(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4186321号
(P4186321)
(45)発行日　平成20年11月26日(2008.11.26)　(24)登録日　平成20年9月19日(2008.9.19)

(51)Int.Cl.　F I
G06F 17/21 (2006.01)　G06F 17/21 550A

請求項の数 7　(全 23 頁)

(21)出願番号　特願平11-212652
(22)出願日　平成11年7月27日(1999.7.27)
(65)公開番号　特開2001-43220(P2001-43220A)
(43)公開日　平成13年2月16日(2001.2.16)
審査請求日　平成18年3月7日(2006.3.7)

(73)特許権者　000002185
ソニー株式会社
東京都港区港南１丁目７番１号
(74)代理人　100067736
弁理士　小池　晃
(74)代理人　100086335
弁理士　田村　榮一
(74)代理人　100096677
弁理士　伊賀　誠司
(72)発明者　長尾　確
東京都品川区東五反田３丁目１４番１３号
株式会社ソニーコンピュータサイエンス
研究所内

審査官　浜岸　広明

最終頁に続く

(54)【発明の名称】文書処理方法及び装置並びに記録媒体

(57)【特許請求の範囲】
【請求項１】
電子文書の形態の文書を処理する文書処理装置の文書処理方法において、
上記文書処理装置が、
上記文書の要約文を作成する要約文作成工程と、
上記作成される要約文中における省略又は置き換えられた語句が存在する場合、これを補完する補完工程とを有し、
上記文書は、複数のエレメントが階層化された内部構造を有し、上記内部構造を示すタグ情報が予め付与されており、
上記要約文作成工程では、上記タグ情報により示される内部構造に基づいた上記エレメントの重要度を算出して上記エレメントに付与し、上記補完工程では、上記タグ情報に基づいて上記補完を行う文書処理方法。
【請求項２】
上記要約文作成工程は、
上記文書の要約文が表示される要約文表示領域の全表示領域内における大きさを可変に設定する設定工程と、
上記設定工程で設定された表示領域の大きさに基づいて上記文書の要約文の長さを決定する決定工程と、
上記決定工程で決定された要約文の長さに基づいて上記要約文表示領域内におさまる長さの上記文書の要約文を作成する工程と

を有して成る請求項１記載の文書処理方法。
【請求項３】
上記補完工程では、
上記タグ情報に基づいて上記作成される要約文中における省略された主語又は目的語が該要約文中に含まれていないとき、元の文書中の対応する主語又は目的語を要約文中に追加するゼロ照応処理を行う請求項１記載の文書処理方法。
【請求項４】
上記ゼロ照応処理では、上記要約文中に、上記省略された主語又は目的語が含まれるか否かを上記タグ情報に基づいて判別し、含まれていないとき上記省略された主語又は目的語を括弧でくくって上記要約文中に追加する請求項３記載の文書処理方法。
【請求項５】
上記補完工程では、
上記タグ情報に基づいて上記作成される要約文中に含まれる被参照エレメントに対応する参照エレメントが上記要約文中に含まれていないときに元の文書中の上記対応する参照エレメントを上記要約文中の被参照エレメントに置き換えを行う請求項１記載の文書処理方法。
【請求項６】
電子文書の形態の文書を処理する文書処理装置において、
上記文書の要約文を作成する要約文作成手段と、
上記作成される要約文中における省略又は置き換えられた語句が存在する場合、これを補完する補完手段とを有し、
上記文書は、複数のエレメントが階層化された内部構造を有し、上記内部構造を示すタグ情報が予め付与されており、
上記要約文作成手段は、上記タグ情報により示される内部構造に基づいた上記エレメントの重要度を算出して上記エレメントに付与し、上記補完工程では、上記タグ情報に基づいて上記補完を行う文書処理装置。
【請求項７】
コンピュータに電子文書の形態の文書を処理する文書処理を実行させる文書処理プログラムが記録されたコンピュータ読み取り可能な記録媒体において、
上記文書処理プログラムは、
上記文書の要約文を作成する要約文作成工程と、
上記作成される要約文中における省略又は置き換えられた語句が存在する場合、これを補完する補完工程とを有し、
上記文書は、複数のエレメントが階層化された内部構造を有し、上記内部構造を示すタグ情報が予め付与されており、
上記要約文作成工程では、上記タグ情報により示される内部構造に基づいた上記エレメントの重要度を算出して上記エレメントに付与し、上記補完工程では、上記タグ情報に基づいて上記補完を行う記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、電子文書を処理する文書処理方法及び装置並びに電子文書を処理する文書処理プログラムが記録された記録媒体に関する。
【０００２】
【従来の技術】
従来、インターネットにおいて、ウィンドウ形式でハイパーテキスト型情報を提供するアプリケーションサービスとしてＷＷＷ（World Wide Web）が知られている。
【０００３】
ＷＷＷは、文書の作成、公開又は共有化の文書処理を実行し、新しいスタイルの文書の在り方を示したシステムである。しかし、文書の実際上の利用の観点からは、文書の内容に

基づいた文書の分類や要約といった、ＷＷＷを越える高度な文書処理が求められている。このような高度な文書処理には、文書の内容の機械的な処理が不可欠である。

【０００４】

しかしながら、文書の内容の機械的な処理は、以下のような理由から依然として困難である。すなわち、第１に、ハイパーテキストを記述する言語であるＨＴＭＬ（Hyper Text Markup Language）は、文書の表現については規定するが、文書の内容についてはほとんど規定しない点、第２に、文書間に構成されたハイパーテキストのネットワークは、文書の読者にとって文書の内容を理解するために必ずしも利用しやすいものではない点、第３に、一般に文章の著作者は読者の便宜を念頭に置かずに著作するが、文書の読者の便宜が著作者の便宜と調整されることはない点が、文書の機械的処理を困難とする理由である。

【０００５】

このように、ＷＷＷは新しい文書の在り方を示したシステムであるが、文書を機械的に処理しないので、高度な文書処理を行うことができなかった。換言すると、高度な文書処理を実行するためには、文書を機械的に処理することが必要となる。

【０００６】

そこで、文書の機械的な処理を目標として、文書の機械的な処理を支援するシステムが自然言語研究の成果に基づいて開発されている。自然言語研究による文書処理として、文書の著作者等による文書の内部構造についての属性情報、いわゆるタグの付与を前提とした、文書に付与されたタグを利用する機械的な文書処理が提案されている。

【０００７】

ところで、近年のコンピュータの普及や、ネットワーク化の進展に伴い、文章処理や、文書の内容に依存した索引などで、テキスト文書の作成、ラベル付け、変更などを行う文書処理の高機能化が求められている。例えば、ユーザの要望に応じた文書の要約や、文書の分類等が望まれる。

【０００８】

すなわち、ユーザは、例えばいわゆるサーチエンジンのような情報検索システムを利用し、インターネットを介して提供される膨大な情報の中から所望の情報を探し出すようにしている。この情報検索システムは、指定されたキーワードに基づいて情報を検索し、検索した情報をユーザに提供するシステムである。ユーザは、提供された情報の中から所望の情報を選択する。

【０００９】

情報検索システムにおいては、このように容易に情報を検索することができるが、ユーザは、検索されて提供された情報を一読して概略を理解し、それが希望する情報であるか否かを判断する必要がある。この作業は、特に、提供された情報の量が多い場合には、ユーザにとって大きな負担となる。そこで、最近、テキスト情報、すなわち文書の内容を自動的に要約するシステムであるいわゆる自動要約文作成システムが注目されている。

【００１０】

自動要約文作成システムは、元の情報、すなわち文書の大意を保持したままテキストの情報の長さや複雑さを減らすことによって、要約文を作成するシステムである。ユーザは、この自動要約文作成システムにより作成された要約文を一読することで、文書の概略を理解することができる。

【００１１】

通常、自動要約文作成システムは、テキスト中の文や単語を１つの単位とし、それに何らかの情報に基づいた重要度を付与して順序付けする。そして、自動要約文作成システムは、上位に順序付けした文や単語を寄せ集め、要約文を作成する。

【００１２】

【発明が解決しようとする課題】

ところで、上述した自動要約文作成システムにおいては、文書から要約文を作成することが可能であるが、作成される要約文の情報量は、文書の情報量等により決定されていた。そのため、自動要約文作成システムにおいては、例えば、作成された要約文が簡略すぎて

ユーザが文書の概略を把握できない場合、ユーザは、より詳細な要約文を参照することができなかった。

【００１３】

また、元の文中の主語の省略されている部分を要約文に取り入れる場合に、省略されている部分が要約文中に含まれていないと、要約文から正確な内容把握が行えないことにもなる。

【００１４】

本発明は、上述の実情に鑑みて提案されたものであり、入力された文書に対して、ユーザの理解が容易で正確な内容の要約文を自動生成し得るような文書処理方法及び装置、並びに文書処理プログラムが記録されてなる記録媒体を提供することを目的とする。

【００１５】

【課題を解決するための手段】

上述の課題を解決するために、本発明は、電子文書の形態の文書を処理する文書処理方法及び装置において、文書の要約文を作成し、作成される要約文中における省略又は置き換えられた語句が存在する場合、これを補完し、文書は、複数のエレメントが階層化された内部構造を有し、内部構造を示すタグ情報が予め付与されており、タグ情報により示される内部構造に基づいたエレメントの重要度を算出してエレメントに付与し、タグ情報に基づいて補完を行うことを特徴としている。

上記補完は、タグ情報に基づいて作成される要約文中における省略された主語又は目的語が該要約文中に含まれていないとき、元の文書中の対応する主語又は目的語を要約文中に追加するゼロ照応処理を行うことが好ましい。

また、上記補完は、タグ情報に基づいて作成される要約文中に含まれる被参照エレメントに対応する参照エレメントが要約文中に含まれていないときに元の文書中の対応する参照エレメントを要約文中の被参照エレメントに置き換えを行うことが好ましい。

