(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2004-126986
(P2004-126986A)
(43)公開日　平成16年4月22日(2004.4.22)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G06F 17/24	G06F	17/24	554N	5B009
G06F 17/21	G06F	17/21	550A	

審査請求 有　請求項の数 8　OL　（全 17 頁）

(21)出願番号　特願2002-290946(P2002-290946)
(22)出願日　平成14年10月3日(2002.10.3)

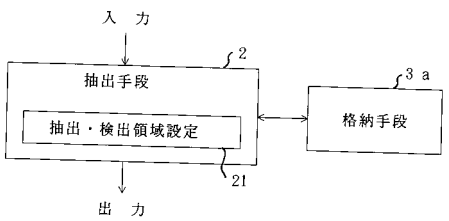
(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(74)代理人　100103827
弁理士　平岡　憲一
(74)代理人　100097836
弁理士　福井　國敏
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

Fターム(参考)　5B009 QA03　QB18　RB32

(54)【発明の名称】文書差分検出装置及びプログラム

(57)【要約】
【課題】文書の特徴や新情報のわかりやすい表示を行うこと。
【解決手段】文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段２１と、情報を格納する格納手段３ａと、抽出手段２とを備え、前記抽出手段２は、入力された文書データの現在の前記検出領域以外の領域から全ての前記抽出単位に相当するものを抽出して前記格納手段３ａに格納し、現在の前記検出領域において、前記格納手段３ａに格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力することを、前記検出領域ごとに繰り返す。
【選択図】　図１

本発明の原理説明図



【特許請求の範囲】
【請求項１】
文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段と、
情報を格納する格納手段と、
抽出手段とを備え、
前記抽出手段は、入力された文書データの現在の前記検出領域以外の領域から全ての前記抽出単位に相当するものを抽出して前記格納手段に格納し、現在の前記検出領域において、前記格納手段に格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力することを、前記検出領域ごとに繰り返すことを特徴とした文書差分検出装置。
【請求項２】
文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段と、
情報を格納する格納手段と、
抽出手段とを備え、
前記抽出手段は、入力された文書データの現在の前記検出領域において、前記格納手段に格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力し、前記強調表示したものを前記格納手段に格納することを、前記検出領域ごとに繰り返すことを特徴とした文書差分検出装置。
【請求項３】
前記格納手段に予め前記強調表示しない前記抽出単位のデータを格納することを特徴とした請求項１又は２記載の文書差分検出装置。
【請求項４】
前記抽出単位として、単語の単位とすることを特徴とした請求項１～３のいずれかに記載の文書差分検出装置。
【請求項５】
前記検出領域の単位として、箇条書きの単位とすることを特徴とした請求項１～４のいずれかに記載の文書差分検出装置。
【請求項６】
前記検出領域の単位として、特許請求の範囲の単位とすることを特徴とした請求項１～４のいずれかに記載の文書差分検出装置。
【請求項７】
文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段と、
入力された文書データの現在の前記検出領域以外の領域から全ての前記抽出単位に相当するものを抽出して格納手段に格納し、現在の前記検出領域において、前記格納手段に格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力することを、前記検出領域ごとに繰り返す抽出手段として、
コンピュータを機能させるためのプログラム。
【請求項８】
文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段と、
入力された文書データの現在の前記検出領域において、格納手段に格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力し、前記強調表示したものを前記格納手段に格納することを、前記検出領域ごとに繰り返す抽出手段として、
コンピュータを機能させるためのプログラム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】

本発明は、文書（又は文章）の差分を検出して、文書の違いを容易に理解できるようにする文書差分検出装置及びプログラムに関する。

【０００２】

【従来の技術】

従来、ｄｉｆｆコマンドを用いて、入力された複数の文書データの差分を検出し、複数の文書データの差分の内で、共通部分は一つを出力し、不一致部分はそれぞれを並べて出力する技術があった。

【０００３】

ここで、ｄｉｆｆ（ディフ）とは、ＵＮＩＸ（ユニックス）（登録商標）のファイル比較ツールｄｉｆｆのことである。このｄｉｆｆコマンドは、与えられた二つのファイルの差分を順序情報を保持したまま行を単位として出力するものである。

【０００４】

ｄｉｆｆコマンドには、－Ｄオプションという便利なオプションがある。このオプションを付けてｄｉｆｆコマンドを使うと差分部分だけでなく共通部分も出力される。つまり、ファイルのマージが実現される。また、差分部分を見やすく表示するため、差分部分の始まり、差分部分の終わり、差分を構成する二つのデータの境界を表す表示を行う。このような、ファイルのマージを行う場合のｄｉｆｆを、Ｍｄｉｆｆ　（エムディフ）と呼ぶ（Ｍ　はｍｅｒｇｅ　の　Ｍである）（例えば、非特許文献１及び特願２００１－３１１３２９参照）。

【０００５】

この技術を用いて、一つの特許の複数の請求項の間の差分を検出する実験を行なった。これは新しい試みである。ある特許の二つの請求項を一行に１個の単語がはいるように変形してから、それらの　Ｍｄｉｆｆをとった（なお、以下の説明では請求項等のすみ付き括弧は「〔」又は「〕」に置き換えてある）。

【０００６】

例１、

〔請求項１７〕　前記プリンタシステムは上位装置を有することを特徴とする請求項１６記載のプリンタシステムの制御方法。

〔請求項１８〕　前記プリンタシステムはプリンタを有することを特徴とする請求項１６記載のプリンタシステムの制御方法。

【０００７】

（上記例１の　Ｍｄｉｆｆ結果）

前記プリンタシステムは

；＝＝＝＝＝ｂｅｇｉｎ＝＝＝＝＝

上位装置

；－－－－－－－－－

プリンタ

；＝＝＝＝＝ｅｎｄ＝＝＝＝＝

を有することを特徴とする請求項１６記載のプリンタシステムの制御方法。

【０００８】

上記例１の請求項１７と請求項１８の　Ｍｄｉｆｆをとった結果から、たいへん容易に請求項１７と請求項１８の違いを理解することができる。即ち、；＝＝＝＝＝ｂｅｇｉｎ＝＝＝＝＝は差分部分の始まり、；＝＝＝＝＝ｅｎｄ＝＝＝＝＝は差分部分の終わり、；－－－－－－－－－は差分を構成する二つのデータの境界を表す。ここで、違いは「上位装置」と「プリンタ」である。しかし、違いがもっとややこしい場合は、Ｍｄｉｆｆ　の結果は見にくいことになる。

