(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2017-126185
(P2017-126185A)
(43)公開日　平成29年7月20日(2017.7.20)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/22	(2006.01)	G06F	17/22	652	5B109
		G06F	17/22	617	
		G06F	17/22	647	

審査請求 未請求　請求項の数 11　ＯＬ　（全 22 頁）

(21)出願番号　特願2016-4797(P2016-4797)
(22)出願日　平成28年1月13日(2016.1.13)

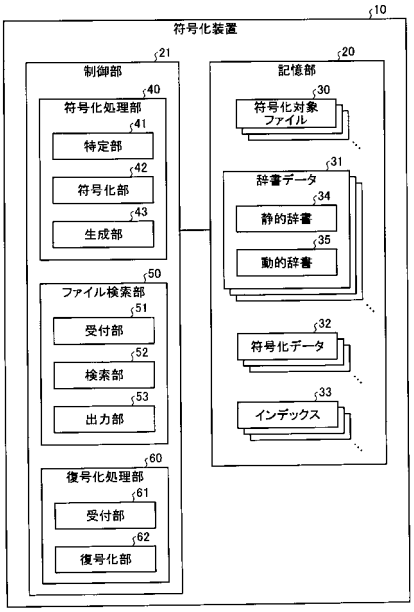
(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　110002147
特許業務法人酒井国際特許事務所
(72)発明者　出内　将夫
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　片岡　正弘
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　田尾　幸資
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

Ｆターム(参考)　5B109 NH20　QA06　SA08

(54)【発明の名称】符号化プログラム、符号化方法、符号化装置、復号化プログラム、復号化方法および復号化装置

(57)【要約】
【課題】書構造に対応した符号化、復号化を行う符号化プログラム、符号化方法、符号化装置、復号化プログラム、復号化方法および復号化装置を提供する。
【解決手段】特定部４１は、構造化された第１の文書の文書構造を特定する。符号化部４２は、文書構造を特定した第１の文書中の特定階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する。
【選択図】図３

符号化装置の構成の一例を示す図



【特許請求の範囲】
【請求項１】
コンピュータに、
構造化された第１の文書の文書構造を特定し、
文書構造を特定した前記第１の文書中の特定階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する
処理を実行させることを特徴とする符号化プログラム。
【請求項２】
前記符号化する処理は、前記第１の文書中の文書構造を規定する文字列を、共通の符号化方式により符号化する
ことを特徴とする請求項１に記載の符号化プログラム。
【請求項３】
前記符号化する処理は、データ属性が類似する階層の文字列を同じ符号化方式により符号化する
ことを特徴とする請求項１または２に記載の符号化プログラム。
【請求項４】
前記符号化する処理は、前記特定階層の文字列を、当該特定階層に出現する文字列の特性に対応した符号化方式により符号化する
ことを特徴とする請求項１～３の何れか１つに記載の符号化プログラム。
【請求項５】
前記符号化する処理は、１またはデータ属性が類似する複数の階層ごとに、出現頻度の高いパターンを短い符号に変換する符号化方式により符号化する
ことを特徴とする請求項１～４の何れか１つに記載の符号化プログラム。
【請求項６】
コンピュータに、
符号化方式ごとに、符号化した文字列に出現したパターンを示したインデックスを生成する
処理をさらに実行させることを特徴とする請求項１～５の何れか１つに記載の符号化プログラム。
【請求項７】
コンピュータが、
構造化された第１の文書の文書構造を特定し、
文書構造を特定した前記第１の文書中の特定階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する
処理を実行することを特徴とする符号化方法。
【請求項８】
構造化された第１の文書の文書構造を特定する特定部と、
前記特定部により文書構造が特定された前記第１の文書中の特定階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する符号化部と、
を有することを特徴とする符号化装置。
【請求項９】
コンピュータに、
構造化された第１の文書が、当該第１の文書の文書構造に対応した階層構造に応じた符号化方式により符号化された符号化データの特定階層の文字列を、当該特定階層の符号化方式により復号化する
処理を実行させることを特徴とする復号化プログラム。
【請求項１０】
コンピュータが、
構造化された第１の文書が、当該第１の文書の文書構造に対応した階層構造に応じた符号化方式により符号化された符号化データの特定階層の文字列を、当該特定階層の符号化

方式により復号化する
処理を実行することを特徴とする復号化方法。
【請求項１１】
構造化された第１の文書が、当該第１の文書の文書構造に対応した階層構造に応じた符号化方式により符号化された符号化データの特定階層の文字列を、当該特定階層の符号化方式により復号化する復号化部
を有することを特徴とする復号化装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、符号化プログラム、符号化方法、符号化装置、復号化プログラム、復号化方法および復号化装置に関する。
【背景技術】
【０００２】
従来から、例えば、ＸＭＬ（Extensible Markup Language）などの構造化された文書データが利用されている。例えば、異なるシステム間でデータを交換する共通フォーマットとして、ＸＭＬが広く普及している。また、ＸＭＬなど構造化された形式で各種の文書データが公開される。この際、保存や通信の際のデータ量を減らすため、構造化された文書データは、例えば、ｚｉｐなどの圧縮形式で全体が圧縮されて保存される。圧縮された文書データを活用する場合は、圧縮された文書データ全体の伸長を行って文書データを復元し、復元した文書データに対して各種の解析が行われる。例えば、文書データの特定の階層に特定の単語を含むかを検索する場合は、復元した文書データに対して字句解析や構造解析が行われる。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開２００５－２１５９５１号公報
【特許文献２】特開２００２－２９７５６８号公報
【特許文献３】特開２００５－１８６７２号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
しかしながら、ｚｉｐなどの圧縮形式で全体が圧縮された文書データを活用する場合、圧縮された文書データ全体を伸長した後、各種の解析を行うため、処理量が多い。圧縮された文書データは、携帯端末など処理能力の低い端末で活用される場合もあり、活用する際の処理量が多いと、処理に時間がかかる。また、構造化された文書において、文書構造の一部のみを利用する場合でも、ｚｉｐなどの圧縮形式で全体が圧縮されている場合は、文書全体を伸長する。
【０００５】
一つの側面では、文書構造に対応した符号化、復号化を行う符号化プログラム、符号化方法、符号化装置、復号化プログラム、復号化方法および復号化装置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】
【０００６】
第１の案では、符号化プログラムは、コンピュータに、構造化された第１の文書の文書構造を特定する処理を実行させる。符号化プログラムは、コンピュータに、文書構造を特定した第１の文書中の特定階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する処理を実行させる。
【発明の効果】
【０００７】