【００１６】

ここで、上記省略された主語又は目的語はゼロ照応エレメントと称され、上記要約文中にこのゼロ照応エレメントが含まれているか否かを判別し、含まれていないときに、当該ゼロ照応エレメントを要約文中に括弧でくくって追加することが好ましい。

また、上記被参照エレメントは例えば代名詞又は限定節であり、上記参照エレメントは例えば先行詞である。上記置き換えの際には、要約文中の被参照エレメントに対応する参照エレメントが要約文中に含まれているか否かを判別し、対応する参照エレメントが要約文中に含まれていないとき要約文中の被参照エレメントを対応する参照エレメントに置き換え、上記対応する参照エレメントが要約文中に含まれているとき上記要約文中の上記被参照エレメントを置き換えないことが好ましい。

【００１７】

これによって、要約文中に省略された主語や目的語が全く含まれないことが回避される。また、要約文中に先行詞なしの代名詞や限定節が含まれることが回避される。

【００１８】

【発明の実施の形態】

以下、図面を参照して、本発明に係る文書処理方法及び装置並びに文書処理プログラムが記録された記録媒体の実施の形態について説明する。

【００１９】

本発明の実施の形態としての文書処理装置は、図１に示すように、制御部１１及びインターフェース１２を備える本体１０と、ユーザからの入力を受けて本体１０に送る入力部２０と、外部からの信号を受信して本体１０に送る通信部２２と、本体１０からの出力を表示する表示部３０と、記録媒体３２に対して情報を記録／再生する記録／再生部３１とを有している。

【００２０】

本体１０は、制御部１１及びインターフェース１２を有し、この文書処理装置の主要な部分を構成している。制御部１１は、この文書処理装置における処理を実行するＣＰＵ１３

と、揮発性のメモリであるＲＡＭ１４と、不揮発性のメモリであるＲＯＭ１５とを有している。ＣＰＵ１３は、例えばＲＯＭ１５に記録されたプログラムの手順にしたがって各処理を実行するための制御を行う。ＲＡＭ１４には、ＣＰＵ１３が各種の処理を実行する上で必要なプログラムやデータが一時的に格納される。インターフェース１２は、制御部１１、入力部２０、通信部２２、表示部３０及び記録／再生部３１に接続されている。このインターフェース１２は、制御部１１の制御の下に、入力部２０及び通信部２２からのデータの入力、表示部３０へのデータの送信、記録／再生部３１に対するデータの送受信について、データを送信するタイミングを調整したり、データの形式を変換したりする。

【００２１】

入力部２０は、この文書処理装置に対するユーザの入力を受ける部分であり、例えばキーボードやマウスにより構成される。ユーザは、この入力部２０を用い、キーボードによりキーワードを入力したり、マウスにより表示部３０に表示されている電子文書のエレメントを選択して入力したりすることができる。なお、以下では電子文書を単に文書と称することにする。ここで、エレメントとは文書を構成する要素であって、例えば文書、文、句及び語が含まれる。

【００２２】

通信部２２は、この文書処理装置に外部から通信路、例えば電話回線を介して送信される信号を受信する部分である。具体的には、通信部２２は、例えば、モデム、ターミナルアダプタ等により構成され、電話回線を介してインターネット２３に接続され、インターネットに接続されているサーバ２４にアクセスし、そこから文書等のデータを受信することができるようにされている。このような通信部２２は、外部から送信された複数の文書等のデータを受信し、受信したデータを本体１０に送る。

【００２３】

表示部３０は、この文書処理装置からの文字や画像情報の出力を表示する。表示部３０は、例えば陰極線管（cathode ray tube;CRT）や液晶表示装置（liquid crystal display;LCD）から構成され、例えば単数又は複数のウィンドウを表示したり、文字、図形、又は画像等を表示したりする。

【００２４】

記録／再生部３１は、例えばフロッピーディスク、光ディスク、光磁気ディスクのような着脱可能な記録媒体３２に対してデータの記録及び／又は再生を行う。記録媒体３２には、文書を処理する文書処理プログラムが記録されている。この記録媒体３２には、文書を処理するための電子文書処理プログラムや処理対象とする文書が記録されている。

【００２５】

ハードディスクドライブ３３は、大容量の磁気記録媒体であるハードディスクに対してデータの記録及び／又は再生を行う。

【００２６】

このような文書処理装置は、以下のようにして所望の文書を受信し、表示部３１に表示する。

【００２７】

文書処理装置においては、まずユーザが入力部２０を操作してインターネット２３を介して通信を行うためのプログラムを起動し、サーバ２４（サーチエンジン）のＵＲＬ（Uniform Resource Locator）を入力すると、制御部１１は、通信部２２を制御し、サーバ２４にアクセスする。

【００２８】

これに応じて、サーバ２４は、インターネット２３を介して、文書処理装置の通信部２２に検索画面のデータを出力する。文書処理装置においてＣＰＵ１３は、このデータをインターフェース１２を介して表示部３０に出力し、表示させる。

【００２９】

文書処理装置においては、ユーザが入力部２０を用いてこの検索画面上でキーワード等を入力して検索を指令すると、通信部２２からインターネット２３を介して、サーチエンジ

ンとしてのサーバ２４に対して検索命令が送信される。
【００３０】
サーバ２４は、検索命令を受信すると、この検索命令を実行し、得られた検索結果をインターネット２３を介して通信部２２に送信する。文書処理装置において制御部１１は、通信部２２を制御し、サーバ２４から送信される検索結果を受信させ、その一部を表示部３０に表示させる。
【００３１】
具体的には、ユーザが入力部２０を用いて例えば「ＴＣＰ」というキーワードを入力して検索を指令した場合には、文書処理装置には、サーバ２４から「ＴＣＰ」のキーワードを含む各種情報が送信され、表示部３０に表示される。
【００３２】
続いて、本実施の形態における文書について説明する。本実施の形態においては、文書処理は、文書に付与された属性情報であるタグを参照して行われる。本実施の形態で用いられるタグには、文書の構造を示す統語論的（syntactic）タグと、多言語間で文書の機械的な内容理解を可能にするような意味的（semantic）・語用論的タグとがある。
【００３３】
統語論的なタグとしては、文書の内部構造を記述するものがある。タグ付けによる内部構造は、図２に示すように、文書、文、語彙エレメント等の各エレメントが、通常リンク、参照・被参照リンクにより関連付けられて構成されている。図中において、白丸“○”はエレメントを示し、最下位の白丸は文書における最小レベルの語に対応する語彙エレメントである。また、実線は文書、文、語彙エレメント等のエレメント間のつながり示す通常リンク（normal link） である。破線は参照・被参照による係り受け関係を示す参照リンク（reference link）である。文書の内部構造は、上位から下位への順序で、文書（document）、サブディビジョン（subdivision） 、段落（paragraph）、文（sentence） 、サブセンテンシャルセグメント（subsentential segment） 、・・・、語彙エレメントから構成される。これらのうち、サブディビジョンと段落とは、例えばオプションとして用いられるものである。
【００３４】
一方、意味論・語用論的なタグ付けとしては、係り受け、例えば代名詞の指示対象等を示す統語構造（syntactic structure） に関するタグ付けや多義語の意味のように意味（semantic）の情報を記述するものがある。本実施の形態におけるタグ付けは、ＨＴＭＬ（Hyper Text Markup Language）と同様なＸＭＬ（eXtensible Markup Language）の形式によるものである。
【００３５】
以下にタグ付けされた文や文書の例を示すが、文書へのタグ付けはこの方法に限定されるものではない。また、以下では英語と日本語の文書の例を示すが、タグ付けによる内部構造の記述は、他の言語にも同様に適用することができることは勿論である。
【００３６】
例えば、“Time flies like an arrow.”という文については、下記のようなタグ付けをすることができる。
【００３７】
＜文＞＜名詞句　語義＝“time０”＞time＜／名詞句＞
＜動詞句＞＜動詞　語義＝“fly１”＞flies＜／動詞＞
＜形容動詞句＞＜形容動詞　語義＝“like０”＞like＜／形容動詞＞
＜名詞句＞an＜名詞　語義＝“arrow０”＞arrow＜／名詞＞＜／名詞句＞
＜／形容動詞句＞＜／動詞句＞.＜／文＞
ここで＜文＞、＜名詞＞、＜名詞句＞、＜動詞＞、＜動詞句＞、＜形容動詞＞、＜形容動詞句＞は、それぞれ文、名詞、名詞句、動詞、動詞句、形容詞を含む前置詞句又は後置詞句／形容詞句、形容詞句／形容動詞句のような文の統語構造（syntactic structure） を表している。タグは、エレメントの先端の直前及び終端の直後に対応して配置される。エ

レメントの終端の直後に配置されるタグは、記号“／”によりエレメントの終端であることを示している。エレメントは統語的構成素、すなわち句、節、及び文を示す。なお、語義（word sense）＝“time０”は、語“time”の有する複数の意味、すなわち複数の語義のうちの第０番目の意味を指している。具体的には、語“time”には少なくとも名詞、形容詞、動詞の意味があるが、ここでは語“time”が名詞であることを示している。同様に、語“オレンジ”は少なくとも植物の名前、色、果物の意味があるが、これらも語義によって区別することができる。

【００３８】

本実施の形態に用いられる文書は、図３に示すように、上記図１の表示部３０のウィンドウ１０１に統語構造を表示することができる。このウィンドウ１０１においては、右半面１０３に語彙エレメントが、左半面１０２に文の内部構造がそれぞれ表示されている。このウィンドウ１０１においては、日本語で記述された文書のみならず、英語等の任意の言語で記述された文書についても、統語構造を表示することができる。

【００３９】

このウィンドウ１０１には、一例として、タグ付けにより内部構造を記述された次に示すような文書「Ａ氏のＢ会が終わったＣ市で、一部の大衆紙と一般紙がその写真報道を自主規制する方針を紙面で明らかにした。」の一部が表示されている。この文書のタグ付けの例を次に示す。