【０００９】

例２、

〔請求項１〕

刃部材の先端の刃部を凹凸に形成し波状刃とするとともに螺旋状に湾曲させ、前記刃部材

に取っ手を取り付けたことを特徴とする草取り鎌。
〔請求項２〕
取っ手の上部及び下部に滑り止め部を設けたことを特徴とする草取り鎌。
【００１０】
（上記例２の　Ｍｄｉｆｆ結果）
；＝＝＝＝＝ｂｅｇｉｎ＝＝＝＝＝
刃部材
；―――――――
取っ手
；＝＝＝＝＝ｅｎｄ＝＝＝＝＝
の
；＝＝＝＝＝ｂｅｇｉｎ＝＝＝＝＝
先端の刃
；―――――――
上部及び下部に滑り止め
；＝＝＝＝＝ｅｎｄ＝＝＝＝＝
部を
；＝＝＝＝＝ｂｅｇｉｎ＝＝＝＝＝
凹凸に形成し波状刃とするとともに螺旋状に湾曲させ、前記刃部材に取っ　手を取り付け
；―――――――
設け
；＝＝＝＝＝ｅｎｄ＝＝＝＝＝
たことを特徴とする草取り鎌。
【００１１】
上記例２の請求項１と請求項２の　Ｍｄｉｆｆをとった結果は、違いがややこしいので、Ｍｄｉｆｆ　の結果は見にくいことになっている。即ち、Ｍｄｉｆｆ　は、順序情報を保存する機構であるため、違いが複雑な場合に、違いがわかりにくく、このままでは問題があることがわかった。
【００１２】
【非特許文献１】
村田真樹，外１名，　ｄｉｆｆと言語処理「言語理解とコミュニケーション」社団法人電子情報通信学会２００１年　７月１７日（ＮＬＣ２００１－２６　）　電子情報通信学会技術研究報告，　ｐ．２９～３６
【００１３】
【発明が解決しようとする課題】
上記従来の　Ｍｄｉｆｆを用いるものは、違いが複雑な場合に、Ｍｄｉｆｆ　の結果が見にくいことになるものであった。
【００１４】
本発明は上記問題点の解決を図り、違いが複雑な場合にもわかりやすい表示を行うことを目的とする。
【００１５】
【課題を解決するための手段】
図１は本発明の原理説明図である。図１中、２は抽出手段、３ａは格納手段、２１は抽出・検出領域設定手段である。
【００１６】
本発明は、前記従来の課題を解決するため次のような手段を有する。
【００１７】
（１）：文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段２１と、情報を格納する格納手段３ａと、抽出手段２とを備え、前記抽出手段２は、入力

された文書データの現在の前記検出領域以外の領域から全ての前記抽出単位に相当するものを抽出して前記格納手段３ａに格納し、現在の前記検出領域において、前記格納手段３ａに格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力することを、前記検出領域ごとに繰り返す。このため、新しい情報である文書の特徴や差分を容易に抽出表示することができる。

【００１８】

（２）：文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段２１と、情報を格納する格納手段３ａと、抽出手段２とを備え、前記抽出手段２は、入力された文書データの現在の前記検出領域において、前記格納手段３ａに格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力し、前記強調表示したものを前記格納手段３ａに格納することを、前記検出領域ごとに繰り返す。このため、新しく出現する抽出単位に相当するもの（例えば単語）を容易に抽出して表示することができる。

【００１９】

（３）：前記（１）又は（２）の文書差分検出装置において、前記格納手段３ａに予め前記強調表示しない前記抽出単位のデータを格納する。このため、予めそれほど重要でない表現を強調表示しないようにでき、見やすくすることができる。

【００２０】

（４）：前記（１）～（３）の文書差分検出装置において、前記抽出単位として、単語の単位とする。このため、新しく出現する単語を抽出表示することができる。

【００２１】

（５）：前記（１）～（４）の文書差分検出装置において、前記検出領域の単位として、箇条書きの単位とする。このため、箇条書き間の違いを容易に理解することができる。

【００２２】

（６）：前記（１）～（４）の文書差分検出装置において、前記検出領域の単位として、特許請求の範囲の単位とする。このため、特許請求の範囲の特徴や違いを容易に理解することができる。

【００２３】

（７）：文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段２１と、入力された文書データの現在の前記検出領域以外の領域から全ての前記抽出単位に相当するものを抽出して格納手段３ａに格納し、現在の前記検出領域において、前記格納手段３ａに格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力することを、前記検出領域ごとに繰り返す抽出手段２として、コンピュータを機能させるためのプログラム又はプログラムを記録したコンピュータ読取可能な記録媒体とする。このため、このプログラムをコンピュータにインストールすることで文書の特徴や差分を容易に抽出表示することができる文書差分検出装置を容易に提供することができる。

【００２４】

（８）：文書データの差分として出力する対象の単位である抽出単位と文書データの差分を検出するために比較する領域の単位である検出領域を設定する抽出・検出領域設定手段２１と、入力された文書データの現在の前記検出領域において、格納手段３ａに格納されていない前記抽出単位に相当するものを強調表示して現在の検出領域の文書を出力し、前記強調表示したものを前記格納手段３ａに格納することを、前記検出領域ごとに繰り返す抽出手段２として、コンピュータを機能させるためのプログラム又はプログラムを記録したコンピュータ読取可能な記録媒体とする。このため、このプログラムをコンピュータにインストールすることで新しく出現する抽出単位に相当するものを抽出して表示することができる文書差分検出装置を容易に提供することができる。

