(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4014160号
(P4014160)
(45)発行日　平成19年11月28日(2007.11.28)　(24)登録日　平成19年9月21日(2007.9.21)

(51)Int.Cl.　　F I
G06F 17/30 (2006.01)　　G06F 17/30 210D
G06F 17/21 (2006.01)　　G06F 17/21 501T

請求項の数 20　(全 36 頁)

(21)出願番号　特願2003-155256(P2003-155256)
(22)出願日　平成15年5月30日(2003.5.30)
(65)公開番号　特開2004-355528(P2004-355528A)
(43)公開日　平成16年12月16日(2004.12.16)
審査請求日　平成16年2月27日(2004.2.27)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
(74)代理人　100086243
弁理士　坂口　博
(74)代理人　100091568
弁理士　市位　嘉宏
(74)代理人　100108501
弁理士　上野　剛史

最終頁に続く

(54)【発明の名称】情報処理装置、プログラム、及び記録媒体

(57)【特許請求の範囲】

【請求項１】

文字列、および、その文字列の表示形式を指定する情報であるタグ情報のそれぞれを文書構成要素として少なくとも含む文書情報について、複数の前記文書構成要素を複数のグループに分類する情報処理装置であって、

前記文書情報における前記複数の文書構成要素のそれぞれを、その文書構成要素がタグ情報ならばその役割を示し、その文書構成要素が文字列ならば文字数およびその文字列に含まれる文字の種類の少なくとも一方を示す情報である要素識別情報に変換する構成要素変換部と、

予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組を前記文書情報から選択し、前記文書情報に含まれるこの要素識別情報の組のそれぞれを、この要素識別情報の組の配列パターンを示す他の要素識別情報に変換する文書内配列パターン変換部と、

前記文書内配列パターン変換部により前記文書情報を繰り返し変換した結果得られた文書情報において、前記文書内配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部と

を更に備える情報処理装置。

【請求項２】

前記構成要素変換部により変換された文書情報における前記要素識別情報を、出現する

頻度の高い順に順次選択して被選択情報とする構成要素選択部と、
前記構成要素選択部により選択された各被選択情報について、その被選択情報と、前記文書情報においてその被選択情報の次に配列されその被選択情報と同一の種類である前記要素識別情報との間に配列された間隙構成要素を検出する間隙構成要素検出部と
を更に備え、
前記文書内配列パターン変換部は、前記構成要素選択部により順次選択される被選択情報について、その被選択情報及びその間隙構成要素を前記要素識別情報の組として検出する
請求項１記載の情報処理装置。
【請求項３】
前記間隙構成要素検出部は、前記文書情報の最後に配列される被選択情報から、前記文書情報の終端までの間に配列された文書構成要素である終端構成要素を更に検出し、
前記文書内配列パターン変換部は、前記文書情報の最後に配列される被選択情報と、前記終端構成要素とを、前記要素識別情報の組として検出する
請求項２記載の情報処理装置。
【請求項４】
前記文書内配列パターン変換部は、既に検出した前記要素識別情報の組より出現頻度が高いことを更に条件として、繰り返し出現する前記要素識別情報の組を検出する
請求項２記載の情報処理装置。
【請求項５】
前記文書内配列パターン変換部は、前記基準頻度以上で繰り返し出現する前記要素識別情報の組であって、変換前の文書構成要素を当該要素識別情報の組に変換するべく前記文書内配列パターン変換部が繰り返す変換の回数が予め定められた基準回数以下の要素識別情報の組を、前記配列パターンを示す要素識別情報に変換する
請求項１記載の情報処理装置。
【請求項６】
前記文書情報は、前記文書構成要素として、表示画面上に表示する画像を識別する画像識別情報を含み、
前記構成要素変換部は、前記画像識別情報を、前記画像識別情報により識別される画像の形状を示す前記要素識別情報に変換する
請求項１記載の情報処理装置。
【請求項７】
前記文書情報は、利用者に表示する表示情報と、当該表示情報に対する指示に応じて表示するべき他の情報の格納位置を示す格納位置情報とを前記文書構成要素として含み、
前記構成要素変換部は、前記格納位置情報を、前記他の情報が格納される前記格納位置の範囲を示す前記要素識別情報に変換する
請求項１記載の情報処理装置。
【請求項８】
前記文書情報は、前記文書構成要素として、表示画面に表示する表示情報と、当該表示情報の表示形式を指定する情報であるタグ情報とを含むタグ付き文書であり、
前記タグ情報は、当該タグ情報により表示形式を指定する表示情報において更に内側タグ情報を含む、外側タグ情報であり、
前記構成要素変換部による変換後の文書情報において、前記外側タグ情報を根ノードとして生成し、前記内側タグ情報を前記根ノードの葉ノードとして生成した、文書構造情報を生成する文書構造情報生成部と、
複数の文書情報のそれぞれについて、前記文書構造情報生成部により生成された前記文書構造情報を比較することにより、一の文書情報が他の文書情報と同一の構造を有するか否かを出力する文書情報同一性出力部と
を更に備える請求項１記載の情報処理装置。
【請求項９】

前記文書情報は、表示画面に表示するべき情報を指示する文書情報であり、更に、前記表示画面に表示されないコメント情報を含み、
前記文書内配列パターン変換部は、当該コメント情報を、一の前記要素識別情報の組と、他の前記要素識別情報の組との境界を示す情報として用いる
請求項１記載の情報処理装置。
【請求項１０】
前記グループ分類部により分類された各グループについて、当該グループに含まれる文書構成要素に対する目次を示す目次情報を出力する目次情報出力部
を更に備える請求項１記載の情報処理装置。
【請求項１１】
前記グループ分類部により分類された各グループについて、そのグループに属する文書構成要素が前記文書情報のどの領域に位置するかを示す配置情報を、前記文書情報とは別体に生成して出力するアノテーション出力部
を更に備える請求項１記載の情報処理装置。
【請求項１２】
前記文書情報に応じて表示画面上に情報を出力する表示部と、
前記文書内配列パターン変換部により変換される対象となる要素識別情報の組について、当該要素識別情報の組の変換元である文書構成要素が、前記表示画面に一度に表示可能かどうかを判断し、前記表示画面に表示可能でない場合に、当該要素識別情報の組を、変換を繰り返し行う対象から除外する反復終了判断部と
を更に備える請求項１記載の情報処理装置。
【請求項１３】
グループに分類する対象である対象文書情報と予め定められた関係を有する関連文書情報を検出する関連文書検出部と、
前記対象文書情報及び前記関連文書情報から、前記文書内配列パターン変換部により変換された要素識別情報を除外した文書について、当該対象文書情報及び当該関連文書情報の双方において出現する前記要素識別情報の組であって、前記対象文書情報及び前記関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組を特定し、前記対象文書情報に含まれるこの要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する文書間配列パターン変換部と
を更に備え、
前記グループ分類部は、更に、前記文書間配列パターン変換部により変換された結果得られた前記対象文書情報において、前記文書間配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類する
請求項１記載の情報処理装置。
【請求項１４】
前記グループ分類部は、分類したグループのそれぞれについて、当該グループ内の文書構成要素が前記文書情報において果たす役割又は当該グループ内の文書構成要素の内容を示すタイトル情報を、更に生成する
請求項１記載の情報処理装置。
【請求項１５】
前記グループ分類部は、当該要素識別情報の組の境界に設けられた、表示画面に表示されないコメント情報に含まれる情報を、前記タイトル情報として生成する
請求項１４記載の情報処理装置。
【請求項１６】
文字列、および、その文字列の表示形式を指定する情報であるタグ情報のそれぞれを文書構成要素として少なくとも含む文書情報について、複数の前記文書構成要素を、複数のグループに分類する情報処理装置であって、

グループに分類する対象である対象文書情報と予め定められた関係を有する関連文書情報を検出する関連文書検出部と、
前記対象文書情報及び前記関連文書情報のそれぞれにおいて、前記複数の文書構成要素のそれぞれを、その文書構成要素がタグ情報ならばその役割を示し、その文書構成要素が文字列ならば文字数およびその文字列に含まれる文字の種類の少なくとも一方を示す情報である要素識別情報に変換する構成要素変換部と、
前記対象文書情報及び前記関連文書情報の双方において出現する前記要素識別情報の組であって、前記対象文書情報及び前記関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組を特定し、前記対象文書情報に含まれるこの要素識別情報の組のそれぞれを、当該文書構成要素の組の配列パターンを示す前記要素識別情報に変換する文書間配列パターン変換部と、
前記文書間配列パターン変換部により変換された結果得られた前記対象文書情報において、前記文書間配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部と
を備える情報処理装置。
【請求項１７】
前記関連文書検出部は、前記対象文書情報が格納される格納位置から予め定められた範囲内に格納されている文書情報を前記関連文書情報として検出する
請求項１６記載の情報処理装置。
【請求項１８】
前記対象文書情報は、当該対象文書情報が生成される以前に存在していた既存文書情報を更新することにより生成され
前記関連文書検出部は、前記既存文書情報を前記関連文書情報として検出する
請求項１６記載の情報処理装置。
【請求項１９】
文字列、および、その文字列の表示形式を指定する情報であるタグ情報のそれぞれを文書構成要素として少なくとも含む文書情報について、複数の前記文書構成要素を、複数のグループに分類する情報処理装置を制御するプログラムであって、
前記情報処理装置を、
前記文書情報における前記複数の文書構成要素のそれぞれを、その文書構成要素がタグ情報ならばその役割を示し、その文書構成要素が文字列ならば文字数およびその文字列に含まれる文字の種類の少なくとも一方を示す情報である要素識別情報に変換する構成要素変換部と、
予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組を前記文書情報から選択し、前記文書情報に含まれるこの要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す他の要素識別情報に変換する文書内配列パターン変換部と、
前記文書内配列パターン変換部により前記文書情報を繰り返し変換した結果得られた文書情報において、前記文書内配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部と
して機能させるプログラム。
【請求項２０】
文字列、および、その文字列の表示形式を指定する情報であるタグ情報のそれぞれを文書構成要素として少なくとも含む文書情報について、複数の前記文書構成要素を、複数のグループに分類する情報処理装置を制御するプログラムであって、
前記情報処理装置を、
グループに分類する対象である対象文書情報と予め定められた関係を有する関連文書情報を検出する関連文書検出部と、

前記対象文書情報及び前記関連文書情報のそれぞれにおいて、前記複数の文書構成要素のそれぞれを、その文書構成要素がタグ情報ならばその役割を示し、その文書構成要素が文字列ならば文字数およびその文字列に含まれる文字の種類の少なくとも一方を示す情報である要素識別情報に変換する構成要素変換部と、

前記対象文書情報及び前記関連文書情報の双方において出現する前記要素識別情報の組であって、前記対象文書情報及び前記関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組を選択し、前記対象文書情報に含まれるこの要素識別情報の組のそれぞれを、当該文書構成要素の組の配列パターンを示す前記要素識別情報に変換する文書間配列パターン変換部と、

前記文書間配列パターン変換部により変換された結果得られた前記対象文書情報において、前記文書間配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部と

して機能させるプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、情報処理装置、プログラム、及び記録媒体に関する。特に本発明は、文書情報の内容を分類する情報処理装置、プログラム、及び記録媒体に関する。

【０００２】

【従来の技術】

従来、インターネット等のワールド・ワイド・ウェブ（ＷＷＷ）システムにおいて、知的障害者・高齢者等に対してウェブページを快適に閲覧させることができる技術として、トランスコーディングが用いられている。トランスコーディングを行う際に、ウェブページに付加されたアノテーション等の情報を用いることにより、ウェブページ中のコンテンツを内容又は種類に応じて並び替えたり、ウェブページ中の目次を作成したりすることができる。これにより、障害者・高齢者等は、ウェブページを利便に閲覧することができる。

【０００３】

また、上記トランスコーディングの為のアノテーションを適切に付加するには、ある程度の知識が必要であり、かつアノテーションの付加に要する作業量が大きい。そこで、アノテーションの付加を支援する技術が提案されている（特許文献１、非特許文献２、及び特許文献３参照。）。例えば、特許文献３に記載の技術は、ウェブページ内のレイアウトを決定するタグの構造及び特徴に基づいて、ウェブページ内を分類することができる。

【０００４】

【特許文献１】

特開２００３－８５０８７号公報

【０００５】

【非特許文献２】

Ｈ．Ｔａｋａｇｉ，Ｃ．Ａｓａｋａｗａ，Ｋ．Ｆｕｋｕｄａ，Ｊ．Ｍａｅｄａ著、「Ｓｉｔｅ－ｗｉｄｅ Ａｎｎｏｔａｔｉｏｎ：Ｒｅｃｏｎｓｔｒｕｃｔｉｎｇ Ｅｘｉｓｔｉｎｇ Ｐａｇｅｓ ｔｏ ｂｅ Ａｃｃｅｓｓｉｂｌｅ」、ＣＳＵＮ２００２

【０００６】

【特許文献３】

特開２００２－２４５０６８号公報

【０００７】

【発明が解決しようとする課題】

しかしながら、上記技術は、構造が動的に変化するウェブページについて、コンテンツを適切に分類することはできない。例えば、ウェブページが、野球のスコアボード、出場選手、及び出場選手の成績表を含む場合、当該ウェブページは逐次更新され動的に変化する場合がある。このような場合、上記技術は、ウェブページのレイアウトが変化したのか、

又は、野球の試合が進行してスコアボードや出場選手のレコード等が変化したのかを区別することができない。

【０００８】

そこで本発明は、上記の課題を解決することのできる情報処理装置、プログラム、及び記録媒体を提供することを目的とする。この目的は特許請求の範囲における独立項に記載の特徴の組み合わせにより達成される。また従属項は本発明の更なる有利な具体例を規定する。

【０００９】

【課題を解決するための手段】

即ち、本発明の第１の形態によると、文字列、および、その文字列の表示形式を指定する情報であるタグ情報のそれぞれを文書構成要素として少なくとも含む文書情報について、複数の前記文書構成要素を複数のグループに分類する情報処理装置であって、文書情報における複数の文書構成要素のそれぞれを、その文書構成要素がタグ情報ならばその役割を示し、その文書構成要素が文字列ならば文字数およびその文字列に含まれる文字の種類の少なくとも一方を示す情報である要素識別情報にそれぞれ変換する構成要素変換部と、予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組を前記文書情報から選択し、前記文書情報に含まれるこの要素識別情報の組のそれぞれを、この要素識別情報の組の配列パターンを示す他の要素識別情報に変換する文書内配列パターン変換部と、文書内配列パターン変換部により文書情報を繰り返し変換した結果得られた文書情報において、文書内配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類するグループ分類部とを備える情報処理装置、当該情報処理装置を制御するプログラムを提供する。

なお上記の発明の概要は、本発明の必要な特徴の全てを列挙したものではなく、これらの特徴群のサブコンビネーションも又発明となりうる。

【００１０】

【発明の実施の形態】

以下、発明の実施の形態を通じて本発明を説明するが、以下の実施形態は特許請求の範囲にかかる発明を限定するものではなく、又実施形態の中で説明されている特徴の組み合わせの全てが発明の解決手段に必須であるとは限らない。

図１は、情報処理装置１０の機能ブロック図を示す。情報処理装置１０は、ＨＴＭＬ等のタグ付き文書である文書情報について、当該文書情報に含まれる複数の文書構成要素をグループに分類する装置である。具体的には、本実施形態において、情報処理装置１０は、ＷＷＷサーバ等から取得したＨＴＭＬ文書について、当該文書情報に含まれるタグ情報等をグループに分類して表示する。これに代えて、情報処理装置１０は、ＷＷＷサーバ等のサーバ装置であって、ウェブブラウザ等からのリクエストに応じてＨＴＭＬ文書内の情報を分類し、分類した当該情報をウェブブラウザに返送してもよい。