【００４０】

＜文書＞＜文＞＜形容動詞句　関係＝“位置”＞＜名詞句＞＜形容動詞句　場所＝“Ｃ市”＞
＜形容動詞句　関係＝“主語”＞＜名詞句　識別子＝“Ｂ会”＞＜形容動詞句　関係＝“所属”＞＜人名　識別子＝“Ａ氏”＞Ａ氏＜／人名＞の＜／形容動詞句＞＜組織名　識別子＝“Ｂ会”＞Ｂ会＜／組織名＞＜／名詞句＞が＜／形容動詞句＞
終わった＜／形容動詞句＞＜地名　識別子＝“Ｃ市”＞Ｃ市＜／地名＞＜／名詞句＞で、＜／形容動詞句＞＜形容動詞句　関係＝“主語”＞＜名詞句　識別子＝“press”　統語＝“並列”＞＜名詞句＞＜形容動詞句＞一部の＜／形容動詞句＞大衆紙＜／名詞句＞と＜名詞＞一般紙＜／名詞＞＜／名詞句＞が＜／形容動詞句＞
＜形容動詞句　関係＝“目的語”＞＜形容動詞句　関係＝“内容”　主語＝“press”＞＜形容動詞句　関係＝“目的語”＞＜名詞句＞＜形容動詞句＞＜名詞　共参照＝“Ｂ会”＞そ＜／名詞＞の＜／形容動詞句＞写真報道＜／名詞句＞を＜／形容動詞句＞
自主規制する＜／形容動詞句＞方針を＜／形容動詞句＞
＜形容動詞句　関係＝“位置”＞紙面で＜／形容動詞句＞
明らかにした。＜／文＞＜／文書＞

この文書においては、「一部の大衆紙と一般紙」は、統語＝“並列”というタグにより並列であることが表されている。並列の定義は、係り受け関係を共有すると言うことである。特に何も指定がない場合は、例えば、＜名詞句　関係＝ｘ＞＜名詞＞Ａ＜／名詞＞＜名詞＞Ｂ＜／名詞＞＜／名詞句＞ はＡがＢに依存関係のあることを表す。関係＝ｘは関係属性を表す。

【００４１】

関係属性は、統語、意味、修辞についての相互関係を記述する。主語、目的語、間接目的語のような文法機能、動作主、被動作者、受益者などのような主題役割、及び理由、結果などのような修辞関係はこの関係属性により記述される。本実施の形態では、主語、目的語、間接目的語のような比較的容易な文法機能について関係属性を記述する。

【００４２】

また、この文書においては、“Ａ氏”、“Ｂ会”、“Ｃ市”のような固有名詞について、地名、人名、組織名等のタグにより属性が記述されている。これら地名、人名、組織名等のタグが付与される語は固有名詞である。

【００４３】

また、このようなタグ付けされた文書においては、代名詞や限定節についての参照、被参

照関係がタグにより表される。例えば、上記文書においては、「その写真報道を」のエレメントの「その」の部分が、「共参照＝“Ｂ会”」という属性を持つことにより、その部分が「識別子＝“Ｂ会”」という属性を持つエレメント（この場合は名詞句）「Ａ氏のＢ会」であることが示されている。従って、上記「その写真報道を」の「その」の部分を置き換えると、「Ａ氏のＢ会の写真報道を」となる。

【００４４】

さらに、このようなタグ付けされた文書においては、省略された主語や目的語等を他の部分によって補うことができる。すなわち、上記文書の例においては、「自主規制する」のエレメントが「主語＝“press”」という属性を持つことにより、その意味上の主語が「識別子＝“press”」という属性を持つエレメント（この場合は名詞句）「一部の一般紙と大衆紙」であることが示されている。従って、主語を補ったものは、「（一部の一般紙と大衆紙が）自主規制する」となる。このように、省略が他の部分によって補われることをゼロ照応と呼ぶ。

【００４５】

以下、本発明に係る実施の形態の文書処理装置の具体的な動作について説明する。本実施の形態の文書処理装置は、上述したようなタグ付けされた文書に対して、自動要約処理を行わせるものであり、この要約文作成の際に、代名詞や限定節の置き換え処理や、省略された主語を補うようなゼロ照応処理を行う。

【００４６】

文書処理装置において文書の要約文を作成する場合には、その文書が図１の表示部３０に文書が表示されている状態で、ユーザが入力部２０を操作し、自動要約モードに切り換える。制御部１１は、この自動要約モードに切り換えられたとき、図４に示すような自動要約文作成プログラムの初期画面を表示して、ユーザによる自動要約文作成の開始操作を待つ。

【００４７】

すなわち、ユーザが上記自動要約文作成モードに切り換え操作したときには、図１の制御部１１は、ハードディスク装置３３に保存されている自動要約文作成プログラムを起動し、表示部３０を制御し、図４に示すような自動要約文作成プログラムの初期画面を表示させる。この例においては、表示部３１に表示されるウィンドウ１９０は、文書の名称が表示される文書名表示部１９１、キーワードが入力されるキーワード入力部１９２、文書の要約文を作成するための実行ボタンである要約文作成実行ボタン１９３等が表示される表示領域２００と、文書が表示される表示領域２１０と、文書の要約文が表示される表示領域２２０とに区分されている。

【００４８】

表示領域２００の文書名表示部１９１には、表示領域２１０に表示される文書の文書名等が表示される。また、キーワード入力部１９２には、例えば入力部２０のキーボード等を用いて文書の要約文を作成するためのキーワードが入力される。要約文作成実行ボタン１９３は、例えば入力部２０のマウス等を用いて押されることによって、表示領域２１０に表示されている文書の要約文作成処理を実行開始するための実行ボタンである。

【００４９】

表示領域２１０には、文書が表示される。表示領域２１０の右端には、スクロールバー２１１と、このスクロールバー２１１を上下に動かすためのボタン２１２，２１３が設けられており、ユーザが例えば入力部２０のマウス等を用いて、スクロールバー２１１を上下に直接動かしたり、ボタン２１２，２１３を押してスクロールバー２１１を上下に動かすことによって、表示領域２１０に表示される表示内容を縦方向にスクロールすることができる。ユーザは、入力部２０を操作することによって、表示領域２１０に表示されている文書の一部を選択して要約させることもでき、文書全体を要約させることもできる。

【００５０】

表示領域２２０には、要約文が表示される。図４においては、要約文がまだ作成されていない状態であるため、この表示領域２２０には、何も表示されていない。ユーザは、入力

部２０を操作することによって、要約文の表示領域２２０の表示範囲（大きさ）を変更することができる。具体的には、ユーザは、同図に示す表示領域２２０の表示範囲（大きさ）を、例えば図５に示すように拡大することができる。

【００５１】

文書処理装置は、ユーザが例えば入力部２０のマウス等を用いて、要約文作成実行ボタン１９３を押してオン状態とすると、ＣＰＵ１３の制御のもとに、図６に示す処理を実行して要約文の作成を開始する。

【００５２】

文書から要約文を作成する処理は、文書の内部構造に関するタグ付けに基づいて実行される。文書処理装置においては、先に図５に示したように、ウィンドウ１９０の表示領域２２０の大きさを変更することができる。文書処理装置は、ＣＰＵ１３の制御のもとに、新たにウィンドウ１９０が表示部３１に描画されるか、又は、表示領域２２０の大きさが変更された後、要約文作成実行ボタン１９３が操作されたときには、表示領域２２０に適合するように、ウィンドウ１９０の表示領域２１０に少なくともその一部が表示されている文書から、要約文を作成する処理を実行する。

【００５３】

まず、文書処理装置は、図６に示すように、ステップＳ２１において、ＣＰＵ１３の制御のもとに、活性拡散と呼ばれる処理を行う。本実施の形態においては、活性拡散により得られた中心活性値を重要度として採用することによって、文書の要約文を行う。すなわち、内部構造に関するタグ付けがされた文書においては、活性拡散を行うことによって、各エレメントに対して、内部構造に関するタグ付けに応じた中心活性値を付与することができる。

【００５４】

ここで、活性拡散は、中心活性値の高いエレメントと関わりのあるエレメントにも高い中心活性値を与えるような処理である。すなわち、活性拡散は、照応（anaphora；共参照（coreference））表現されたエレメントとその先行詞との間で中心活性値が等しくなり、それ以外では各中心活性値が同じ値に収束していく。この中心活性値は、文書の内部構造に関するタグ付けに応じて決定されるため、内部構造を考慮した文書の分析に利用することができる。

【００５５】

文書処理装置は、図７に示す一連の工程を経ることによって、活性拡散を実行する。

【００５６】

まず、文書処理装置は、図７に示すように、ステップＳ４１において、ＣＰＵ１３の制御のもとに、各エレメントの初期化を行う。文書処理装置は、語彙エレメントを除いた全てのエレメントと語彙エレメントとに対して中心活性値の初期値を割り当てる。例えば、文書処理装置は、中心活性値の初期値として、語彙エレメントを除いた全てのエレメントに対しては“１”を、語彙エレメントに対しては“０”を割り当てる。また、文書処理装置は、各エレメントの中心活性値の初期値に均一ではない値を予め割り当てることによって、活性拡散の結果得られた中心活性値に、初期値の偏りを反映させることができる。例えば、文書処理装置は、ユーザが関心を有するエレメントに対しては、中心活性値の初期値を高く設定することによって、ユーザの関心を反映した中心活性値を得ることができる。

【００５７】

エレメント間で参照・被参照による係り受けの関係にあるリンクである参照・被参照リンクと、それ以外のリンクである通常リンクとに関しては、エレメントを連結するリンクの端点の端点活性値を“０”に設定する。文書処理装置は、このようにして付与した端点活性値の初期値を例えばＲＡＭ１４に記憶させる。

【００５８】

ここで、エレメントとエレメントの連結構造の一例を図８に示す。同図においては、文書を構成するエレメントとリンクの構造の一部として、エレメントE_i及びエレメントE_jが示されている。エレメントE_iとエレメントE_jとは、それぞれ、中心活性値e_i，e_jを有

