(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2005-352888
(P2005-352888A)
(43)公開日　平成17年12月22日(2005.12.22)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	320D	5B009
G06F 17/21	G06F 17/30	170J	5B075
	G06F 17/30	350C	
	G06F 17/21	550K	

審査請求 未請求　請求項の数 15　OL　（全 16 頁）

(21)出願番号　特願2004-174516 (P2004-174516)
(22)出願日　平成16年6月11日 (2004.6.11)

（特許庁注：以下のものは登録商標）
1．WINDOWS

(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100091096
弁理士　平木　祐輔
(72)発明者　大井　洋子
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　今一　修
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内
(72)発明者　丹羽　芳樹
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社日立製作所中央研究所内

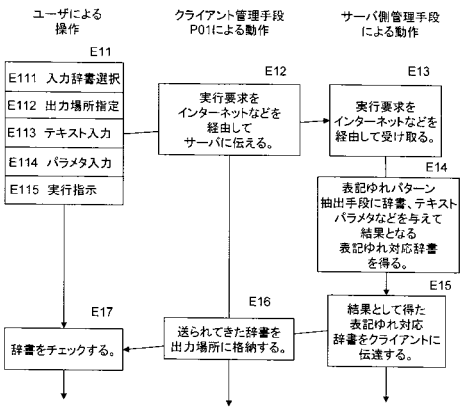
Fターム(参考) 5B009 QA15　QA16　VA02

最終頁に続く

(54)【発明の名称】表記揺れ対応辞書作成システム

(57)【要約】
【課題】着目する用語を中心として文書に起こりうる表記揺れを効率的に漏れなく収集する。
【解決手段】大規模な用語集合の中から表記揺れ候補と考えられる用語を予め選別しておき、表記揺れ候補となった用語に対してコストを調整した編集距離を測ることにより、表記揺れの候補となった用語の中から表記揺れと考えられる用語を収集する。
【選択図】図４

図4



【特許請求の範囲】
【請求項１】
検索語として与えられた用語の表記揺れを抽出するシステムにおいて、
テキスト文書から用語の集合を収集する用語収集部と、
前記用語収集部によって収集された用語の集合の中から前記検索語に類似した用語群を検索する類似用語検索部と、
前記類似用語検索部によって検索された用語群の中から前記検索語の表記揺れを抽出する表記揺れ検索部とを備え、
前記類似用語検索部は、１文字ずつずらした隣接する所定長の部分文字列の共有度を基準にして、比較する２つの用語の類似度を判定し、
前記表記揺れ検索部は、前記検索語との編集距離の総コストが与えられた閾値より小さい用語を前記検索語の表記揺れとして抽出することを特徴とする表記揺れ抽出システム。
【請求項２】
請求項１記載の表記揺れ抽出システムにおいて、前記表記揺れ検索部は、２つの用語の編集距離を、文字の置換、挿入、削除に対して割り当てられたコストを用いて計算することを特徴とする表記揺れ抽出システム。
【請求項３】
請求項１記載の表記揺れ抽出システムにおいて、前記類似用語検索部は、前記検索語との文字列数の相違が許容値内にある用語を対象として前記検索語に類似した用語群を検索することを特徴とする表記揺れ抽出システム。
【請求項４】
請求項１記載の表記揺れ抽出システムにおいて、前記検索語の文字列を１文字ずつずらした部分文字列の索引を作成する索引作成部を有することを特徴とする表記揺れ抽出システム。
【請求項５】
請求項３記載の表記揺れ抽出システムにおいて、前記文字列数の相違の許容値を入力する入力部を有することを特徴とする表記揺れ抽出システム。
【請求項６】
請求項１記載の表記揺れ抽出システムにおいて、前記部分文字列の長さ、前記編集距離の総コストの閾値を入力する入力部を有することを特徴とする表記揺れ抽出システム。
【請求項７】
請求項１記載の表記揺れ抽出システムにおいて、前記検索語として１つの辞書の複数の見出し語を与え、前記辞書に対する表記揺れ辞書を構築することを特徴とする表記揺れ抽出システム。
【請求項８】
コンピュータを用いて検索語として入力された用語の表記揺れを抽出する方法において、
コンピュータが、
指定されたテキスト文書から用語の集合を収集する工程と、
前記収集された用語の集合の中から、１文字ずつずらした隣接する所定長の部分文字列の共有度を基準にして、比較する２つの用語の類似度を判定し、前記検索語に類似した用語群を検索する類似用語検索工程と、
検索された前記用語群の中から、前記検索語との編集距離の総コストが与えられた閾値より小さい用語を前記検索語の表記揺れとして抽出し、前記検索語の表記揺れを抽出する表記揺れ検索工程と、
を実行する表記揺れ抽出方法。
【請求項９】
請求項８記載の表記揺れ抽出方法において、前記表記揺れ検索工程では、２つの用語の編集距離を、文字の置換、挿入、削除に対して割り当てられたコストを用いて計算することを特徴とする表記揺れ抽出方法。