【００１１】

情報処理装置１０は、関連文書検出部１００と、構成要素変換部１０５と、構成要素選択部１１０と、間隙構成要素検出部１２０と、文書内配列パターン変換部１３０と、変換指示入力部１３５と、反復終了判断部１４０と、文書間配列パターン変換部１６０と、グループ分類部１７０と、並替出力部１８０と、目次情報出力部１９０と、グループ識別情報生成部２００と、アノテーション出力部２０２と、表示部２０５と、文書構造情報生成部２１０と、文書情報同一性出力部２２０とを備える。

【００１２】

関連文書検出部１００は、グループ分けを行う対象の対象文書情報を取得すると、当該対象文書情報と予め定められた関係を有する関連文書情報を検出する。例えば、関連文書検出部１００は、対象文書情報が格納される格納位置から予め定められた範囲内に格納されている文書情報を、関連文書情報として検出する。対象文書情報が格納される格納位置から予め定められた範囲内とは、例えば、対象文書情報と同一のディレクトリ内であってもよいし、同一のサイト内であってもよい。また、関連文書検出部１００は、複数の関連文

書情報を検出することが好ましい。そして、関連文書検出部１００は、対象文書情報及び関連文書情報を順次構成要素変換部１０５に送る。

【００１３】

構成要素変換部１０５は、関連文書検出部１００から文書情報を受け取ると、当該文書情報における複数の文書構成要素のそれぞれを当該文書構成要素の種類又は役割を示す要素識別情報にそれぞれ変換し、変換結果の文書情報を構成要素選択部１１０に送る。なお、要素識別情報とは、文書構成要素そのものであってもよい。即ち、構成要素変換部１０５は、箇条書きの始点を示す文書構成要素である＜ｌｉ＞を、同様に箇条書きの始点を示す要素識別情報である＜ｌｉ＞に変換してもよい。また、好ましくは、構成要素変換部１０５は、変換前の文書構成要素及び変換後の要素識別情報を対応付けた情報をメモリ等に格納しておく。

【００１４】

また、構成要素変換部１０５は、文書内配列パターン変換部１３０又は文書間配列パターン１６０からの通知を受けて、当該構成要素変換部１０５により新たに変換するべき文書構成要素の組合せを示す新規登録要素集合を、当該新規登録要素集合の変換先とするべき要素識別情報に対応付けて、メモリ等から取得する。そして、構成要素変換部１０５は、文書情報における新規登録要素集合のそれぞれを、対応する要素識別情報に変換する。

【００１５】

構成要素選択部１１０は、構成要素変換部１０５により変換された文書情報又は反復終了判断部１４０から受け取った文書情報について、当該文書情報における要素識別情報の何れかを、例えば、出現する頻度の高い順に被選択情報として選択し、選択結果を間隙構成要素検出部１２０に送る。ここで、出現する頻度とは、例えば、一の文書情報における要素識別情報の個数である。これに代えて、出現する頻度とは、文書情報のサイズ当たりの、要素識別情報の個数であってもよい。

【００１６】

間隙構成要素検出部１２０は、構成要素選択部１１０により選択された各被選択情報について、当該被選択情報と、文書情報において当該被選択情報の次に配列され当該被選択情報と同一の種類である情報との間に配列された間隙構成要素を検出する。そして、間隙構成要素検出部１２０は、間隙構成要素を被選択情報毎に文書内配列パターン変換部１３０に送る。また、好ましくは、間隙要素検出部１２０は、文書情報において複数の被選択情報の後に配列される終端構成要素を更に検出する。この場合、間隙要素検出部１２０は、更に、終端構成要素を文書内配列パターン変換部１３０に送る。

【００１７】

文書内配列パターン変換部１３０は、関連文書検出部１００により変換された文書情報において、予め定められた基準頻度以上で繰り返し出現する要素識別情報の組のそれぞれを変換元の候補として選択する。そして、文書内配列パターン変換部１３０は、変換元の候補である当該要素識別情報の組の配列パターンを示す要素識別情報を変換先の候補として生成する。そして、文書内配列パターン変換部１３０は、変換元の候補及び変換先の候補を変換指示入力部１３５に送る。

【００１８】

文書内配列パターン変換部１３０は、変換指示入力部１３５から変換指示を受けると、変換元の候補である要素識別情報の組のそれぞれを、変換先の候補である要素識別情報に変換し、変換結果を反復終了判断部１４０に送る。要素識別情報の組を検出する具体例として、文書内配列パターン変換部１３０は、以下の処理を行う。まず、文書内配列パターン変換部１３０は、間隙構成要素を被選択情報毎に、間隙要素検出部１２０から受け取る。また、文書内配列パターン変換部１３０は、終端構成要素を、複数の被選択情報のうち文書情報の最後に配列される被選択情報に対応付けて受け取る。

【００１９】

続いて、文書内配列パターン変換部１３０は、被選択情報及び当該被選択情報に対応した間隙構成要素を、要素識別情報の組として検出し、当該要素識別情報の組が基準頻度以上

で繰り返し出現する場合に、各要素識別情報の組を変換する。また、文書内配列パターン検出部１３０は、複数の被選択情報のうち文書情報の最後に配列される被選択情報と、終端構成要素とを、要素識別情報の組として検出する。これにより、文書情報の終端に配列された終端構成要素を、文書内配列パターン変換部１３０による変換の対象に含めることができる。

なお、好ましくは、文書内配列パターン変換部１３０は、変換前の要素識別情報の組及び変換後の要素識別情報を対応付けてメモリ等に記録しておく。

【００２０】

更に、文書内配列パターン変換部１３０は、配列パターンを示す要素識別情報に変換した変換元である複数の文書構成要素を、変換先の要素識別情報に対応付けて、構成要素変換部１０５による新たな変換対象である新規登録要素集合としてメモリ等に登録し、構成要素変換部１０５に通知する。また、文書内配列パターン変換部１３０は、当該新規登録要素集合を登録するべきか否かを判断するべく、新規登録要素集合の候補となる情報を利用者等に対して出力してもよい。そして、文書内配列パターン変換部１３０は、利用者からの入力に基づいて、当該新規登録要素集合を登録してもよい。

【００２１】

変換指示入力部１３５は、文書内配列パターン変換部１３０における変換元の候補である要素識別情報の組及び変換先の候補である要素識別情報を利用者に対して出力することにより、文書内配列パターン変換部１３０による要素識別情報への変換を行うか否かを入力させ、入力結果に応じて変換指示を文書内配列パターン変換部１３０に送る。また、変換指示入力部１３５は、文書内配列パターン変換部１３０による要素識別情報への変換を行うか否かを利用者に入力させなくともよい。この場合、文書内配列パターン変換部１３０は、変換処理の毎に利用者に問い合わせることなく、予め定めた規則に従い、文書構成要素の組を変換する。

【００２２】

反復終了判断部１４０は、文書情報を文書内配列パターン変換部１３０から受け取ると、文書内配列パターン変換部１３０によって変換された文書情報に対して、更に、文書内配列パターン変換部１３０による変換を繰り返し行わせるべく、受け取った文書情報を構成要素選択部１１０に送る。また、反復終了判断部１４０は、予め定められた終了条件を満たした場合、例えば、文書内配列パターン変換部１３０によって変換対象の要素識別情報の組が検出されなかった場合に、繰り返し処理を終了するべく、文書情報を文書間配列パターン変換部１６０に送る。

【００２３】

文書間配列パターン変換部１６０は、文書情報、例えば、対象文書情報及び関連文書情報を反復終了判断部１４０から順次受け取る。そして、文書間配列パターン変換部１６０は、対象文書情報及び関連文書情報から、文書内配列パターン変換部１３０により変換された要素識別情報を除外した文書について、当該対象文書情報及び当該関連文書情報の双方において出現する要素識別情報の組を特定する。続いて、文書間配列パターン変換部１６０は、特定した当該要素識別情報の組のうち、当該対象文書情報及び当該関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する要素識別情報の組を検出する。そして、文書間配列パターン変換部１６０は、基準頻度以上で繰り返し出現する要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換し、変換結果の対象文書情報及び関連文書情報をグループ分類部１７０に送る。

【００２４】

なお、文書間配列パターン変換部１６０は、変換前の要素識別情報の組及び変換後の要素識別情報を対応付けてメモリ等に記録しておく。更に、文書間配列パターン変換部１６０は、配列パターンを示す要素識別情報に変換した変換元である複数の文書構成要素を、変換先の要素識別情報に対応付けて、構成要素変換部１０５による新たな変換対象である新規登録要素集合としてメモリ等に登録し、構成要素変換部１０５に通知する。また、文書

間配列パターン変換部１６０は、当該新規登録要素集合を登録するべきか否かを判断するべく、新規登録要素集合の候補となる情報を利用者等に対して出力してもよい。そして、文書間配列パターン変換部１６０は、利用者からの入力に基づいて、当該新規登録要素集合を登録してもよい。

【００２５】

グループ分類部１７０は、文書間配列パターン変換部１６０から受け取った対象文書情報及び関連文書情報のそれぞれにおいて、文書内配列パターン変換部１３０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する。同様に、グループ分類部１７０は、文書間配列パターン変換部１６０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する。そして、グループ分類部１７０は、分類結果と伴に、対象文書情報及び関連文書情報を並替出力部１８０、目次情報出力部１９０、グループ識別情報生成部２００、アノテーション出力部２０２、及び文書構造情報生成部２１０に送る。

【００２６】

更に、グループ分類部１７０は、分類したグループのそれぞれについて、当該グループ内の文書構成要素が文書情報において果たす役割又は当該グループ内の文書構成要素の内容を示すタイトル情報を生成し、目次情報出力部１９０、及びアノテーション出力部２０２に出力してもよい。より詳しくは、当該グループに分類された要素識別情報の組が、コメント情報を境界として検出されたものである場合には、グループ分類部１７０は、当該コメント情報に含まれる情報をタイトル情報として生成してもよい。

【００２７】

また、グループ分類部１７０は、分類したグループのそれぞれについて、文書情報における当該グループの重要度を示す重要度情報を更に生成し、並替出力部１８０、目次情報出力部１９０、及びアノテーション出力部２０２に出力してもよい。具体的には、グループ分類部１７０は、当該要素識別情報の組が文書情報のどこに位置するかを示す配置情報、当該要素識別情報の組に変換された文書構成要素の色の情報、当該文書構成要素の大きさの情報、及び当該文書構成要素が文字列である場合の文字列の内容に基づいて、重要度情報を生成してもよい。

【００２８】

更に、グループ分類部１７０は、グループに属する文書構成要素を利用者に対して出力し、当該グループに分類するべきか否かを示す指示を受け取ってもよい。この場合、グループ分類部１７０は、当該指示に応じてグループに分類するか否かを判断する。また、グループ分類部１７０は、グループに属する文書構成要素を利用者に対して出力し、当該グループに対応付けて生成するべきタイトル情報を指示する役割指定指示を受け取ってもよい。この場合、グループ分類部１７０は、役割指定指示に応じてグループ情報を生成してもよい。

【００２９】

並替出力部１８０は、対象文書情報において、複数の文書構成要素を、グループ分類部１７０により分類されたグループ毎に並び替える。更に、並替出力部１８０は、グループの内容に応じて、複数のグループを並び替えてもよい。例えば、並替出力部１８０は、文字や画像を含むグループを重要度の高いグループとしてより先に並び替え、リンクリスト、ヘッダ、フッタ、及び広告を含むグループを重要度の低いグループとしてより後に並び替えてもよい。即ち、並替出力部１８０は、複数のグループ中の文書構成情報を、当該グループの重要度情報が高い順に並び替えてもよい。そして、並替出力部１８０は、並び替えた結果の対象文書情報を表示部２０５に送る。

【００３０】

処理の具体例としては、並替出力部１８０は、構成要素変換部１０５、文書内配列パターン変換部１３０、及び文書間配列パターン変換部１６０によりメモリ等に格納されてきた情報に基づいて、各グループに分類された文書構成要素を特定し、当該文書構成要素をグ

ループ毎に並び替える。なお、並替出力部１８０は、更に、グループ分類部１７０から受け取った関連文書情報の分類結果に基づいて、対象文書情報及び関連文書情報のそれぞれから、タイトル情報が同一であるグループを選択してもよい。この場合、並替出力部１８０は、選択したこれらのグループに属する文書構成要素を表示部２０５に送ってもよい。

【００３１】

目次情報出力部１９０は、グループ分類部１７０により分類された各グループについて、当該グループに含まれる文書構成要素に対する目次を示す目次情報を、受け取ったタイトル情報に基づいて生成し表示部２０５に出力する。また、目次情報出力部１９０は、グループ分類部１７０により分類された各グループについて、当該グループに含まれる文書構成要素が、文書情報のどの部分に位置するかを示す情報を更に出力してもよい。より具体的には、この情報とは、ＨＴＭＬ文書におけるアンカーを用いた情報であってもよい。更に、目次情報出力部１９０は、各グループの位置を示すこの情報に、グループのタイトル又はグループの重要度情報を対応付けて出力してもよい。

【００３２】

グループ識別情報生成部２００は、グループ分類部１７０により分類された各グループに対応付けて、当該グループを識別するグループ識別情報を文書情報内に生成し、表示部２０５に出力する。例えば、グループ識別情報生成部２００は、文書情報内のグループの境界に、グループの境界であることを明示するための画像を生成してもよい。これらを受けて、表示部２０５は、対象文書情報及び関連文書情報に応じて、表示画面上に情報を出力する。

【００３３】

また、文書情報がＨＴＭＬ文書である場合には、グループ識別情報生成部２００は、以下の処理を行ってもよい。例えば、グループ識別情報生成部２００は、グループ識別情報を音声で出力させるための出力指示情報を、文書情報内の所定のタグにおけるａｌｔオプションのパラメータとして生成してもよい。出力指示情報とは、例えば、透明かつ非常に小さな画像など、表示画面上では視覚的に識別することが困難な画像などである。この結果、通常のブラウザ画面での表示に与える影響を非常に小さなものに抑える一方、音声ブラウザ等は、当該画像ファイルの内容等に基づいて、グループ識別情報を音声として出力する。これにより、通常のブラウザを使用する健常者の利便性を保ちつつ、音声ブラウザを使用する障害者等の利便性を高めることができる。

【００３４】

アノテーション出力部２０２は、グループ分類部１７０により分類された各グループを識別する情報を、文書情報とは別体に生成して出力する。例えば、アノテーション出力部２０２は、複数のグループのそれぞれについて、当該グループに属する文書構成要素が対象文書情報のどの領域に位置するかを示す配置情報であるアノテーション情報を、各グループを識別する情報として出力する。一例としては、アノテーション出力部２０２は、ＸＰａｔｈ又はＸＰｏｉｎｔｅｒ等の技術により配置情報を生成し出力してもよい。更に、アノテーション出力部２０２は、受け取ったタイトル情報及び重要度情報を配置情報に更に対応付けた情報を、アノテーション情報として出力してもよい。

【００３５】

また、情報処理装置１０は、アノテーション情報の作成者による作業を支援することができる。例えば、アノテーション情報の作成者は、情報処理装置１０により出力されたアノテーション情報に基づき、当該アノテーション情報を修正又は変更することにより、所望のアノテーション情報を作成することができる。この結果、アノテーション情報の作成者は、初めからアノテーション情報を作成する場合と比較して、所望のアノテーション情報を効率的に作成することができる。

【００３６】

文書構造情報生成部２１０は、グループ分類部１７０から受け取った対象文書情報及び関連文書情報のそれぞれについて、当該文書情報の構造を示す文書構造情報を生成し、当該文書構造情報を、受け取った分類結果に対応付けて文書情報同一性出力部２２０に送る。