し、リンクL_{ij}にて接続されている。リンクL_{ij}のエレメントE_iに接続する端点は、T_{ij}であり、エレメントE_jに接続する端点は、T_{ji}である。エレメントE_iは、リンクL_{ij}により接続されるエレメントE_jの他に、リンクL_{ik}，L_{il}及びL_{im}により図示しないエレメントE_k，E_l及びE_mにそれぞれ接続している。エレメントE_jは、リンクL_{ji}により接続されるエレメントE_iの他に、リンクL_{jp}，L_{jq}及びL_{jr}により図示しないエレメントE_p，E_q及びE_rにそれぞれ接続している。

【００５９】

続いて、文書処理装置は、図７中のステップＳ４２において、ＣＰＵ１３の制御のもとに、文書を構成するエレメントE_iを計数するカウンタの初期化を行う。すなわち、文書処理装置は、エレメントを計数するカウンタのカウンタ値ｉを“１”に設定する。このことにより、カウンタは、第１番目のエレメントE_1を参照していることになる。

【００６０】

続いて、文書処理装置は、ステップＳ４３において、ＣＰＵ１３の制御のもとに、カウンタが参照するエレメントについて、新たな中心活性値を計算するリンク処理を実行する。このリンク処理については、さらに後述する。

【００６１】

続いて、文書処理装置は、ステップＳ４４において、ＣＰＵ１３の制御のもとに、文書中の全てのエレメントについて新たな中心活性値の計算が完了したか否かを判断する。

【００６２】

ここで、文書処理装置は、文書中の全てのエレメントについて新たな中心活性値の計算が完了したことを判断した場合には、ステップＳ４５へと処理を移行し、一方、文書中の全てのエレメントについて新たな中心活性値の計算が完了していないことを判断した場合には、ステップＳ４７へと処理を移行する。

【００６３】

具体的には、文書処理装置は、ＣＰＵ１３の制御のもとに、カウンタのカウンタ値ｉが、文書が含むエレメントの総数に達したか否かを判断する。そして、文書処理装置は、カウンタのカウンタ値ｉが、文書が含むエレメントの総数に達したことを判断した場合には、全てのエレメントが計算済みであるものとして、ステップＳ４５へと処理を移行する。一方、文書処理装置は、カウンタのカウンタ値ｉが、文書が含むエレメントの総数に達していないことを判断した場合には、全てのエレメントについて計算が終了していないものとしてステップＳ４７へと処理を移行する。

【００６４】

文書処理装置は、カウンタのカウンタ値ｉが、文書が含むエレメントの総数に達していないことを判断した場合には、ステップＳ４７において、ＣＰＵ１３の制御のもとに、カウンタのカウント値ｉを“１”だけインクリメントさせ、カウンタのカウント値を“ｉ＋１”とする。このことにより、カウンタは、ｉ＋１番目のエレメント、すなわち次のエレメントを参照する。そして、文書処理装置は、ステップＳ４３へと処理を移行し、端点活性値の計算及びこれに続く一連の行程が、次のｉ＋１番目のエレメントについて実行される。

【００６５】

また、文書処理装置は、カウンタのカウンタ値ｉが、文書が含むエレメントの総数に達したことを判断した場合には、ステップＳ４５において、ＣＰＵ１３の制御のもとに、文書に含まれる全てのエレメントの中心活性値の変化分、すなわち新たに計算された中心活性値の元の中心活性値に対する変化分について平均値を計算する。

【００６６】

文書処理装置は、ＣＰＵ１３の制御のもとに、例えばＲＡＭ１４に記憶された元の中心活性値と新たに計算した中心活性値を、文書に含まれる全てのエレメントについて読み出す。文書処理装置は、新たに計算した中心活性値の元の中心活性値に対するそれぞれの変化分の総和を文書に含まれるエレメントの総数で除することにより、全てのエレメントの中心活性値の変化分の平均値を計算する。文書処理装置は、このように計算した全てのエレ

メントの中心活性値の変化分の平均値を、例えばＲＡＭ１４に記憶させる。

【００６７】

そして、文書処理装置は、ステップＳ４６において、ＣＰＵ１３の制御のもとに、ステップＳ４５で計算した全てのエレメントの中心活性値の変化分の平均値が、予め設定された閾値以内であるか否かを判断する。そして、文書処理装置は、この変化分が閾値以内であると判断した場合には、この一連の行程を終了する。一方、文書処理装置は、変化分が閾値以内でないと判断した場合には、ステップＳ４２へと処理を移行し、カウンタのカウント値ｉを“１”に設定して文書のエレメントの中心活性値を計算する一連の行程を再び実行する。文書処理装置においては、これらのステップＳ４２乃至ステップＳ４６のループが繰り返される毎に、変化分は、徐々に減少する。

【００６８】

文書処理装置は、このようにして活性拡散を行うことができる。つぎに、この活性拡散を行うためにステップＳ４３において実行されるリンク処理について図９を参照して説明する。なお、同図に示すフローチャートは、１つのエレメントE_iに対する処理を示したものであるが、この処理は、全てのエレメントに対して行われるものである。

【００６９】

まず、文書処理装置は、図９に示すように、ステップＳ５１において、ＣＰＵ１３の制御のもとに、文書を構成する１つのエレメントE_iと一端が接続されたリンクを計数するカウンタの初期化を行う。すなわち、文書処理装置は、リンクを計数するカウンタのカウント値ｊを“１”に設定する。このカウンタは、エレメントE_iと接続された第１番目のリンクL_{ij}を参照することになる。

【００７０】

続いて、文書処理装置は、ステップＳ５２において、ＣＰＵ１３の制御のもとに、エレメントE_iとE_jを接続するリンクL_{ij}について、関係属性のタグを参照することによって、そのリンクL_{ij}が通常リンクであるか否かを判断する。文書処理装置は、リンクL_{ij}が、語に対応する語彙エレメント、文に対応する文エレメント、段落に対応する段落エレメント等の間の関係を示す通常リンクと、参照・被参照による係り受けの関係を示す参照リンクのいずれであるかを判断する。文書処理装置は、リンクL_{ij}が通常リンクであると判断した場合には、ステップＳ５３へと処理を移行し、リンクL_{ij}が参照リンクであると判断した場合には、ステップＳ５４へと処理を移行する。

【００７１】

文書処理装置は、リンクL_{ij}が通常リンクであると判断した場合には、ステップＳ５３において、エレメントE_iの通常リンクL_{ij}に接続された端点T_{ij}の新たな端点活性値を計算する処理を行う。

【００７２】

このステップＳ５３では、ステップＳ５２における判別により、リンクL_{ij}が通常リンクであることが明らかになっている。エレメントE_iの通常リンクL_{ij}に接続される端点T_{ij}の新たな端点活性値t_{ij}は、エレメントE_jの端点活性値のうち、リンクL_{ij}以外のリンクに接続する全ての端点T_{jp}，T_{jq}，T_{jr}の端点活性値t_{jp}、t_{jq}，t_{jr}と、エレメントE_iがリンクL_{ij}により接続されるエレメントE_jの中心活性値e_jとを加算し、この加算で得た値を文書に含まれるエレメントの総数で除することにより求められる。

【００７３】

文書処理装置は、ＣＰＵ１３の制御のもとに、例えばＲＡＭ１４から必要な端点活性値及び中心活性値を読み出す。文書処理装置は、読み出された端点活性値及び中心活性値について、上述のようにその通常リンクと接続された端点の新たな端点活性値を計算する。そして、文書処理装置は、このように計算した新たな端点活性値を、例えばＲＡＭ１４に記憶させる。

【００７４】

一方、文書処理装置は、リンクL_{ij}が通常リンクでないと判断した場合には、ステップＳ５４において、エレメントE_iの参照リンクに接続された端点T_{ij}の端点活性値を計算す

る処理を行う。

【００７５】

このステップＳ５４では、ステップＳ５２における判別により、リンクL_{ij}が参照リンクであることが明らかになっている。エレメントE_iの参照リンクL_{ij}に接続される端点T_{ij}の端点活性値t_{ij}は、エレメントE_jの端点活性値のうち、リンクL_{ij}を除いたリンクに接続される全ての端点T_{jp}，T_{jq}，t_{jr}の端点活性値t_{jp}，t_{jq}，t_{jr}と、エレメントE_iがリンクL_{ij}により接続されるエレメントE_jの中心活性値e_jとを加算することにより求められる。

【００７６】

文書処理装置は、ＣＰＵ１３の制御のもとに、例えばＲＡＭ１４に記憶された端点活性値及び中心活性値から、必要な端点活性値及び中心活性値を読み出す。文書処理装置は、読み出された端点活性値及び中心活性値を用いて、上述のように参照リンクと接続された新たな端点活性値を計算する。そして、文書処理装置は、このように計算した端点活性値を、例えばＲＡＭ１４に記憶させる。

【００７７】

これらのステップＳ５３における通常リンクの処理及びステップＳ５４における参照リンクの処理は、ステップＳ５２からステップＳ５５に至り、ステップＳ５７を介してステップＳ５２に戻るループに示すように、カウント値ｉにより参照されているエレメントE_iに接続される全てのリンクL_{ij}に対して実行される。なお、ステップＳ５７では、エレメントE_iに接続されるリンクを計数するカウント値ｊをインクリメントしている。

【００７８】

文書処理装置は、これらのステップＳ５３又はステップＳ５４の処理を行った後、ステップＳ５５において、ＣＰＵ１３の制御のもとに、エレメントE_iに接続される全てのリンクについて端点活性値が計算されたか否かを判別する。そして、文書処理装置は、全てのリンクについて端点活性値が計算されていると判断した場合には、ステップＳ５６の処理へと移行し、全てのリンクについて端点活性値が計算されていないと判断した場合には、ステップＳ５７へと処理を移行する。

【００７９】

ここで、文書処理装置は、全てのリンクについて端点活性値が計算されていると判断した場合には、ステップＳ５６において、ＣＰＵ１３の制御のもとに、エレメントE_iの中心活性値e_iの更新を実行する。