【請求項１０】
請求項８記載の表記揺れ抽出方法において、前記類似用語検索工程では、前記検索語との文字列数の相違が許容値内にある用語を対象として前記検索語に類似した用語群を検索することを特徴とする表記揺れ抽出方法。
【請求項１１】
請求項８記載の表記揺れ抽出方法において、前記検索語の文字列を１文字ずつずらした部分文字列の索引を作成する索引作成工程を有することを特徴とする表記揺れ抽出方法。
【請求項１２】
請求項１０記載の表記揺れ抽出方法において、前記文字列数の相違の許容値の入力を受け付ける工程を有することを特徴とする表記揺れ抽出方法。
【請求項１３】
請求項８記載の表記揺れ抽出方法において、前記部分文字列の長さ、前記編集距離の総コストの閾値の入力を受け付ける工程を有することを特徴とする表記揺れ抽出方法。
【請求項１４】
請求項８記載の表記揺れ抽出方法において、前記検索語として１つの辞書の複数の見出し語に対して前記工程を順次実行し、前記辞書に対する表記揺れ辞書を構築することを特徴とする表記揺れ抽出方法。
【請求項１５】
コンピュータに請求項８記載の表記揺れ抽出方法を実行させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、文書中で使用される用語の表記揺れを抽出する方法に関し、特に大規模な医学生物学文献等から漏れなく専門用語を抽出するのに有用な方法に関するものである。
【背景技術】
【０００２】
用語（ここでは、単語や複合語を意味する）を書き言葉として使用するとき、用語に表記の揺れが生じる場合がある。例えば、“leucocyte”と“leukocyte”、“sulphate”と“sulfate”などである。同一の事物を示す用語にこのような表記揺れが生じているとき、それを考慮せずに検索や情報抽出などを行うと、結果に漏れが生じる。例えば、ユーザの要求に合った情報を文書から抽出し提供するようなシステムでは、ユーザにとって興味ある分野について予め用語辞書（例えば、生物用語辞書など）を用意しておき、文書からその用語辞書に合致する部分を抜き出し、GUIを通してユーザが指定した要求に見合った情報を提供する。ユーザは、興味ある分野について効率の良い情報収集が行えるわけである。しかしながら、このようなシステムにおいて、ある一表記しか保持していない用語辞書を用いて情報抽出を行った場合には、表記揺れが生じている部分が抽出結果から漏れるという問題が生じる。例えば、“leukocyte”が用語辞書に登録されていて、その表記揺れである“leucocyte”が文書に表れた場合には、“leucocyte”と“leukocyte”は、同一の事物を指すにも関わらず、“leucocyte”で書かれた情報は抽出されず漏れてしまうことになる。
【０００３】
このような問題に対応するには、表記揺れに対応した辞書を構築し、表記揺れに対応した辞書を備えた情報検索や情報抽出のシステムにする必要がある。表記揺れに対応した辞書は、表記揺れの用語を事前に元の用語の同義語として登録しておき、表記揺れに対応した辞書を備えたシステムでは、情報抽出する際に表記揺れの用語も一緒に使って抽出を行う。前記の例で言えば、“leucocyte”を“leukocyte”の同義語として登録しておき、“leucocyte”の入力に対して“leucocyte”と“leukocyte”で抽出を行う。
【０００４】
表記揺れに対応した辞書は、一般に手作業あるいは計算機を使って辞書の見出し語と表記揺れの用語との対応付けを行い、得られた表記ゆれの用語を辞書へ登録することにより

作成する。計算機を使って見出し語と表記揺れの用語との対応を取る特開平7-73197号公報「異表記語辞書作成支援装置」では、索引語内の用語同士の類似性を判断することによって表記揺れの用語を収集している。また、特開2003-288366号公報「類似テキスト検索装置」では、用語それぞれのＮグラム要素の一致を求める方法によって類似度の計算を行い、表記揺れを吸収した形で用語のマッチングをさせている。ここで、Ｎグラムとは、用語の連接する部分文字列のデータ形式（用語の索引）である。部分文字列の文字数をN(自然数)で指定する。３グラムの場合、例えば、用語“NICAA”に対して、用語を連接する3文字ずつの要素に分け、“NIC”，“ICA”，“CAA”という用語の索引が作成される。また、Ｎグラムによる類似度の計算とは、両方の文字列に共通して含まれるＮ文字の部分文字列を求める。次に、共通する部分文字列に対して重みを設定する。そしてこの重みをすべての一致する部分に関して加算する。この加算して得られた総和が、Ｎグラムによる類似度となる。

【０００５】

【特許文献１】特開平7-73197号公報

【特許文献２】特開2003-288366号公報

【発明の開示】

【発明が解決しようとする課題】

【０００６】

表記揺れに対応した辞書の作成を手作業で行う方法では、見出し語に対応する表記揺れをすべて見つけだし登録するのは困難である。一方、特開平7-73197号公報の方法は、検索のための用語を集めた索引語の中から順に用語を取り出し、索引語の残りの用語を比べて類似度を算出し、類似度が予め定めた値以上のものを表記揺れの用語（異表記語）として取り出す。ここでは、文字列同士の対応付けをＬＣＳ（Longest Common Subsequence: 最長共通部分文字列）法や、ヘッケル(Heckel)法などによって行い、対応付けの終わった文字列の対に対して、一致文字列長、不一致文字列長、一致区分数などにより、一致文字列長が長いほど、不一致文字列長が短いほど、類似度が高いなどの評価を行い、二つの文字列の類似度を数値化している。しかし、このような類似度の算出方法は、索引語の量が増えると文字列同士の組合せ数が増え、用語の文字列長が長くなると文字列同士の対応付けが複雑になり、どちらの場合も計算量が膨大になり、計算時間の観点から現実的ではない。また、文字列同士の長さの差が大きいと異表記とはいえないため、類似文字列の長さが違いすぎるものは除去する方法を取っているが、類似文字列を探し出した後での絞り込みは非効率的である。

【０００７】

特開2003-288366号公報では、テキストの類似度を算出するために、テキストそれぞれのＮグラムの要素一致度を演算し、一致度の高いものを類似テキストとしている。例えば、見出し語“windows”に対して、“winodws”と“windows2000”という２つの用語があった場合、表記揺れと考えられるものは“winodws”の文字列であるが、この方法では“windows”に対しては“win”，“ind”，“ndo”，“dow”，“ows”という３グラム要素の索引、“winodws”に対しては“win”，“ino”，“nod”，“odw”，“dws”という3グラム要素の索引、“windows2000”に対しては“win”，“ind”，“ndo”，“dow”，“ows”，“ws2”，“s20”，“200”,“000”という３グラム要素の索引を生成し、“windows”については類似度1、“windows2000”には類似度5を与える。従って、“winodws”より“windows2000”の文字列の類似度が高く出てしまう。

【０００８】

本発明は、着目する用語（例えば、辞書の見出し語）を中心として文書に起こりうる表記揺れを効率的に漏れなく収集する手段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】

【０００９】

本発明においては、大規模な用語集合の中から表記揺れ候補と考えられる用語を予め選別しておき、表記揺れ候補となった用語に対してコストを調整した編集距離を測ることに

より、表記揺れの候補となった用語の中から表記揺れと考えられる用語を収集する。

【００１０】

検索語として与えられた用語の表記揺れを抽出する本発明のシステムは、テキスト文書から用語の集合を収集する用語収集部と、用語収集部によって収集された用語の集合の中から検索語に類似した用語群を検索する類似用語検索部と、類似用語検索部によって検索された用語群の中から検索語の表記揺れを抽出する表記揺れ検索部とを備え、類似用語検索部は、１文字ずつずらした隣接する所定長の部分文字列の共有度を基準にして、比較する２つの用語の類似度を判定し、表記揺れ検索部は、検索語との編集距離の総コストが与えられた閾値より小さい用語を検索語の表記揺れとして抽出する。