具体的には、文書情報は、文書構成要素として、表示画面に表示する表示情報と、当該表示情報の表示形式を指定するタグ情報とを含むタグ付き文書である。そして、タグ情報は、当該タグ情報により表示形式を指定する表示情報において更に内側タグ情報を含む、外側タグ情報である。この場合、文書構造情報生成部２１０は、構成要素変換部１０５による変換後の文書情報において、外側タグ情報を根ノードとして生成し、内側タグ情報を当該根ノードの葉ノードとして生成した情報を、文書構造情報として生成する。

【００３７】

文書情報同一性出力部２２０は、文書構造情報生成部２１０から受け取った文書情報のそれぞれについて、文書構造情報生成部２１０により生成された文書構造情報を比較することにより、一の文書情報が他の文書情報と同一の構造を有するか否かを判断し、判断結果を出力する。例えば、文書情報同一性出力部２２０は、文書構造情報の葉ノードのそれぞれに対して予めハッシュ値を定めておき、当該ハッシュ値が異なるか否かを判断するＤＯＭＨＡＳＨ技術を用いてもよい。更に、文書情報同一性出力部２２０は、分類情報を用いて同一性を判断してもよい。

【００３８】

このように、情報処理装置１０は、文書情報に含まれる複数の文書構成要素をグループに分類する。更に、情報処理装置１０は、分類したグループ毎に文書構成要素を並び替え、目次を作成する等の処理を行うことができる。

【００３９】

図２は、情報処理装置１０のフローチャートを示す。関連文書検出部１００は、グループ分けを行う対象の対象文書情報を取得すると、当該対象文書情報と予め定められた関係を有する関連文書情報を検出する（Ｓ２００）。例えば、関連文書検出部１００は、対象文書情報が格納される格納位置から予め定められた範囲内に格納されている文書情報を、関連文書情報として検出する。

【００４０】

これに代えて、対象文書情報が、当該対象文書情報が生成される以前に存在していた既存文書情報を更新することにより生成された場合には、関連文書検出部１００は、当該既存文書情報を関連文書情報として検出してもよい。より詳細には、関連文書検出部１００は、対象文書情報と同一のファイル名、パス名、又はＵＲＬを有する文書を、関連文書情報として検出してもよい。

また、関連文書検出部１００は、対象文書情報が格納される格納位置から予め定められた範囲内に格納されている文書情報及び対象文書情報が生成される以前に存在していた既存文書情報の双方を、関連文書情報として検出してもよい。

【００４１】

構成要素変換部１０５は、文書情報（例えば、対象文書情報又は関連文書情報）における複数の文書構成要素のそれぞれを当該文書構成要素の種類又は役割を示す要素識別情報にそれぞれ変換する（Ｓ２１０）。そして、文書内配列パターン変換部１３０は、変換後の文書情報において、予め定められた基準頻度以上で繰り返し出現する要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する（Ｓ２２０）。

【００４２】

文書間配列パターン変換部１６０は、対象文書情報及び関連文書情報から、文書内配列パターン変換部１３０により変換された要素識別情報を除外した文書について、当該対象文書情報及び当該関連文書情報の双方において出現する要素識別情報の組を特定する（Ｓ２３０）。続いて、文書間配列パターン変換部１６０は、特定した当該要素識別情報の組のうち、当該対象文書情報及び当該関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する要素識別情報の組を検出する。そして、文書間配列パターン変換部１６０は、基準頻度以上で繰り返し出現する要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する。

【００４３】

グループ分類部１７０は、文書間配列パターン変換部１６０から受け取った対象文書情報及び関連文書情報のそれぞれにおいて、文書内配列パターン変換部１３０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する（Ｓ２４０）。同様に、グループ分類部１７０は、文書間配列パターン変換部１６０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する。

【００４４】

文書構造情報生成部２１０は、文書間配列パターン変換部１６０から受け取った対象文書情報及び関連文書情報のそれぞれについて、当該文書情報の構造を示す文書構造情報を生成する（Ｓ２５０）。そして、文書情報同一性出力部２２０は、対象文書情報及び関連文書情報のそれぞれが同一の構造を有するか否かを、文書構造情報生成部２１０により生成された文書構造情報を比較することにより判断し、判断結果を出力する。

【００４５】

なお、文書構造情報生成部２１０及び文書構造情報生成部２１０による上記処理は、本実施形態における必須の構成とはならない。即ち、文書構造情報生成部２１０は、文書構造情報を生成しなくともよい。また、文書情報同一性出力部２２０は、同一性の判断結果を出力しなくともよい。

【００４６】

また、情報処理装置１０は、Ｓ２２０及びＳ２３０のうち一方の処理を行わなくともよい。即ち、文書内配列パターン変換部１３０は、基準頻度以上で繰り返し出現する要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換しなくともよい。また、文書間配列パターン変換部１６０は、対象文書情報及び関連文書情報から、文書内配列パターン変換部１３０により変換された要素識別情報を除外した文書について、当該対象文書情報及び当該関連文書情報の双方において出現する要素識別情報の組を特定しなくともよい。

【００４７】

図３は、図２のＳ２１０におけるフローチャートを示す。構成要素変換部１０５は、文書情報における各文書構成要素について以下の処理を繰り返し実行する。文書情報は、文書構成要素として、表示画面上に表示する画像を識別する画像識別情報を有している。構成要素変換部１０５は、文書構成要素がこの画像識別情報である場合に（Ｓ３００：ＹＥＳ）、当該画像識別情報を、当該画像識別情報により識別される画像の形状、データサイズ、又は格納位置を示す要素種別情報に変換する（Ｓ３１０）。

【００４８】

例えば、構成要素変換部１０５は、当該画像識別情報により識別される画像における、縦の長さ及び横の長さの比率が所定値より大きい場合に、当該画像識別情報を、文章等の区切り画像を示す要素識別情報である＜ｄｉｖｉｄｅｒ　／＞に変換してもよい。また、構成要素変換部１０５は、透明又は単一色、かつ所定サイズより小さな画像であって、縦の長さ及び横の長さの比率が所定値より大きい領域に配置された画像を識別する画像識別情報を、＜ｄｉｖｉｄｅｒ　／＞に変換してもよい。

【００４９】

一方、文書構成要素が画像識別情報でない場合に（Ｓ３００：ＮＯ）、構成要素変換部１０５は、文書構成要素がテキストデータであるか否かを判断する（Ｓ３２０）。文書構成要素がテキストデータである場合に（Ｓ３２０：ＹＥＳ）、構成要素変換部１０５は、当該テキストデータを、当該テキストデータの内容又はデータサイズを示す要素種別情報に変換する（Ｓ３３０）。

【００５０】

文書構成要素がテキストデータでない場合に（Ｓ３２０：ＮＯ）、構成要素変換部１０５は、文書構成要素が、利用者からの指示を受けて他の情報を表示させるリンク情報であるか否かを判断する（Ｓ３４０）。文書構成要素がリンク情報である場合に（Ｓ３４０：ＹＥＳ）、構成要素変換部１０５は、当該リンク情報のリンク先に基づく要素識別情報に変

換する（Ｓ３５０）。
【００５１】
構成要素変換部１０５は、以上の処理を各文書構成要素に適用した後、次の処理を行う。構成要素変換部１０５は、文書構成要素が所定の規則に適合するか否かを判断する（Ｓ３６０）。例えば、構成要素変換部１０５は、文書内の索引を示すリンクリスト情報を検出するべく、当該リンクリスト情報を形成する文書構成要素の配列パターンを予め格納しておき、文書構成要素の配列が当該配列パターンに適合するか否かを判断する。
【００５２】
変換後の文書情報において、要素識別情報が所定の規則に適合する場合に（Ｓ３６０：ＹＥＳ）、構成要素変換部１０５は、当該文書構成要素を、当該規則の内容を示す要素識別情報に変換する（Ｓ３７０）。例えば、構成要素変換部１０５は、リンクリストを形成する文書構成要素の組を、リンクリストである旨を示す要素識別情報に変換する。他の例としては、構成要素変換部１０５は、野球等のスポーツのスコアボードを形成する文書構成要素の組を、スコアボードを示す要素識別情報に変換する。また、構成要素変換部１０５は、所定の広告用サイトへのリンク情報であって、リンク先の情報を表示するべく指示を受ける表示情報が画像である場合に、当該リンク情報及び画像を、広告を示す要素識別情報である＜ａｄ　／＞に変換する。
【００５３】
また、構成要素変換部１０５は、互いに異なる複数の文書構成要素を同一の要素識別情報に変換してもよい。例えば、構成要素変換部１０５は、所定サイズ以下の画像を示す文書構成要素である＜ｉｍｇ　ｓｒｃ＝”ａｒｒｏｗ．ｇｉｆ”＞と、箇条書きの開始点等を示す点である「・」とを、同一の要素識別情報である＜ｂｕｌｌｅｔ　／＞に変換してもよい。このように、構成要素変換部１０５は、互いに異なる複数の文書構成要素であっても、同一の種類であれば、当該文書構成要素を、当該種類を識別する要素識別情報に変換する。
【００５４】
また、構成要素変換部１０５は、互いに同一な複数の文書構成要素を、互いに異なる複数の要素識別情報に変換してもよい。例えば、構成要素変換部１０５は、箇条書きを示す文書構成要素である「・今日のニュース」を、それぞれが要素識別情報である＜ｂｕｌｌｅｔ　／＞及び＜ｓｈｏｒｔｔｅｘｔ　／＞に変換する。一方、構成要素変換部１０５は、文書構成要素である「今日のテストは算数・理科です」を、要素識別情報である＜ｓｈｏｒｔｔｅｘｔ　／＞に変換する。即ち、構成要素変換部１０５は、互いに同一な文書構成要素である点「・」であっても、当該文書構成要素の役割に応じて互いに異なる要素識別情報に変換する。
【００５５】
また、構成要素変換部１０５は、文書内配列パターン変換部１３０又は文書間配列パターン１６０からの通知を受けて、新規登録要素集合を、当該新規登録要素集合の変換先とするべき要素識別情報に対応付けて、メモリ等から取得する。そして、構成要素変換部１０５は、文書情報における新規登録要素集合のそれぞれを、対応する要素識別情報に変換する。即ち、情報処理装置１０は、以前に変換されたことのある文書構成要素の組については、文書内配列パターン変換部１３０又は文書間配列パターン１６０に代えて、構成要素変換部１０５により変換処理を行う。これにより、情報処理装置１０は、変換処理を効率化できる。
【００５６】
このように、構成要素変換部１０５は、文書情報における複数の文書構成要素のそれぞれを、文書構成要素の種類又は役割を示す要素識別情報にそれぞれ変換する。また、構成要素変換部１０５は、所定の条件を満たす要素識別情報の組を、当該条件の内容を示す要素識別情報に更に変換する。この結果、情報処理装置１０は、例えば、ある文書構成要素の組が、利用者に対して表形式で表示するべく配置されたものであるか、或いは、文書情報のレイアウトを整えるべく表形式を利用したものであるのかを区別することができる。更

に、情報処理装置１０は、一部の文書構成要素の組については、当該文書構成要素の組の果たす役割、例えば、スコアボードである旨又は広告である旨を特定することができる。

【００５７】

また、構成要素変換部１０５による処理の手順は、本図の例に限定されない。例えば、構成要素変換部１０５は、Ｓ３６０における所定の規則が、要素識別情報でなく文書構成要素を対象とする場合には、画像、テキスト、及びリンクの変換（Ｓ３１０、Ｓ３３０、及びＳ３５０）に先だって、所定の規則に適合するか否かを判断してもよい（Ｓ３６０）。

【００５８】

図４は、図２のＳ２２０におけるフローチャートを示す。構成要素選択部１１０は、構成要素変換部１０５により変換された各文書情報における要素識別情報を、出現する頻度の高い順に、順次被選択情報として選択する（Ｓ４００）。そして、間隙構成要素検出部１２０は、構成要素選択部１１０により選択された当該被選択情報と、文書情報において当該被選択情報の次に配列され当該被選択情報と同一の種類である情報との間に配列された間隙構成要素を検出する（Ｓ４１０）。

【００５９】

ここで、間隙構成要素は、被選択情報と同一の種類の要素識別情報を含まない。即ち、間隙構成要素検出部１２０は、被選択情報の次に配列される要素識別情報を順次検出し、被選択情報と同一の種類の要素識別情報を検出するまでに検出された要素識別情報を、間隙構成要素とする。

【００６０】

このように、文書内配列パターン変換部１３０は、被選択情報及び間隙構成要素を、要素識別情報の組として検出する。これに代えて、対象文書情報が、対象文書中に配列され表示画面に表示されないコメント情報を含む場合、文書内配列パターン変換部１３０は、当該コメント情報を、一の要素識別情報の組と、他の要素識別情報の組との境界を示す情報として用いてもよい。例えば、対象文書情報が、コメント情報として、＜！－ＳＴＡＲＴ　ＡＡ＞及び＜！－ＥＮＤ　ＡＡ＞を含む場合、文書内配列パターン変換部１３０は、これらのコメント情報の間に配列される情報を、要素識別情報の組として検出してもよい。更に、グループ分類部１７０は、当該コメント情報に含まれる情報、例えば、「ＡＡ」の部分の文字列に基づいて、グループの果たす役割を特定してもよい。更に、文書内配列パターン変換部１３０は、ＳＴＡＲＴ等の所定の文字列がコメント情報内に含まれていない場合には、当該コメント情報を要素識別情報の組の境界として用いなくともよい。

【００６１】

また、更に他の例として、文書内配列パターン変換部１３０は、所定以上のサイズの空白又は改行を示す文書構成要素を、一の要素識別情報の組と、他の要素識別情報の組との境界を示す情報として用いてもよい。

【００６２】

文書内配列パターン変換部１３０は、構成要素選択部１１０により選択された被選択情報及び間隙構成要素検出部１２０により検出された間隙構成要素の組が、構成要素選択部１１０及び間隙構成要素検出部１２０により既に検出されていた要素識別情報の組より出現頻度が高い場合に（Ｓ４２０：ＹＥＳ）、構成要素選択部１１０により選択された被選択情報及び間隙構成要素検出部１２０により検出された間隙構成要素の組を変換対象の候補として選択する（Ｓ４３０）。一方、既に検出されていた要素識別情報の組より出現頻度が低い場合に（Ｓ４２０：ＮＯ）、情報処理装置１０は、Ｓ４４０に処理を移す。

【００６３】

間隙要素検出部１２０が検出する間隙構成要素とは、文書構成要素及び要素識別情報の何れも含まない空集合であってもよい。この場合、文書内配列パターン変換部１３０は、同一の種類の被選択情報が複数連続して配列される場合に、当該複数の被選択情報を、変換対象の候補として選択する。

【００６４】

構成要素選択部１１０は、文書情報における何れかの要素識別情報について未だ被選択情

報として選択していない場合に（Ｓ４４０：ＮＯ）、Ｓ４００に処理を戻すことにより、次に出現頻度の高い要素識別情報を被選択情報として選択する。一方、構成要素選択部１１０は、文書情報における全ての要素識別情報を被選択情報として選択した場合に（Ｓ４４０：ＹＥＳ）、Ｓ４３０において変換対象の候補として選択した要素識別情報の組を、所定の判断に基づき、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する処理を行う（Ｓ４５０）。

【００６５】

反復終了判断部１４０は、予め定められた終了条件を満たしていない場合に（Ｓ４６０：ＮＯ）、変換後の文書情報について更に変換処理を行うべく、Ｓ４００に処理を戻す（Ｓ４７０）。ここで、反復終了判断部１４０は、文書内配列パターン変換部１３０により変換される対象となる要素識別情報の組について、当該要素識別情報の変換元である文書構成要素の合計サイズが、予め定められた条件を満たす場合、例えば一つの表示画面に一度に表示可能な情報のサイズに達した場合に、当該要素識別情報の組を、以降の処理において文書内配列パターン変換部１３０による変換を繰り返し行わせる対象から除外する。例えば、反復終了判断部１４０は、当該要素識別情報の組を、以降の変換を停止する旨を示す所定の要素識別情報に変換してもよい。これにより、各グループにおける文書構成要素の合計サイズを所定サイズ以下に抑え、各グループを表示画面内に表示させることができる。