【００８０】

エレメントE_iの中心活性値e_iの新たな値、すなわち更新値は、エレメントE_iの現在の中心活性値e_iと、エレメントE_iの全ての端点の新たな端点活性値との和である$e_i' = e_i + \Sigma t_j'$をとることにより求められる。ここで、プライム“’”は、新たな値という意味である。このように、新たな中心活性値は、そのエレメントの元の中心活性値に、そのエレメントの端点の新たな端点活性値の総和に加えることにより得られる。

【００８１】

文書処理装置は、ＣＰＵ１３の制御のもとに、例えばＲＡＭ１４に記憶された端点活性値及び中心活性値から必要な端点活性値を読み出す。文書処理装置は、上述したような計算を実行し、そのエレメントE_iの中心活性値e_iを算出する。そして、文書処理装置は、計算した新たな中心活性値e_iを例えばＲＡＭ１４に記憶させる。

【００８２】

このようにして、文書処理装置は、文書中の各エレメントについて、新たな中心活性値を計算する。そして、文書処理装置は、このようにして図６中のステップＳ２１における活性拡散を実行する。

【００８３】

続いて、文書処理装置は、図６中のステップＳ２２において、ＣＰＵ１３の制御のもとに、先に図４に示した表示部３０に表示されているウィンドウ１９０の表示領域２２０の大きさ、すなわちこの表示領域２２０に表示可能な最大文字数をW_sと設定する。また、文

書処理装置は、ＣＰＵ１３の制御のもとに、要約文Ｓを初期化して初期値S_0＝””と設定する。これは、要約文に何も文字列が存在していないことを示す。文書処理装置は、このように設定した、表示領域２２０に表示可能な最大文字数W_s及び要約文Ｓの初期値S_0を、例えばＲＡＭ１４に記憶させる。

【００８４】

続いて、文書処理装置は、ステップＳ２３において、ＣＰＵ１３の制御のもとに、要約文の骨格の順次での作成をカウントするカウンタのカウント値ｉを“１”に設定する。すなわち、文書処理装置は、カウント値について、ｉ＝１と設定する。文書処理装置は、このように設定したカウント値ｉを例えばＲＡＭ１４に記憶させる。

【００８５】

続いて、文書処理装置は、ステップＳ２４において、ＣＰＵ１３の制御のもとに、カウンタのカウント値ｉについて、要約文作成対照の文章からｉ番目に平均中心活性値の高い文の骨格を抽出する。ここで、平均中心活性値とは、１つの文を構成する各エレメントの中心活性値を平均したものである。文書処理装置は、例えばＲＡＭ１４に記憶させた要約文S_{i-1}を読み出し、この要約文S_{i-1}に対して抽出した文の骨格の文字列を加えて、要約文S_iとする。そして、文書処理装置は、このようにして得た要約文S_iを、例えばＲＡＭ１４に記憶させる。同時に、文書処理装置は、文の骨格に含まれないエレメントの中心活性値順のリストl_iを作成し、このリストl_iを例えばＲＡＭ１４に記憶させる。

【００８６】

すなわち、このステップＳ２４においては、文書処理装置は、ＣＰＵ１３の制御のもとに、活性拡散の結果を用いて、平均中心活性値の大きい順に文を選択し、選択された文の骨格を抽出する。文の骨格は、文から抽出した必須エレメントにより構成される。必須エレメントになり得るものは、エレメントの主辞（head）と、主語（subject）、目的語（object）、間接目的語（indirect object）、所有者（posessor）、原因（cause）、条件（condition）又は比較（comparison）の関係属性を有するエレメントと、等位構造とされた関連するエレメントが必須エレメントのときには、その等位構造に直接含まれるエレメントとである。文書処理装置は、文の必須エレメントをつなげて文の骨格を生成し、要約文に加える。

【００８７】

続いて、文書処理装置は、ステップＳ２５において、ＣＰＵ１３の制御のもとに、要約文S_iの長さ、すなわち文字数が、ウィンドウ１９０の表示領域２２０の最大文字数W_sよりも多いか否かを判断する。

【００８８】

ここで、文書処理装置は、要約文S_iの文字数が最大文字数W_sよりも多いと判断した場合には、ステップＳ３０において、ＣＰＵ１３の制御のもとに、要約文S_{i-1}を最終的な要約文として設定し、一連の処理を終了する。なお、この場合には、要約文S_i＝S_0＝“”を出力するため、要約文は、表示領域２２０に表示されないことになる。

【００８９】

一方、文書処理装置は、要約文S_iの文字数が最大文字数W_sよりも多くないと判断した場合には、ステップＳ２６の処理へと移行し、ＣＰＵ１３の制御のもとに、ｉ＋１番目に平均中心活性値が高い文の平均中心活性値と、ステップＳ２４で作成したリストl_iのエレメントの中で最も中心活性値が高いエレメントの中心活性値とを比較する。そして、文書処理装置は、ｉ＋１番目に平均中心活性値が高い文の平均中心活性値が、リストl_iのエレメントの中で最も中心活性値が高いエレメントの中心活性値よりも高いと判断した場合には、ステップＳ２７へと処理を移行する。一方、文書処理装置は、ｉ＋１番目に平均中心活性値が高い文の平均中心活性値が、リストl_iのエレメントの中で最も中心活性値が高いエレメントの中心活性値よりも高くないと判断した場合には、ステップＳ２８へと処理を移行する。

【００９０】

文書処理装置は、ｉ＋１番目に平均中心活性値が高い文の平均中心活性値が、リストl_i

のエレメントの中で最も中心活性値が高いエレメントの中心活性値よりも高いと判断した場合には、ステップＳ２７において、ＣＰＵ１３の制御のもとに、カウンタのカウント値ｉを“１”だけインクリメントさせ、ステップＳ２４へと処理を戻す。

【００９１】

また、文書処理装置は、ｉ＋１番目に平均中心活性値が高い文の平均中心活性値が、リストl_iのエレメントの中で最も中心活性値が高いエレメントの中心活性値よりも高くないと判断した場合には、ステップＳ２８において、ＣＰＵ１３の制御のもとに、リストl_iのエレメントの中で最も中心活性値の高いエレメントｅを要約文S_iに加えてSS_iを生成し、さらに、エレメントｅをリストl_iから削除する。そして、文書処理装置は、このようにして生成した要約文SS_iを例えばＲＡＭ１４に記憶させる。

【００９２】

続いて、文書処理装置は、ステップＳ２９において、ＣＰＵ１３の制御のもとに、要約文SS_iの文字数がウィンドウ１９０の表示領域２２０の最大文字数W_sよりも多いか否かを判別する。文書処理装置は、要約文SS_iの文字数が最大文字数W_sよりも多くないと判別した場合には、ステップＳ２６からの処理を繰り返す。一方、文書処理装置は、要約文SS_iの文字数が最大文字数W_sよりも多いと判別した場合には、ステップＳ３１において、ＣＰＵ１３の制御のもとに、要約文S_iを最終的な要約文として設定し、表示領域２２０に表示して一連の処理を終了する。このようにして、文書処理装置は、最大文字数W_sよりも多くならないように要約文を生成する。

【００９３】

文書処理装置は、このような一連の処理を行うことによって、タグ付けされた文書を要約して要約文を作成することができる。文書処理装置は、例えば図４に示した文書を要約した場合には、図１０に示すような要約文を作成し、表示範囲の表示領域２２０に表示する。

【００９４】

すなわち、文書処理装置は、「TCP/IPの歴史はARPANETを抜きにして語ることはできない。ARPANETは1969年北米西海岸の４個所の大学、研究機関のホストコンピュータを50kbpsの回線で結んだ小規模なネットワークからARPANETは出発した。当時は1964年にメインフレームの汎用コンピュータシリーズが開発された。この時代背景を考えると、将来のコンピュータ通信の最盛を見越したこのようなプロジェクトは、まさに米国ならではのものであったといえるだろう。」という要約文を作成し、表示領域２２０に表示する。

【００９５】

文書処理装置においては、ユーザは、文書の全文章を一読する代わりに、この要約文を読むことで、文章の概要を理解し、この文章が所望する情報であるか否かを判定することができる。

【００９６】

なお、文書処理装置においては、文書中のエレメントに対して重要度を付与する方法としては、必ずしも上述したような活性拡散を用いる必要はなく、例えば、文書中に出現する単語の重みの総和を文書の重要度とする方法でもよい。また、重要度の付与方法は、これらの方法以外のものを利用することもできる。さらに、表示領域２００のキーワード入力部１９２にキーワードを入力することによって、そのキーワードに基づいた重要度の設定を行うこともできる。

【００９７】

さて、文書処理装置は、先に図５に示したように、表示部３１に表示されるウィンドウ１９０の表示領域２２０の表示範囲を拡大することができるが、作成した要約文が表示領域２２０に表示されている状態において、表示領域２２０の表示範囲を変更すると、その表示範囲に応じて、要約文の情報量を変更することができる。

【００９８】

この場合、文書処理装置は、ＣＰＵ１３の制御のもとに、ユーザが入力部２０を操作することに対応して、表示部３１に表示されたウィンドウ１９０の表示領域２２０の表示範囲

が変更されるまで待機する。そして、文書処理装置は、表示領域２２０の表示範囲が変更されると、ＣＰＵ１３の制御のもとに、先に図６に示した一連の処理と同様の処理を行い、表示領域２２０の表示範囲に対応した要約文を作成する。

【００９９】

文書処理装置は、このようにして、表示領域２２０の表示範囲に応じた要約文を新たに作成することができる。例えば、文書処理装置は、ユーザが入力部２０のマウスをドラッグ操作することにより表示領域２２０の表示範囲を拡大すると、より詳細な要約文を新たに作成し、図１１に示すように、新たな要約文をウィンドウ１９０の表示領域２２０に表示する。