【発明の効果】

【００１１】

本発明によると、表記揺れを収集する作業を少ない労力で精度良く（漏れなく）行うことができ、この表記揺れまで含めて情報抽出を行うと抽出結果に表記揺れが存在した場合にでも漏れなく情報を集めることができる。

【発明を実施するための最良の形態】

【００１２】

本発明は、表記揺れ辞書を構築する際に有効であるが、本発明の用途はこれに限定されないので、ここでは本発明のコアとなる部分について詳細を述べ、実施例にて手段の用途を説明することにする。

【００１３】

本発明では、着目する用語に対する表記揺れの候補をまず収集し、収集された候補の中から更に表記揺れを選別する。具体的には、次に示す通りである。説明のために、“iccar”という用語に対する表記揺れを収集する場合を例に取る。

【００１４】

まず初めに、表記揺れを探し出す対象となる用語を用意する。この場合は、上記にも述べたように“iccar”を用意する。次に、着目する用語がよく現れる分野の文書データから予め既存の方法を用いて、文書データ中の用語を切り出しておく。ここで既存の方法で切り出される用語とは、一例を挙げれば文書中に現れる名詞などが考えられる。例を使って説明すると、“iccar”は生物学の分野でよく現れるので、生物学の分野の文書から用語を切り出してきて、“ICCAR”，“ICAA”，“aar”，“Schaar”，“CaARN1”，“alphalaAR”などを収集する。

【００１５】

次に、切り出された用語の集合から、着目する用語に類似した用語（表記揺れの候補の用語）を収集する。この際、類似している順に、ユーザがパラメタkに設定した数だけ収集する。表記揺れの候補の用語を収集するための、類似度の計算方法としては、着目する用語と既存の方法によって切り出された用語のそれぞれに対して、文字列長による索引を含んだＮグラムによる類似度の計算方法を用いる。

【００１６】

ここで、類似度の計算方法として特開2003-288366号公報の方法と異なるのは、単なるＮグラムではなく、文字列長による索引を含んだＮグラムを用いる点である。文字列長による索引を含んだＮグラムとは、図７に示すようなものである。例えば、用語“ICCAR”について、３グラムの索引、“[IC”，“ICA”，“CAA”，“AAR”“AR]”と文字列長の索引“%5”を持つ。ここで“[“ ,“]”は文字列の先頭と末尾を示す記号である。

【００１７】

類似度の計算方法は、共通する索引に対して重みを設定する。そしてこの重みをすべての一致する部分に関して加算する。この加算して得られた総和が、文字列の類似度となる。重みを１として計算すると“ICCAR”と“ICCA8”の類似度は、3と文字列長の類似度１となる。ここでNグラムが一致した場合の重みを、例では１として考えたが、例えば特異的な文字を含むNグラムの索引が一致する場合には、重みを高くするようなことも考えられる。つまりシステムでどのような文字列をより類似すると考えるかによって重みを変え

ることも考えられる。

【００１８】

表記揺れの候補の用語としては、着目する用語の文字数±mの文字数を持つ用語を収集する。パラメタmはユーザによって設定することができる。長さによる制約を行う方法とは次のようである。着目する用語に文字列数の許容度による索引（例えば、文字数4に対して±2の許容度の索引を作る場合には、%2，%3，%4，%5，%6）を生成させ、既存の方法によって切り出された用語にも文字列長の索引（例えば、文字数４であれば、%4の索引）を生成させる。Ｎグラムを用いた類似度の計算と同様に共通する索引を持つ場合には重みを与え、文字列同士の重みを加算することによって文字列長の類似度の計算を行う。文字列長の許容範囲内に収まる用語であればこの文字列長の類似度は“１”となる。そこで、文字列の類似度が高く、文字列長の類似度も1となるものを集めることによって長さの制約を充たし、着目する用語に類似した用語を収集できる。例えば“iccar”について、３グラムかつ文字列数の許容度2で索引を生成させると、“[ic”，“icc”，“cca”，“car”，“ar]”，“%3”，“%4”，“%5”，“%6”，“%7”が生成され、切り出した語の中にある用語の“car”との間で類似度を測ると、索引は“[ca”，“car”，“ar]”，“%3”なので、類似度は2と文字列長の類似度１となる。

【００１９】

長さによる制約をつけて類似文字列の候補を収集する理由としては、表記揺れによって増減する文字は高々数個と考えられるからで、長さによる制約によって特開2003-288366号公報で問題となるような、表記揺れではないが類似する用語まで収集してしまう可能性を排除することができる。

【００２０】

このようにして、類似度の計算を行い、パラメタkによって設定された数だけ、文字列長の類似度が１かつ類似度の高いものから順に表記揺れの候補の用語を集める。集められた表記揺れの候補の用語には、表記揺れの用語と表記揺れではなく単に類似した単語とが混じっている。そこで、表記揺れの候補の用語から、更に表記揺れだけに絞込みをかけるために、見出し語と表記揺れの候補の用語の間で編集距離を測る。

【００２１】

編集距離とは、通常、一方の文字列から他方の文字列を得るために行う、文字の操作（挿入・削除・置換）の回数を指す。ところが、ある文字列の置換によって全く別の事物を指すことが起こったり、記号によっては挿入されても事物は変わらないなど、文字や操作の種類による重要度の違いが見られるため、表記揺れを収集する際には、このような文字や操作の種類によってコストを変えた編集距離を用いたほうが、表記揺れの場合の編集距離を低くすることができ、表記揺れをクローズアップさせることができる。

【００２２】

そこで、本発明では操作の重みを、表記揺れと考えられる文字の置換や挿入、削除については低く設定し、表記揺れと考えられない操作については高く設定する。コストの設定については、例えば図１０に示したように、文字列間で数字の置換があった場合には表記揺れとは考えにくいので高めのコストである100を与えたり、大文字小文字の置換については、表記揺れと考えられるので低めのコスト、例えば10を与えて計算する。このことによって表記揺れの候補の用語のうち、表記揺れによって生じた用語については、編集距離の総コストが低くなる。

