

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3682535号

(P3682535)

(45) 発行日 平成17年8月10日(2005.8.10)

(24) 登録日 平成17年6月3日(2005.6.3)

(51) Int. Cl.⁷

F I

G06F 17/24

G06F 17/24 554N

G06F 17/21

G06F 17/21 550A

請求項の数 8 (全 18 頁)

(21) 出願番号	特願2002-290946 (P2002-290946)	(73) 特許権者	301022471
(22) 出願日	平成14年10月3日(2002.10.3)		独立行政法人情報通信研究機構
(65) 公開番号	特開2004-126986 (P2004-126986A)		東京都小金井市貫井北町4-2-1
(43) 公開日	平成16年4月22日(2004.4.22)	(74) 代理人	100103827
審査請求日	平成14年10月3日(2002.10.3)		弁理士 平岡 憲一
		(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
		審査官	和田 財太

最終頁に続く

(54) 【発明の名称】 文書差分検出装置及びプログラム

(57) 【特許請求の範囲】

【請求項1】

情報の入力を行う入力手段と、

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とが前記入力手段により設定される抽出・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】

情報の入力を行う入力手段と、

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とが前記入力手段により設定される抽出・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、

10

20

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とが抽出・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し、前記強調表示したもの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とが抽出・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中で、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

10

20

30

40

50

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）。

10

【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

。

20

【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。

30

【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

；

40

50

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。

【 0 0 1 1 】

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。

【 0 0 1 2 】

【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

【 0 0 1 3 】

【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。

【 0 0 1 4 】

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。

【 0 0 1 5 】

【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、3 a は格納手段、2 1 は抽出・検出領域設定手段である。

【 0 0 1 6 】

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。

【 0 0 1 7 】

(1) : 情報の入力を行う入力手段と、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とが前記入力手段により設定される抽出・検出領域設定手段 2 1 と、情報を格納する格納手段 3 a 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 3 a に格納し、現在の前記検出領域において、前記格納手段 3 a 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。

【 0 0 1 8 】

(2) : 情報の入力を行う入力手段と、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とが前記入力手段により設定される抽出・検出領域設定手段 2 1 と、情報を格納する格納手段 3 a 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 3 a 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 3 a 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

10

20

30

40

50

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、
10 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とが抽出・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
20 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とが抽出・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
30 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
40 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ht
tp://chasen.aist-nara.ac.jp/index.html.jpで公開されている）。
50

【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

E O S

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）

10

【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。

20

【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。

30

【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。

40

【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

50

段 4 に出力する。

【 0 0 3 8 】

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。

【 0 0 3 9 】

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。 10

〔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。

【 0 0 4 0 】

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。 20

【 0 0 4 1 】

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。

【 0 0 4 2 】

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。

【 0 0 4 3 】

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。 30

〔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。

【 0 0 4 4 】

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。

【 0 0 4 5 】

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 S 1 1 ~ S 1 2 - 2 に従って説明する。 40

【 0 0 4 6 】

S 1 1 : 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。

【 0 0 4 7 】

S 1 2 : 入力手段 1 から処理 S 1 1 で定めた検出領域ごとに入力データが入力され、抽出 50

手段 2 は、以下の処理 S 1 2 - 1 と処理 S 1 2 - 2 を繰り返す。

【 0 0 4 8 】

S 1 2 - 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。

【 0 0 4 9 】

S 1 2 - 2 : 処理 S 1 2 - 1 で強調表示した表現を抽出物記憶装置 3 に格納する。

【 0 0 5 0 】

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。 10

【 0 0 5 1 】

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 20

【 0 0 5 2 】

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。

【 0 0 5 3 】

・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。

【 0 0 5 4 】

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 50

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。

《〔0007〕》

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》〕

10

20

30

40

50

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。

10

【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ら行なった。

【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《make their 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.》

20

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.》 These extracted differences can be used as《synonym phrases》 because the definition sentences in the same entry have the same《meaning.》 In《another situation,》 we used aligned《spoken-language》 and《written-language texts》 and extracted the differences between them. 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 related texts by using Diff. In this《study,》 we used this function for the《alignment》 of a patent claim and its《embodiment (working》《example). Finally,》

30

40

50

》 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.》

【 0 0 6 1 】

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。

【 0 0 6 2 】

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。

10

【 0 0 6 3 】

(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。

【 0 0 6 4 】

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。

20

【 0 0 6 5 】

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 S 2 1 ~ S 2 3 - 2 に従って説明する。

30

【 0 0 6 6 】

S 2 1 : 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。

【 0 0 6 7 】

S 2 2 : 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。

【 0 0 6 8 】

S 2 3 : 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 S 2 1 で定めた検出領域ごとに以下の処理 S 2 3 - 1 と処理 S 2 3 - 2 を繰り返す。

40

【 0 0 6 9 】

S 2 3 - 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。

【 0 0 7 0 】

S 2 3 - 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。

【 0 0 7 1 】

50

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】

・ユーザー辞書を用いない場合、以下のようなテキスト例（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

10

20

30

40

50

「《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、フロッピー、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】