本発明の１実施態様によれば、文書構造に対応した符号化を行うことができるという効果を奏する。

【図面の簡単な説明】

【０００８】

【図１】図１は、符号化処理の流れを概略的に示した図である。

【図２Ａ】図２Ａは、検索処理の流れを概略的に示した図である。

【図２Ｂ】図２Ｂは、インデックスが生成されない場合の検索処理の流れを概略的に示した図である。

【図３】図３は、符号化装置の構成の一例を示す図である。

【図４】図４は、符号の割当ての一例を示す図である。

【図５】図５は、符号の割当ての一例を示す図である。

【図６】図６は、スキーマの概略的構成を示した図である。

【図７Ａ】図７Ａは、タグにより文書構造を示した文書の一例を示す図である。

【図７Ｂ】図７Ｂは、文書の一部にタグによりメタデータを付与した文書の一例を示す図である。

【図８Ａ】図８Ａは、符号化の一例を示す図である。

【図８Ｂ】図８Ｂは、符号化の一例を示す図である。

【図９】図９は、符号化の流れを概略的に示した図である。

【図１０Ａ】図１０Ａは、検索の一例を示した図である。

【図１０Ｂ】図１０Ｂは、検索の一例を示した図である。

【図１１】図１１は、符号化処理の手順の一例を示すフローチャートである。

【図１２】図１２は、検索処理の手順の一例を示すフローチャートである。

【図１３】図１３は、検索処理の手順の一例を示すフローチャートである。

【図１４】図１４は、復号化処理の手順の一例を示すフローチャートである。

【図１５】図１５は、符号の割当ての一例を示す図である。

【図１６】図１６は、符号化プログラムを実行するコンピュータの一例を示す図である。

【図１７】図１７は、検索プログラムを実行するコンピュータを示す図である。

【図１８】図１８は、復号化プログラムを実行するコンピュータの一例を示す図である。

【発明を実施するための形態】

【０００９】

以下に、本願の開示する符号化プログラム、符号化方法、符号化装置、復号化プログラム、復号化方法および復号化装置の実施例を図面に基づいて詳細に説明する。なお、この実施例によりこの権利範囲が限定されるものではない。各実施例は、処理内容を矛盾させない範囲で適宜組み合わせることが可能である。

【実施例１】

【００１０】

［符号化処理］

最初に、図１を用いて符号化処理の概要について説明する。図１は、符号化処理の流れを概略的に示した図である。以下では、構造化された文書が記憶された符号化対象ファイル３０を符号化する場合を例に説明する。

【００１１】

符号化対象ファイル３０には、例えば、ＸＭＬにより、構造化された文書が記憶されている。ＸＭＬでは、テキストにより文書が記述され、タグにより要素が区切られ、構造化されている。図１の例では、病院で患者のカルテのデータをＸＭＬにより、構造化された文書とした場合を示している。図１の例では、要素名「体温」のタグにより、体温が３６．０と記録されている。また、要素名「概要」のタグにより、患者に関する概要「ＸＸＸ・・・」が記録されている。なお、符号化対象ファイル３０は、構造化された文書であれば、何れであってもよい。

【００１２】

符号化対象ファイル３０の符号化を行う場合、符号化装置１０の符号化処理部４０は、

符号化対象ファイル３０に記憶された文書を読み出し、文書の文書構造を特定する（図１（１））。符号化処理部４０は、例えば、符号化対象ファイル３０に対応してＸＭＬのスキーマ（Schema）が定義されている場合、符号化対象ファイル３０に対応するスキーマから文書構造を特定してもよく、文書を解析して文書構造を特定してもよい。

【００１３】

符号化処理部４０は、読み出した文書を、文書構造に対応した階層構造に応じた符号化方式により符号化し、符号化したデータを符号化データ３２として格納する（図１（２））。

【００１４】

例えば、符号化処理部４０は、文書構造を規定する文字列を、共通の符号化方式により符号化する。図１の例では、文書構造を示すタグ「＜体温＞」を符号Ａ１に符号化し、タグ「＜／体温＞」を符号Ａ２に符号化し、タグ「＜概要＞」を符号Ａ３に符号化し、タグ「＜／概要＞」を符号Ａ４に符号化している。終了タグの符号は、開始タグの符号と別な符号としてもよく、タグの終了を示す符号と開始タグの符号を組み合わせてもよい。

【００１５】

また、例えば、符号化処理部４０は、階層ごとに、当該階層に出現する文字列の特性に対応した符号化方式により文字列を符号化する。符号化の際に割り当てる符号は、符号化方式ごとにユニークであればよい。このように、階層ごとの符号化方式により当該階層に出現する文字列を符号化する場合、階層ごとに、符号がユニークであればよい。このため、階層が異なると異なる文字列に同じ符号を割当てることができる。図１の例では、「体温」の階層の文字列「３６．０」を符号Ｂ１に符号化し、「概要」の階層の文字列「ＸＸＸ・・・」の「ＸＸＸ」を符号Ｂ１に符号化している。例えば、符号化処理部４０は、階層ごとに、出現頻度の高い文字や単語などのパターンを短い符号に変換する符号化方式により文字列を符号化する。これにより、階層ごとに、文字列に含まれる各種のパターンのうち、出現頻度が高いパターンを短い符号に変換できるため、符号化対象ファイル３０全体を短い符号に変換できる。

【００１６】

符号化処理部４０は、符号化方式ごとに、変換した文字列と当該文字列に対応する符号を辞書データ３１に記憶する。図１の例では、文字列「３６．０」と符号Ｂ１とが対応付けて辞書データ３１Ａに記憶され、文字列「ＸＸＸ」と符号Ｂ１とが対応付けて辞書データ３１Ｂに記憶されている。

【００１７】

符号化処理部４０は、符号化方式ごとに、符号化した文字列に出現したパターンを示したインデックス３３を生成する（図１（３））。インデックスとは、パターンが含まれるファイルを示したデータである。例えば、インデックスには、パターンおよびファイルに１つのビットが対応付け、ビットの値により、パターンが出現したか否かを記憶するビットマップ型のインデックスがある。また、インデックスには、パターンおよびファイルに複数のビットが対応付け、複数のビットによりパターンの出現回数の情報を保持するカウントマップ型のインデックスがある。図１の例では、符号化処理部４０は、カウントマップ型のインデックス３３Ａ、３３Ｂを生成する。インデックス３３Ａは、「体温」の階層の文字列に出現したパターンの出現回数の情報を保持する。インデックス３３Ｂは、「概要」の階層の文字列に出現したパターンの出現回数の情報を保持する。図１の例では、インデックス３３Ａ、３３Ｂには、符号化対象ファイル３０のファイル番号「１」および符号Ｂ１に対応付けて、出現回数がそれぞれ複数ビットで記憶されている。なお、本実施例では、符号化処理部４０は、符号化の際にインデックス３３Ａ、３３Ｂを生成する場合を説明するが、これらに限定されるものではなく、適宜変更可能である。例えば、符号化処理部４０は、インデックス３３Ａ、３３Ｂを生成しなくてもよい。