【００２３】

“iccar”と“ICC-u”の編集距離の計算を図１０のコスト表を用いて行うと、90となる。図１１に編集距離の計算の動きを説明する。行列の$C_{0\cdots|x|,0\cdots|y|}$にコストが入力されている。|x|は文字列の長さを表し、x_iはi番目の文字を示している。C_{ij}は$x_{1\cdots i}$と$y_{1\cdots j}$の間で計算される最小コストが入力されている。cは図１０に表されているような操作に関連するコストを示している。

【００２４】

【数１】

$$C_{i,0} = i*50$$

$$C_{0,j} = j*50$$

$$C_{i,j} = if(x_i = y_j) \quad then \quad C_{i-1,j-1} \quad else \quad c + \min(C_{i-1,j}, C_{i,j-1}, C_{i-1,j-1})$$

【００２５】

行列上で右下に得られたコストが編集距離の総コストとなる。

予め設定した閾値よりも総コストが低くなった場合に、その用語を着目する用語の表記揺れとする。閾値はユーザによって設定される。

【実施例１】

【００２６】

表記揺れに対応した辞書を構築する際の実施例を示す。ユーザによって、表記揺れを収集する対象となるマスター辞書や表記揺れを収集するテキストやパラメタが設定され、出力として表記揺れに対応した辞書が生成される。辞書の見出し語についてそれぞれ、テキストから表記揺れを収集し、辞書へ表記揺れを登録していき、全体として表記揺れに対応した辞書とする。

【００２７】

図１に、表記揺れ対応辞書作成システムの全体のシステム構成例を示す。本システムは、クライアント側計算装置C、サーバ側計算装置S、および通信ネットワークNより構成される。クライアント側計算装置とサーバ側計算装置が同一計算装置であって通信ネットワークを必ずしも用いない構成も可能である。必要に応じて印刷装置Prnも用いる。

【００２８】

クライアント側計算装置Cは演算手段C1と主記憶手段C2、補助記憶手段C3、入力手段としてのキーボードC41やマウスC42、更に表示手段C5などから構成される。主記憶手段C2では、クライアント管理手段P01が稼動し、表示手段C5上にGUIが表示されるとともに、クライアント側計算装置Cにおける処理全体を統括する。

【００２９】

サーバ側計算装置S側も同様に演算手段S1、主記憶手段S2、補助記憶手段S3、キーボードS41、マウスS42および表示手段S5などから構成される。サーバ側計算装置Sの主記憶手段S2では、以下に詳細を示す処理手段群が稼動する。これらの処理は、一時的なデータ2として、検索要求21、パラメタ22を主記憶手段S2上に動的もしくは固定的に確保して利用する。

【００３０】

サーバ側計算装置Sの補助記憶手段S3には、1次データ3となるべきテキストデータ31や辞書32、それから加工されて各種処理で参照される、２次加工データ群4が格納される。また、更に加工されて各種処理で参照されるデータが３次加工データ群5として格納される。２次加工データ群4には、テキスト31から切り出した用語41が含まれる。３次加工データ群5には、用語41から生成されるＮグラムデータ（用語と用語のＮグラムのデータ）51などが含まれる。

【００３１】

図２は、辞書構築などの要求、パラメタ設定を行うユーザインターフェースの一例である。図１におけるクライアント側計算装置のGUIの主画面11は、表記揺れを探し出す元となる見出し語が格納されている、マスター辞書を入力（指定）する入力部111、表記揺れに対応した辞書を出力する場所を指定する出力辞書格納指定部112、テキスト指定部（表記揺れを抽出する文書を指定する部分）113、表記揺れの候補数などのパラメタ設定部114、実行ボタン115から構成される。パラメタ設定部114では、表記揺れ候補の文字列長が見出し語の文字列長に対してどれくらい相違してもよいかを表す文字列長の許容度、表記揺

れの候補数、Ｎグラムを生成する際にテキストを連接する何文字ずつの要素に分けるかの指定、編集距離の総コストの閾値等を指定する。

【００３２】

図３は、サーバ側計算装置における処理手段全体の構成例を示す図である。サーバ側計算装置Sにおける処理の全体を統括するのがサーバ側管理手段P02であり、それから直接呼び出されるのが、テキストデータ31から用語を収集する用語収集手段P11、部分文字列の索引を作成する索引作成手段P12、部分文字列の共有度によって類似文字列を検索する類似文字列検索手段P13、文字列間編集距離によって表記揺れを検索する表記揺れ検索手段P14である。更にその下の要素的な処理手段として、文字列長による制約部P21、部分文字列による共有度をスコア付けし文字列を序列化する文字列序列化部P22、文字列間編集距離計算部P23を備える。索引作成手段P12によって図７のようなデータ51が生成される。

【００３３】

図４により表記揺れ収集処理を説明する。左のラインがユーザ操作の流れ、中央のラインがクライアント側計算装置での処理の流れ、右のラインがサーバ側計算装置での処理の流れを示している。はじめにユーザの操作として、主画面の入力辞書指定部111（図２）において辞書の選択操作E111を行い、出力辞書の格納場所指定部112において辞書の出力場所の設定操作E112を行い、続いて表記揺れを収集するテキストを選択する操作E113をテキスト指定部113において行い、パラメタ設定部114において検索数などのパラメタ値の設定操作E114を行い、実行ボタン115を押して表記揺れ収集の実行指示E115を行う。

【００３４】

それをうけてクライアント側管理手段P01は、辞書、テキスト、パラメタ類等をLANやインターネットなどの通信ネットワークN（図１）を通じてサーバ側計算装置Sで稼動しているサーバ側管理手段P02に伝える（E12）。クライアント側計算装置とサーバ側計算装置が同一の場合にはプロセス間通信手段などによって伝える。

【００３５】

サーバ側管理手段P02（図３）は受け取った作業要求に基づき、表記揺れ抽出手段Pにテキスト、辞書、パラメタ類を与える。表記揺れ抽出手段では、受け取ったテキストデータ31から、用語収集手段P11によって用語を収集し、２次加工データ41を生成する。次に、P12の索引作成手段によって、２次加工データ41を更に加工し、用語-索引データ51を生成させる。次に、辞書32の各見出し語に対して、類似文字列検索手段P13によって、用語-索引データ51を参照しながら、部分文字列共有度によって類似文字列を検索する。その際、文字列長による制約部P21で文字列長による制約を行うことによって、ユーザが設定した文字列長の許容度の範囲で、類似文字列を検索する。文字列序列化部P22によって部分文字列による共有度をスコア付けして文字列を序列化し、類似度の高いものを表記揺れの候補とする。各見出し語に対して表記揺れの候補として得られたものを更に表記揺れ検索手段P14によって文字列間編集距離を参照して表記揺れを選択する。表記揺れとして得られたものを各見出し語の表記揺れとして辞書へ登録し、結果となる表記揺れ対応辞書を得る（E13,E14）。