10

20

30

40

50

(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 10
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。 20

【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】

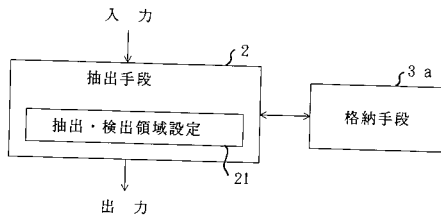
2 抽出手段

3 a 格納手段

2 1 抽出・検出領域設定手段

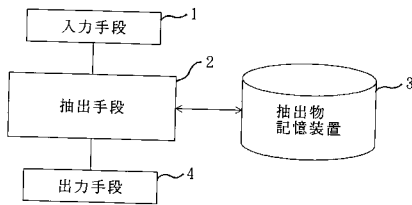
【図1】

本発明の原理説明図



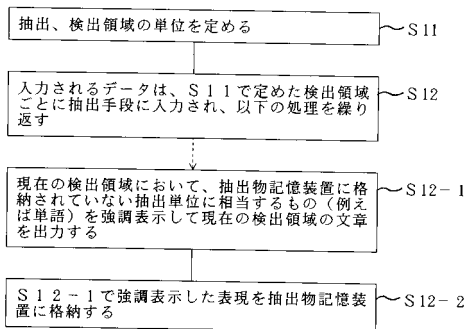
【図2】

文書差分検出装置の説明図



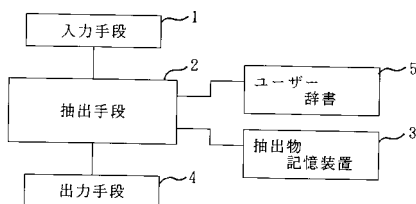
【図4】

手法2の文書差分検出処理フローチャート



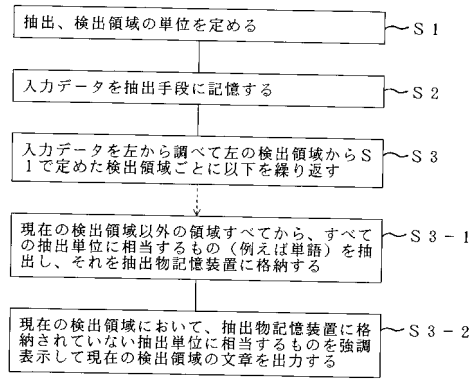
【図5】

ユーザー辞書を設ける文書差分検出装置の説明図



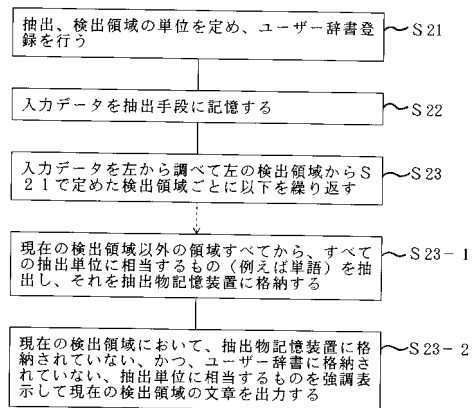
【図3】

手法1の文書差分検出処理フローチャート



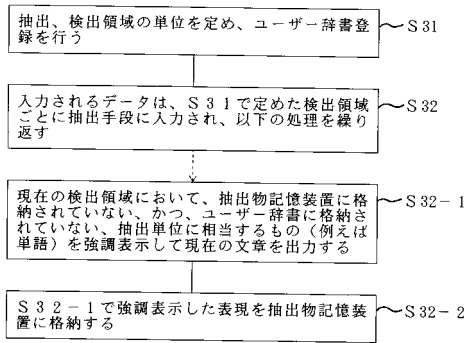
【図6】

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



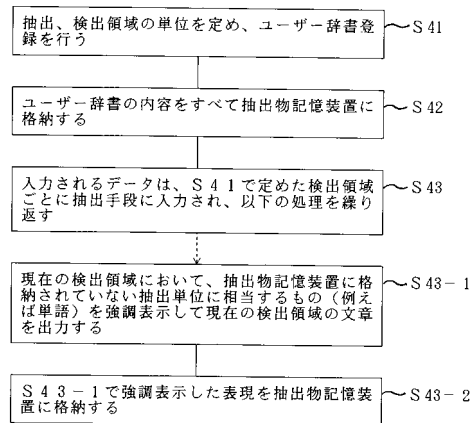
【 図 7 】

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



フロントページの続き

- (56)参考文献 特開2000-112934(JP,A)
特開2002-024211(JP,A)
村田真樹・井佐原均, diffと言語処理, 電子情報通信学会技術研究報告NLC2001-26, 日本, 2001年 7月10日, Vol.101, No.190, p.29-p36
村田真樹, diffを用いた言語処理 - 便利な差分検出ツールmdiffの利用 -, 情報処理学会研究報告2001-NL-144, 日本, 2001年 7月17日, Vol.2001, No.69, p.127-p.134

- (58)調査した分野(Int.Cl.⁷, DB名)
G06F17/21-17/30