【０１００】

すなわち、文書処理装置は、「TCP/IPの歴史はARPANETを抜きにして語ることはできない。ARPANETはアメリカ国防省DODの国防高等研究計画局がスポンサーとなって構築されてきた、実験および研究用のパケット交換ネットワークである。1969年北米西海岸の４個所の大学、研究機関のホストコンピュータを50kbpsの回線で結んだきわめて小規模なネットワークからARPANETは出発した。当時は1945年に世界初のコンピュータであるENIACがペンシルバニア大学で開発され、1964年にはじめてICを理論素子として実装したメインフレームの汎用コンピュータシリーズが開発され、やっとコンピュータが産声をあげたばかりあった。この時代背景を考えると、将来のコンピュータ通信の最盛を見越したこのようなプロジェクトは、まさに米国ならではのものであったといえるだろう。」という要約文を作成し、表示領域２２０に表示する。

【０１０１】

このように、文書処理装置においては、表示された要約文が簡略すぎて文書の概略を把握することができない場合、ユーザは、表示領域２２０の表示範囲を拡大することで、より多くの情報量を有するより詳細な要約文を参照することができる。

【０１０２】

ここで、文書処理装置は、このようにして文書の要約文を作成する際に、その要約文中に、代名詞や限定節が要約文に含まれる場合には置き換えを行い、また主語や目的語等が省略されている場合には対応する主語や目的語等を補うような上述したゼロ照応の処理を行っている。

【０１０３】

先ず、この代名詞等の置き換えやゼロ照応の具体例について、次のような文書を参照しながら説明する。

【０１０４】

「仕事について。

【０１０５】

わたしは今の仕事があまり好きではない。しかし、それをやらなければならない。」

この文書のタグ付けの例は、次のようになる。

【０１０６】

＜文書＞

＜タイトル＞＜形容動詞句　関係＝“目的語”＞＜名詞句＞仕事＜／名詞句＞に＜／形容動詞句＞ついて＜／タイトル＞

＜段落＞

＜文＞＜形容動詞句　関係＝“主語”＞＜名詞句　識別子＝“識別子１”＞わたし＜／名詞句＞は＜／形容動詞句＞＜形容動詞句　関係＝“目的語”＞＜名詞句
識別子＝“識別子０”＞＜形容動詞句＞今の＜／形容動詞句＞仕事＜／名詞句＞が＜／形容動詞句＞＜動詞＞＜形容動詞句　関係＝“程度”＞あまり＜／形容動詞句＞＜動詞＞好きではない＜／動詞＞＜／動詞＞。＜／文＞＜文＞＜形容動詞句＞しかし、＜／形容動詞句＞＜動詞＞＜形容動詞句　関係＝“目的語”＞＜名詞句　参照＝“識別子０”＞それ＜／名詞句＞を＜／形容動詞句＞＜動詞　主語＝“識別子１”＞やらなければならない＜／動詞＞＜／動詞＞。＜／文＞

＜／段落＞
＜／文書＞
【０１０７】
この文書において、「それ」という名詞句は、「参照＝“識別子０”」という属性を有し、「識別子＝“識別子０”」を含むエレメントである「今の仕事」という名詞句を参照している。すなわち、被参照エレメントである代名詞「それ」に対応する先行詞となる参照エレメントが「今の仕事」である。従って、要約文中に名詞句「それ」が含まれるにも拘わらず名詞句「今の仕事」が含まれていない場合には、要約文中の「それ」を「今の仕事」に置き換えるものである。
【０１０８】
また、上記文書において、「やらなければならない」という動詞は、「主語＝“識別子１”」という属性を有することから、その意味上の主語は、「識別子＝“識別子１”」という属性を有するエレメントである「わたし」という名詞句であることが分かる。すなわち、エレメント「やらなければならない」のゼロ照応エレメントが「わたし」である。従って、要約文中に「やらなければならない」が含まれているにも拘わらず意味上の主語「わたし」が含まれていない場合には、要約文中で、「（わたしが）やらなければならない」のようにゼロ照応エレメントを補うものである。
【０１０９】
このような代名詞等の置き換え処理やゼロ照応処理は、上述した要約文作成に続いて、あるいは要約文作成と同時に行われるものであり、これらの処理内容の具体例について、図１２及び図１３のフローチャートを参照しながら説明する。
【０１１０】
すなわち、図１２は、代名詞や限定節が要約文に含まれる場合の処理を説明するためのフローチャートであり、この図１２に示す処理は、例えば上記図６のステップＳ３０，Ｓ３１に続いて行われる。この図１２に示す処理において、文書中の参照・被参照関係における代名詞などの被参照エレメントをリストアップするために被参照リストRBListを用いており、要約用の語彙エレメントの配列のｉ番目の要素をt_iとし、この語彙エレメントt_iの参照エレメントをr_iとしている。
【０１１１】
図１２の最初のステップＳ７１において、文書処理装置は、図１のＣＰＵ１３の制御のもとに、上記被参照リストRBListを空にする。次のステップＳ７２で、文書処理装置は、要約用の語彙エレメントを配列順にカウントするカウンタのカウント値ｉを１に設定する（ｉ＝１）。
【０１１２】
次のステップＳ７３で、文書処理装置は、要約用の語彙エレメントの配列のｉ番目のエレメントt_iに関して、該語彙エレメントt_iの被参照エレメント集合を被参照リストRBListに加える。また、語彙エレメントt_iの参照エレメントをr_iとする。このステップＳ７３での処理は、当該エレメントt_iを他の代名詞等が参照している場合には他の代名詞等を被参照リストRBListに加え、当該エレメントt_iが代名詞等であって他のエレメント（先行詞）を参照している場合には参照しているエレメント（先行詞）を参照エレメントr_iとするものである。
【０１１３】
次のステップＳ７４で、文書処理装置は、語彙エレメントt_iの参照エレメントr_iが存在するか否かを判別する。文書処理装置は、このステップＳ７４でＹＥＳ、すなわちr_iが存在する、と判別されたときはステップＳ７５に進み、ＮＯのときはステップＳ７６に進む。すなわち、ステップＳ７４での判別により、当該エレメントt_iが代名詞等であって参照エレメントr_iが存在しているときのみ、ステップＳ７５に進む。
【０１１４】
ステップＳ７５で、文書処理装置は、語彙エレメントt_iが上記被参照リストRBListの要素であるか否かを判別し、ＹＥＳのときはステップＳ７６に進み、ＮＯのときはステップＳ

７７に進む。ステップＳ７６で、文書処理装置は、語彙エレメントt_iを要約文に追加し、ステップＳ７９に進む。ステップＳ７７で、文書処理装置は、t_iの参照エレメントr_iの語彙列を要約文に追加して、ステップＳ７８に進み、r_iの被参照エレメント集合を被参照リストRBListに加えた後、ステップＳ７９に進む。

【０１１５】

これらのステップＳ７５～Ｓ７８での処理は、エレメントt_iが代名詞等であって他のエレメントを参照している場合に、当該エレメントt_iが被参照リストRBListの要素であれば、すなわち既に先行詞が要約文中に含まれていれば、エレメントt_iを先行詞で置き換えることなくそのまま要約文に加え、エレメントt_iが被参照リストRBListの要素でなければ、先行詞がまだ要約文中にふくまれていないことから、当該エレメントt_iを先行詞である参照エレメントr_iで置き換えて要約文に加えるものである。

【０１１６】

ステップＳ７９で、文書処理装置は、要約用の語彙エレメントの配列の全てについてステップＳ７３以降の処理が終了したか否かを判別し、ＮＯのときはステップＳ８０にて上記カウント値ｉを１だけインクリメント（ｉ＝ｉ＋１）した後、ステップＳ７３に戻り、ＹＥＳのときは処理を終了する。

【０１１７】

以上のような置き換え処理により、要約文中に代名詞や限定節等の被参照エレメントが存在するにも拘わらず、対応する先行詞としての参照エレメントが要約文中に含まれていない場合に、最初に現れた被参照エレメントt_iが参照エレメントr_iで置き換えられると共に、この参照エレメントr_iの被参照エレメント集合が被参照リストRBListに加えられるから、その後の同じ参照エレメントr_iに対応する被参照エレメントについては、置き換えされずにそのまま要約文に加えられることになる。

【０１１８】

なお、図１２に示す具体的な置き換え処理の例については、上記図６に示す要約文作成の処理に続いて行うものとして説明しているが、要約文作成と同時に行わせてもよい。

【０１１９】

次に、図１３は、要約文に省略された主語や目的語等を含む文が存在する場合の前述したようなゼロ照応処理を説明するためのフローチャートであり、この図１３に示す処理は、例えば上記図６のステップＳ３０，Ｓ３１に続いて、上記図１２の処理の前、後、あるいは同時に行われる。この図１３に示す処理において、文書中の省略された主語や目的語等のゼロ照応エレメントをリストアップするためにゼロ照応リストZAListを用いており、要約用の語彙エレメントの配列のｉ番目の要素をt_iとし、この語彙エレメントt_iのゼロ照応エレメントをz_iとしている。

【０１２０】

図１３の最初のステップＳ８１において、文書処理装置は、図１のＣＰＵ１３の制御のもとに、上記ゼロ照応リストZAListを空にする。次のステップＳ８２で、文書処理装置は、要約用の語彙エレメントを配列順にカウントするカウンタのカウント値ｉを１に設定する（ｉ＝１）。

【０１２１】

次のステップＳ８３で、文書処理装置は、要約用の語彙エレメントの配列のｉ番目の要素をt_iに関して、該語彙エレメントt_iのゼロ照応エレメントが存在すればそれをr_iとする。次のステップＳ８４で、文書処理装置は、語彙エレメントt_iのゼロ照応エレメントz_iが存在するか否かを判別する。このステップＳ８４でＹＥＳ（ゼロ照応エレメントz_iが存在）と判別されたときはステップＳ８５に進み、ＮＯのときはステップＳ８６に進む。