【００３６】

それを再び、ネットワークやプロセス間通信によりクライアント管理手段P01に伝える（E15）。クライアント管理手段P01では返ってきた辞書を出力辞書格納指定部112で指定された格納場所へ格納する（E16）。

【００３７】

図５は、用語収集手段P11が行う処理の詳細である。処理はテキスト31から用語収集手段P11によって用語の収集を行い、２次加工データである用語の集合41として格納する。ここで、テキストデータ31から収集された用語の集合とは、一例を挙げれば文書中に現れる名詞の集合である。

【００３８】

図６は、テキストから切り出された用語の集合41から索引作成手段P12が行う処理である。用語の集合41から索引作成手段P12によって３次加工データである用語-索引のデータ

51が作られる。図７は部分文字列による索引のデータ例であり、Ｎグラムのパラメタを３とした場合の部分文字列の索引を示している。例えば、用語“ICAA”に対して、テキストを連接する3文字ずつの要素に分け、“[IC”,“ICA”,“CAA”,“AA]”という索引が作成されている。ここで“[“ ,“]”は文字列の先頭と末尾を示す記号である。また、“％”の後に文字列長を付加した索引を持つ。この文字列長の索引を持つことがデータの特徴となっている。

【００３９】

図８は、類似文字列検索手段P13が行う処理の詳細である。辞書の用語32を入力として、その用語に対する部分文字列の索引を索引作成手段P12を用いて生成する。表記揺れで増減する文字列は高々±mであるので、その文字列長±mの索引を生成する。mはユーザによって指定される。文字列長が5である文字列“iccar”に対して３グラムで許容度±1の索引を生成すると、“[ic”、“icc”、“cca”、“car”、“ar]”、“％4”、“％6”となる。次に、３次加工データの用語－索引データ51を参照し、見出し語とテキストデータ31から切り出された用語41との類似度を計算する。類似度の計算方法は、共通する索引に対して重みを設定し、この重みをすべての一致する部分に関して加算する。この加算して得られた総和が、文字列長による索引を含んだＮグラムによる類似度となる。“ICCAR”と“ICCA8”の類似度は、3で文字列長の類似度は１となる。

【００４０】

そして、文字列長の類似度が１で、文字列の類似度の高い方から順に上位k個を類似文字列として出力する。kはユーザによって指定される。これらの処理は辞書の各見出し語に対して行われる。

【００４１】

図９は、表記揺れ検索手段P14による文字列間編集距離を用いた処理の詳細である。類似文字列を入力として、入力辞書の用語との文字列間編集距離を測る。編集距離の計算では、表記揺れと考えられる文字列の挿入、置換、削除についてはコストを低く設定するような重み付きの編集距離を用いる。編集距離が近かった文字列で編集距離の総コストがある閾値以下の用語を、入力辞書の用語の表記揺れの文字列として得る。これらの処理は辞書の各見出し語に対して行われる。

【００４２】

図１０は編集距離計算のコストの一例を示すテーブルである。本例では、ハイフンの挿入・削除、大文字小文字の置換については、表記揺れと考えられるのでコストを低く設定し、数字の置換や-x-（ハイフン、文字、ハイフン）の置換・挿入・削除については、表記揺れとは考えられないのでコストを高く設定してある。

【００４３】

図１２は、表記揺れを収集した例を示す図である。本例は、図１２（ａ）に示すように、見出し語“iccar”に対して３グラムと4グラムの部分文字列の索引を作り、文字列の許容度m=1で、表記揺れの候補k=4とし、編集距離の閾値を60として表記揺れ“ICCAR”を収集した例を示す。テキストから収集された用語は、“ICCAR”,“ICAA”,“aar”,“Schaar”,“CaARN1”,“alphalaAR”であるとした。テキストから収集された各用語に対して文字列長の索引を与え、３グラムと４グラムの共有度から類似度を計算すると、図１２（ｂ）に示すようになる。ここで表記揺れの候補として、文字列長の類似度が１で類似度の高いものから順に４個の用語を選択すると、図１２（ｃ）のようになる。これら４個の用語に対して、図１０に示したコストを用いて編集距離を計算する。編集距離の閾値を60以下という条件を満たす用語“ICCAR”が表記揺れとして抽出される。

【実施例２】

【００４４】

文書を検索する場合には、ユーザが興味ある事柄に関連する用語（検索語）を入力し、文書に付加されている索引語とユーザの入力した用語とを照合し、索引語と検索語が合致した場合、その索引語を持つ文書を結果として提示する。その際、文書に付加されている索引語とユーザ入力の用語の間に表記揺れがあると、検索結果に漏れが生じる。文書につ

いている索引語とユーザ入力の用語の間で本発明の手段を用い、ユーザ入力の用語の表記揺れとして考えられる索引語が付加されている文書も、検索結果として出力するシステムについて説明する。

【００４５】

全体の構成は図１の構成と同様であるが、サーバ側の補助記憶手段Ｓ3には一次データ群としてテキストデータ33が格納されており、２次加工データ群としてテキストデータの索引語42が格納されていて、３次加工データ群として索引語のＮグラムデータ52が格納される。

【００４６】

図１３は、検索要求、パラメタ設定を行うユーザインターフェースの一例である。クライアント側計算装置のGUIの主画面11は、検索語を入力する部分211、表記揺れの候補数などのパラメタ設定部212、実行ボタン213、出力結果表示部214から構成される。パラメタ設定部212では、表記揺れの候補の文字列長を見出し語に対してどれくらいの許容度を持たせるかを表す文字列長の許容度、表記揺れの候補数、Ｎグラムを生成する場合に、テキストを連接する何文字ずつの要素に分けるかを指定できるようにしておく。また、編集距離の総コストの閾値も指定できるようにしておく。