【００１８】

［検索処理］

次に、図２Ａを用いて、実施例１にかかる符号化装置１０が実施する検索処理の概要に

ついて説明する。図２Ａは、検索処理の流れを概略的に示した図である。図２Ａの例では、図１により符号化された符号化データ３２と、辞書データ３１Ａ、３１Ｂと、インデックス３３Ａ、３３Ｂが示されている。なお、図２Ａの例では、圧縮データ３２に符号化された文字列を識別しやすくするため、符号の後に括弧記号「（）」で囲んで符号化された文字列を記載している。

【００１９】

符号化装置１０のファイル検索部５０は、検索条件の入力を受け付ける。例えば、図２Ａの例では、ファイル検索部５０は、階層「概要」、文字列「ＸＸＸ」との検索条件を受け付ける。

【００２０】

ファイル検索部５０は、検索条件を満たすファイルを検索する。例えば、ファイル検索部５０は、階層「概要」の文字列を変換した際の辞書データ３１Ｂを参照して、文字列「ＸＸＸ」に対応する符号Ｂ１を特定する（図２Ａ（１））。ファイル検索部５０は、階層「概要」の文字列を変換した際に生成したインデックス３３Ｂを参照して、符号Ｂ１が出現したファイルのファイル番号を特定する（図２Ａ（２））。図２Ａの例では、インデックス３３Ｂに、符号化対象ファイル３０のファイル番号「１」および符号Ｂ１に対応付けて、出現回数が記憶されているため、ファイル番号「１」の符号化対象ファイル３０が検索条件を満たすと検索される。このように、符号化装置１０は、符号化された符号化データ３２に対して文字列の検索を行う場合、符号化データ３２を復号化することなく文字列を検索できるため、活用する際の処理量を減らすことができる。

【００２１】

なお、上述したように、インデックス３３Ａ、３３Ｂは、必ずしも生成されなくてもよい。図２Ｂは、インデックスが生成されない場合の検索処理の流れを概略的に示した図である。図２Ｂの例では、図１により符号化された符号化データ３２と、辞書データ３１が示されている。なお、図２Ｂの例でも、符号化データ３２に符号化された文字列を識別しやすくするため、符号の後に括弧記号「（）」で囲んで符号化された文字列を記載している。

【００２２】

ファイル検索部５０は、検索条件の入力を受け付ける。例えば、図２Ｂの例では、ファイル検索部５０は、階層「概要」、文字列「ＸＸＸ」との検索条件を受け付ける。

【００２３】

ファイル検索部５０は、検索条件を満たすファイルを検索する。例えば、ファイル検索部５０は、共通の符号化方式により符号化されたタグを復号化する。そして、ファイル検索部５０は、ファイル検索部５０は、階層「概要」の文字列を変換した際の辞書データ３１Ｂを参照して、階層「概要」の部分の符号を復号化する（図２Ｂ（１））。そして、ファイル検索部５０は、復号化された部分から文字列「ＸＸＸ」を検索する（図２Ｂ（２））。この場合でも、ファイル検索部５０は、階層「概要」の部分の符号を復号化するのみで検索を行えるため、符号化データ全体を復号化する場合と比較して、活用する際の処理量を減らすことができる。

【００２４】

［装置構成］

次に、符号化装置１０の構成について説明する。図３は、符号化装置の構成の一例を示す図である。符号化装置１０は、構造化された文書の圧縮などの符号化を行う装置である。符号化装置１０は、例えば、パーソナルコンピュータ、サーバコンピュータなどのコンピュータや、タブレット端末、スマートフォンなどの情報処理装置である。符号化装置１０は、１台のコンピュータとして実装してもよく、また、複数台のコンピュータによるクラウドとして実装することもできる。なお、本実施例では、符号化装置１０を１台のコンピュータとした場合を例として説明する。図３に示すように、符号化装置１０は、記憶部２０と、制御部２１とを有する。なお、符号化装置１０は、コンピュータや情報処理装置が有する上記の機器以外の他の機器を有してもよい。また、本実施例では、符号化装置１

０により符号化およびファイル検索を行う場合を例として説明するが、符号化とファイル検索は別な装置で行ってもよい。

【００２５】

記憶部２０は、ハードディスク、ＳＳＤ（Solid State Drive）、光ディスクなどの記憶装置である。なお、記憶部２０は、ＲＡＭ（Random Access Memory）、フラッシュメモリ、ＮＶＳＲＡＭ（Non Volatile Static Random Access Memory）などのデータを書き換え可能な半導体メモリであってもよい。

【００２６】

記憶部２０は、制御部２１で実行されるＯＳ（Operating System）や各種プログラムを記憶する。例えば、記憶部２０は、後述する符号化処理や検索処理を行うプログラムを記憶する。さらに、記憶部２０は、制御部２１で実行されるプログラムで用いられる各種データを記憶する。例えば、記憶部２０は、符号化対象ファイル３０と、辞書データ３１と、符号化データ３２と、インデックス３３とを記憶する。

【００２７】

符号化対象ファイル３０は、符号化対象のテキストデータが記憶されたデータである。例えば、符号化対象ファイル３０には、ＸＭＬにより、構造化された文書が記憶されている。

【００２８】

辞書データ３１は、データの符号化および復号化に用いる辞書のデータである。

【００２９】

ここで、本実施例では、構造化された文書を符号化する際に、構造や属性に応じて符号化方式を切り替える。辞書データ３１は、辞書を用いて符号化する符号化方式で用いる辞書のデータである。辞書データ３１は、辞書を用いて符号化する符号化方式ごとに設けられる。例えば、辞書データ３１は、階層化された文書の階層のうち、辞書を用いて符号化を行う階層ごと、または、辞書を用いて符号化を行う階層でデータ属性が類似する階層ごとに、設けられている。辞書データ３１は、静的辞書３４と、動的辞書３５とを有する。

【００３０】

静的辞書３４は、文書の構造や属性に応じて出現頻度の高いパターンに対応する符号を保持したデータである。動的辞書３５は、文書の構造や属性に応じて出現頻度の低いパターンに対応する符号を保持したデータである。静的辞書３４は、予め設けられる。動的辞書３５は、必要に応じて動的に生成される。