【００２５】

【発明の実施の形態】
（１）：文書差分検出装置の説明
図２は文書差分検出装置の説明図である。図２において、文書差分検出装置には、入力手段１、抽出手段２、抽出物記憶装置３、出力手段４が設けてある。入力手段１は、キーボード、マウス、読み取り装置等の情報の入力を行うものである。抽出手段２は、入力された文書の差分を抽出するものである。抽出物記憶装置３は、単語、漢字、名詞句などの抽出物を格納する抽出物記憶手段である。出力手段４は、表示装置、プリンタ等の情報の出力を行うものである。
【００２６】
▲１▼：形態素解析システムの説明
日本語を単語に分割するために、抽出手段２が行う形態素解析システムが必要になる。ここではＣｈａＳｅｎについて説明する（奈良先端大で開発されている形態素解析システム茶筌ｈｔｔｐ：／／ｃｈａｓｅｎ．ａｉｓｔ－ｎａｒａ．ａｃ．ｊｐ／ｉｎｄｅｘ．ｈｔｍｌ．ｊｐ　で公開されている）。
【００２７】
これは、日本語文を分割し、さらに、各単語の品詞も推定してくれる。例えば、「学校へ行く」を入力すると以下の結果を得ることができる。
【００２８】
学校　　ガッコウ　　学校　　名詞－一般
へ　　へ　　へ　　助詞－格助詞－一般
行く　　イク　　行く　　動詞－自立　　五段・カ行促音便　　基本型
ＥＯＳ
このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。本発明の抽出手段２では、この機能のうち単語を分割する部分（形態素解析手段）だけを利用する。
【００２９】
▲２▼：英語のｓｔｅｍｍｅｒ　（ステマー）の説明
抽出手段２で単語を抽出するには、英語では単語はわかち書きされているので、単語を基本形式に戻すｓｔｅｍｍｉｎｇをするだけでよい。このｓｔｅｍｍｉｎｇをするアルゴリズムとしては有名なＰｏｒｔｅｒのものがある（　Ｐｏｒｔｅｒ，　Ｍ．Ｆ．，　１９８０，　Ａｎ　ａｌｇｏｒｉｔｈｍ　ｆｏｒ　ｓｕｆｆｉｘ　ｓｔｒｉｐｐｉｎｇ，　Ｐｒｏｇｒａｍ，　１４（３）　：１３０－１３７　参照）。
【００３０】
▲３▼：抽出単位、検出領域の説明
文字、段落、文、箇条書の項目などは、文書の形式から機械的に認識できる。例えば文字ならば、１バイトや２バイトコードで認識できる。段落ならば、字下げ、改行により認識できる。文ならば、句点やピリオドの存在により認識できる。箇条書は、字下げ、改行、箇条書項目の先頭の記号などにより認識できる。単語の認識については先にあげた形態素解析システムや　ｓｔｅｍｍｅｒにより認識される。前記認識は、例えば、それぞれの認識手段を抽出手段内に設けて行うことができる。
【００３１】
（２）：差分検出の説明
本発明の差分を検出するやり方には二つの手法（方法）がある。これらの手法は　Ｄｉｆｆ　コマンドを使わない。以下、この二つの手法をフローチャートにより説明する。
【００３２】
▲１▼：手法１
図３は手法１の文書差分検出処理フローチャートである。以下、図３の処理Ｓ１～Ｓ３－２に従って説明する。
【００３３】
Ｓ１：入力手段１等により、予め抽出の単位（抽出単位）、検出領域の単位を定める。抽

出単位とは、差分として出力する対象の単位である。抽出単位には、「単語」「漢字」「名詞句」などが考えられる。検出領域の単位とは、差分を検出するために比較する領域の単位のことである。検出領域の単位には、「文字」「単語」「文」「箇条書の項目」「段落」「特許の請求項」などが考えられる。

【００３４】

Ｓ２：抽出手段２は、すべての入力データを記憶手段（抽出手段２内の）に記憶させる。

【００３５】

Ｓ３：抽出手段２は、入力されたデータを左から調べて左の検出領域から処理Ｓ１で定めた検出領域ごとに以下の処理Ｓ３－１と処理Ｓ３－２を繰り返す。

【００３６】

Ｓ３－１：抽出手段２は、現在の検出領域以外の領域すべてから、すべての抽出単位に相当するもの（例えば単語）を抽出し、それを抽出物記憶装置３に格納する。

【００３７】

Ｓ３－２：抽出手段２は、現在の検出領域において、抽出物記憶装置３に格納されていない抽出単位に相当するもの（例えば単語）を強調表示して現在の検出領域の文章を出力手段４に出力する。

【００３８】

▲２▼：手法１の例によるの説明

手法１の例を特許明細書の請求項（検出領域）を例に抽出単位を単語とした説明をする。現在分析している請求項以外の請求項すべてからすべての単語を抽出し、現在分析している請求項において他の請求項に現れない単語を特定する。その結果を以下の例３に示す。

【００３９】

例３

〔請求項１〕《刃部材》の《先端》の《刃》部を《凹凸》に《形成し波状刃》とする《とともに螺旋状》に《湾曲させ、前記刃部材》に取っ手を《取り付け》たことを特徴とする草取り鎌。

〔請求項２〕取っ手の《上部及び下部》に《滑り止め》部を《設け》たことを特徴とする草取り鎌。

【００４０】

上記例３は、他の請求項に現れなかった単語は「《」と「》」の括弧で囲われている（強調表示）。この結果は例２の　Ｍｄｉｆｆの結果よりもはるかに見やすい。この例３から大変容易に〔請求項２〕の特徴が「上部及び下部の滑り止め部」であると理解できる。もし、請求項２の特徴が「滑り止め部」であると理解できたならば、この用語「滑り止め部」を含む実施の形態、実施例中の段落を抜き出すことで、容易に請求項２に対応する実施の形態、実施例を抽出することもできる。

【００４１】

このようにこの手法は、特徴や差分を抽出するのに大変役に立つ。また、ある請求項に対応する実施の形態、実施例の抽出、即ち、請求項と実施の形態、実施例の対応づけにも役立つのである。

【００４２】

次にこの手法１を、三つの請求項を持つ他の例に使ってみた。この場合、以下の例４のような結果を得た。

【００４３】

例４

〔請求項１〕《刃部材》の《先端》の《刃》部を《凹凸》に《形成し波状刃》とする《とともに螺旋状》に《湾曲させ、前記刃部材》に取っ手を《取り付け》たことを特徴とする草取り鎌。