【００４７】

図１４により、処理の流れを説明する。左のラインがユーザ操作の流れ、中央のラインがクライアント側計算装置での処理の流れ、右のラインがサーバ側計算装置での処理の流れを示している。はじめにユーザの操作として主画面の検索語入力部211（図１３）において、検索語の入力E211を行い、パラメタ設定部212においてパラメタ値の設定操作E212を行い、実行ボタン213を押して表記揺れ収集の実行指示E213を行う。

【００４８】

それをうけてクライアント側管理手段P01は、辞書、テキスト、パラメタ類等をLANやインターネットなどの通信ネットワークN（図１）を通じてサーバ側計算装置Sで稼動しているサーバ側管理手段P02に伝える（E22）。クライアント側計算装置とサーバ側計算装置が同一の場合にはプロセス間通信手段などによって伝える。サーバ側管理手段P02は受け取った作業要求に基づき、表記揺れ抽出手段に検索語、パラメタ類を与えて、表記揺れ抽出手段では、受け取ったテキストデータ32から、用語収集手段P11によって、索引語を収集し、２次加工データ42を生成する。次に、索引作成手段P12によって、２次加工データ42を更に加工し、索引語-索引データ52を生成させる。次に、検索語に対して、類似文字列検索手段P13によって、用語-索引データ52を参照しながら、部分文字列共有度に基づいて類似文字列を検索する。その際、文字列長による制約部P21で文字列長による制約を行うことによって、ユーザが設定した文字列長の許容度の範囲で、類似文字列を検索する。文字列序列化部P22では、部分文字列による共有度をスコア付けし、類似度の高いものを表記揺れの候補とする。表記揺れの候補として得られたものを更に表記揺れ検索手段P14では、文字列間編集距離に基づいて表記揺れを選択する。表記揺れとして得られた用語が索引語となっている文書を検索結果とする（E23,E24）。それを再び、ネットワークやプロセス間通信によりクライアント管理手段P01に伝える（E25）。クライアント管理手段P01では、返ってきた結果を出力結果表示部214へ表示する（E26）。ユーザは結果をチェックする（E27）。

【００４９】

図１５は、用語収集手段P11が行う処理の詳細である。用語収集手段P11はテキスト32から用語の収集を行い、２次加工データである索引語の集合42として格納する。

【００５０】

図１６は、テキストから索引語のデータ42から索引作成手段P12が行う処理である。索引語の集合42から索引作成手段P12によって３次加工データである索引語-索引52のデータが作られる。

【００５１】

図１７は、類似文字列検索手段P13が部分文字列共有度を用いて行う処理の詳細である

。検索語を入力として、その用語に対する部分文字列の索引を索引作成手段P12を用いて生成する。表記揺れで増減する文字列は高々±mであるので、その文字列長±mの索引を生成する。mはユーザによって指定される。文字列長が5である文字列“iccar”に対して許容度±1の索引を生成すると、“[ic”、“icc”、“cca”、“car”、“ar]”、“%4”、“%6”となる。次に、３次加工データの索引語-索引データ52を参照し、検索語と索引語42との類似度を計算する。類似度の計算方法は、共通する索引に対して重みを設定し、この重みをすべての一致する部分に関して加算する。この加算して得られた総和が、文字列長による索引を含んだＮグラムによる類似度となる。“ICCAR”と“ICCA8”の類似度は、3で文字列長の類似度は１となる。そして、文字列長の類似度が１で類似度の高い順に上位k個を類似文字列として出力する。kはユーザによって指定される。

【００５２】

図１８は、表記揺れ検索手段P14による文字列間編集距離を用いた処理の詳細である。類似文字列を入力として、検索語との文字列間編集距離を測る。編集距離の計算では、表記揺れと考えられる文字列の挿入、置換、削除についてはコストを低く設定するような重み付きの編集距離を用いる。編集距離が近かった文字列で編集距離の総コストがある閾値以下の用語を検索語の表記揺れの文字列として得る。