【００６６】

これに代えて、反復終了判断部１４０は、文書構成要素の合計サイズが、２つの表示画面に表示可能な情報のサイズに達した場合に、当該要素識別情報の組を、以降の処理において文書内配列パターン変換部１３０による変換を繰り返し行わせる対象から除外してもよい。即ち、表示画面に表示可能な情報のサイズとは、利用者が情報を閲覧するべく表示画面をスクロールする量が予め定めた量以下となる情報のサイズであってもよい。

【００６７】

一方、予め定められた終了条件を満たした場合に（Ｓ４６０：ＹＥＳ）、反復終了判断部１４０は、処理を終了する。例えば、反復終了判断部１４０は、Ｓ４５０において変換対象の候補の何れも変換されなかった場合に、予め定められた終了条件を満たしたと判断してもよい。

【００６８】

図５は、図４のＳ４５０におけるフローチャートを示す。文書内配列パターン変換部１３０は、変換対象の候補となる要素識別情報の組のそれぞれについて、以下の処理を行う。文書内配列パターン変換部１３０は、変換対象の候補となる要素識別情報の組が、文書情報において基準頻度以上出現していない場合に（Ｓ５００：ＮＯ）、当該要素識別情報の組を変換対象から除外するべく処理を終了する。

【００６９】

一方、変換対象の候補となる要素識別情報の組が、文書情報において基準頻度以上出現した場合に（Ｓ５００：ＹＥＳ）、変換対象の候補となる要素識別情報の組における変換の階層が、所定の値以下であるか否かを判断する（Ｓ５１０）。ここで、要素識別情報の組における変換の階層とは、例えば、変換前の文書構成要素を当該要素識別情報の組に変換するべく文書内配列パターン変換部１３０が文書情報に対して繰り返す変換の回数をいう。

【００７０】

変換対象の候補となる要素識別情報の組における変換の階層が、所定の値を超える場合に（Ｓ５１０：ＮＯ）、文書内配列パターン変換部１３０は、当該要素識別情報の組を変換対象から除外するべく処理を終了する。一方、変換対象の候補となる要素識別情報の組における変換の階層が、所定の値以下である場合に（Ｓ５１０：ＹＥＳ）、文書内配列パターン変換部１３０は、変換対象の候補となる要素識別情報の組について、当該要素識別情報の組に変換される前の文書構成要素の合計サイズが、予め定められた基準サイズより小さいか否かを判断する（Ｓ５２０）。

【００７１】
当該要素識別情報の組に変換される前の文書構成要素の合計サイズが、予め定められた基準サイズ以上の場合に（Ｓ５２０：ＮＯ）、文書内配列パターン変換部１３０は、当該要素識別情報の組を変換対象から除外するべく処理を終了する。一方、当該要素識別情報の組に変換される前の文書構成要素の合計サイズが、予め定められた基準サイズより小さい場合に（Ｓ５２０：ＹＥＳ）、文書内配列パターン変換部１３０は、変換対象の候補となる要素識別情報の組を、変換対象と決定する。即ち、文書内配列パターン変換部１３０は、文書情報に出現する当該要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する（Ｓ５３０）。
【００７２】
このように、文書内配列パターン変換部１３０は、基準頻度以上で繰り返し出現する要素識別情報の組のうち、当該要素識別情報の組に変換される前の文書構成要素の合計サイズが予め定められた基準サイズより小さい要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する。これにより、グループ内の文書構成要素のサイズが大きくなりすぎるのを防ぐことができる。例えば、情報処理装置１０がＰＤＡ等の携帯情報通信装置であり、表示画面の大きさが限定される場合であっても、情報処理装置１０は、当該表示画面に表示可能なグループに分類することができる。
【００７３】
また、文書内配列パターン変換部１３０は、基準頻度以上で繰り返し出現する要素識別情報の組であって、変換前の文書構成要素を当該要素識別情報の組に変換するべく文書内配列パターン変換部１３０が文書情報に対して繰り返す変換の回数が基準回数以下の要素識別情報の組を、配列パターンを示す要素識別情報に変換する。これにより、グループ内の文書構成要素のサイズが大きくなりすぎるのを防ぐことができる。更に、文書内配列パターン変換部１３０は、分類するべきグループにおける変換の階層が予め特定できている等の場合には、要素識別情報の組を適切に変換することができる。
【００７４】
なお、本図に示した処理の順序は一例であり、情報処理装置１０は、本図の例に代えて、他の順序で判断を行ってもよい。また、情報処理装置１０は、本図に示した判断の少なくとも一部を行わなくともよい。
【００７５】
例えば、当該要素識別情報の組に変換される前の文書構成要素の合計サイズが、予め定められた基準サイズ以上の場合であっても（Ｓ５２０：ＮＯ）、文書内配列パターン変換部１３０は、当該要素識別情報の組を変換対象としてもよい。この場合、文書情報全体が一のグループとして分類されることとなる。これにより、より精度が高くかつ柔軟なグループ分けを実現することができる場合がある。以下に例を示す。
【００７６】
情報処理装置１０は、文書内配列パターン変換部１３０による変換の経過を示す情報を利用者に出力する。例えば、情報処理装置１０は、変換処理のそれぞれについて、変換元である要素識別情報の組と、変換先である要素識別情報とを対応付けて、変換の経過を示す情報として階層構造により出力する。利用者は、当該階層構造の中から、実際にグループとして分割することを希望する要素識別情報を選択する。
【００７７】
これにより、利用者による入力が必要となるものの、基準サイズによりグループを決定する場合と比較して、より適切かつ柔軟にグループ分けを実現できる。また、基準サイズ以下のグループにおいて当該グループのタイトル情報が特定できなかった場合であっても、基準サイズ以上のグループにおいて当該グループのタイトル情報が特定できた場合には、このタイトル情報を利用して基準サイズ以下のグループのタイトル情報を特定し、グループ分けの精度を高めることができる場合がある。
【００７８】
また、情報処理装置１０は、本図に示した判断に対して、更に他の判断を追加して行って

もよい。例えば、文書内配列パターン変換部１３０は、要素識別情報の組として検出するべきでない配列パターンを、利用者からの入力等に基づいて定めてもよい。即ち、文書内配列パターン変換部１３０は、当該配列パターンと一致する要素識別情報の組を、変換対象から除外する。

【００７９】

要素識別情報の組として検出するべきでない配列パターンとは、例えば、表示領域の区切りように用いられる所定の形状の画像を含む配列パターンであってもよいし、所定以上の面積を有する空白を含む配列パターンであってもよい。また、文書内配列パターン変換部１３０は、要素識別情報の組として検出するべきでない配列パターンか否かを、変換処理の毎に利用者に問い合わせてもよいし、予め利用者に入力させておいてもよい。

【００８０】

また、文書内配列パターン変換部１３０は、図４及び図５に示した方法に代えて、他の方法を用いて、要素識別情報の組を検出し、当該要素識別情報の組の配列パターンを示す要素識別情報に変換してもよい。即ち、文書内配列パターン変換部１３０が要素識別情報の組を検出する方法は、基準頻度以上出現する要素識別情報の組を検出する方法であればよい。

【００８１】

図６は、図２のＳ２３０におけるフローチャートを示す。文書間配列パターン変換部１６０は、対象文書情報及び関連文書情報のそれぞれから、文書内配列パターン変換部１３０により変換された要素識別情報を除外した文書を検出対象とする（Ｓ６００）。

【００８２】

より具体的には、文書間配列パターン変換部１６０は、文書内配列パターン変換部１３０により変換された要素識別情報を、その旨を示す要素識別情報である＜ｇｒｏｕｐ　／＞に変換してもよい。これにより、文書間配列パターン変換部１６０は、以降の処理で、文書内配列パターン変換部１３０により変換された要素識別情報を適切に検出し、除外することができる。

【００８３】

続いて、文書間配列パターン変換部１６０は、特定した当該要素識別情報の組のうち、当該対象文書情報及び当該関連文書情報を併せた文書において、要素識別情報を、出現する頻度の高い順に、順次被選択情報として選択する（Ｓ６１０）。ここで、文書間配列パターン変換部１６０は、当該対象文書情報及び当該関連文書情報を併せた文書に対して、当該対象文書情報及び当該関連文書情報の境界を示す情報を挿入する。そして、文書間配列パターン変換部１６０は、被選択情報を含む要素識別情報の組を、変換対象の候補として選択する（Ｓ６２０）。

【００８４】

具体的には、文書間配列パターン変換部１６０は、被選択情報及び間隙構成要素を、要素識別情報の組として選択する。ここで、文書間配列パターン変換部１６０は、文書内配列パターン変換部１３０により既に要素識別情報に変換された旨を示す＜ｇｒｏｕｐ　／＞を、一の文書構成要素の組と、他の文書構成要素の組との境界として用いてもよい。また、文書間配列パターン変換部１６０は、対象文書情報及び関連文書情報の境界を示す情報を、一の文書構成要素の組と、他の文書構成要素の組との境界として用いてもよい。

【００８５】

変換対象の候補である要素識別情報の組が、予め定められた基準頻度以上出現する場合に（Ｓ６３０：ＹＥＳ）、文書間配列パターン変換部１６０は、当該要素識別情報の組を、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する（Ｓ６４０）。続いて、対象文書情報及び関連文書情報を併せた文書において未だ被選択情報として選択していない要素識別情報がある場合に（Ｓ６５０：ＮＯ）、文書間配列パターン変換部１６０は、Ｓ６１０に処理を戻すことにより、次に出現頻度の高い要素識別情報を被選択情報として選択する。対象文書情報及び関連文書情報を併せた文書における全ての要素識別情報を被選択情報として選択した場合に（Ｓ６５０：ＹＥＳ）、文書間配列パターン変換部１

６０は、処理を終了する。

【００８６】

なお、文書間配列パターン変換部１６０が用いる基準頻度は、文書内配列パターン変換部１３０が用いる基準頻度とは異なっていてもよい。即ち、文書間配列パターン変換部１６０は、文書内配列パターン変換部１３０が判断に用いる基準頻度である第１頻度とは異なる第２頻度以上、変換対象の候補である要素識別情報の組が出現する場合に、要素識別情報の組を、当該要素識別情報の組の配列パターンを示す要素識別情報に変換してもよい。

【００８７】

図７は、図２のＳ２４０におけるフローチャートを示す。グループ分類部１７０は、文書内配列パターン変換部１３０により対象文書情報を繰り返し変換した結果得られた文書情報において、文書内配列パターン変換部１３０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する（Ｓ７００）。同様に、グループ分類部１７０は、文書間配列パターン変換部１６０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する。

【００８８】

更に、グループ分類部１７０は、分類したグループのそれぞれについて、当該グループ内の文書構成要素が文書情報において果たす役割を示すタイトル情報を生成し、目次情報出力部１９０、グループ識別情報生成部２００、及びアノテーション出力部２０２に出力してもよい。より詳しくは、当該グループに分類された要素識別情報の組が、コメント情報を境界として検出されたものである場合には、グループ分類部１７０は、当該コメント情報に含まれる情報をタイトル情報として生成してもよい。

【００８９】

また、グループ分類部１７０は、分類したグループのそれぞれについて、文書情報における当該グループの重要度を示す重要度情報を更に生成し、並替出力部１８０、グループ識別情報生成部２００、及びアノテーション出力部２０２に出力してもよい。具体的には、グループ分類部１７０は、当該要素識別情報の組が文書情報のどこに位置するかを示す配置情報、当該要素識別情報の組に変換された文書構成要素の色の情報、当該文書構成要素の大きさの情報、及び当該文書構成要素が文字列である場合の文字列の内容に基づいて、重要度情報を生成してもよい。

【００９０】

更に、グループ分類部１７０は、文書情報のうちグループに分類した文書構成要素を除外した部分、即ち、グループとして分類できなかった部分を、その旨を示すグループに分類する。この場合、グループ分類部１７０は、グループとして分類できなかった部分を、更に、文書情報内の位置、表示画面上での位置、及び表示画面上での背景色に応じて、更に、複数のグループに分類してもよい。

【００９１】

並替出力部１８０は、対象文書情報及び関連文書情報のそれぞれにおいて、複数の文書構成要素を、グループ分類部１７０により分類されたグループ毎に並び替える（Ｓ７１０）。そして、目次情報出力部１９０は、グループ分類部１７０により分類された各グループについて、当該グループに含まれる文書構成要素に対する目次を示す目次情報を生成する（Ｓ７２０）。更に、グループ識別情報生成部２００は、グループ分類部１７０により分類された各グループに対応付けて、当該グループを音声により識別することを容易にする為の情報を文書情報内に生成する（Ｓ７３０）。

【００９２】

アノテーション出力部２０２は、グループ分類部１７０により分類された各グループを識別するアノテーション情報を出力する（Ｓ７４０）。例えば、アノテーション出力部２０２は、複数のグループのそれぞれについて、当該グループに属する文書構造情報が、対象文書情報のどの領域に位置するかを示す配置情報を、アノテーション情報として出力する。一例としては、アノテーション出力部は、ＸＰａｔｈ又はＸＰｏｉｎｔｅｒ等の技術に

より、配置情報を生成し出力してもよい。

【００９３】

これにより、アノテーション情報を入力とする他の装置等は、対象文書情報中の情報の並び替え、目次作成、又はグループ識別情報の出力のみならず、対象文書情報のダイジェストを作成する等の、他のトランスコーディング技術を用いた処理を行うことができる。

【００９４】

また、情報処理装置１０は、アノテーション情報の作成者による作業を支援することができる。例えば、アノテーション情報の作成者は、情報処理装置１０により出力されたアノテーション情報に基づき、当該アノテーション情報を修正又は変更することにより、所望のアノテーション情報を作成することができる。この結果、アノテーション情報の作成者は、初めからアノテーション情報を作成する場合と比較して、効率的に所望のアノテーション情報を作成することができる。

【００９５】

図８は、構成要素変換部１０５が文書構成要素を要素識別情報に変換する一例を示す。本図に示すように、構成要素変換部１０５は、文書構成要素を、当該文書構成要素の内容、データサイズ、又は格納位置等を示す要素識別情報に変換する。具体的には、構成要素変換部１０５は、文書構成要素が数値である場合に、当該文書構成要素を、数値を示す要素識別情報である＜ｄｉｇｉｔ　／＞に変換する。なお、好ましくは、構成要素変換部１０５は、変換先の要素識別情報として、ＸＭＬ（ｅＸｔｅｎｓｉｂｌｅ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）の規格に準拠した形式のタグ情報を生成する。これにより、ＸＭＬ文書を操作する既存のプログラムを利用することができる。

【００９６】

また、構成要素変換部１０５は、文書構成要素が１００文字以上の文字列である場合に、当該文書構成要素を、長い文字列を示す要素識別情報である＜ｌｏｎｇｔｅｘｔ　／＞に変換する。同様に、構成要素変換部１０５は、３０から１００文字の文字列、２から３０文字の文字列、及び１文字の文字のそれぞれを、要素識別情報である＜ｍｉｄｔｅｘｔ　／＞、＜ｓｈｏｒｔｔｅｘｔ　／＞、及び＜ｌｅｔｔｅｒ　／＞のそれぞれに変換する。

【００９７】

これに代えて、構成要素変換部１０５は、文字列を表すタグ情報に、当該文字列の長さを示す属性を対応付けた情報を、要素識別情報として生成してもよい。例えば、構成要素変換部１０５は、＜ｌｏｎｇｔｅｘｔ　／＞に代えて、所定長より長くかつ数値でない文字列を示す要素識別情報である＜ｔｅｘｔ　ｌｅｎｇｔｈ＝“ｌｏｎｇ“　ｉｓ＿ｄｉｇｉｔ＝”ｎｏ”＞を生成してもよい。更に、構成要素変換部１０５は、文字列の長さを示す情報を属性としたタグ情報を、要素識別情報として生成してもよい。