〔請求項２〕取っ手の上部に滑り止め部を設けたことを特徴とする草取り鎌。

〔請求項３〕取っ手の上部《及び下部》に滑り止め部を設けたことを特徴とする草取り鎌。

【００４４】
上記例４の結果では、請求項２と請求項３の特徴である「滑り止め部」を抽出することができなかった。この問題を解決するために二つ目の新しい手法（手法２）を考えた。
【００４５】
▲２▼：手法２
図４は手法２の文書差分検出処理フローチャートである。以下、図４の処理Ｓ１１～Ｓ１２－２に従って説明する。
【００４６】
Ｓ１１：入力手段１等により、予め抽出の単位（抽出単位）、検出領域の単位を定める。抽出単位とは、差分として出力する対象の単位である。抽出単位には、「単語」「漢字」「名詞句」などが考えられる。検出領域の単位とは、差分を検出するために比較する領域の単位のことである。検出領域の単位には、「文字」「単語」「文」「箇条書の項目」「段落」「特許の請求項」などが考えられる。
【００４７】
Ｓ１２：入力手段１から処理Ｓ１１で定めた検出領域ごとに入力データが入力され、抽出手段２は、以下の処理Ｓ１２－１と処理Ｓ１２－２を繰り返す。
【００４８】
Ｓ１２－１：抽出手段２は、現在の検出領域において、抽出物記憶装置３に格納されていない抽出単位に相当するもの（例えば単語）を強調表示して現在の検出領域の文章を出力手段４に出力する。ただし、抽出物記憶装置３は最初は空である。
【００４９】
Ｓ１２－２：処理Ｓ１２－１で強調表示した表現を抽出物記憶装置３に格納する。
【００５０】
▲２▼：手法２の例によるの説明
・手法２の例を特許明細書の請求項を例に抽出単位を単語とした説明をする。二つ目の新しい手法は、今分析している請求項よりも上のすべての請求項からすべての単語を取り出し、今分析している請求項において今分析している請求項よりも上のすべての請求項にあらわれない単語を特定する。その結果を、以下の例５に示す。
【００５１】
例５
〔請求項１〕《刃部材の先端の刃部を凹凸に形成し波状刃とするとともに螺旋状に湾曲させ、前記刃部材に取っ手を取り付けたことを特徴とする草取り鎌。》
〔請求項２〕取っ手の《上部》に《滑り止め》部を《設け》たことを特徴とする草取り鎌。
〔請求項３〕取っ手の上部《及び下部》に滑り止め部を設けたことを特徴とする草取り鎌。
【００５２】
この場合、請求項２と請求項３の特徴である「滑り止め部」を抽出することができた。この方法により、新しく出現する単語を差分として抽出することができる。
【００５３】
・手法２を用いた普通の文の例を説明する。ここで、抽出の単位、検出領域の単位とも単語である。
【００５４】
例６
《本研究の目的は，日本語》の《受け身文》，《使役》文《を能動》文《に変換する際》に《変更され》《るべき格助詞》を《機械学習》を《用いて自動》変換する《ことである．》日本語の受け身文，使役文の《例》を《図１　と》図《２　》に《あげる》．図１の文の日本語の《接尾辞「》れ《た」》は《受動態》を《示す助動詞》で《あり》，《この》文は受け身文である．図２　の文の日本語の接尾辞「《せ》た」は使役を示す助動詞であり，この文は使役文である．《これら》の文に《対》《応》する能動文を図《３　》

に示す．図１　の文《が》能動文に変換さ《れるとき》は，《（ｉ）　》格助詞「に」は格助詞「が」に《（ｉｉ）》格助詞「が」は格助詞「を」に変換される．図２　の文が能動文に変換されるときは，（ｉ）　格助詞「が」の《部分》「《彼》が」の《文節》が《消去》され，（ｉｉ）格助詞「に」が格助詞「が」に変換され，《（ｉｉｉ）　》格助詞「を」は変換され《ず》に《そのまま残る》．本研究では，これらの格助詞の変換《（　》例《：》格《助》《詞》「に」の格助詞「が」《へ》の変換《）　》と，《不要》部分の消去（　例：「彼が」の消去）　を，研究の《対象》とする．（　《以降》，《本稿》では《便宜上》「彼が」《など》の消去の部分《も》格助詞の変換と《呼ぶ》．）
受け身文，使役文の能動文への変換は，文《生成》，《言い換え》，文の《平易化／言語》《運用支援》，《自然》言語文《から》の《知識獲得や情報抽出》，《質問応答システム》と《多く》の研究《分野》で《役に立つもの》である．《例えば》，質問応答システムでは，質問文が《能》《動》文で《答え》が《受動》文で《書か》れて《いる場合》，質問文と答えを《含む》文で，文の《構造》が《異なるため》に，質問の答えを《取り出す》のが《困難な》場合がある．この《よう》な《問》《題》も受け身文，使役文の能動文への変換が《できる》ように《なる》と《解決》する《のであ》る．このように受け身文，使役文の能動文への変換は，自然言語《処理》で《重要》なものである．

【００５５】

この例６の表示により、第二段落は、《生成》《言い換え》《平易化／言語》などの話が新たに生じていることなどがわかる。また、第二段落では、《役に立つもの》《困難な》《できる》《重要》などの評価する際に用いる言語表現が多く用いられていることから、手法の妥当性や有用性を記述していることも容易に理解できる。

【００５６】

・手法２を用いた発明の詳細な説明文の例を説明する。ここで、抽出の単位、検出領域の単位とも単語である。

【００５７】

例７

《次に、本発明について図面を参照して説明する。図１は》本発明《である草取り鎌の正面》図、図《２》は本発明である草取り鎌の《背面》図、図《３》は、本発明である草取り鎌の《右側面》である。

《〔００００７〕》

本草取り鎌１は、図３《に示すよう》に、《刃部材》２の刃《部》２《ｂ》は《当該先端》の《一面が波状》の波状刃《５》に《形成され》て《いるとともに》背面が《平坦》に形成されている刃部材２《と》、《取っ手》３《から構成》されている。

〔《０００８》〕

刃部材２は、図１、図２《及び》図３に示すように、《延長》部２《ａ》が《あり》取っ手３の《約》２《倍程》の《長》さがある。波状刃５の刃部２ｂは《一方向》に《湾曲》している。

〔《０００９》〕

図《４》は本発明の草取り鎌の刃部の正面《拡大》図である。図に示すように、《雑草》を《刈り取る》刃部２ｂは、《凸》部５ａと《凹》部５ｂが《交互》に《存在》し波状と《なっ》ている。

〔《００１０》〕

図５は本発明である草取り鎌の刃部の拡大図である。刃部２ｂを構成する凸部５ａの先端は《やや左方向》に《傾い》ている。《これ》は、雑草を《より引っ掛け》て《刈り取り易く》する《ため》である。

