施態様に従う上記分割は有用である。

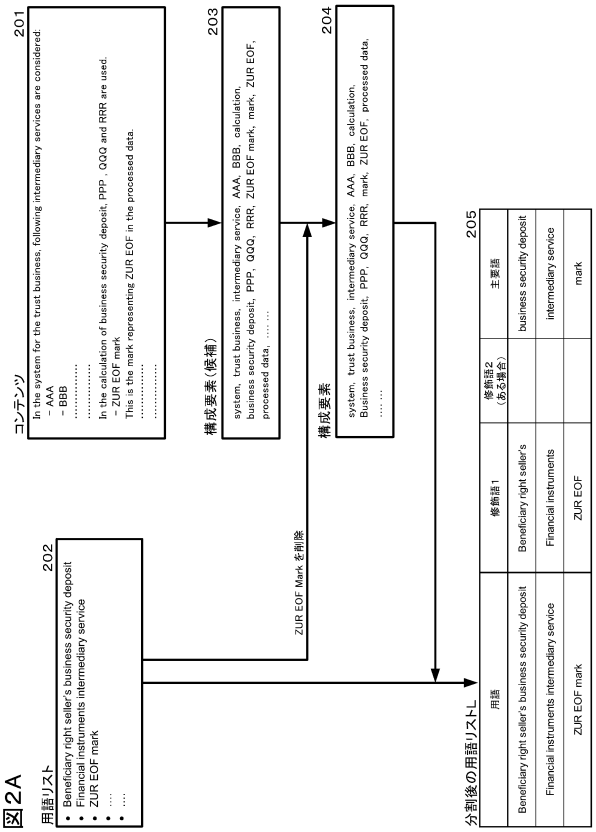
【０１６６】

（２）用語の見直しをするに際して、当該用語がデータ項目の場合には、当該用語がデータベースのカラム名になっていたり、プログラム中の変数となっていたりする為に、構造ルールを設けるがある。このような場合に、本発明の実施態様に従いコンテンツから抽出した構成要素で用語を分割することによって当該用語が適切な粒度で分割される為に当該分割された結果に基づいて、上記構造ルールを設けることが可能となる、当該構造ルールを設けることで、用語の構造を共通理解でき、さらには用語が意味する概念についての理解が一意に決まることが期待でき、また、不明確な用語を作成する必要がないという点で、本発明の実施態様に従う上記分割は有用である。