【００９８】

このように、構成要素変換部１０５は、文字列である文書構成要素を、当該文字列が数字であるか数字以外であるかに応じて定まる要素識別情報に変換する。更に、構成要素変換部１０５は、文字列である文書構成要素を、当該文字列に含まれる語句に応じて定まる要素識別情報に変換してもよい。

【００９９】

また、構成要素変換部１０５は、表示画面に表示する画像を識別する画像識別情報であって、当該画像が対象文書情報と同一のサイト内にあり、かつ当該画像のサイズが３００×３００ピクセル以上である画像識別情報を、要素識別情報である＜ｉｎ－ｌａｒｇｅｉｍｇ　／＞に変換する。一例として、対象文書情報がＨＴＭＬ文書である場合の画像識別情報とは、＜ｉｍｇ　ｓｒｃ＝”ＡＡＡ．ＪＰＧ”＞等のイメージタグ情報である。なお、構成要素変換部１０５は、画像の表示サイズを特定するべく、表示するべき画像をレンダリング及び／又は解析してもよい。より具体的には、構成要素変換部１０５は、画像ファイルのヘッダ部分に記録されたサイズ情報を解析してもよい。

【０１００】

更に、構成要素変換部１０５は、レンダリング等により、表示画面において画像が表示さ

れる位置又は画像の色彩等の情報を解析してもよい。この場合、構成要素変換部１０５は、画像識別情報を、当該画像の表示位置又は色彩等を示す要素識別情報に変換してもよい。

【０１０１】

同様に、構成要素変換部１０５は、画像識別情報であって、当該画像識別情報における画像が対象文書情報と同一サイト内にあり、かつ当該画像のサイズが１００×１００から３００×３００ピクセルである画像識別情報を、要素識別情報である＜ｉｎ－ｍｉｄｉｍｇ　／＞に変換する。また、構成要素変換部１０５は、画像識別情報であって、当該画像識別情報における画像が対象文書情報と同一サイト内にあり、かつ当該画像のサイズが１００×１００ピクセル以下である画像識別情報を、要素識別情報である＜ｉｎ－ｓｍａｌｌ　ｉｍｇ　／＞に変換する。同様に、構成要素変換部１０５は、対象文書情報と異なるサイト内にある画像を識別する画像識別情報を、当該画像の大きさに応じて、＜ｏｕｔ－ｌａｒｇｅｉｍｇ　／＞、＜ｏｕｔ－ｍｉｄｉｍｇ＞、又は＜ｏｕｔ－ｓｍａｌｌｉｍｇ　／＞に変換する。

【０１０２】

これに代えて、構成要素変換部１０５は、画像を表すタグ情報に、当該画像の大きさを示す属性を対応付けた情報を、要素識別情報として生成してもよい。例えば、構成要素変換部１０５は、＜ｉｎ－ｌａｒｇｅｉｍｇ　／＞に代えて、所定範囲内のデータサイズかつサイト内に存在する画像を示す要素識別情報である＜ｉｍａｇｅ　ｓｉｚｅ＝”ｌａｒｇｅ”　ｌｏｃａｔｉｏｎ＝”ｉｎ”　／＞を生成してもよい。

【０１０３】

このように、構成要素変換部１０５は、文書情報における各文書構成要素を、当該文書構成要素のデータサイズに応じた要素識別情報に変換する。また、好ましくは、構成要素変換部１０５は、文字列又は画像の種類に応じて異なる要素識別情報に変換する。例えば、構成要素変換部１０５は、文字列の文字数が同一であっても、文字列において数字、漢字、ひらがな、カタカナ、及びアルファベットのそれぞれが占める割合に応じて、異なる要素識別情報に変換してもよい。また、構成要素変換部１０５は、画像のデータサイズが同一であっても、画像が縦長又は横長であるかに応じて異なる要素識別情報に変換してもよいし、画像のデータ形式に応じて異なる要素識別情報に変換してもよい。

【０１０４】

また、文書情報は、利用者に表示する表示情報と、当該表示情報に対する利用者からの指示に応じて表示するべき他の情報の格納位置を示す格納位置情報とを文書構成要素として含んでいる。例えば、文書情報がＨＴＭＬ文書である場合の、表示情報及び格納位置情報とは、利用者からの指示を受けて他の情報を表示させるハイパーリンクである。一例として、格納位置情報とは、＜ａ　ｈｒｅｆ＝”ＢＢＢ．ＨＴＭＬ”＞等のタグ情報及び＜／ａ＞等のタグ情報であり、表示情報とは、当該＜ａ＞タグ及び＜／ａ＞タグの間に配列された情報である。

【０１０５】

構成要素変換部１０５は、当該格納位置情報を、当該格納位置情報が示す格納位置の範囲を示す要素識別情報に変換する。例えば、構成要素変換部１０５は、リンク先の他の情報が、対象文書情報と同一のディレクトリ内に格納されている場合に、当該他の情報の格納位置を示す格納位置情報を、格納位置の範囲を示す要素識別情報である＜ｓａｍｅｄｉｒ－ｌｉｎｋ＞及び＜／ｓａｍｅｄｉｒ－ｌｉｎｋ＞に変換する。より具体的には、構成要素変換部１０５は、格納位置情報であると伴に表示情報の開始点を示すタグ情報である＜ａ　ｈｒｅｆ＝”ＢＢＢ．ＨＴＭＬ”＞等を＜ｓａｍｅｄｉｒ－ｌｉｎｋ＞に変換する。また、構成要素変換部１０５は、表示情報の終了点を示す＜／ａ＞を、＜／ｓａｍｅｄｉｒ－ｌｉｎｋ＞に変換する。

【０１０６】

これに代えて、構成要素変換部１０５は、リンクを表すタグ情報に、当該リンクのリンク先を示す属性を対応付けた情報を、要素識別情報として生成してもよい。例えば、構成要

素変換部１０５は、＜ｓａｍｅｄｉｒ－ｌｉｎｋ＞及び＜／ｓａｍｅｄｉｒ－ｌｉｎｋ＞に代えて、＜ｌｉｎｋ　ｌｏｃａｔｉｏｎ＝”ｓａｍｅ”　ｔａｒｇｅｔ＝”ｎｏｎｅ”　／＞を生成してもよい。

【０１０７】

また、構成要素変換部１０５は、＜ａ　ｈｒｅｆ＝”ＢＢＢ．ＨＴＭＬ”＞及び＜／ａ＞間に配列された表示情報を変換しなくともよい。また、構成要素変換部１０５は、当該表示情報に付いては、別途他の種類を示す要素識別情報に変換してもよい。

【０１０８】

また、構成要素変換部１０５は、リンク先の他の情報が、対象文書情報と同一サイト内に格納されている場合に、当該他の情報の格納位置を示す格納位置情報を、格納位置の範囲を示す要素識別情報である＜ｉｎ－ｌｉｎｋ＞及び＜／ｉｎ－ｌｉｎｋ＞に変換する。

【０１０９】

また、構成要素変換部１０５は、リンク先の他の情報が、対象文書情報と異なるサイト内に格納されている場合に、当該他の情報の格納位置を示す格納位置情報を、格納位置の範囲を示す要素識別情報である＜ｏｕｔ－ｌｉｎｋ＞及び＜／ｏｕｔ－ｌｉｎｋ＞に変換する。

【０１１０】

このように、構成要素変換部１０５は、文書構成要素の種類又は役割を、タグ情報における属性のパラメータに応じて検出してもよい。そして、構成要素変換部１０５は、当該文書構成要素を、当該文書構成要素のパラメータの種類等を示す要素識別情報に変換する。

【０１１１】

このように、構成要素変換部１０５は、文書情報における複数の文書構成要素のそれぞれを、当該文書構成要素の種類又は役割を示す要素識別情報にそれぞれ変換する。

【０１１２】

図９は、構成要素変換部１０５が所定の要素識別情報の組を要素識別情報に変換する例を示す。構成要素変換部１０５は、図８に示した変換に続いて、予め定めた条件に適合した要素識別情報の組を、要素識別情報に変換する。例えば、本図は、構成要素変換部１０５が、野球等のスポーツにおけるスコアボードを形成する文書構成要素の組を検出し、スコアボードを示す要素識別情報、例えば、＜ｂａｓｅｂａｌｌ－ｓｃｏｒｅｂｏａｒｄ　／＞に変換する例を示している。

【０１１３】

構成要素変換部１０５は、文書情報において、縦３列×横１３行以上の表を構成し、かつ以下の各条件を満たす文書構成要素の組を検出する。まず、構成要素変換部１０５は、１行目の２列目から数値を示す要素識別情報である＜ｄｉｇｉｔ　／＞が９個以上連続し、かつそれ以降に＜ｓｈｏｒｔｔｅｘｔ　／＞又は＜ｌｅｔｔｅｒ　／＞が３個連続することを条件とする。更に、構成要素変換部１０５は、２行目及び３行目の一列目は、＜ｓｈｏｒｔｔｅｘｔ　／＞若しくは＜ｌｅｔｔｅｒ　／＞又はリンク情報に対応付けられた＜ｓｈｏｒｔｔｅｘｔ　／＞若しくは＜ｌｅｔｔｅｒ　／＞であることを条件とする。

【０１１４】

更に、構成要素変換部１０５は、２行目及び３行目のそれぞれにおいて、＜ｄｉｇｉｔ　／＞が２列目から１２個以上連続することを条件とする。但し、＜ｄｉｇｉｔ　／＞が連続する個数は、１行目において連続する＜ｄｉｇｉｔ　／＞の個数に３を加えた数とすることを条件とする。なお、構成要素変換部１０５は、野球のコールドゲーム又は途中経過等を示すスコアボードを検出するべく、他の条件を用いてもよい。

【０１１５】

このように、構成要素変換部１０５は、データの内容の詳細が日々変化する文書構成要素の組であっても、予め定めたパターンに適合するか否かを判定することにより、当該文書構成要素の組が示すデータの種類を適切に判別し、文書構成要素に変換することができる。

【０１１６】

図１０（ａ）は、文書情報により表示された表示画面の一例を示す。図１０（ｂ）は、図１０（ａ）に示す表示画面を表示した文書情報に対して、文書内配列パターン変換部１３０による変換を行った結果を示す。

【０１１７】

表示部２０５は、箇条書きのアイテムの開始点を示す画像と、箇条書きの内容を示す文字列とを対応付けて表示する。例えば、表示部２０５は、社会、スポーツ、金融・経済、及び政治という文字列のそれぞれを、画像に対応付けて表示している。更に、これらの文字列は、ハイパーリンクとなっており、即ち、これらの文字列に対して利用者から指示を受けた場合には、表示部２０５は、他の情報を表示する。

【０１１８】

構成要素変換部１０５は、箇条書きのアイテムの開始点を示す画像を識別する画像識別情報を、箇条書きのアイテムの開始点を示す要素識別情報である＜ｂｕｌｌｅｔ　／＞に変換する。また、構成要素変換部１０５は、表示画面に表示するべき文字列を、当該文字列が２文字以上２０文字以内であることを示す要素識別情報である＜ｓｈｏｒｔｔｅｘｔ　／＞に変換する。更に、構成要素変換部１０５は、当該文字列に対して利用者から指示を受けた場合のリンク先を示す格納位置情報を、当該格納位置情報の示す格納位置が対象文書情報と同一サイト内である旨を示す要素識別情報である＜ｉｎ－ｌｉｎｋ＞及び＜／ｉｎ－ｌｉｎｋ＞に変換する。また、構成要素変換部１０５は、当該文字列に続く文字列を次の行に表示することを示す文書構成要素、例えば、＜ｂｒ＞タグを、その旨を示す要素識別情報である＜ｎｅｗ－ｌｉｎｅ　／＞に変換する。

【０１１９】

なお、要素識別情報の一例である＜ｂｕｌｌｅｔ　／＞とは、箇条書きのアイテムの開始点を示す情報であればよく、本例に示した所定サイズ以下の画像の他、特定の条件を満たす＜ｌｅｔｔｅｒ　／＞又は＜ｄｉｇｉｔ　／＞等から構成されていてもよい。

【０１２０】

構成要素変換部１０５により変換された対象文書情報について、文書内配列パターン変換部１３０は、要素識別情報の組である＜ｂｕｌｌｅｔ　／＞、＜ｉｎ－ｌｉｎｋ＞、＜ｓｈｏｒｔｔｅｘｔ　／＞、＜／ｉｎ－ｌｉｎｋ＞、及び＜ｎｅｗ－ｌｉｎｅ　／＞を検出する。そして、文書内配列パターン変換部１３０は、当該要素識別情報の組が基準頻度以上で繰り返し出現する場合に、当該要素識別情報の組の配列パターンを示す要素識別情報、例えば、箇条書きのアイテムを示す要素識別情報である＜ｉｔｅｍｉｚｅｄｌｉｎｋ　／＞等に変換する。

【０１２１】

グループ分類部１７０は、文書内配列パターン変換部１３０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループとして分類する。例えば、グループ分類部１７０は、＜ｉｔｅｍｉｚｅｄｌｉｎｋ　／＞に変換された文書構成要素を、箇条書きを示すグループとして分類する。好ましくは、グループ分類部１７０は、分類したグループのそれぞれについて、当該グループの果たす役割を特定し、当該グループに対応付けて出力する。例えば、グループ分類部１７０は、２文字以上所定の文字数（例えば、３０）以下の文字列を示す要素識別情報、即ち、＜ｓｈｏｒｔｔｅｘｔ　／＞から構成されるリンク情報のグループについて、当該グループがリンクリストを示すと判断する。

【０１２２】

このように、文書内配列パターン変換部１３０は、要素識別情報の出現頻度及び配列順に基づいて、要素識別情報の組を変換することができる。

【０１２３】

なお、本図の例においては、文書内配列パターン変換部１３０が、基準頻度より高い頻度で出現する文書構成要素の組を検出したが、これに代えて、構成要素変換部１０５が、これらの文書構成要素の組を、予め定めた条件に適合したと判断して変換してもよい。例えば、構成要素変換部１０５は、＜ｂｕｌｌｅｔ　／＞、＜ｉｎ－ｌｉｎｋ＞、＜ｓｈｏｒ

ｔｔｅｘｔ　／＞、＜／ｉｎ－ｌｉｎｋ＞、及び＜ｂｒ＞がこの順に配列されており、かつこの配列が連続している場合に、予め定めた条件に適合したと判断してもよい。この場合、構成要素変換部１０５は、これらの要素識別情報の組を、リンクリストを示す＜ｓｉｔｅ－ｉｎｄｅｘ　／＞に変換する。本条件を含む条件の一例を正規表現として以下に示す。

【０１２４】

（リンクリストを示す正規表現）

［［＜ｂｕｌｌｅｔ　／＞］？　（＜ｓａｍｅｄｉｒ－ｌｉｎｋ＞＜ｓｈｏｒｔｔｅｘｔ　／＞＜／ｓａｍｅｄｉｒ－ｌｉｎｋ＞）｜（＜ｉｎ－ｌｉｎｋ＞＜ｓｈｏｒｔｔｅｘｔ　／＞＜／ｉｎ－ｌｉｎｋ＞）　［＜ｂｒ＞｜＜ｐ＞］＋］

【０１２５】

このように、構成要素変換部１０５は、上記の条件を満たす要素識別情報の組を、リンクリストを示す要素識別情報である＜ｓｉｔｅ－ｉｎｄｅｘ　／＞に変換する。

【０１２６】

更に他の例としては、構成要素変換部１０５は、＜ｂｕｌｌｅｔ　／＞、＜ｉｎ－ｌｉｎｋ＞、＜ｓｈｏｒｔｔｅｘｔ　／＞、及び＜／ｉｎ－ｌｉｎｋ＞から構成されるテーブル又はリストを、リンクリストを示す＜ｓｉｔｅ－ｉｎｄｅｘ
／＞に変換してもよい。