〔《００１１》〕

図《６》は本発明である草取り鎌の刃部の湾曲《状態》を《示した一部》拡大図である。図に示すように、刃部２ｂの延長部２ａより刃部２ｂの先端２《ｃ》は《垂直線》６からより湾曲している。

〔《００１２》〕

図《７》は、図《中》の《Ａ－》Ａ線に《沿っ》た《断面》図である。刃部２ｂの《上面》７は《傾斜》し、凸部５ａの先端５ｃは《尖っ》ている。《そして》、刃部２ｂ《自体》が湾曲するとともに《螺旋》している。

〔《００１３》〕

図《８》は、本発明である草取り鎌の《他》の《実施例》の正面図、図《９》は本発明である草取り鎌の他の実施例の背面図、図《１０》は本発明である草取り鎌の他の実施例の右側面図、図《１１》は、本発明である草取り鎌の他の実施例の一部拡大図である。

〔《００１４》〕

本例の草取り鎌１ａは、刃部材２の延長部２ａが《短い》とともに刃部２ｂの《部分》がやや《大きく》形成してある。

〔《００１５》〕

《また》、取っ手３が《長く》、《握り》部３ｂの《上》に、握り部３ｂの《径》よりやや《大きい》径の《上滑り止め》部３ａを《設ける》とともに、《下》に《も同様》に握り部３ｂより《大》径の下《滑り》止め部３《ｃｂ》を《設け》てある。

〔《００１６》〕

図１０に示すように、本例の草取り鎌１ａの刃部２ｂも図１から図７《まで》に示した草取り鎌１と同様に螺旋《状》に湾曲している。

〔《００１７》〕

《この》ように、先端部が螺旋状に湾曲さ《せること》により、《芝生等》に《生え》ている雑草を《根こそぎ取り除く》ことが《容易》と《なる》。

〔《００１８》〕

【００５８】

この例７では、段落番号００１２で、「螺旋」がここで初出とわかる。段落番号００１５で、「滑り止め部」が重要とわかる。また、段落番号００１７で、「根こそぎ取り除く」という面白い表現がここで初出とわかる。

【００５９】

・手法２による英語のテキストでの例を説明する。ここで、抽出の単位、検出領域の単位ともに単語である。また、ｓｔｅｍｍｉｎｇはせず、単語の認識はスペースで区切られているかで行なった。

【００６０】

例８

《In the PATENT task of NTCIR－3, we participated in 》the《optional task,》《where 》the《participants can perform any kind 》of《research related to 》《patents. We think that》 in《a 》PATENT《attempt,》 the optional task《is very 》《interesting, because》 we《have already heard》 that《some》 participants in《previous contests wanted》 to《maketheir studies as freely》 as《they》《wanted. Various new ideas or》 new《topics will come up 》in《an》 optional《task. These attempts would be novel and valuable. 》In the《other 》《contests, too,》 we《hope》 that《such》 attempts will be《made. 》In《this contest, 》we《made》 the《following three 》studies《for 》the optional task of《PATENT. 》We《extracted rewriting rules using data》 of patents. We《aligned 》the《claim 》of a《patent》

and《its embodiment.》We extracted《differences among plural claims》in a《patent.》《The first two》topics《were given by organizers》of PATENT as《examples》of the optional task. We《consider these》studies to be very《interesting.》The《last topic》is《our idea.》We《sometimes write》a《patent,》and《had》the《experience》of《wanting》to《know》the《difference》of《claims. So,》we《did》this《study.》We have《been studying natural language processing》using the《Unix》《command Diff.》We《previously proposed ways》to《use Diff》in natural language《processing.》The Diff command is very《suitable》for《doing》the《above》three《studies.》We have already extracted rewriting rules by using Diff in some research《topics. For example,》we《used》a《pair》of《definition sentences having》the《same word entry》in two《different》《dictionaries》and extracted the differences《between them.》Theseextracted differences can be used as《synonym phrases》because the definition sentences in the same entry have thesame《meaning.》In《another situation,》we used aligned《spoken−language》and《written−language texts》and extracted the differences betweenthem. These extracted differences can be used as rewriting rules《transforming》spoken−language sentences《into》written−language sentences or transforming written−language sentences into spoken−language《sentences.》Diff can《also》be used for《alignment.》Diff《has》a《function》of《merging》data《like》a《DP−matching algorithm. So》we can《align》two relatedtexts by using Diff. In this《study,》we used this function for the《alignment》of a patent claim and its《embodiment (working》《example). Finally,》we used Diff for《extracting》the differences of patent claims.《Extracting》differences is an《original》function of Diff. Extracting differences between claims《enables us》to《understand》the claims of a patent《more deeply.》

【００６１】

この例８では、真ん中あたりの箇条書で、箇条書部での主要ワードがそれぞれ強調されて

いる。即ち、（《ｅｘｔｒａｃｔｅｄ ｒｅｗｒｉｔｉｎｇ ｒｕｌｅｓ ｕｓｉｎｇ ｄａｔａ》や《ａｌｉｇｎｅｄ 》 ｔｈｅ《ｃｌａｉｍ 》ｏｆ ａ《ｐａｔｅｎｔ》 ａｎｄ《ｉｔｓ ｅｍｂｏｄｉｍｅｎｔ． 》や《ｄｉｆｆｅｒｅｎｃｅｓ ａｍｏｎｇ ｐｌｕｒａｌ ｃｌａｉｍｓ 》）容易に各箇条書の要点が理解できる。

【００６２】

最後の段落では、Ｄｉｆｆの話が始まったとわかる。また，《ｄｅｆｉｎｉｔｉｏｎ ｓｅｎｔｅｎｃｅｓ》《ｓｙｎｏｎｙｍ ｐｈｒａｓｅｓ 》《ｓｐｏｋｅｎ－ｌａｎｇｕａｇｅ 》《ｗｒｉｔｔｅｎ－ｌａｎｇｕａｇｅ ｔｅｘｔｓ》《ＤＰ－ｍａｔｃｈｉｎｇ ａｌｇｏｒｉｔｈｍ 》などの主要なキーワード（キーフレーズ）がすぐに目に入る。内容理解等に便利である。

【００６３】

（３）：ユーザー辞書を設ける文書差分検出装置の説明

予め各ユーザーは、ユーザー辞書なるものをもっておき、その辞書にあるものは強調しないようにするものである。これにより、重要でない表現を予め強調しないようにし、見やすくすることができる。