【００３１】

静的辞書３４は、対応する階層に出現する文字列の特性に対応して、文字列に対応する符号が記憶されている。例えば、静的辞書３４は、対応する階層に標準的に出現する文字列や数字などのパターンに対応する符号が記憶されている。また、静的辞書３４は、対応する階層で出現頻度の高いパターンに短い符号が対応付けて記憶されている。例えば、人間の体温は、通常、３５．０℃～４２．０℃の範囲に収まり、３６．０℃前後の頻度が高い。そこで、例えば、体温の階層に対応する静的辞書３４には、３５．０～４２．０の数値に対して符号が対応付けて記憶されており、３６．０前後に対して短い符号が割り当てられて記憶されている。また、本実施例では、概要に出現する文字列を単語の単位で符号化する。例えば、本実施例では、単語を、一般的な文書を解析して、出現頻度が相対的に高い高頻度単語と、出現頻度が相対的に低い低頻度単語とに分けている。例えば、出現頻度の高い順に所定の順位までの基礎単語を高頻度単語とし、所定の順位以降の基礎単語を低頻度単語とする。高頻度単語については、短い符号を予め割り当てて、割り当てた符号と高頻度単語を対応付けて静的辞書３４に記憶させる。例えば、高頻度単語については、予め２バイト（１６ビット）の符号を割り当て、割り当てた符号を静的辞書３４に予め記憶させる。低頻度単語については、出現した際に符号を動的に割り当てて、割り当てた符号を動的辞書３５に記憶させる。すなわち、符号は、高頻度単語については予め登録され、低頻度単語については動的に割り当てられて動的辞書３５に記憶される。なお、概要に出現する文字列や数字などのパターンが特定のパターンに定まる場合は、概要の階層に対

応する静的辞書３４には、特定のパターンと符号を対応付けて予め記憶させてもよい。
【００３２】
動的辞書３５は、対応する階層に出現する文字列の特性に対応して、動的に割り当てられた符号に関する各種の情報を保持したデータである。例えば、概要の階層に対応した動的辞書３５には、低頻度単語など出現頻度の低いパターンに動的に割り当てられた符号が記憶される。
【００３３】
図４は、符号の割当ての一例を示す図である。図４には、２バイト（１６ビット）の符号に対する割当ての一例が示されている。上部の横方向の項目は、最初の１バイト目を０～Ｆの１６進表記で示しており、「＊」は、２バイト目を示している。例えば、「１＊ｈ」は、１バイト目が２進数表記で「００００００００１」であることを示す。左側の縦方向の項目は、２バイト目を０～Ｆの１６進表記で示しており、「＊」は、１バイト目を示している。例えば、「＊２ｈ」は、２バイト目が２進数表記で「００００００１０」であることを示す。
【００３４】
図４では、縦方向の項目と横方向の項目に対応する領域に、符号に対応させるパターンを示す。例えば、「０＊ｈ」、「１＊ｈ」の符号については、各階層とも、同じ制御コードに同じ符号を対応付けている。また、「２＊ｈ」～「５＊ｈ」の符号については、各階層とも、同じタグに同じ符号を対応付けている。また、「６＊ｈ」～「Ｆ＊ｈ」の符号については、各階層でそれぞれ個別にパターンに符号を割当て可能としている。例えば、文字列を単語の単位で符号化する場合、「６＊ｈ」～「９＊ｈ」の符号については、予め定めた高頻度単語に対して割当ている。「Ａ＊ｈ」～「Ｆ＊ｈ」の符号については、低頻度単語が出現した際に符号を動的に割り当てる。「Ｅ＊ｈ」、「Ｆ＊ｈ」は、符号の不足に対応するため、３バイトの符号としている。
【００３５】
辞書データ３１は、辞書を用いて符号化を行う階層ごと、または、辞書を用いて符号化を行う階層でデータ属性が類似する階層ごとに設けられ、「６＊ｈ」～「Ｆ＊ｈ」の符号について、階層に出現する文字列の特性に対応して文字列と符号を対応付けて記憶する。
【００３６】
なお、辞書データ３１は、タグに対して動的に符号を割当て可能としてもよい。図５は、符号の割当ての一例を示す図である。図５の例では、１バイト目が「５＊ｈ」の符号について、特定階層のタグとして動的に符号を割当て可能としている。
【００３７】
図３に戻り、符号化データ３２は、符号化対象ファイル３０をそれぞれ符号化したデータである。インデックス３３は、符号化した文字列に出現したパターンの出現回数を記憶したデータである。例えば、インデックス３３は、符号化方式ごとに設けられ、符号化した文字列に出現したパターンの出現回数と出現したファイルのファイル番号を対応付けて記憶される。
【００３８】
制御部２１は、符号化装置１０を制御するデバイスである。制御部２１としては、ＣＰＵ（Central Processing Unit）、ＭＰＵ（Micro Processing Unit）等の電子回路や、ＡＳＩＣ（Application Specific Integrated Circuit）、ＦＰＧＡ（Field Programmable Gate Array）等の集積回路を採用できる。制御部２１は、各種の処理手順を規定したプログラムや制御データを格納するための内部メモリを有し、これらによって種々の処理を実行する。制御部２１は、各種のプログラムが動作することにより各種の処理部として機能する。例えば、制御部２１は、符号化処理部４０と、ファイル検索部５０と、復号化処理部６０とを有する。
【００３９】
符号化処理部４０は、符号化対象ファイル３０に記憶された構造化された文書を読み出し、読み出した文書を、文書構造に対応した階層構造に応じた符号化方式により符号化し

た符号化データ３２を生成する。符号化処理部４０は、特定部４１と、符号化部４２と、生成部４３とを有する。

【００４０】

特定部４１は、各種の特定を行う。例えば、特定部４１は、符号化対象ファイル３０に記憶されたＸＭＬの文書の文書構造を特定する。例えば、特定部４１は、符号化対象ファイル３０に対応してＸＭＬのスキーマが定義されている場合、符号化対象ファイル３０に対応するスキーマから文書構造を特定する。

【００４１】

図６は、スキーマの概略的構成を示した図である。ＸＭＬの文書では、文書構造を示すＸＭＬスキーマ７０が定義される。ＸＭＬスキーマ７０には、ＸＭＬの文書の文書構造の定義や、末端要素の型や制約の定義が、スキーマ言語により記述されている。図６の例では、構造定義として、文書構造を示すタグの入れ子の関係や、タグの制約などが記述される。また、図６の例では、末端要素の型や制約として、格納される文字列のデータ型、数値の最大、最小、ｓｔｒｉｎｇ（文字列）の長さ、使える文字、ｓｔｒｉｎｇを、例えば、Ｍａｌｅ、Ｆｅｍａｌｅ等の選択型として使用しているかが記述される。符号化対象ファイル３０には、ＸＭＬスキーマ７０の定義に対応して、ＸＭＬで文書が記憶される。図６の例では、ＸＭＬの文書には、文字コードが１行目に記載され、ＸＭＬスキーマ７０に対応した文書構造で文書が記憶される。なお、ＸＭＬスキーマ７０は、文書構造を柔軟に定義でき、符号化対象ファイル３０ごとにタグの個数を変更可能な定義もできる。例えば、符号化対象ファイル３０Ａは、Ｘタグの配下にＹタグが１０個存在し、符号化対象ファイル３０Ｂは、Ｘタグの配下にＹタグが２０個存在する文書構成とすることもできる。