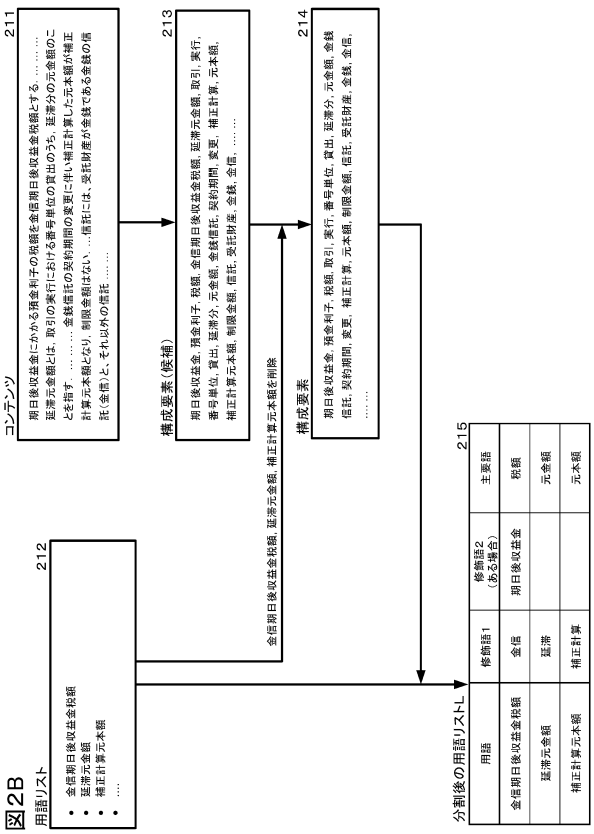
【図１】



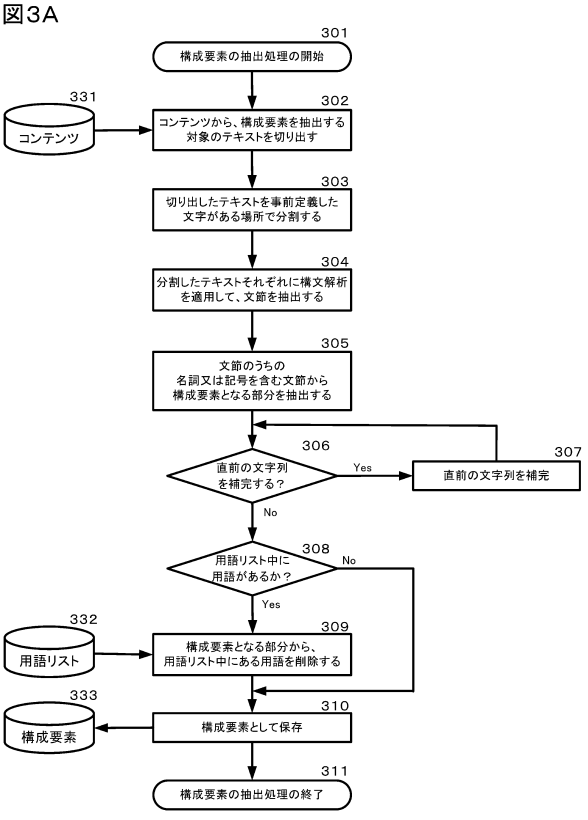
【図２Ａ】



【図２Ｂ】

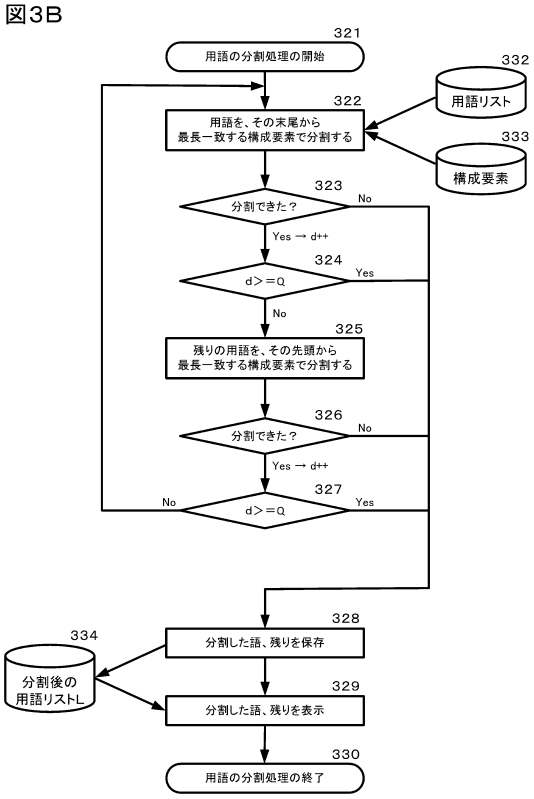


【図３Ａ】



【図３Ｂ】

図3B


321
用語の分割処理の開始
332
用語リスト
322
用語を、その末尾から
最長一致する構成要素で分割する
333
構成要素
323
分割できた？
No
Yes → d++
324
d＞＝Q
Yes
No
325
残りの用語を、その先頭から
最長一致する構成要素で分割する
326
分割できた？
No
Yes → d++
327
No
d＞＝Q
Yes
334
分割後の
用語リストL
328
分割した語、残りを保存
329
分割した語、残りを表示
330
用語の分割処理の終了

【図４】

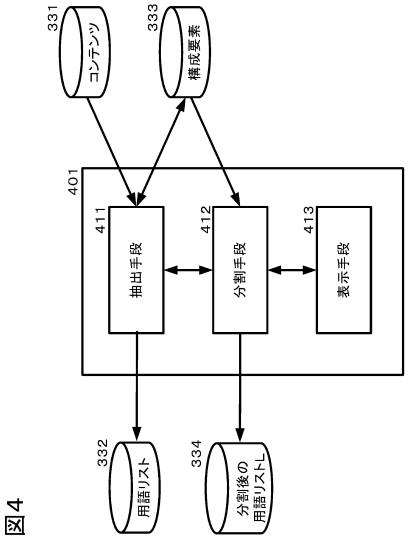

331
コンテンツ
333
構成要素
401
411
抽出手段
412
分割手段
413
表示手段
332
用語リスト
334
分割後の
用語リストL

図4

フロントページの続き

(74)代理人 100112690
弁理士 太佐 種一
(72)発明者 竹内 広宜
東京都江東区豊洲五丁目６番５２号 ＮＢＦ豊洲キャナルフロント 日本アイ・ビー・エム株式会社 東京基礎研究所内
(72)発明者 中村 大賀
東京都江東区豊洲五丁目６番５２号 ＮＢＦ豊洲キャナルフロント 日本アイ・ビー・エム株式会社 東京基礎研究所内
(72)発明者 十河 維
東京都中央区日本橋箱崎町１９番地２１ 日本アイ・ビー・エム株式会社内

審査官 長 由紀子

(56)参考文献 特開２０１１－０９６２４５（ＪＰ，Ａ）
特開平１０－２６０８２４（ＪＰ，Ａ）
特開２００３－０１５８６９（ＪＰ，Ａ）
特開平０４－２１５１８２（ＪＰ，Ａ）
特開平０６－２３１１８８（ＪＰ，Ａ）
特開平０７－０６５００８（ＪＰ，Ａ）
特開２００２－２５９３７０（ＪＰ，Ａ）
特開２００５－２９３５８２（ＪＰ，Ａ）
高田 京児，ビジネスの視点でデータを整理 内部統制にも対応できる統合データベース設計の研究，ＤＢ Ｍａｇａｚｉｎｅ，日本，株式会社翔泳社，２００９年 2月 1日，第18巻第10号，p.98-110

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／２０－２８