【０１２７】

図１１（ａ）は、文書情報により表示された表示画面の他の例を示す。図１１（ｂ）は、図１１（ａ）に示す表示画面を表示した文書情報に対して、文書内配列パターン変換部１３０による変換を行った結果を示す。

【０１２８】

表示部２０５は、箇条書きのアイテムの開始点を示す画像と、箇条書きの内容を示す文字列とを対応付けて表示する。例えば、表示部２０５は、新聞記事の見出しを示す文字列のそれぞれを、箇条書きを示す記号（例えば、「・」）に対応付けて表示している。更に、これらの文字列は、ハイパーリンクとなっており、これらの文字列に対して利用者から指示を受けた場合には、表示部２０５は、他の情報を表示する。

【０１２９】

構成要素変換部１０５は、箇条書きのアイテムの開始点を示す黒点を、箇条書きのアイテムの開始点を示す要素識別情報である＜ｂｕｌｌｅｔ　／＞に変換する。また、構成要素変換部１０５は、表示画面に表示するべき文字列を、当該文字列が３０文字以上１００文字以内であることを示す要素識別情報である＜ｍｉｄｔｅｘｔ　／＞に変換する。更に、構成要素変換部１０５は、当該文字列に対して利用者から指示を受けた場合のリンク先を示す格納位置情報を、当該格納位置情報の示す格納位置が対象文書情報と同一サイト内である旨を示す要素識別情報である＜ｉｎ－ｌｉｎｋ＞及び＜／ｉｎ－ｌｉｎｋ＞に変換する。また、構成要素変換部１０５は、当該文字列に続く文字列を次の行に表示することを示す文書構成要素、例えば、＜ｂｒ＞タグを、その旨を示す要素識別情報である＜ｎｅｗ－ｌｉｎｅ　／＞に変換する。

【０１３０】

構成要素変換部１０５により変換された対象文書情報について、文書内配列パターン変換部１３０は、要素識別情報の組である＜ｂｕｌｌｅｔ　／＞、＜ｉｎ－ｌｉｎｋ＞、＜ｍｉｄｔｅｘｔ　／＞、＜／ｉｎ－ｌｉｎｋ＞、及び＜ｎｅｗ－ｌｉｎｅ　／＞を検出する。そして、文書内配列パターン変換部１３０は、当該要素識別情報の組が基準頻度以上で繰り返し出現する場合に、当該要素識別情報の組の配列パターンを示す要素識別情報、例えば、箇条書きのアイテムを示す要素識別情報である＜ｌｏｎｇｉｔｅｍｉｚｅｄｌｉｎｋ　／＞等に変換する。

【０１３１】

グループ分類部１７０は、文書内配列パターン変換部１３０により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の文書構成要素をグループと

して分類する。例えば、グループ分類部１７０は、＜ｌｏｎｇｉｔｅｍｉｚｅｄｌｉｎｋ　／＞に変換された文書構成要素を、箇条書きを示すグループとして分類する。好ましくは、グループ分類部１７０は、分類したグループのそれぞれについて、当該グループの果たす役割を特定し、当該グループに対応付けて出力する。例えば、グループ分類部１７０は、文字数が所定範囲内（例えば、３０から１００文字）の文字列を示す要素識別情報、即ち、＜ｍｉｄｔｅｘｔ　／＞から構成されるリンク情報のグループであって、当該文字列の終端に時刻を示す文字列が配列されているグループについて、当該グループが新聞等のヘッドラインを示すと判断する。

【０１３２】

図１２は、変換対象である対象文書情報の表示例を示す。図１３は、図１２に示す対象文書情報において、グループ分類部１７０による分類の一例を示す。図１２における対象文書情報は、新聞記事を示すＨＴＭＬ文書である。図１３において実線の矩形のそれぞれは、文書構成要素を示している。また、丸印は、要素識別情報を示している。また、点線で囲まれた領域は、グループ分類部１７０により分類されるグループを示している。また、図１３においては関連文書情報の一部についても、グループ分類部１７０による分類の一例を図示する。

【０１３３】

情報処理装置１０は、対象文書情報として、新聞の社会面を示すオブジェクトを表示させる文書構成要素の組１２００と、それぞれが社会面の記事の詳細に対するリンクでありかつ当該記事のヘッドラインを示す文書構成要素の組１２１０、文書構成要素の組１２２０、文書構成要素の組１２３０、文書構成要素の組１２４０、及び文書構成要素の組１２５０とを取得する。

【０１３４】

例えば、文書構成要素の組１２１０は、箇条書きを示す文書構成要素である「・」と、記事のヘッドラインを示す文字列である文書構成要素と、当該文字列から新聞記事へのリンク情報を示す文書構成要素、例えば、＜ａ＞タグと、文字列の改行を示す文書構成要素、例えば、＜ｂｒ＞タグとを含む。文書構成要素の組１２２０から文書構成要素の組１２５０のそれぞれは、文書構成要素の組１２１０と略同一であるので説明を省略する。

【０１３５】

更に、情報処理装置１０は、対象文書情報として、新聞の政治面を示すオブジェクトを表示させる文書構成要素の組１２７０と、政治面の記事の詳細に対するリンクでありかつ当該記事のヘッドラインを示す文書構成要素の組と、新聞の各面へのリンクの集合を表示させる文書構成要素の組１２９０と、当該新聞のロゴを示す文書構成要素１２９２と、当該新聞記事における著作権の帰属を示す文書構成要素１２９５とを取得する。

【０１３６】

構成要素変換部１０５は、文書構成要素の組１２００を、新聞記事のジャンルを示す要素識別情報に変換する。そして、構成要素変換部１０５は、文書構成要素の組１２１０を、箇条書きの先頭記号を示す要素識別情報である＜ｂｕｌｌｅｔ　／＞と、サイズが所定範囲内の文字列を示す要素識別情報である＜ｍｉｄｔｅｘｔ　＞と、新聞記事へのリンク、例えば、サイト内のリンクを示す要素識別情報である＜ｉｎ－ｌｉｎｋ　／＞と、改行を示す要素識別情報である＜ｎｅｗ－ｌｉｎｅ＞とに変換する。

【０１３７】

続いて、文書内配列パターン変換部１３０は、これらの要素識別情報の組を、箇条書きのアイテムを示す要素識別情報である＜ｉｔｅｍｉｚｅｄｌｉｎｋ　／＞に変換する。また、文書構成要素の組１２１０から文書構成要素の組１２５０が連続して配置されているので、文書内配列パターン変換部１３０は、＜ｉｔｅｍｉｚｅｄｌｉｎｋ　／＞の組を、ヘッドラインの集合を示す要素識別情報１２６０に変換する。更に、文書内配列パターン変換部１３０は、要素識別情報１２６０及び新聞記事のジャンルを示す要素識別情報を、ジャンル毎の新聞記事のヘッドラインを示す要素識別情報１２６５に変換する。

【０１３８】

新聞の政治面の記事についても同様に、構成要素変換部１０５は、「・」、文字列、リンク、及び改行等の文書構成要素を、＜ｂｕｌｌｅｔ　／＞、＜ｍｉｄｔｅｘｔ　＞、＜ｉｎ－ｌｉｎｋ　／＞、及び＜ｎｅｗ－ｌｉｎｅ＞等の要素識別情報に変換する。そして、文書内配列パターン変換部１３０は、これらの要素識別情報を＜ｉｔｅｍｉｚｅｄｌｉｎｋ　／＞に変換する。

【０１３９】

続いて、文書内配列パターン変換部１３０は、箇条書きのアイテムを示す要素識別情報、例えば、＜ｉｔｅｍｉｚｅｄｌｉｎｋ　／＞の組を、ヘッドライン記事の集合を示す要素識別情報１２８０に変換する。また、文書内配列パターン変換部１３０は、要素識別情報１２８０及び新聞記事のジャンルを示す要素識別情報を、ジャンル毎の新聞記事のヘッドラインを示す要素識別情報１２８５に変換する。

【０１４０】

グループ分類部１７０は、要素識別情報１２６５及び要素識別情報１２８５のそれぞれに変換された複数の文書構成要素を、ジャンル毎の新聞記事のヘッドラインを示すグループ１３００及びグループ１３１０に分類する。

【０１４１】

また、情報処理装置１０は、サイト内リンクを表示させる文書構成要素の組１２９０を、サイト内リンクを示すグループ１３２０に分類するべく以下の処理を行う。まず、構成要素変換部１０５は、それぞれが文書構成要素の組を形成する、「社会」、「政治」、及び途中を一部省略して「ひと」を検出する。文書構成要素の組のそれぞれは、文書構成要素である文字列と、文書構成要素であるリンク情報とを有している。そして、構成要素変換部１０５は、文字列を、所定の文字数以下の文字列を示す＜ｓｈｏｒｔｔｅｘｔ　／＞に変換し、リンク情報を、同一サイト内へのリンクを示す＜ｉｎ－ｌｉｎｋ＞及び＜／ｉｎ－ｌｉｎｋ＞に変換し、文字列を区切る記号「｜」を、＜ｂｕｌｌｅｔ　／＞に変換する。

【０１４２】

文書内配列パターン変換部１３０は、＜ｉｎ－ｌｉｎｋ＞、＜ｓｈｏｒｔｔｅｘｔ　／＞、＜／ｉｎ－ｌｉｎｋ＞、及び＜ｂｕｌｌｅｔ　／＞から構成される要素識別情報の組が、対象文書情報において基準頻度以上で繰り返し出現すると判断する。これにより、グループ分類部１７０は、リンクリストを示す文字列から構成される文書構成要素の組１２９５、例えば、「社会｜政治｜…｜ひと｜」のように配列されたリンク情報の組を、リンクリストを示すグループとして分類することができる。

【０１４３】

また、情報処理装置１０は、フッタを示す文字列及び画像を、フッタ部を示すグループ１３３０ａに分類するべく以下の処理を行う。まず、構成要素変換部１０５は、文書構成要素１２９５を、所定のキーワード、例えば、「著作権」を含む文字列を示す要素識別情報である＜ｃｏｐｙｒｉｇｈｔ　／＞に変換する。そして、構成要素変換部１０５は、文書構成要素１２９２を、サイズが所定範囲内の画像を示す要素識別情報である＜ｍｉｄｉｍａｇｅ　／＞に変換する。

【０１４４】

文書間配列パターン変換部１６０は、＜ｃｏｐｙｒｉｇｈｔ　／＞及び＜ｍｉｄｉｍａｇｅ　／＞から構成される要素識別情報の組が、対象文書情報及び関連文書情報を併せた文書において基準頻度以上で繰り返し出現すると判断する。そして、文書間配列パターン変換部１６０は、この要素識別情報の組を、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する。

【０１４５】

グループ分類部１７０は、文書間配列パターン変換部１６０により変換された要素識別情報をグループとして分類する。例えば、グループ分類部１７０は、対象文書情報における文書構成要素１２９２及び文書構成要素１２９５を、フッタ部を示すグループ１３３０ａに分類する。また、グループ分類部１７０は、関連文書情報における文書構成要素１３９

２及び文書構成要素１３９５を、フッタ部を示すグループ１３３０ｂに分類してもよい。更に、グループ分類部１７０は、これらのグループのタイトル情報が、フッタ部である旨を、当該グループに含まれる要素識別情報、例えば、＜ｃｏｐｙｒｉｇｈｔ　／＞に基づいて特定してもよい。

【０１４６】

図１４は、情報処理装置１０のハードウェア構成の一例を示す。情報処理装置１０は、ホストコントローラ１０８２により相互に接続されるＣＰＵ１０００、ＲＡＭ１０２０、グラフィックコントローラ１０７５、及び表示装置１０８０を有するＣＰＵ周辺部と、入出力コントローラ１０８４によりホストコントローラ１０８２に接続される通信インターフェイス１０３０、ハードディスクドライブ１０４０、及びＣＤ－ＲＯＭドライブ１０６０を有する入出力部と、入出力コントローラ１０８４に接続されるＲＯＭ１０１０、フレキシブルディスクドライブ１０５０、及び入出力チップ１０７０を有するレガシー入出力部とを備える。

【０１４７】

ホストコントローラ１０８２は、ＲＡＭ１０２０と、高い転送レートでＲＡＭ１０２０をアクセスするＣＰＵ１０００及びグラフィックコントローラ１０７５とを接続する。ＣＰＵ１０００は、ＲＯＭ１０１０及びＲＡＭ１０２０に格納されたプログラムに基づいて動作し、各部の制御を行う。グラフィックコントローラ１０７５は、ＣＰＵ１０００等がＲＡＭ１０２０内に設けたフレームバッファ上に生成する画像データを取得し、表示装置１０８０上に表示させる。これに代えて、グラフィックコントローラ１０７５は、ＣＰＵ１０００等が生成する画像データを格納するフレームバッファを、内部に含んでもよい。

【０１４８】

入出力コントローラ１０８４は、ホストコントローラ１０８２と、比較的高速な入出力装置である通信インターフェイス１０３０、ハードディスクドライブ１０４０、及びＣＤ－ＲＯＭドライブ１０６０を接続する。通信インターフェイス１０３０は、ネットワークを介して他の装置と通信する。ハードディスクドライブ１０４０は、情報処理装置１０が使用するプログラム及びデータを格納する。ＣＤ－ＲＯＭドライブ１０６０は、ＣＤ－ＲＯＭ１０９５からプログラム又はデータを読み取り、ＲＡＭ１０２０を介して入出力チップ１０７０に提供する。

【０１４９】

また、入出力コントローラ１０８４には、ＲＯＭ１０１０と、フレキシブルディスクドライブ１０５０や入出力チップ１０７０等の比較的低速な入出力装置とが接続される。ＲＯＭ１０１０は、情報処理装置１０の起動時にＣＰＵ１０００が実行するブートプログラムや、情報処理装置１０のハードウェアに依存するプログラム等を格納する。フレキシブルディスクドライブ１０５０は、フレキシブルディスク１０９０からプログラム又はデータを読み取り、ＲＡＭ１０２０を介して入出力チップ１０７０に提供する。入出力チップ１０７０は、フレキシブルディスク１０９０や、例えばパラレルポート、シリアルポート、キーボードポート、マウスポート等を介して各種の入出力装置を接続する。

【０１５０】

情報処理装置１０に提供されるプログラムは、フレキシブルディスク１０９０、ＣＤ－ＲＯＭ１０９５、又はＩＣカード等の記録媒体に格納されて利用者によって提供される。プログラムは、記録媒体から読み出され、入出力チップ１０７０を介して情報処理装置１０にインストールされ、情報処理装置１０において実行される。

【０１５１】

情報処理装置１０にインストールされて実行されるプログラムは、関連文書検出モジュールと、構成要素変換モジュールと、構成要素選択モジュールと、間隙構成要素検出モジュールと、文書内配列パターン変換モジュールと、変換是非入力モジュールと、反復処理モジュールと、文書間配列パターン変換モジュールと、グループ分類モジュールと、並替出力モジュールと、目次情報出力モジュールと、音声出力命令生成モジュールと、表示モジュールと、文書構造情報生成モジュールと、文書情報同一性出力モジュールとを含む。各

モジュールが情報処理装置１０に働きかけて行わせる動作は、図１から図１３において説明した情報処理装置１０における、対応する部材の動作と同一であるから、説明を省略する。

【０１５２】

以上に示したプログラム又はモジュールは、外部の記憶媒体に格納されてもよい。記憶媒体としては、フレキシブルディスク１０９０、ＣＤ－ＲＯＭ１０９５の他に、ＤＶＤやＰＤ等の光学記録媒体、ＭＤ等の光磁気記録媒体、テープ媒体、ＩＣカード等の半導体メモリ等を用いることができる。また、専用通信ネットワークやインターネットに接続されたサーバシステムに設けたハードディスク又はＲＡＭ等の記憶装置を記録媒体として使用し、ネットワークを介してプログラムを情報処理装置１０に提供してもよい。