【００６４】

図５はユーザー辞書を設ける文書差分検出装置の説明図である。図５において、文書差分検出装置には、入力手段１、抽出手段２、抽出物記憶装置３、出力手段４、ユーザー辞書５が設けてある。入力手段１は、キーボード、マウス、読み取り装置等の情報の入力を行うものである。抽出手段２は、入力された文書の差分を抽出するものである。抽出物記憶装置３は、単語、漢字、名詞句などの抽出物を格納する抽出物記憶手段である。出力手段４は、表示装置、プリンタ等の情報の出力を行うものである。ユーザー辞書５は、予め各ユーザーが登録しておく辞書である。

【００６５】

▲１▼：ユーザー辞書を設ける手法１の説明

図６はユーザー辞書を設ける手法１の文書差分検出処理フローチャートである。以下、図６の処理Ｓ２１～Ｓ２３－２に従って説明する。

【００６６】

Ｓ２１：入力手段１等により、予め抽出の単位（抽出単位）、検出領域の単位を定め、ユーザー辞書５登録を行う。抽出単位とは、差分として出力する対象の単位である。抽出単位には、「単語」「漢字」「名詞句」などが考えられる。検出領域の単位とは、差分を検出するために比較する領域の単位のことである。検出領域の単位には、「文字」「単語」「文」「箇条書の項目」「段落」などが考えられる。

【００６７】

Ｓ２２：抽出手段２は、すべての入力データを（抽出手段２内の）記憶手段に記憶させる。

【００６８】

Ｓ２３：抽出手段２は、入力されたデータを左から調べて左の検出領域からＳ２１で定めた検出領域ごとに以下の処理Ｓ２３－１と処理Ｓ２３－２を繰り返す。

【００６９】

Ｓ２３－１：抽出手段２は、現在の検出領域以外の領域すべてから、すべての抽出単位に相当するもの（例えば単語）を抽出し、それを抽出物記憶装置３に格納する。

【００７０】

Ｓ２３－２：抽出手段２は、現在の検出領域において、抽出物記憶装置３に格納されていない、かつ、ユーザー辞書５に格納されていない抽出単位に相当するもの（例えば単語）を強調表示して現在の検出領域の文章を出力手段４に出力する。

【００７１】

▲２▼：ユーザー辞書を設ける手法２の説明

図７はユーザー辞書を設ける手法２の文書差分検出処理フローチャートである。以下、図７の処理Ｓ３１～Ｓ３２－２に従って説明する。

【００７２】
Ｓ３１：入力手段１等により、予め抽出の単位（抽出単位）、検出領域の単位を定め、ユーザー辞書５登録を行う。抽出単位とは、差分として出力する対象の単位である。抽出単位には、「単語」「漢字」「名詞句」などが考えられる。検出領域の単位とは、差分を検出するために比較する領域の単位のことである。検出領域の単位には、「文字」「単語」「文」「箇条書の項目」「段落」などが考えられる。
【００７３】
Ｓ３２：入力手段１から処理Ｓ３１で定めた検出領域ごとに入力データが入力され、抽出手段２は、以下の処理Ｓ３２－１と処理Ｓ３２－２を繰り返す。
【００７４】
Ｓ３２－１：抽出手段２は、現在の検出領域において、抽出物記憶装置３に格納されていない、かつ、ユーザー辞書に格納されていない、抽出単位に相当するもの（例えば単語）を強調表示して現在の検出領域の文章を出力手段４に出力する。ただし、抽出物記憶装置３は最初は空である。
【００７５】
Ｓ３２－２：処理Ｓ３２－１で強調表示した表現を抽出物記憶装置３に格納する。
【００７６】
▲３▼：ユーザー辞書を設ける手法２（他の実現法）の説明
図８はユーザー辞書を設ける手法２（他の実現法）の文書差分検出処理フローチャートである。以下、図８の処理Ｓ４１～Ｓ４３－２に従って説明する。
【００７７】
Ｓ４１：入力手段１等により、予め抽出の単位（抽出単位）、検出領域の単位を定め、ユーザー辞書５登録を行う。抽出単位とは、差分として出力する対象の単位である。抽出単位には、「単語」「漢字」「名詞句」などが考えられる。検出領域の単位とは、差分を検出するために比較する領域の単位のことである。検出領域の単位には、「文字」「単語」「文」「箇条書の項目」「段落」などが考えられる。
【００７８】
Ｓ４２：抽出手段２は、ユーザー辞書５の内容をすべて抽出物記憶装置３に格納する。
【００７９】
Ｓ４３：入力手段１から処理Ｓ４１で定めた検出領域ごとに入力データが入力され、抽出手段２は、以下の処理Ｓ４３－１と処理Ｓ４３－２を繰り返す。
【００８０】
Ｓ４３－１：抽出手段２は、現在の検出領域において、抽出物記憶装置３に格納されていない抽出単位に相当するもの（例えば単語）を強調表示して現在の検出領域の文章を出力手段４に出力する。
【００８１】
Ｓ４３－２：処理Ｓ４３－１で強調表示した表現を抽出物記憶装置３に格納する。
【００８２】
・ユーザ辞書を用いない場合、以下のようになったテキスト例（例９）をとってみる。ここで、手法２を用い、抽出単位、検出領域の単位ともに単語である。
【００８３】
例９
《本研究の目的は，日本語》の《受け身文》，《使役》文《を能動》文《に変換する際》に《変更され》《るべき格助詞》を《機械学習》を《用いて自動》変換する《ことである．》日本語の受け身文，使役文の《例》を《図１　と》図《２　》に《あげる》．図１　の文の日本語の《接尾辞「》れ《た」》は《受動態》を《示す助動詞》で《あり》，《この》文は受け身文である．図２　の文の日本語の接尾辞「《せ》た」は使役を示す助動詞であり，この文は使役文である．《これら》の文に《対》《応》する能動文を図《３　》に示す．図１　の文《が》能動文に変換さ《れるとき》は，《（ｉ）　》格助詞「に」は格助詞「が」に《（ｉｉ）》格助詞「が」は格助詞「を」に変換される．図２　の文が能