【００４２】

特定部４１は、符号化対象ファイル３０に対応してＸＭＬスキーマ７０が定義されている場合、ＸＭＬスキーマ７０から文書構造を特定する。なお、特定部４１は、符号化対象ファイル３０に記憶された文書を解析して文書構造を特定してもよい。

【００４３】

符号化部４２は、符号化対象ファイル３０に記憶された文書の符号化を行う。例えば、符号化部４２は、特定部４１により文書構造を特定したＸＭＬの文書を符号化対象ファイル３０から読み出す。そして、符号化部４２は、読み出した文書を、文書構造に対応した階層構造に応じた符号化方式により符号化する。例えば、符号化部４２は、読み出した文書に出現するタグに対して順に符号を割当てて、符号化する。なお、文書構造に出現するタグが定まっている場合、タグと符号を対応付けたタグ用の辞書データを予め記憶させ、符号化部４２は、タグ用の辞書データを用いて、読み出した文書に出現するタグを符号化してもよい。また、文書構造に出現する頻度の高い一部のタグについてタグ用の辞書データに記憶させ、符号化部４２は、頻度の高い一部のタグをタグ用の辞書データを用いて符号化し、それ以外のタグに順に符号を割当てて符号化してもよい。

【００４４】

ここで、構造化された文書には、タグにより文書の要素を区分けして文書構造を示した文書と、文書の一部にタグによりメタデータを付与した文書がある。

【００４５】

図７Ａは、タグにより文書構造を示した文書の一例を示す図である。図７Ａの例では、例１の文書は、タグにより「概要」、「本文」が定義されている。また、例１の文書は、タグにより「概要」、「本文」と区分された部分にそれぞれ内容に応じた文字列（テキスト）が記憶された文書が示されている。例２の文書は、タグにより「特許」が定義され、「特許」の下層に「名称」、「課題」、「効果」が定義されている。また、例２の文書は、タグにより「名称」、「課題」、「効果」と区分けされた部分にそれぞれ内容に応じた文字列が記憶された文書が示されている。

【００４６】

符号化部４２は、タグを共通の符号化方式により符号化する。例１の文書は、「概要」、「本文」のタグを共通の符号化方式により符号化する。例２の文書は、「特許」、「名

称」、「課題」、「効果」のタグを共通の符号化方式により符号化する。

【００４７】

また、符号化部４２は、タグにより区分けされた部分の文字列を、それぞれ階層に応じた符号化方式により符号化する。例えば、符号化部４２は、タグにより区分けされた部分の文字列を、それぞれ階層に対応した辞書データ３１を用いて符号化する。例えば、符号化部４２は、文字列に出現した単語が、階層に対応した辞書データ３１の静的辞書３４または動的辞書３５に登録されている場合、出現した単語を静的辞書３４または動的辞書３５に登録された符号に符号化する。また、符号化部４２は、文字列に出現した単語が、階層に対応した辞書データ３１の静的辞書３４または動的辞書３５に登録されていない場合、符号を動的に割り当て、出現した単語を割り当てた符号に符号化する。符号化部４２は、出現した単語と割り当てた符号を対応付けて動的辞書３５に登録する。これにより、以降、動的辞書３５に登録された単語は、出現した際に、動的辞書３５を用いて同じ符号に符号化される。なお、符号化部４２は、データ属性が類似する階層の文字列を同じ符号化方式により符号化してもよい。これにより、符号化部４２は、データ属性が類似する階層の文字列を同じ辞書データ３１で符号化できる。

【００４８】

図７Ｂは、文書の一部にタグによりメタデータを付与した文書の一例を示す図である。図７Ｂの例では、例３の文書は、「ＡＡＡへのリンクはこちら」の文書の「リンク」部分にタグによりリンク先のＵＲＬをメタデータとして付与した場合を示している。例４の文書は、「ＢＢＢを訴えたため、ＣＣＣを疑い、ＤＤＤを投与した」の文書の「ＢＢＢ」部分が病状を示し、「ＣＣＣ」部分が病名を示し、「ＤＤＤ」部分が薬名を示すことをタグによりメタデータとして付与した場合を示している。例５の文書は、「２０１５／３／６に、大阪で鈴木に会う」の文書の「２０１５／３／６」部分が日時を示し、「大阪」部分が地名を示し、「鈴木」部分が人名を示すことをタグによりメタデータとして付与した場合を示している。

【００４９】

符号化部４２は、タグを共通の符号化方式により符号化する。例３の文書は、「リンク」のタグを共通の符号化方式により符号化する。例４の文書は、「病状」、「病名」、「薬名」のタグを共通の符号化方式により符号化する。例５の文書は、「日時」、「地名」、「人名」のタグを共通の符号化方式により符号化する。また、符号化部４２は、タグにより区分けされた部分の文字列を、それぞれ階層に応じた符号化方式により符号化する。例えば、符号化部４２は、タグにより区分けされた部分の文字列を、それぞれ階層に対応した辞書データ３１を用いて符号化する。

【００５０】

図８Ａは、符号化の一例を示す図である。図８Ａの例は、タグ「Ａ」の下層にタグ「Ｂ」が定義された文字列のデータを符号化した一例を示している。図８Ａの例では、タグ「Ａ」、タグ「Ｂ」の符号の間に、文字列のデータを符号化したコードが記憶される。なお、図８Ａの例では、タグ「Ａ」、タグ「Ｂ」の終了タグの符号を、タグの終了を示す符号と開始タグの符号を組み合わせたものとしている。

【００５１】

図８Ｂは、符号化の一例を示す図である。図８Ｂの例は、「大阪で鈴木に会う」の文書の「大阪」部分が地名を示し、「鈴木」部分が人名を示すことをタグによりメタデータとして付与された文字列のデータを符号化した一例を示している。図８Ｂの例では、地名の開始符号「２５ｈ」と終了符号「２０ｈ」、「２５ｈ」の間で、「大阪」が「Ｂ０ｈ」と符号されている。また、人名の開始符号「２６ｈ」と終了符号「２０ｈ」、「２６ｈ」の間で、「鈴木」が「Ｂ０ｈ」と符号されている。