【０１５３】

以上、本実施形態から明らかなように、情報処理装置１０は、文書情報に含まれる複数の文書構成要素を、文書構成要素の出現頻度に基づいて、グループに分類する。更に、情報処理装置１０は、関連する文書情報との類似性に基づいて、文書構成要素を分類する。これにより、情報処理装置１０は、構造が動的に変化する文書情報、例えば、日々更新されるウェブページについて、文書構成要素を適切に分類することができる。

【０１５４】

以上に示した実施形態によると、以下の各項目に示す情報処理装置、プログラム、及び記録媒体が実現される。

【０１５５】

（項目１）　文書情報に含まれる複数の文書構成要素を、複数のグループに分類する情報処理装置であって、前記文書情報における前記複数の文書構成要素のそれぞれを、当該文書構成要素の種類又は役割を示す要素識別情報にそれぞれ変換する構成要素変換部と、前記構成要素変換部により変換された前記文書情報において、予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す前記要素識別情報に変換する文書内配列パターン変換部と、前記文書内配列パターン変換部により前記文書情報を繰り返し変換した結果得られた文書情報において、前記文書内配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部とを備える情報処理装置。

（項目２）　前記構成要素変換部により変換された前記文書情報において、複数の前記要素識別情報の何れかを選択する構成要素選択部と、前記構成要素選択部により選択された被選択情報と、前記文書情報において前記被選択情報の次に配列され前記被選択情報と同一の種類である前記要素識別情報との間に配列された間隙構成要素を検出する間隙構成要素検出部とを更に備え、前記文書内配列パターン検出部は、前記被選択情報及び前記間隙構成要素を前記要素識別情報の組として検出する項目１記載の情報処理装置。

【０１５６】

（項目３）　前記間隙構成要素検出部は、前記文書情報において複数の前記被選択情報の後に配列される終端構成要素を更に検出し、前記文書内配列パターン検出部は、前記複数の被選択情報のうち前記文書情報の最後に配列される被選択情報と、前記終端構成要素とを、前記要素識別情報の組として検出する項目２記載の情報処理装置。

（項目４）　前記構成要素選択部は、前記構成要素変換部により変換された文書情報における前記要素識別情報を、出現する頻度の高い順に、順次前記被選択情報として選択し、前記間隙構成要素検出部は、前記構成要素選択部により選択された各被選択情報について、前記間隙構成要素を検出し、前記文書内配列パターン変換部は、前記構成要素選択部により順次選択される被選択情報について、前記被選択情報及び前記間隙構成要素を前記要素識別情報の組として検出する項目２記載の情報処理装置。

（項目５）　前記文書内配列パターン変換部は、既に検出した前記要素識別情報の組より出現頻度が高いことを更に条件として、繰り返し出現する前記要素識別情報の組を検出する項目４記載の情報処理装置。

【０１５７】
（項目６）　前記文書内配列パターン変換部は、前記配列パターンを示す前記要素識別情報に変換した変換元である複数の文書構成要素を、変換先の前記要素識別情報に対応付けて、前記構成要素変換部による新たな変換対象である新規登録要素集合として登録し、前記構成要素変換部は、更に、前記文書情報における前記新規登録要素集合のそれぞれを、前記文書内配列パターン変換部による変換先の前記要素識別情報に変換する項目１記載の情報処理装置。
（項目７）　前記文書内配列パターン変換部は、前記基準頻度以上で繰り返し出現する前記要素識別情報の組であって、変換前の文書構成要素を当該要素識別情報の組に変換するべく前記文書内配列パターン変換部が繰り返す変換の回数が予め定められた基準回数以下の要素識別情報の組を、前記配列パターンを示す要素識別情報に変換する項目１記載の情報処理装置。
（項目８）　前記文書内配列パターン変換部は、前記基準頻度以上で繰り返し出現する前記要素識別情報の組のうち、当該要素識別情報の組に変換される前の文書構成要素の合計サイズが予め定められた基準サイズより小さい前記要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す前記要素識別情報に変換する項目１記載の情報処理装置。
【０１５８】
（項目９）　前記文書内配列パターン変換部における変換元の要素識別情報の組及び変換先の要素識別情報を利用者に対して出力することにより、前記文書内配列パターン変換部による要素識別情報への変換を行うか否かを入力させる変換指示入力部を更に備える項目１記載の情報処理装置。
（項目１０）　前記構成要素変換部は、前記文書構成要素を、当該文書構成要素のデータサイズに応じた前記要素識別情報に変換する項目１記載の情報処理装置。
（項目１１）　前記文書情報は、前記文書構成要素として、表示画面上に表示する画像を識別する画像識別情報を含み、前記構成要素変換部は、前記画像識別情報を、前記画像識別情報により識別される画像の形状を示す前記要素識別情報に変換する項目１記載の情報処理装置。
（項目１２）　前記文書情報は、利用者に表示する表示情報と、当該表示情報に対する指示に応じて表示するべき他の情報の格納位置を示す格納位置情報とを前記文書構成要素として含み、前記構成要素変換部は、前記格納位置情報を、前記他の情報が格納される前記格納位置の範囲を示す前記要素識別情報に変換する項目１記載の情報処理装置。
【０１５９】
（項目１３）　前記文書情報は、前記文書構成要素として、表示画面に表示する表示情報と、当該表示情報の表示形式を指定するタグ情報とを含むタグ付き文書であり、前記タグ情報は、当該タグ情報により表示形式を指定する表示情報において更に内側タグ情報を含む、外側タグ情報であり、前記構成要素変換部による変換後の文書情報において、前記外側タグ情報を根ノードとして生成し、前記内側タグ情報を前記根ノードの葉ノードとして生成した、文書構造情報を生成する文書構造情報生成部と、複数の文書情報のそれぞれについて、前記文書構造情報生成部により生成された前記文書構造情報を比較することにより、一の文書情報が他の文書情報と同一の構造を有するか否かを出力する文書情報同一性出力部とを更に備える項目１記載の情報処理装置。
（項目１４）　前記文書情報は、表示画面に表示するべき情報を指示する文書情報であり、更に、前記表示画面に表示されないコメント情報を含み、前記文書内配列パターン検出部は、当該コメント情報を、一の前記要素識別情報の組と、他の前記要素識別情報の組との境界を示す情報として用いる項目１記載の情報処理装置。
（項目１５）　前記グループ分類部により分類された各グループについて、当該グループに含まれる文書構成要素に対する目次を示す目次情報を出力する目次情報出力部を更に備える項目１記載の情報処理装置。
【０１６０】

（項目１６）　前記複数の文書構成要素を、前記グループ分類部により分類されたグループ毎に並び替える並替出力部を更に備える項目１記載の情報処理装置。
（項目１７）　前記文書内配列パターン変換部は、前記基準頻度以上で繰り返し出現する要素識別情報を、当該要素識別情報の組の配列パターンを示す要素識別情報として、当該文書情報における当該要素識別情報の組の重要度を示す重要度情報に変換し、前記並替出力部は、更に、複数の前記グループ中の文書構成情報を、当該グループに含まれる重要度情報の高い順に並び替える項目１６記載の情報処理装置。
（項目１８）　前記グループ分類部により分類された各グループに対応付けて、当該グループを識別する情報を、前記文書情報内に生成するグループ識別情報生成部を更に備える項目１記載の情報処理装置。
（項目１９）　前記グループ分類部により分類された各グループを識別する情報を、前記文書情報とは別体に生成して出力するアノテーション出力部を更に備える項目１記載の情報処理装置。
（項目２０）　前記文書情報に応じて表示画面上に情報を出力する表示部と、前記文書内配列パターン変換部により変換される対象となる要素識別情報の組について、当該要素識別情報の変換元である文書構成要素の合計サイズが、前記表示画面に表示可能な情報のサイズに達した場合に、当該要素識別情報の組を、変換を繰り返し行う対象から除外する反復終了判断部とを更に備える項目１記載の情報処理装置。
【０１６１】
（項目２１）　グループに分類する対象である対象文書情報と予め定められた関係を有する関連文書情報を検出する関連文書検出部と、前記対象文書情報及び前記関連文書情報から、前記文書内パターン変換部により変換された要素識別情報を除外した文書について、当該対象文書情報及び当該関連文書情報の双方において出現する前記要素識別情報の組であって、前記対象文書情報及び前記関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す要素識別情報に変換する文書間配列パターン変換部とを更に備え、前記グループ分類部は、更に、前記文書間配列パターン変換部により変換された結果得られた前記対象文書情報において、前記文書間配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類する項目１記載の情報処理装置。
（項目２２）　前記グループ分類部は、分類したグループのそれぞれについて、当該グループ内の文書構成要素が前記文書情報において果たす役割又は当該グループ内の文書構成要素の内容を示すタイトル情報を、更に生成する項目１記載の情報処理装置。
（項目２３）　前記グループ分類部は、当該要素識別情報の組の境界に設けられた前記コメント情報に含まれる情報を、前記タイトル情報として生成する項目２２記載の情報処理装置。
【０１６２】
（項目２４）　前記グループ分類部は、分類したグループのそれぞれについて、前記文書情報における当該グループの重要度を示す重要度情報を更に生成する項目１記載の情報処理装置。
（項目２５）　文書情報に含まれる複数の文書構成要素を、複数のグループに分類する情報処理装置であって、グループに分類する対象である対象文書情報と予め定められた関係を有する関連文書情報を検出する関連文書検出部と、前記対象文書情報及び前記関連文書情報のそれぞれにおいて、前記複数の文書構成要素のそれぞれを、当該文書構成要素の種類又は役割を示す要素識別情報に変換する構成要素変換部と、前記対象文書情報及び前記関連文書情報の双方において出現する前記要素識別情報の組であって、前記対象文書情報及び前記関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組のそれぞれを、当該文書構成要素の組の配列パターンを示す前記要素識別情報に変換する文書間配列パターン変換部と、前記文書間配列パターン変換部により変換された結果得られた前記対象文書情報において、前記文書間配列パターン変換

部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部とを備える情報処理装置。

【０１６３】

（項目２６）　前記関連文書検出部は、前記対象文書情報が格納される格納位置から予め定められた範囲内に格納されている文書情報を前記関連文書情報として検出する項目２５記載の情報処理装置。

（項目２７）　前記対象文書情報は、当該対象文書情報が生成される以前に存在していた既存文書情報を更新することにより生成され前記関連文書検出部は、前記既存文書情報を前記関連文書情報として検出する項目２５記載の情報処理装置。

（項目２８）　文書情報に含まれる複数の文書構成要素を、複数のグループに分類する情報処理装置を制御するプログラムであって、前記情報処理装置を、前記文書情報における前記複数の文書構成要素のそれぞれを、当該文書構成要素の種類又は役割を示す要素識別情報にそれぞれ変換する構成要素変換部と、前記構成要素変換部により変換された前記文書情報において、予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組のそれぞれを、当該要素識別情報の組の配列パターンを示す前記要素識別情報に変換する文書内配列パターン変換部と、前記文書内配列パターン変換部により前記文書情報を繰り返し変換した結果得られた文書情報において、前記文書内配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部として機能させるプログラム。

【０１６４】

（項目２９）　文書情報に含まれる複数の文書構成要素を、複数のグループに分類する情報処理装置を制御するプログラムであって、前記情報処理装置を、グループに分類する対象である対象文書情報と予め定められた関係を有する関連文書情報を検出する関連文書検出部と、前記対象文書情報及び前記関連文書情報のそれぞれにおいて、前記複数の文書構成要素のそれぞれを、当該文書構成要素の種類又は役割を示す要素識別情報に変換する構成要素変換部と、前記対象文書情報及び前記関連文書情報の双方において出現する前記要素識別情報の組であって、前記対象文書情報及び前記関連文書情報を併せた文書において予め定められた基準頻度以上で繰り返し出現する前記要素識別情報の組のそれぞれを、当該文書構成要素の組の配列パターンを示す前記要素識別情報に変換する文書間配列パターン変換部と、前記文書間配列パターン変換部により変換された結果得られた前記対象文書情報において、前記文書間配列パターン変換部により変換された要素識別情報のそれぞれについて、当該要素識別情報に変換された複数の前記文書構成要素をグループとして分類するグループ分類部として機能させるプログラム。

（項目３０）　項目２８又は項目２９記載のプログラムを記録した記録媒体。

【０１６５】

以上、本発明を実施形態を用いて説明したが、本発明の技術的範囲は上記実施形態に記載の範囲には限定されない。上記実施形態に、多様な変更または改良を加えることができる。そのような変更または改良を加えた形態も本発明の技術的範囲に含まれ得ることが、特許請求の範囲の記載から明らかである。

【０１６６】

【発明の効果】

上記説明から明らかなように、本発明によれば文書中の情報を適切に分類することができる。

【図面の簡単な説明】

【図１】図１は、情報処理装置１０の機能ブロック図を示す。

【図２】図２は、情報処理装置１０のフローチャートを示す。

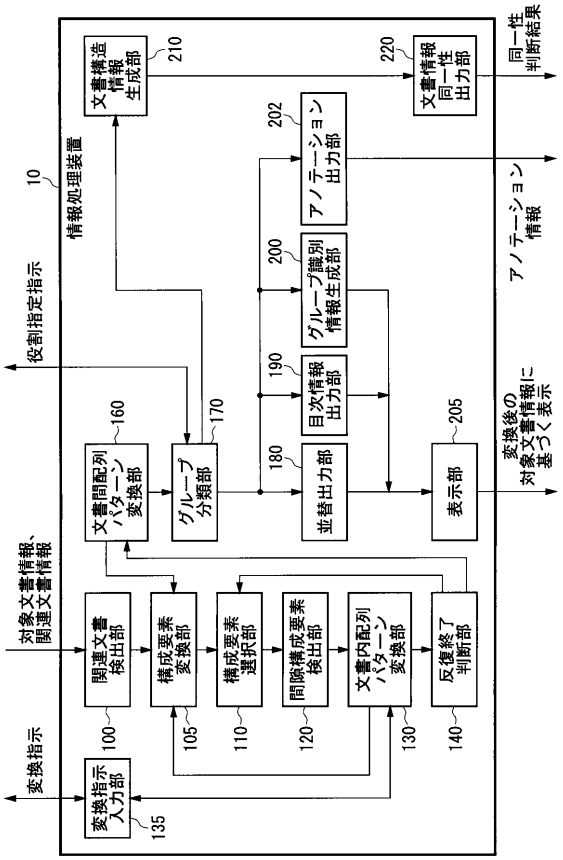
【図３】図３は、図２のＳ２１０におけるフローチャートを示す。

【図４】図４は、図２のＳ２２０におけるフローチャートを示す。

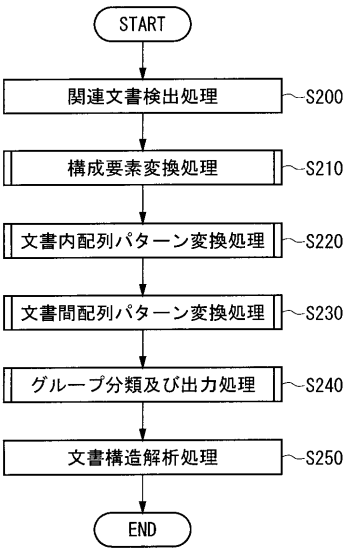
【図５】図５は、図４のＳ４５０におけるフローチャートを示す。

【図６】図６は、図２のＳ２３０におけるフローチャートを示す。
【図７】図７は、図２のＳ２４０におけるフローチャートを示す。
【図８】図８は、構成要素変換部１０５が文書構成要素を要素識別情報に変換する一例を示す。
【図９】図９は、構成要素変換部１０５が所定の要素識別情報の組を要素識別情報に変換する例を示す。
【図１０】図１０（ａ）は、文書情報により表示された表示画面の一例を示す。図１０（ｂ）は、図１０（ａ）に示す表示画面を表示した文書情報に対して、文書内配列パターン変換部１３０による変換を行った結果を示す。
【図１１】図１１（ａ）は、文書情報により表示された表示画面の他の例を示す。図１１（ｂ）は、図１１（ａ）に示す表示画面を表示した文書情報に対して、文書内配列パターン変換部１３０による変換を行った結果を示す。
【図１２】図１２は、変換対象である対象文書情報の表示例を示す。
【図１３】図１３は、図１２に示す対象文書情報において、グループ分類部１７０による分類の一例を示す。
【図１４】図１４は、情報処理装置１０のハードウェア構成の一例を示す。
【符号の説明】
１０　情報処理装置
１００　関連文書検出部
１０５　構成要素変換部
１１０　構成要素選択部
１２０　間隙構成要素検出部
１３０　文書内配列パターン変換部
１３５　変換指示入力部
１４０　反復終了判断部
１６０　文書間配列パターン変換部
１７０　グループ分類部
１８０　並替出力部
１９０　目次情報出力部
２００　グループ識別情報生成部
２０２　アノテーション出力部
２０５　表示部
２１０　文書構造情報生成部
２２０　文書情報同一性出力部