【０１２２】

ステップＳ８５で、文書処理装置は、語彙エレメントt_iが上記ゼロ照応リストZAListの要素であるか否かを判別し、ＹＥＳのときはステップＳ８６に進み、ＮＯのときはステップＳ８７に進む。ステップＳ８６では、語彙エレメントt_iを要約文に追加し、ステップＳ

８９に進む。文書処理装置は、ステップＳ８７では、ゼロ照応エレメントz_iが要約文中に既に含まれて存在しているか否かを判別し、ＹＥＳのときは上記ゼロ照応リストZAListにゼロ照応エレメントz_iを加えてステップＳ８６に進み、ＮＯのときはステップＳ８９に進む。ステップＳ８９で、文書処理装置は、語彙エレメントt_iのゼロ照応エレメントz_iの語彙列を括弧でくくり、語彙エレメントt_iと共に要約文に追加して、ステップＳ９０に進み、語彙エレメントt_iとゼロ照応エレメントz_iとをゼロ照応リストZAListに加えた後、ステップＳ９１に進む。ステップＳ８９、Ｓ９０においては、文書処理装置は、ゼロ照応エレメントz_iが主語ならばその語彙列に「が」を付加して括弧でくくり、目的語ならばその語彙列に「を」を付加して括弧でくくって、語彙エレメントt_iの前あるいは後に配置する。日本語の場合には、ゼロ照応エレメントは語彙エレメントの前に配置する。

【０１２３】

文書処理装置は、ステップＳ９１で、要約用の語彙エレメントの配列の全てについてステップＳ８３以降の処理が終了したか否かを判別し、ＮＯのときはステップＳ９２にて上記カウント値ｉを１だけインクリメント（ｉ＝ｉ＋１）した後、ステップＳ８３に戻り、ＹＥＳのときは処理を終了する。

【０１２４】

以上のようなゼロ照応処理により、要約文中に主語や目的語等が省略されたエレメントt_iが存在するにも拘わらず、対応するゼロ照応エレメントz_iが要約文中に含まれていない場合に、当該エレメントt_iにゼロ照応エレメントz_iが括弧でくくられて付加されると共に、このエレメントt_iとゼロ照応エレメントz_iとがゼロ照応リストZAListに加えられるから、その後の同じゼロ照応エレメントz_iに対応するエレメントについては、ゼロ照応エレメントz_iが付加されずにそのまま要約文に加えられることになる。

【０１２５】

ところで、これらの図１２に示す代名詞等の置き換え処理や、図１３に示すゼロ照応処理を、上述した図６に示した要約文作成に続いて行う場合には、要約文中の文字数が変化し、上述した要約文の最大文字数、すなわち上記要約文表示領域の大きさに応じて決まる表示可能な最大文字数を超えてしまったり、最大文字数よりも少なくなってしまうことがある。そこで、要約文中の文字数を上記最大文字数以内の最大の文字数にするために、最終的な文字数の調整作業が必要である。これは、上記代名詞等の置き換え処理やゼロ照応処理を行うことにより要約文中の文字数が最大文字数を超えた場合には、要約文中の重要度の低いエレメント、すなわち上記中心活性値の低いエレメントから順次削除して、要約文中の文字数が上記最大文字数以内に収まるようにする。また、文字数が最大文字数よりも少なくなった場合には、元の文書中の要約文に含まれないエレメントの内の最も中心活性値が高いエレメントから順に要約文中に付加して行き、上記最大文字数を超える直前でエレメントの付加を停止することで、上記最大文字数に最も近く、最大文字数以内の文字数の要約文を得ることができる。

【０１２６】

なお、本実施の形態においては、文書へのタグ付けの方法の一例を示したが、本発明がこのタグ付けの方法に限定されないことは勿論である。また、本実施の形態においては、文書処理装置の通信部２２に外部から電話回線を介して文書が送信されるとしたが、本発明はこれに限定されない。例えば、衛星等を介して文書が送信される場合にも適用でき、また、記録／再生部３１において記録媒体３２から読み出されたり、文書処理装置のＲＯＭ１３に文書が書き込まれていたりしてもよい。

【０１２７】

また、本発明の実施の形態においては、上記図１の記録媒体３２として、上述した文書処理プログラムが書き込まれたディスク状記録媒体やテープ状記録媒体等を提供することも容易に実現できる。さらに、上述した文書処理プログラムについては、通信回線等の伝送媒体を介して供給することも容易に実現できる。

【０１２８】

また、上述の実施の形態においては、文書処理装置の表示部３０に表示された文書から所望のエレメントを選択するデバイスとしてマウスを例示したが、本発明がこれに限定されないことはいうまでもない。文書処理装置におけるエレメントの入力には、タブレット、ライトペン等の他のデバイスを利用することができる。

【０１２９】

さらに、上述の実施の形態においては日本語の文章を例示したが、本発明は、日本語に限定されず、英語、ドイツ語、フランス語、ロシア語、イタリア語、スペイン語、中国語、韓国語等の種々の言語に適用できることはいうまでもない。

【０１３０】

【発明の効果】

以上の説明からも明らかなように、本発明によれば、文書の要約文を作成し、作成される要約文中における省略された主語又は目的語が該要約文中に含まれていないとき、元の文書中の対応する主語又は目的語を要約文中に追加することにより、ゼロ照応エレメントが要約文中に必ず１回は現れることになり、ユーザの理解が容易で正確な内容の要約文を自動生成することができる。

【図面の簡単な説明】

【図１】本実施の形態を適用した文書処理装置の概略構成を示すブロック図である。

【図２】文書のタグ付けによる内部構造の一例を示す図である。

【図３】文書のタグ付けによる内部構造を表示したウィンドウを示す図である。

【図４】文書を表示したウィンドウを示す図である。

【図５】文書を表示したウィンドウを示す図であって、要約文を表示する表示領域が図１３に示す表示領域よりも拡大された様子を示す図である。

【図６】要約文を作成する際の一連の処理を説明するフローチャートである。

【図７】活性拡散を行う際の一連の処理を説明するフローチャートである。

【図８】活性拡散の処理を説明するためのエレメントの連結構造を示す図である。

【図９】活性拡散のリンク処理を行う際の一連の処理を説明するフローチャートである。

【図１０】文書とその要約文を表示したウィンドウを示す図である。

【図１１】文書とその要約文を表示したウィンドウを示す図であって、図５に示すウィンドウに要約文を表示した様子を示す図である。

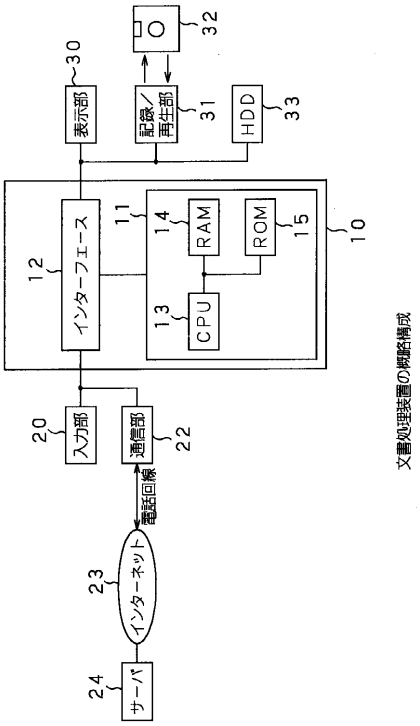
【図１２】被参照エレメントが要約文に含まれる場合の参照エレメントでの置き換え処理を説明するためのフローチャートである。

【図１３】要約文中でのゼロ照応処理を説明するためのフローチャートである。

【符号の説明】

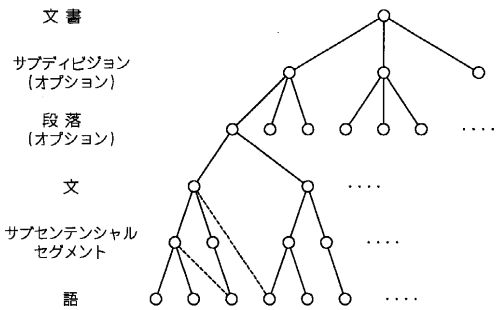
１０　文書処理装置の本体、　１１　制御部、　１２　インターフェース、　１３　ＣＰＵ、　２０　入力部、　２２　通信部、　３０　表示部、　３１　記録／再生部、　３２　記録媒体、　３３　ハードディスク装置