【００５２】

符号化部４２は、階層が異なると異なる文字列に同じ符号を割当てることができるため、階層ごとに文字列を短い符号に変換できる。例えば、図８Ｂの例では、「大阪」と「鈴木」が共に同じ「Ｂ０ｈ」に変換されている。このように、符号化部４２は、階層ごとに

、文字列を短い符号に変換できるため、符号化対象ファイル３０全体を短い符号に変換できる。

【００５３】

なお、符号化部４２は、タグにより区分けされた部分の文字列の属性や範囲によっては、辞書データ３１を用いずに当該文字列を符号化してもよい。例えば、タグにより区分けされた部分の文字列が「０」～「２５５」の範囲の数値を示す文字列である場合、符号化部４２は、「０」～「２５５」の範囲の数値を示す文字列を１バイトの整数型（例えば、ｉｎｔ型）の符号に符号化してもよい。すなわち、符号化部４２は、文字列が数値を示す場合、当該数値の範囲に対応したデータ型の符号に符号化してもよい。数値を表す文字列を数値のデータ型の符号に符号化すると、符号化した状態でも数値の比較や集計など各種の演算を行うことができる。

【００５４】

ここで、ＸＭＬなど構造化された文書では、タグによってコンテキストが規定される。構造化された文書は、タグによってコンテキストが規定され、コンテキストによりデータの処理に関わる要素が定まる。例えば、データの型や値の範囲、文書の構成要素（言語であれば日本語の単語、英語の単語、他言語の単語）など辞書に関わる要素が決まる。また、例えば、テキストであれば検索やマイニング、数値であれば平均値や集計値、頻度分布など、データの内容がどのように活用が可能かの活用分野が定まる。また、図７Ａに示すような、タグにより文書構造を示した文書では、単一タグだけでなく、階層構造を上位から順に辿ると判明するコンテキストがある。例えば、＜Ａ＞＜合計＞Ｔ＜／合計＞＜／Ａ＞では、「Ｔ」がＡに関する合計を示している。また、例えば、図７Ｂに示すような、文書の一部にタグによりメタデータを付与した文書では、単一タグで囲まれた領域で、階層構造に加え、追加されるコンテキストがある。例えば、<地名>大阪</地名>は、「大阪」が地名であることを示している。よって、符号化部４２は、タグにより区分けされた部分の文字列を、タグによって規定されるコンテキストに適した符号化方式で符号化することにより、活用する際の処理量を減らすことができる。

【００５５】

符号化部４２は、符号化対象ファイル３０に記憶された文書の符号化したデータを符号化データ３２として格納する。

【００５６】

生成部４３は、符号化方式ごとに、符号化した文字列に出現したパターンを示したインデックス３３を生成する。例えば、生成部４３は、符号化した符号化対象ファイル３０に対して、順にファイル番号を付与する。そして、生成部４３は、符号化した符号化対象ファイル３０のファイル番号に対応付けて、符号化対象ファイル３０に出現した数値や単語などのパターンの出現回数を格納したインデックス３３を生成する。

【００５７】

ここで、符号化の流れを説明する。図９は、符号化の流れを概略的に示した図である。符号化装置１０の符号化処理部４０は、符号化対象ファイル３０に記憶された文書を読み出し、文書の文書構造を特定する。符号化処理部４０は、読み出した文書を、文書構造に対応した階層構造に応じた符号化方式により符号化する。例えば、符号化処理部４０は、タグや文字列に出現した単語が、静的辞書３４または動的辞書３５に登録されている場合、出現した単語を静的辞書３４または動的辞書３５に登録された符号に符号化する。また、符号化処理部４０は、タグや文字列に出現した単語が、静的辞書３４または動的辞書３５に登録されていない場合、符号を動的に割り当て、タグや出現した単語を割り当てた符号に符号化する。符号化処理部４０は、タグや出現した単語と割り当てた符号を対応付けて動的辞書３５に登録する。

【００５８】

符号化処理部４０は、符号化対象ファイル３０に記憶された文書の符号化したデータを符号化データ３２として格納する。図９の例では、「概要」と「本文」のタグの階層の文字列がそれぞれ符号化されている。また、符号化処理部４０は、符号化対象ファイル３０

のファイル番号に対応付けて、符号化対象ファイル３０に出現した数値や単語などのパターンの出現回数を格納したインデックス３３を生成する。図９の例では、「概要」と「本文」のタグの階層に対応して、出現回数の集計結果としてインデックス３３Ａ、３３Ｂが生成されている。符号化装置１０は、符号化データ３２を他の装置へ移動させる場合、符号化データ３２に対応して生成された動的辞書３５およびインデックス３３Ａ、３３Ｂも移動させる。

【００５９】

図３に戻り、ファイル検索部５０は、指定された検索条件に従い、ファイルを検索する。ファイル検索部５０は、受付部５１と、検索部５２と、出力部５３とを有する。以下、ファイル検索部５０の各構成について詳細に説明する。

【００６０】

受付部５１は、検索条件を受け付ける。例えば、受付部５１は、検索条件とするキーワードや階層の入力を受け付ける操作画面などの入力インタフェースを提供しており、検索条件とする文字列や階層の入力を受け付ける。

【００６１】

検索部５２は、検索条件を満たすファイルを検索する。例えば、検索部５２は、検索条件の階層に対応する辞書データ３１の静的辞書３４および動的辞書３５を参照して、検索条件のキーワードに対応する符号を特定する。そして、検索部５２は、検索条件の階層に対応するインデックス３３を参照して、特定された符号が出現したファイルのファイル番号を特定する。なお、検索条件のキーワードが単語や数値を複数含む場合、検索部５２は、キーワードを単語や数値に分解して符号化し、それぞれの単語や数値ごとに対応する符号を特定する。検索部５２は、検索条件の階層に対応するインデックス３３を参照して、それぞれの単語や数値ごとに対応する符号が出現したファイルのファイル番号を特定する。ここで、インデックス３３では、検索条件の文字列に含まれる複数の単語や数値の出現順が正しいかを確認できない。そこで、例えば、検索部５２は、特定したファイル番号の符号化対象ファイル３０に検索条件の文字列が含まれるかを検索する。なお、検索部５２は、特定したファイル番号に対応する符号化データ３２の検索条件の階層を復号化して、検索条件の文字列が含まれるかを検索してもよい。