動文に変換されるときは，（ｉ）　格助詞「が」の《部分》「《彼》が」の《文節》が《消去》され，（ｉｉ）格助詞「に」が格助詞「が」に変換され，《（ｉｉｉ）　》格助詞「を」は変換され《ず》に《そのまま残る》．本研究では，これらの格助詞の変換《（》例《：》格《助》《詞》「に」の格助詞「が」《へ》の変換《）　》と，《不要》部分の消去（　例：「彼が」の消去）　を，研究の《対象》とする．（　《以降》，《本稿》では《便宜上》「彼が」《など》の消去の部分《も》格助詞の変換と《呼ぶ》．）
受け身文，使役文の能動文への変換は，文《生成》，《言い換え》，文の《平易化／言語》《運用支援》，《自然》言語文《から》の《知識獲得や情報抽出》，《質問応答システム》と《多く》の研究《分野》で《役に立つもの》である．《例えば》，質問応答システムでは，質問文が《能》《動》文で《答え》が《受動》文で《書か》れて《いる場合》，質問文と答えを《含む》文で，文の《構造》が《異なるため》に，質問の答えを《取り出す》のが《困難な》場合がある．この《よう》な《問》《題》も受け身文，使役文の能動文への変換が《できる》ように《なる》と《解決》する《のであ》る．このように受け身文，使役文の能動文への変換は，自然言語《処理》で《重要》なものである．

【００８４】

・ユーザー辞書としては、発明者の他の論文で出現頻度の高かった語を登録する。
（ユーザー辞書の登録例）
の，　を，　，　，　．，　で，　は，　と，　に，　が，　て，　こと，　し，　する，　た，　よう，　部分，　な，　データ，　差分，　ある，　この，　村田，　いる，　」，　「，　研究，　できる，ｄｉｆｆ，），対応，　も，　システム，　処理，　言語，（，　また，　ファイル，　用い，　もの
といった語を登録する。なお、ユーザー辞書の登録例での単語の区切りは「，　」で表してある。

【００８５】

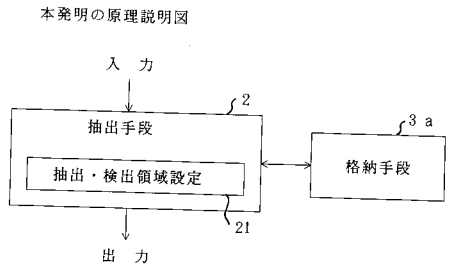
この場合、前記例９は、以下のような結果となる。
《本》研究の《目的》は，《日本語》の《受け身文》，《使役》文を《能動》文に《変換》する《際》に《変更され》《るべき格助詞》を《機械学習》を用いて《自動》変換することである．日本語の受け身文，使役文の《例》を《図１　》と図《２　》に《あげる》．図１　の文の日本語の《接尾辞》「れた」は《受動態》を《示す助動詞》で《あり》，《この》文は受け身文である．図２　の文の日本語の接尾辞「《せ》た」は使役を示す助動詞であり，この文は使役文である．《これら》の文に《対》《応》する能動文を図《３　》に示す．図１　の文が能動文に変換さ《れるとき》は，《（ｉ）　》格助詞「に」は格助詞「が」に《（ｉｉ）》格助詞「が」は格助詞「を」に変換される．図２　の文が能動文に変換されるときは，（ｉ）　格助詞「が」の部分「《彼》が」の《文節》が《消去》され，（ｉｉ）格助詞「に」が格助詞「が」に変換され，《（ｉｉｉ）　》格助詞「を」は変換され《ず》に《そのまま残る》．本研究では，これらの格助詞の変換（　例《：》格《助》《詞》「に」の格助詞「が」《へ》の変換）　と，《不要》部分の消去（　例：「彼が」の消去）　を，研究の《対象》とする．（　《以降》，《本稿》では《便宜上》「彼が」《など》の消去の部分も格助詞の変換と《呼ぶ》．）
受け身文，使役文の能動文への変換は，文《生成》，《言い換え》，文の《平易化／》言語《運用支援》，《自然》言語文《から》の《知識獲得や情報抽出》，《質問応答》システムと《多く》の研究《分野》で《役に立つ》ものである．《例えば》，質問応答システムでは，質問文が《能》《動》文で《答え》が《受動》文で《書か》れている《場合》，質問文と答えを《含む》文で，文の《構造》が《異なるため》に，質問の答えを《取り出す》のが《困難》な場合がある．このような《問》《題》も受け身文，使役文の能動文への変換ができるように《なる》と《解決》する《のであ》る．このように受け身文，使役文の能動文への変換は，自然言語処理で《重要》なものである．

【００８６】

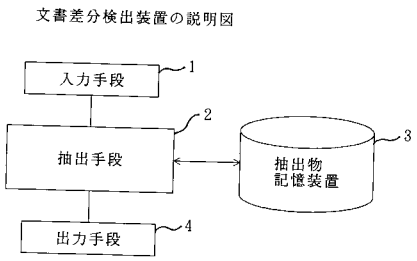
上記結果は、それほど大きな変化はないが、例えば、最初の「研究」や「ことである．」などのそれほど重要でない表現がとられなくなり、少々は見やすくなる。より多くの重要