【図１】



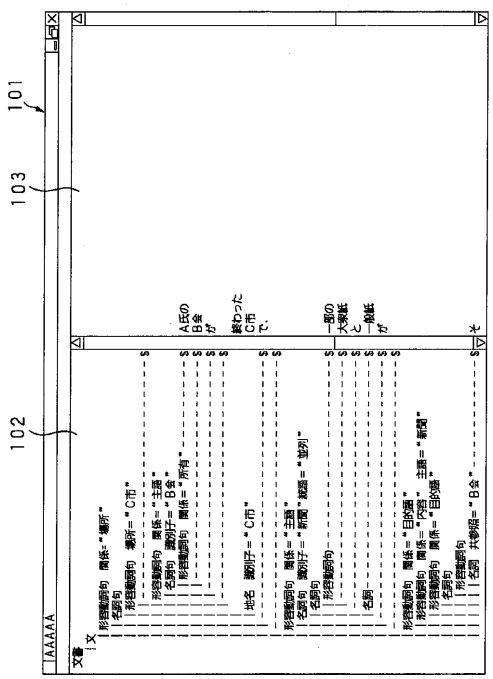
文書処理装置の概略構成

【図２】



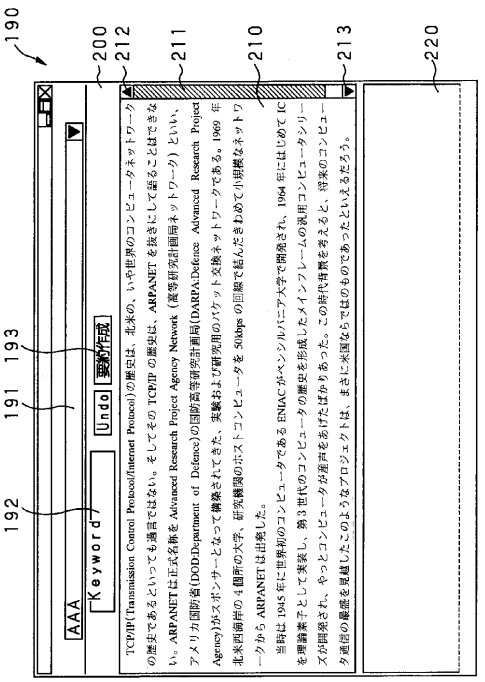
文書の内部構造の説明図

【図３】



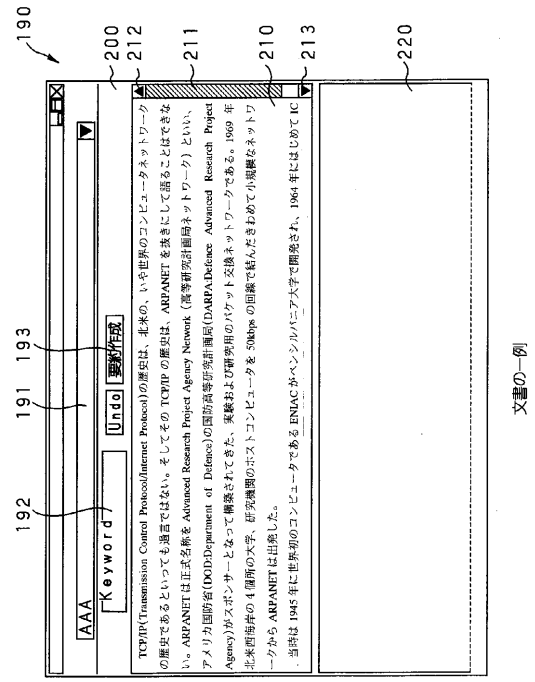
表示部の表示内容の説明図

【図４】



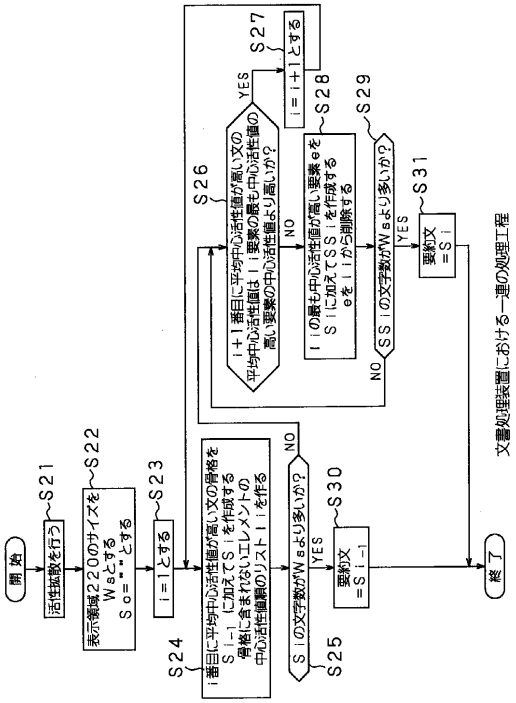
文書の一例

【図５】

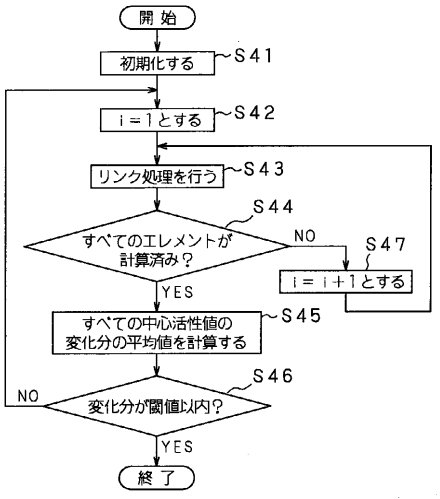


文書の一例

【図６】

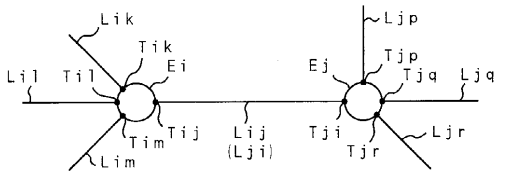


【図７】



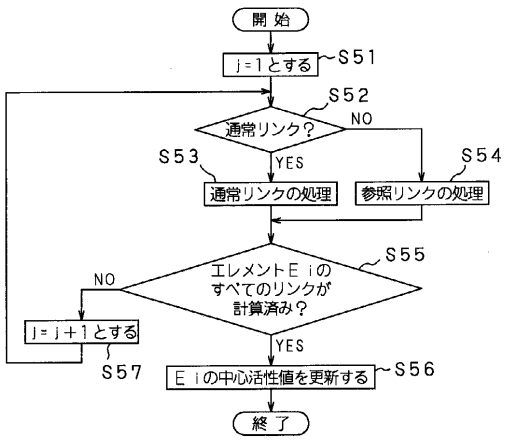
文書処理装置における一連の処理工程

【図８】



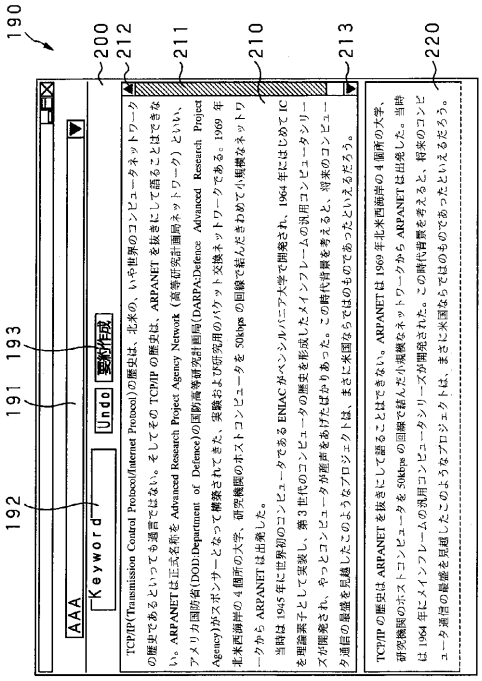
エレメントの連結構造の一例

【図９】

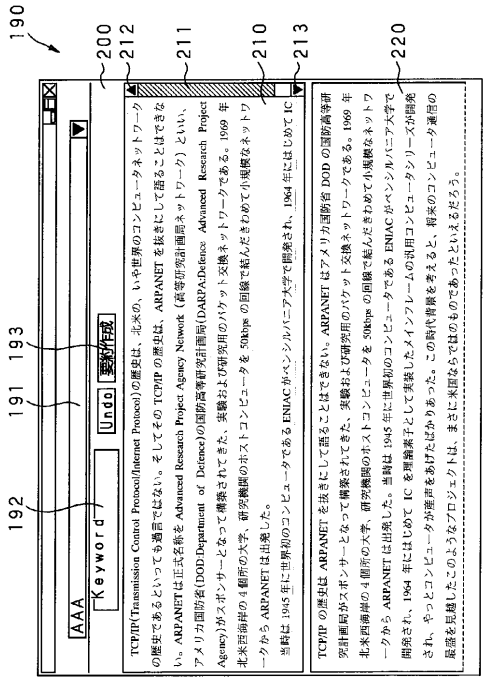


文書処理装置における一連の処理工程

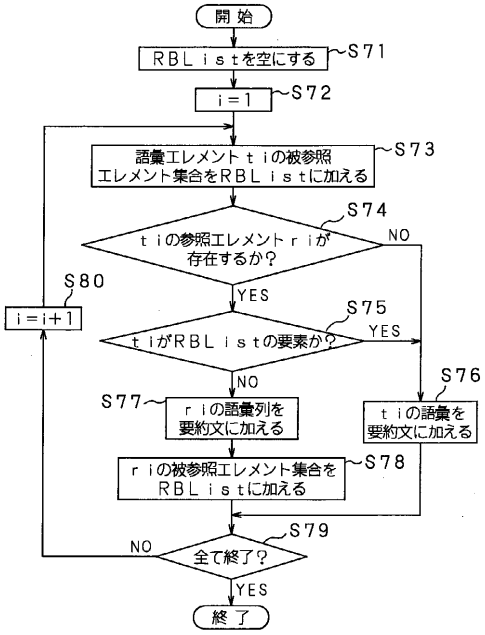
【図１０】



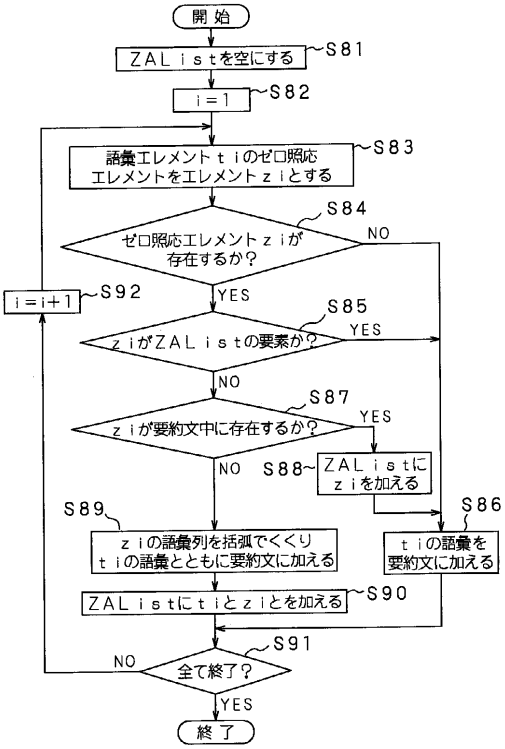
【図１１】



【図１２】



【図１３】



フロントページの続き

(56)参考文献 特開平１０－２０７８９１（ＪＰ，Ａ）
特開平０８－２３５２１９（ＪＰ，Ａ）
特開平０９－１７９８７５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/20-17/26
G06F 17/30