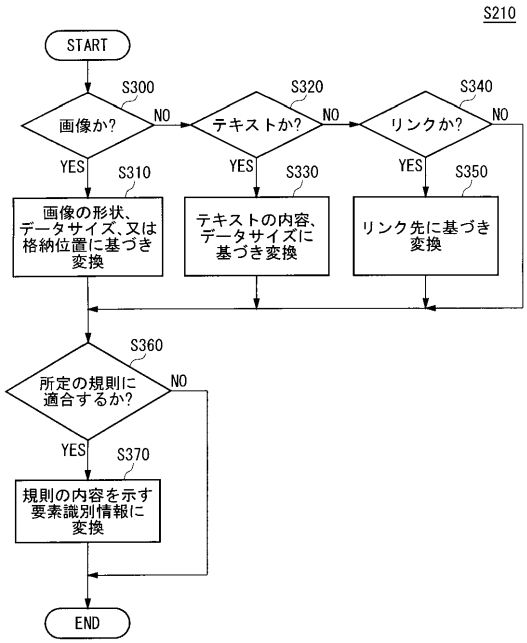
【図１】


情報処理装置
10
文書構造情報生成部
210
文書情報同一性出力部
220
同一性判断結果
アノテーション出力部
202
アノテーション情報
グループ識別情報生成部
200
目次情報出力部
190
並替出力部
180
表示部
205
変換後の対象文書情報に基づく表示
役割指定指示
文書間配列パターン変換部
160
グループ分類部
170
対象文書情報、関連文書情報
関連文書検出部
100
構成要素変換部
105
構成要素選択部
110
間隙構成要素検出部
120
文書内配列パターン変換部
130
反復終了判断部
140
変換指示
変換指示入力部
135

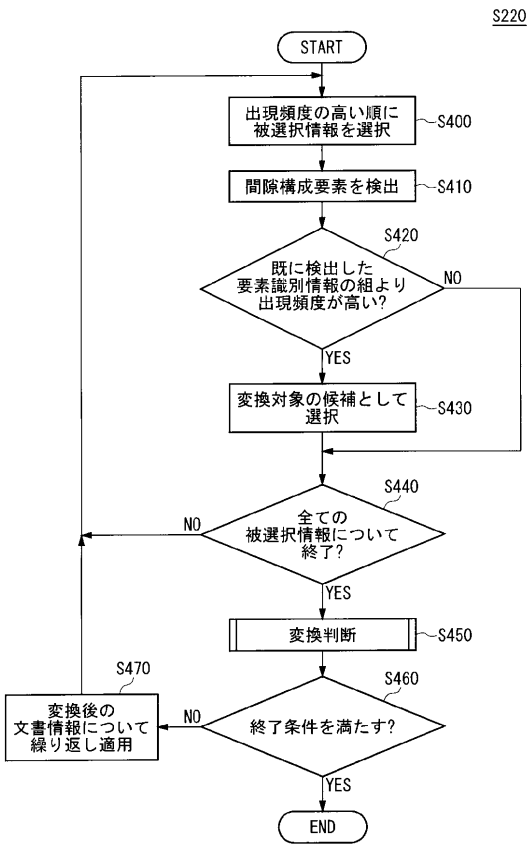
【図２】


START
関連文書検出処理
S200
構成要素変換処理
S210
文書内配列パターン変換処理
S220
文書間配列パターン変換処理
S230
グループ分類及び出力処理
S240
文書構造解析処理
S250
END

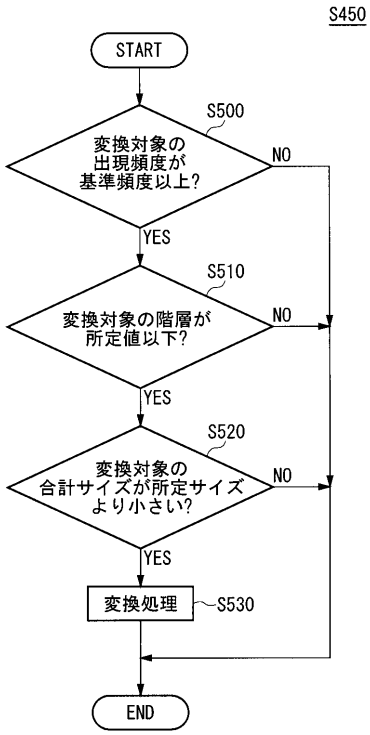
【図３】


S210
START
画像か?
S300
NO
YES
テキストか?
S320
NO
YES
リンクか?
S340
NO
YES
画像の形状、データサイズ、又は格納位置に基づき変換
S310
テキストの内容、データサイズに基づき変換
S330
リンク先に基づき変換
S350
所定の規則に適合するか?
S360
NO
YES
規則の内容を示す要素識別情報に変換
S370
END

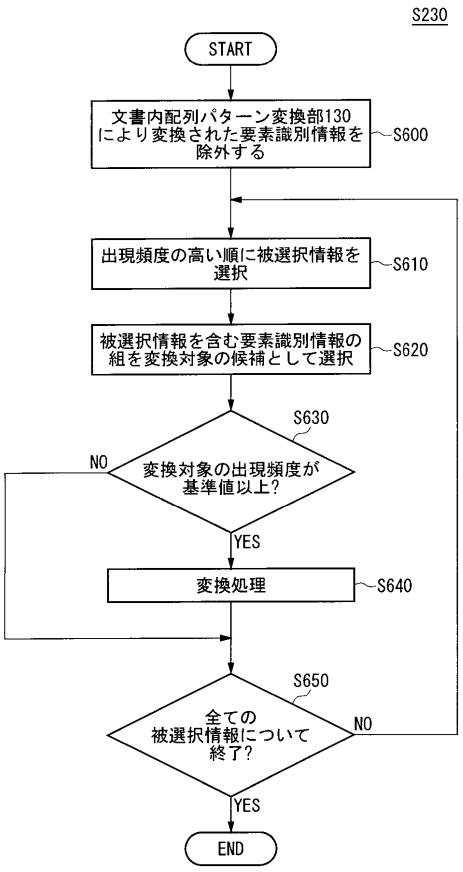
【図４】


S220
START
出現頻度の高い順に被選択情報を選択
S400
間隙構成要素を検出
S410
既に検出した要素識別情報の組より出現頻度が高い?
S420
NO
YES
変換対象の候補として選択
S430
全ての被選択情報について終了?
S440
NO
YES
変換判断
S450
終了条件を満たす?
S460
NO
YES
変換後の文書情報について繰り返し適用
S470
END

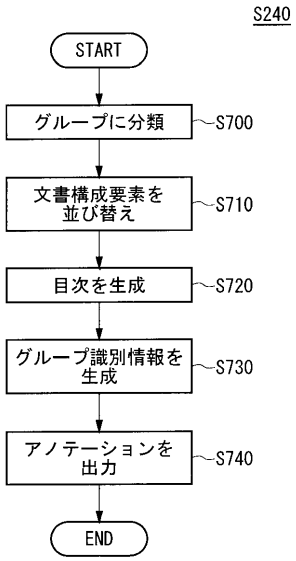
【図５】



【図６】



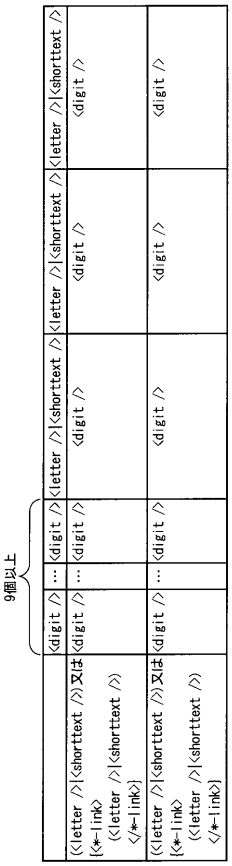
【図７】



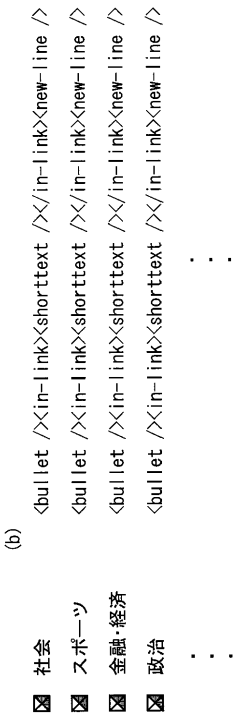
【図８】

要素識別情報	文書構成要素
<digit />	数値
<longtext />	100文字以上の文字列
<midtext />	30～100文字の文字列
<shorttext />	2～30文字の文字列
<letter />	1文字の文字
<in-largeimg />	対象文書情報と同一サイト内にあり、300×300ピクセル以上の画像
<in-midimg />	対象文書情報と同一サイト内にあり、100×100～300×300の画像
<in-smallimg />	対象文書情報と同一サイト内にあり、100×100ピクセル以下の画像
<out-largeimg />	対象文書情報と異なるサイト内にあり、300×300ピクセル以上の画像
<out-midimg />	対象文書情報と異なるサイト内にあり、100×100～300×300の画像
<out-smallimg />	対象文書情報と異なるサイト内にあり、100×100ピクセル以下の画像
<samedir-link>…</samedir-link>	対象文書情報と同一のディレクトリ内へのリンク
<in-link>…</in-link>	対象文書情報と同一サイトへのリンク
<out-link>…</out-link>	対象文書情報と異なるサイトへのリンク

【図９】



【図１０】



【図１１】

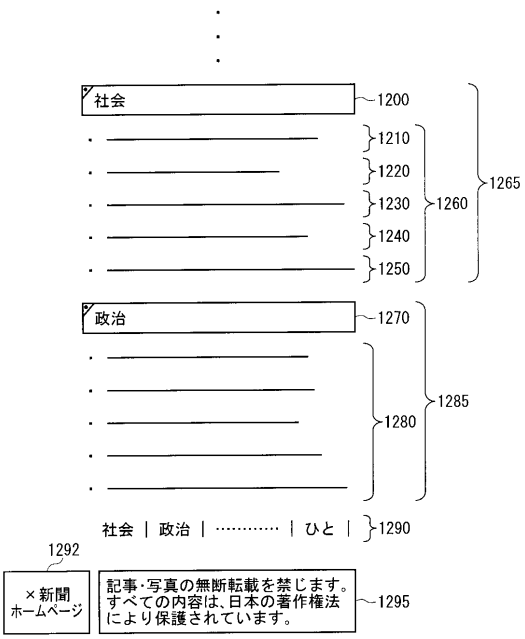
(a)

- ～道路で死亡事故、××町－○○市間が一時通行止めに(21:56)
- AA市内の川に邦人遺体　事件・事故の両面で捜査(21:18)
- BB知事、××社への貸付金○○○億円の債権放棄(20:13)

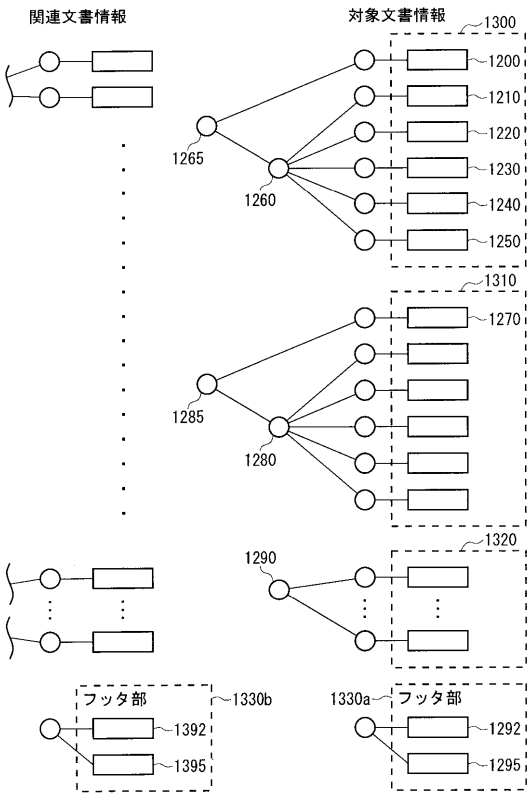
(b)

<bullet /><in-link><midtext /></in-link><new-line />
<bullet /><in-link><midtext /></in-link><new-line />
<bullet /><in-link><midtext /></in-link><new-line />

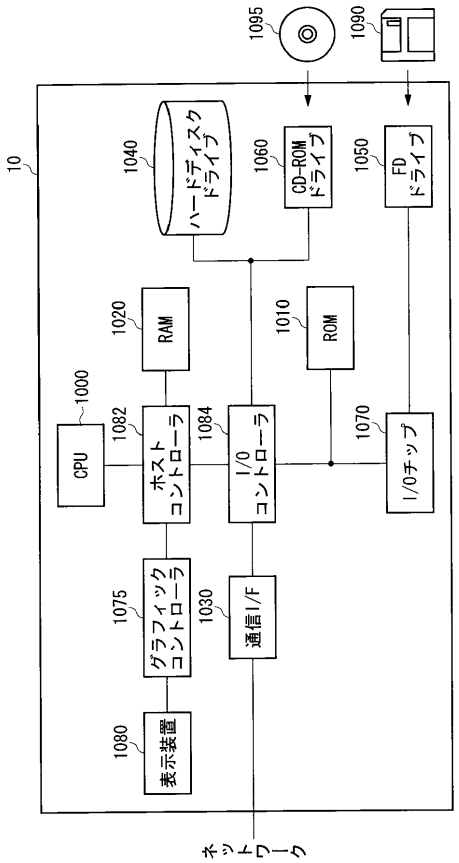
【図１２】



【図１３】


関連文書情報
対象文書情報
1300
1200
1210
1220
1230
1240
1250
1265
1260
1310
1270
1285
1280
1320
1290
フッタ部
1330b
1392
1395
1330a
フッタ部
1292
1295

【図１４】


1095
1090
10
1040
ハードディスクドライブ
1060
CD-ROMドライブ
1050
FDドライブ
1020
RAM
1010
ROM
1000
CPU
1082
ホストコントローラ
1084
I/Oコントローラ
1070
I/Oチップ
1075
グラフィックコントローラ
1030
通信I/F
1080
表示装置
ネットワーク

フロントページの続き

(72)発明者　福田　健太郎
神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　東京基礎研究所内

審査官　深津　始

(56)参考文献　特開２００２－３２４０７７（ＪＰ，Ａ）
特開２００２－３１２３７９（ＪＰ，Ａ）
米国特許出願公開第２００２／００２２９５６（ＵＳ，Ａ１）
特開２００２－３３４０７０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30
G06F 17/21