【図面の簡単な説明】

【００５３】

【図１】表記揺れ対応辞書作成システムのシステム構成例を示す図。

【図２】表記揺れ対応辞書作成を行うユーザインターフェースの例を示す図。

【図３】サーバ側計算装置における処理手段の全体構成例を示す図。

【図４】表記揺れ対応辞書作成の処理の流れを示すフローチャート。

【図５】用語収集手段が行う処理の詳細を示す図。

【図６】索引作成手段が行う処理の詳細を示す図。

【図７】部分文字列の索引作成手段で生成されるデータの例を示す図。

【図８】類似文字列検索手段が行う処理の詳細を示す図。

【図９】表記揺れ検索手段による処理の詳細を示す図。

【図１０】文字列間編集距離の操作に対するコストの例を示す図。

【図１１】文字列間編集距離の計算推移例を示す図。

【図１２】表記揺れの収集例を示す図。

【図１３】ユーザインターフェースの一例を示す図。

【図１４】表記揺れ収集処理の説明図。

【図１５】用語収集手段が行う処理の詳細を示す図。

【図１６】索引作成手段が行う処理の詳細を示す図。

【図１７】類似文字列検索手段が行う処理の詳細を示す図。

【図１８】表記揺れ検索手段が行う処理の詳細を示す図。

【符号の説明】

【００５４】

C：クライアント側計算装置

S：サーバ側計算装置

N：通信ネットワーク

P11：用語収集手段

P12：索引作成手段

P13：類似文字列検索手段

P14：表記揺れ検索手段

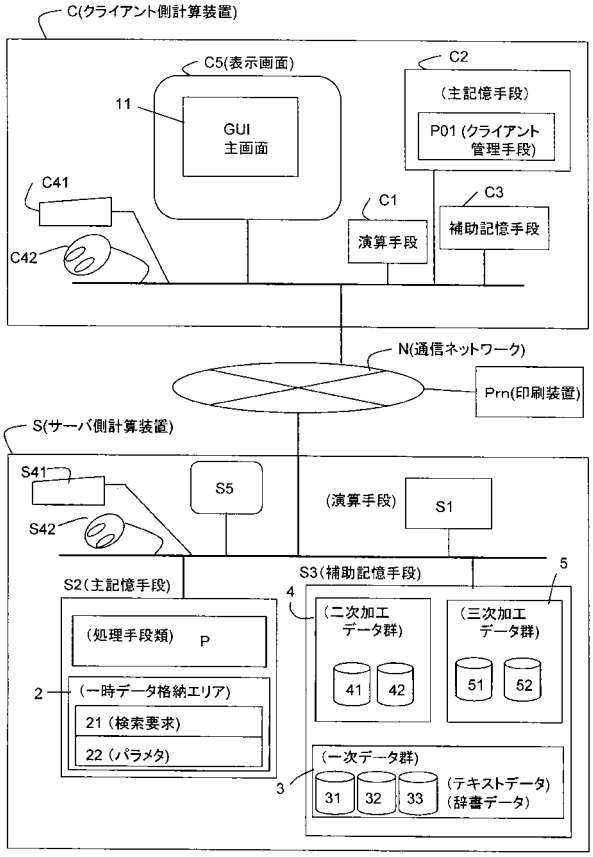
P21：文字列長による制約部

P22：文字列序列化部

P23：文字列間編集距離計算部

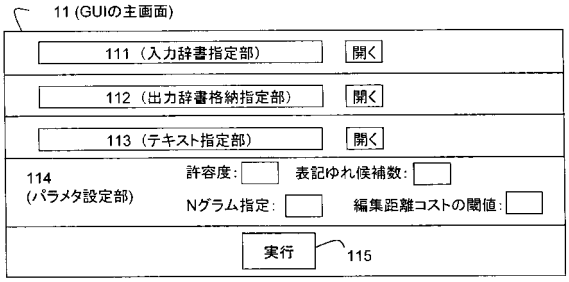
【図１】

図1



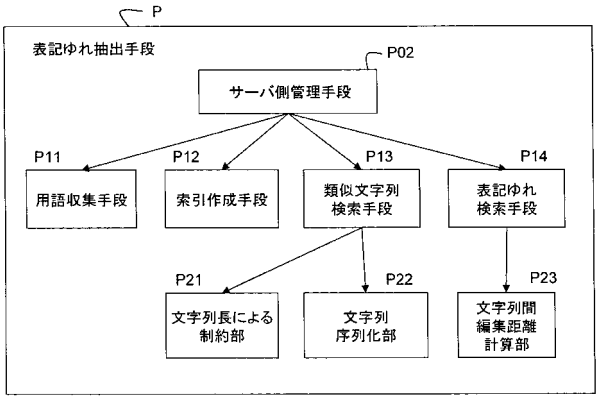
【図２】

図2



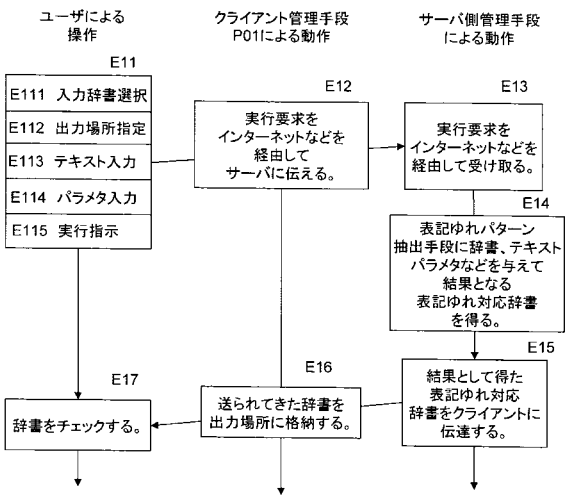
【図３】

図3



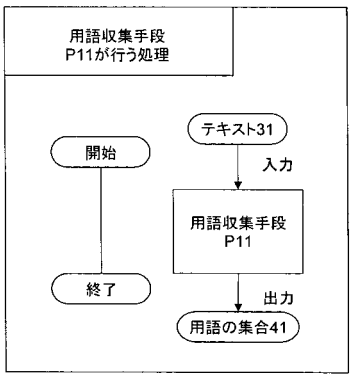
【図４】

図4



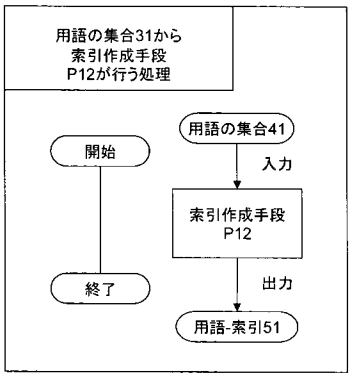
【図５】

図5



【図６】

図6



【図７】

図7

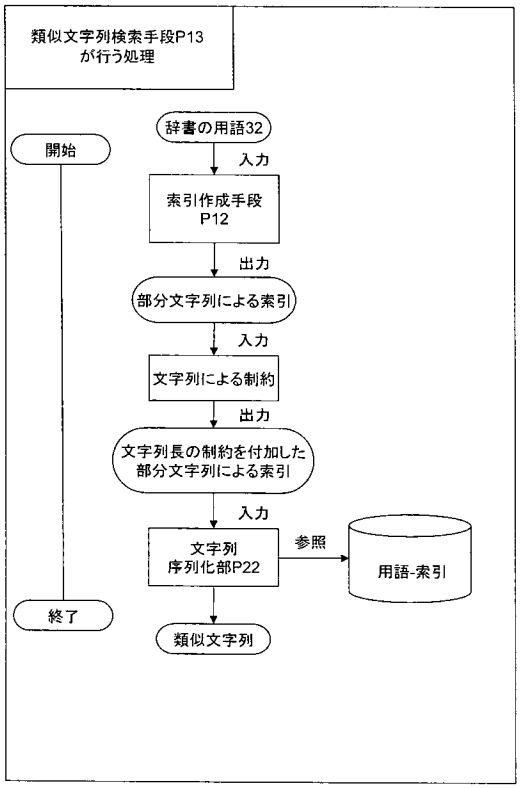
51

用語-索引

用語	部分文字列の索引	文字列長の索引
ICAAR	[IC ICA CAA AAR AR]	%5
ICAA	[IC ICA CAA AA]	%4
aar	[aa aar ar]	%3
Schaar	[Sc Sch cha haa aar ar]	%6
CaARN1	[Ca CaA aAR ARN RN1 N1]	%6
alpha1aAR	[al alp lph pha ha1 a1a 1aA aAR AR]	%9