でない単語をユーザー辞書５に登録することでさらに見やすくすることができる。
【００８７】
なお、前記実施の形態では、強調表示として、２重山括弧で囲む説明をしたが、下線、色分け、背景の変更、字体の変更、点滅等他の強調表示を行うこともできる。
【００８８】
また、このような手法は、照応解析における新情報と旧情報の問題の考察に使うことができる。この「照応解析における新情報と旧情報の問題」に使える手法は、「手法２」の方だけで手法１は使えない。次に、「手法２」の場合、新規に出現した表現が強調表示されるが、言語学的にはこのような文章中に新たに出現した事物は「新情報」と呼ばれる。従って、新規に出現した表現を強調表示する手法２は、言語学でいうところの新情報を抽出していることになっていて、手法２の結果は、言語学でいうところの新情報の考察にも使うことができる。ただし、言語表現の場合、同じ事物を異なる言語表現で言い表す場合もある。その場合、旧情報であっても新しい言語表現であるので、手法２で強調表示する可能性がある。即ち、すべての「新情報」と「旧情報」を正しく区別するわけではない。それでも、手法２は「新情報」と「旧情報」の考察に役立つものである。
【００８９】
更に、抽出単位を漢字とすることで、学校教育等で新しい漢字の出現を容易に理解することができる。漢字の場合は、漢字コードで比較できるため単語のように形態素解析手段が不要となる。
【００９０】
（４）：プログラムインストールの説明
入力手段１、抽出手段２、抽出物記憶装置３、出力手段４、ユーザー辞書５、抽出・検出領域設定手段２１等は、プログラムで構成でき、主制御部（ＣＰＵ）が実行するものであり、主記憶に格納されているものである。このプログラムは、一般的な、コンピュータで処理されるものである。このコンピュータは、主制御部、主記憶、ファイル装置、表示装置、キーボード等の入力手段である入力装置などのハードウェアで構成されている。このコンピュータに、本発明のプログラムをインストールする。このインストールは、フロッピィ、光磁気ディスク等の可搬型の記録（記憶）媒体に、これらのプログラムを記憶させておき、コンピュータが備えている記録媒体に対して、アクセスするためのドライブ装置を介して、或いは、ＬＡＮ等のネットワークを介して、コンピュータに設けられたファイル装置にインストールされる。そして、このファイル装置から処理に必要なプログラムステップを主記憶に読み出し、主制御部が実行するものである。
【００９１】
【発明の効果】
以上説明したように、本発明によれば、次のような効果がある。
【００９２】
（１）：抽出手段で、入力された文書データの現在の検出領域以外の領域から全ての抽出単位に相当するものを抽出して格納手段に格納し、現在の検出領域において、前記格納手段に格納されていない抽出単位に相当するものを強調表示して現在の検出領域の文書を出力することを、検出領域ごとに繰り返すため、新しい情報である文書の特徴や差分を容易に抽出表示することができる。
【００９３】
（２）：抽出手段で、入力された文書データの現在の検出領域において、格納手段に格納されていない抽出単位に相当するものを強調表示して現在の検出領域の文書を出力し、前記強調表示したものを前記格納手段に格納することを、検出領域ごとに繰り返すため、新しく出現する抽出単位に相当するもの（例えば単語）を容易に抽出して表示することができる。
【００９４】
（３）：前記格納手段に予め強調表示しない抽出単位のデータを格納するため、予めそれほど重要でない表現を強調表示しないようにでき、見やすくすることができる。

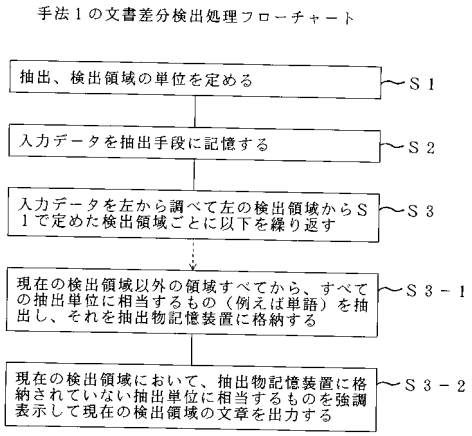
【００９５】
（４）：前記抽出単位として、単語の単位とするため、新しく出現する単語を抽出表示することができる。
【００９６】
（５）：前記検出領域の単位として、箇条書きの単位とするため、箇条書き間の違いを容易に理解することができる。
【００９７】
（６）：前記検出領域の単位として、特許請求の範囲の単位とするため、特許請求の範囲の特徴や違いを容易に理解することができる。
【図面の簡単な説明】
【図１】本発明の原理説明図である。
【図２】実施の形態における文書差分検出装置の説明図である。
【図３】実施の形態における手法１の文書差分検出処理フローチャートである。
【図４】実施の形態における手法２の文書差分検出処理フローチャートである。
【図５】実施の形態におけるユーザー辞書を設ける文書差分検出装置の説明図である。
【図６】実施の形態におけるユーザー辞書を設ける手法１の文書差分検出処理フローチャートである。
【図７】実施の形態におけるユーザー辞書を設ける手法２の文書差分検出処理フローチャートである。
【図８】実施の形態におけるユーザー辞書を設ける手法２（他の実現法）の文書差分検出処理フローチャートである。
【符号の説明】
２　抽出手段
３ａ　格納手段
２１　抽出・検出領域設定手段
【図１】



【図２】

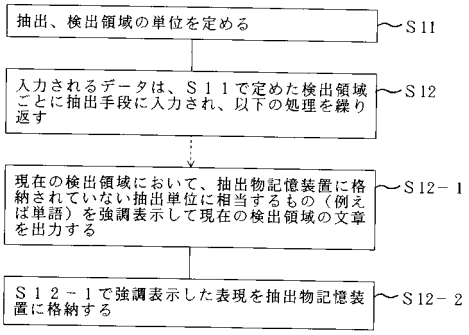


【図３】



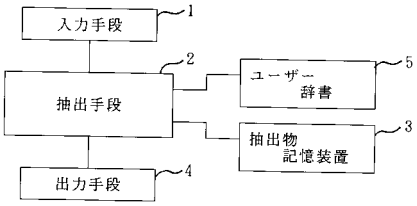
【図４】

手法２の文書差分検出処理フローチャート



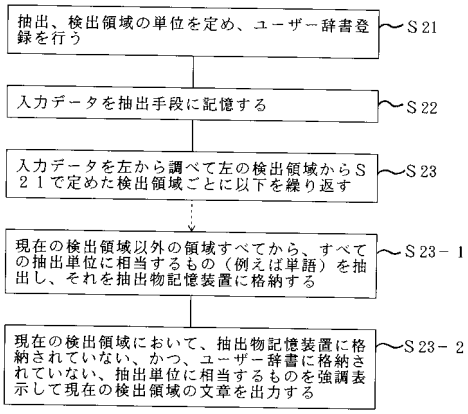
【図５】

ユーザー辞書を設ける文書差分検出装置の説明図



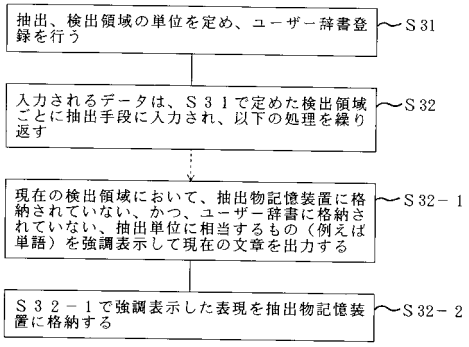
【図６】

ユーザー辞書を設ける手法１の文書差分検出処理フローチャート



【図７】

ユーザー辞書を設ける手法２の文書差分検出処理フローチャート



【図８】

ユーザー辞書を設ける手法２（他の実現法）の文書差分検出処理フローチャート