【００６２】

図１０Ａは、検索の一例を示した図である。図１０Ａの例は、指定されたファイルが「概要」に「ＸＸＸ」というキーワードを含み、「本文」に「ＹＹＹ」というキーワードを含むかを検索する場合を示している。検索部５２は、「概要」の階層に対応する辞書データ３１の静的辞書３４および動的辞書３５を参照して、「ＸＸＸ」に対応する符号を特定する。検索部５２は、「概要」の階層に対応するインデックス３３を参照して、指定されたファイルのファイル番号に「ＸＸＸ」に対応する符号が出現したことが記録されているかを特定する。また、検索部５２は、「本文」の階層に対応する辞書データ３１の静的辞書３４および動的辞書３５を参照して、「ＹＹＹ」に対応する符号を特定する。そして、検索部５２は、「本文」の階層に対応するインデックス３３を参照して、指定されたファイルのファイル番号に「ＹＹＹ」に対応する符号が出現したことが記録されているかを特定する。そして、検索部５２は、指定されたファイルのファイル番号に「ＸＸＸ」に対応する符号と「ＹＹＹ」に対応する符号が出現した記録がある場合、「概要」に「ＸＸＸ」というキーワードを含み、「本文」に「ＹＹＹ」というキーワードを含むかを検索する。

【００６３】

図１０Ｂは、検索の一例を示した図である。図１０Ｂの例は、「概要」に「ＺＺＺ」というキーワードを含むファイルを検索する場合を示している。検索部５２は、「概要」の階層に対応する辞書データ３１の静的辞書３４および動的辞書３５を参照して、「ＺＺＺ」に対応する符号を特定する。そして、検索部５２は、「概要」の階層に対応するインデックス３３を参照して、「ＺＺＺ」に対応する符号が出現したファイルのファイル番号を特定する。

【００６４】

このように、ファイル検索部５０は、符号化データ３２を復号化せずに検索を行えるため、検索の際の処理量が減らすことができ、検索の処理時間を短縮できる。

【００６５】

なお、インデックス３３が生成されない場合、ファイル検索部５０は、指定された階層のみ復号化して、指定された文字列を検索する。この場合でも、ファイル検索部５０は、指定された階層を復号化するのみで検索を行えるため、符号化データ全体を復号化する場合と比較して、活用する際の処理量を減らすことができ、検索の処理時間を短縮できる。

【００６６】

出力部５３は、検索結果の出力を行う。例えば、検索部５２によりファイル番号が特定された場合、出力部５３は、検索結果として、特定されたファイル番号のファイルのファイル名を出力する。一方、検索部５２によりファイル番号が特定されない場合、出力部５３は、検索結果として、該当ファイルなしを出力する。

【００６７】

図３に戻り、復号化処理部６０は、符号化データ３２を復号化する。復号化処理部６０は、受付部６１と、復号化部６２とを有する。以下、復号化処理部６０の各構成について詳細に説明する。

【００６８】

受付部６１は、復号化の指示を受け付ける。例えば、受付部６１は、復号化する対象の符号化データ３２の指定を受け付ける操作画面などの入力インタフェースを提供しており、復号化する対象の符号化データ３２の指定を受け付ける。なお、受付部６１は、復号化する対象の符号化データ３２と共に、符号化する階層の指定を受け付けてもよい。

【００６９】

復号化部６２は、指定された符号化データ３２を復号化する。例えば、復号化部６２は、復号化部６２は、符号化データ３２のそれぞれの階層の符号データを、当該階層の符号化方式により復号化する。例えば、復号化部６２は、符号化データ３２のそれぞれの階層の符号データを、当該階層に対応する辞書データ３１の静的辞書３４および動的辞書３５を用いて、文字列に復号化する。例えば、復号化部６２は、タグの符号データを、共通の符号化方式により復号化する。そして、復号化部６２は、タグで区切られた各階層の符号データを、当該階層に対応する辞書データ３１の静的辞書３４および動的辞書３５を参照して、文字列に復号化する。なお、受付部６１で符号化する階層の指定を受け付けた場合、復号化部６２は、指定された階層の符号データのみを復号化してもよい。

【００７０】

［処理の流れ］

本実施例に係る符号化装置１０が符号化対象ファイル３０を符号化する符号化処理の流れについて説明する。図１１は、符号化処理の手順の一例を示すフローチャートである。この符号化処理は、所定のタイミング、例えば、符号化対象ファイル３０を指定して符号化開始を指示する所定操作が行われたタイミングで実行される。

【００７１】

図１１に示すように、特定部４１は、符号化対象ファイル３０に記憶された構造化された文書の文書構造を特定する（Ｓ１０）。符号化部４２は、文書構造を特定した文書の各階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する（Ｓ１１）。例えば、符号化部４２は、タグを共通の符号化方式により符号化する。また、符号化部４２は、タグにより区分けされた部分の文字列を、それぞれ階層に応じた符号化方式により符号化する。符号化部４２は、符号化したデータを符号化データ３２に格納する（Ｓ１２）。生成部４３は、符号化方式ごとに、符号化した文字列に出現したパターンを示したインデックス３３を生成し（Ｓ１３）、処理を終了する。

【００７２】

次に、本実施例に係る符号化装置１０が検索条件を満たすファイルを検索する検索処理の流れについて説明する。最初に、検索条件に階層が指定されない場合の検索処理の流れを説明する。図１２は、検索処理の手順の一例を示すフローチャートである。この検索処

理は、所定のタイミング、例えば、検索条件を指定して検索開始を指示する所定操作が行われたタイミングで実行される。

【００７３】

図１２に示すように、検索部５２は、辞書データ３１の静的辞書３４および動的辞書３５を参照して、検索条件のキーワードに対応する符号が存在するか判定する（Ｓ２０）。符号が存在しない場合（Ｓ２０否定）、検索部５２は、キーワードを単語や数値に分解してそれぞれ符号化し、それぞれの単語や数値ごとに対応する符号を特定する（Ｓ２１）。検索部５２は、各インデックス３３を参照して、それぞれの単語や数値ごとに対応する符号が出現したファイルのファイル番号を特定する（Ｓ２２）。検索部５２は、特定したファイル番号の符号化対象ファイル３０に検索条件の文字列が含まれるかを検索する（Ｓ２３）。

【００７４】

一方、符号が存在する場合（Ｓ２０肯定）、検索部５２は、インデックス３３を参照して、特定された符号が出現したファイルのファイル番号を特定する（Ｓ２４）。

【００７５】

出力部５３は、検索結果を出力し、処理を終了する（Ｓ２５）。例えば、出力部５３は、検索条件の文字列を含む符号化対象ファイル３０が検索された場合や、検索部５２により符号化対象ファイル３０のファイル番号が特定された場合、符号化対象ファイル３０のファイル名を出力する。