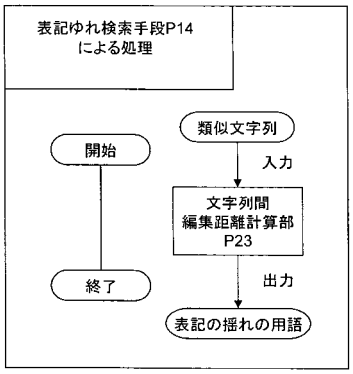
【図８】

図8



【図９】

図9



【図１０】

図10

編集距離計算のコスト

操作	コスト
置換、挿入、削除	50
-x-の置換、挿入、削除	100
-の置換、挿入、削除	10
数字の置換	100
大文字小文字の置換	10

＊ここで、-はスペースやハイフン、コンマなど

【図１１】

図11

		i	c	c	a	r
	0	50	100	150	200	250
I	50	10	60	110	160	210
C	100	60	20	70	120	170
C	150	110	70	30	80	130
-	160	120	80	40	40	90
u	210	170	130	90	90	90

【図１２】

図12

(a)

ターム	文字列長	パラメタm=1とした場合の文字列長を付加した索引
icaar	5	[ic ica caa aar ar] [ica icaa caar aar] %4 %5 %6

文字列長の索引

(b)

ターム	タームと文字列長の索引	類似度	文字列長の類似度
ICAAR	ICAAR %5	9	1
ICAA	ICAA %4	5	1
aar	aar %3	2	0
Schaar	Schaar %6	3	1
CaARN1	CaARN1 %6	3	1
alpha1aAR	alpha1aAR %9	3	0

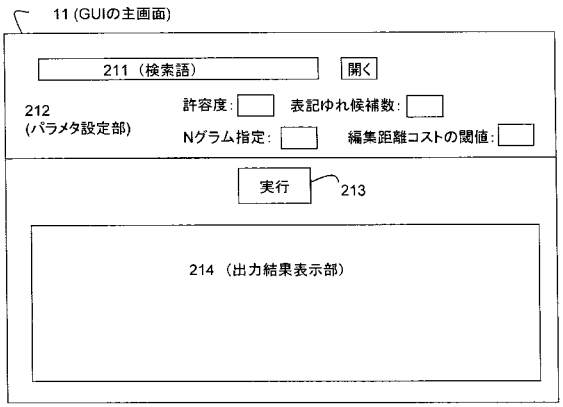
(c)

ターム	編集距離
ICAAR	50
ICAA	90
Schaar	100
CaARN1	180

→表記ゆれとして抽出

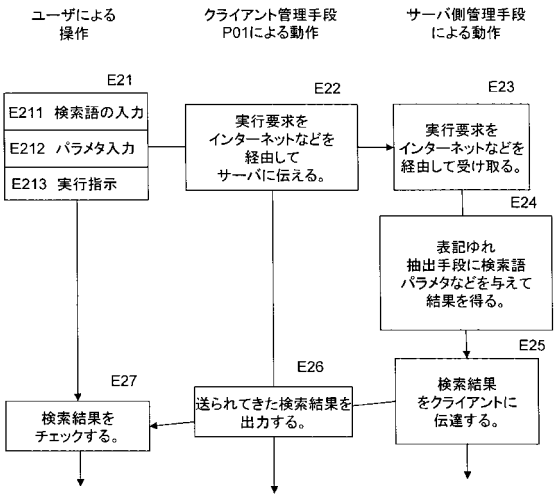
【図１３】

図13



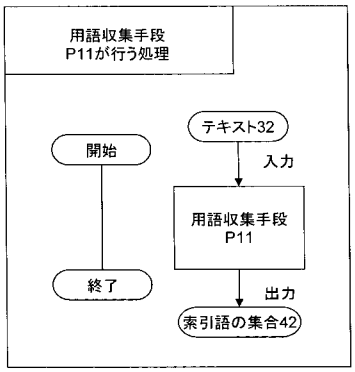
【図１４】

図14



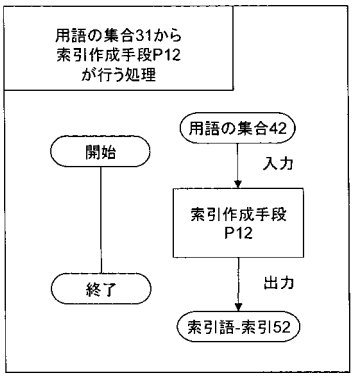
【図１５】

図15



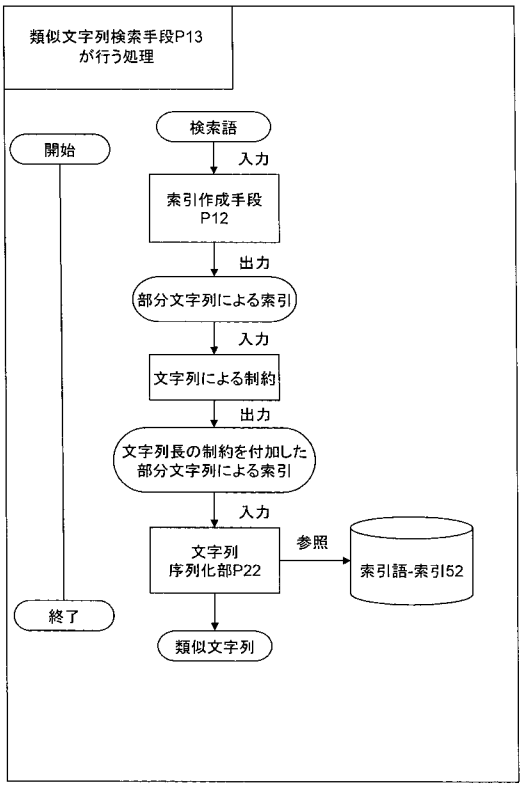
【図１６】

図16



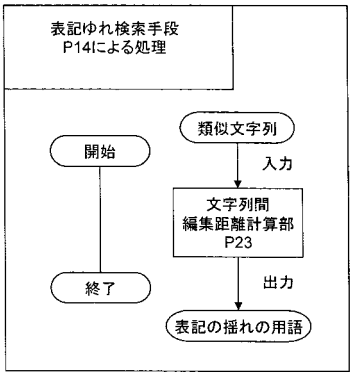
【図１７】

図17



【図１８】

図18



フロントページの続き

Ｆターム(参考) 5B075 KK33 KK37 ND03 NK35 QM02 QM05 UU06