【００７６】

次に、検索条件に階層が指定された場合の検索処理の流れを説明する。図１３は、検索処理の手順の一例を示すフローチャートである。この検索処理は、所定のタイミング、例えば、検索条件を指定して検索開始を指示する所定操作が行われたタイミングで実行される。

【００７７】

図１３に示すように、検索部５２は、辞書データ３１の静的辞書３４および動的辞書３５を参照して、検索条件のキーワードに対応する符号が存在するか判定する（Ｓ３０）。符号が存在しない場合（Ｓ３０否定）、検索部５２は、キーワードを単語や数値に分解してそれぞれ符号化し、それぞれの単語や数値ごとに対応する符号を特定する（Ｓ３１）。検索部５２は、指定された階層のインデックス３３を参照して、それぞれの単語や数値ごとに対応する符号が出現したファイルのファイル番号を特定する（Ｓ３２）。検索部５２は、特定したファイル番号の符号化対象ファイル３０に検索条件の文字列が含まれるかを検索する（Ｓ３３）。

【００７８】

一方、符号が存在する場合（Ｓ３０肯定）、検索部５２は、指定された階層のインデックス３３を参照して、特定された符号が出現したファイルのファイル番号を特定する（Ｓ３４）。

【００７９】

出力部５３は、検索結果を出力し、処理を終了する（Ｓ３５）。例えば、出力部５３は、検索条件の文字列を含む符号化対象ファイル３０が検索された場合や、検索部５２により符号化対象ファイル３０のファイル番号が特定された場合、符号化対象ファイル３０のファイル名を出力する。

【００８０】

次に、本実施例に係る符号化装置１０が符号化データ３２を復号化する復号化処理の流れについて説明する。図１４は、復号化処理の手順の一例を示すフローチャートである。この復号化処理は、所定のタイミング、例えば、復号化する対象の符号化データ３２を指定して符号化開始を指示する所定操作が行われたタイミングで実行される。

【００８１】

復号化部６２は、指定された符号化データ３２から符号データを読み出す（Ｓ４０）。復号化部６２は、読み出した符号データを、階層に対応する辞書データ３１の静的辞書３

４および動的辞書３５を用いて、文字列に復号化する（Ｓ４１）。復号化部６２は、符号化データ３２の読み出しが完了したか否かを判定する（Ｓ４２）。読み出しが完了していない場合は（Ｓ４２否定）、Ｓ４０へ移行する。一方、読み出しが完了した場合は（Ｓ４２肯定）、処理を終了する。

【００８２】

［効果］

上述してきたように、本実施例に係る符号化装置１０は、構造化された文書の文書構造を特定する。符号化装置１０は、文書構造を特定した文書中の特定階層の文字列を、当該文書構造に対応した階層構造に応じた符号化方式により符号化する。これにより、符号化装置１０は、特定階層の部分の符号のみを復号化できるため、活用する際の処理量を減らすことができる。

【００８３】

また、本実施例に係る符号化装置１０は、文書中の文書構造を規定する文字列を、共通の符号化方式により符号化する。これにより、符号化装置１０は、共通の符号化方式で復号化することで、文書中の文書構造を規定する文字列を同じ符号化方式で復元できるため、文書構造を速やかに特定でき、特定の階層のデータを速やかに抽出できる。

【００８４】

また、本実施例に係る符号化装置１０は、データ属性が類似する階層の文字列を同じ符号化方式により符号化する。これにより、符号化装置１０は、データ属性が類似する階層の文字列を同じ辞書データ３１で符号化できる。

【００８５】

また、本実施例に係る符号化装置１０は、特定階層の文字列を、当該特定階層に出現する文字列の特性に対応した符号化方式により符号化する。これにより、符号化装置１０は、特定階層の文字列を特性に対応した符号化方式で符号化できる。

【００８６】

また、本実施例に係る符号化装置１０は、１またはデータ属性が類似する複数の階層ごとに、出現頻度の高いパターンを短い符号に変換する符号化方式により符号化する。これにより、符号化装置１０は、符号化対象ファイル３０を高い圧縮率で符号化できる。

【００８７】

また、本実施例に係る符号化装置１０は、符号化した文字列に出現したパターンを示したインデックス３３を生成する。これにより、符号化装置１０は、インデックス３３からパターンが出現した符号化対象ファイル３０を特定できる。

【実施例２】

【００８８】

さて、これまで開示の装置に関する実施例について説明したが、開示の技術は上述した実施例以外にも、種々の異なる形態にて実施されてよいものである。そこで、以下では、本発明に含まれる他の実施例を説明する。

【００８９】

例えば、上記の実施例では、出現頻度の高いパターンに対応する符号を辞書データ３１の静的辞書３４に予め記憶させる場合について説明したが、これに限定されない。例えば、文書の階層ごとに、文字列で単語や数字など出現するパターンごとの出現頻度を解析により求めて、出現頻度の高いパターンから短い符号を割り当て符号化してもよい。辞書データ３１は、出現したパターンと割り当てた符号を対応付けて記憶させてもよい。

【００９０】

また、上記の実施例では、符号を層構造単位で辞書データ３１に記憶する場合について説明したが、これに限定されない。例えば、共通の辞書データ３１を用いてもよい。また、一部の符号を層構造単位の辞書データ３１で共通に登録して管理してもよい。図１５は、符号の割当ての一例を示す図である。図１５には、一部の符号を層構造単位の辞書データ３１で共通に登録して管理する場合の符号の割当ての一例が示されている。「８＊ｈ」～「Ａ＊ｈ」の符号については、各階層で符号を共通に登録して管理する。例えば、共通

の辞書データ３１により、ファイル全体で符号を管理した方が効率がよい符号がある。例えば、数値情報のＮＡ（未入力）やｎｕｌｌ値（値なし、文字列や数値共通）が、別の値で表現されている場合がある。この場合、共通の辞書データ３１で管理することにより、符号を統一して管理できる。なお、符号を統一して管理する場合でも、ある数値では０．０をＮＡにし、他の値では－９９．９をＮＡに割り当ててもよい。また、文書内全体で登場するような文字列は、符号を統一して管理することが好ましい。例えば、電子書籍の小説において小説の主人公の名前が概要、本文、講評でも登場する場合、主人公の名前の符号を統一して管理することが好ましい。一方、階層構造単位で符号を管理した方が効率がよい符号がある。例えば、階層構造単位で適切な範囲が定まる場合は、階層構造単位で符号を管理した方がよい。適切な範囲から外れた場合は、ＮＡやＮＵＬＬとして符号化する。例えば、人の体温の辞書として、３５．０～４２．０の範囲で辞書データ３１を用意する。体温として３４．８が出現した場合は、ＮＡやＮＵＬＬ、または、動的に符号を割当てて符号化する。また、人の身長の辞書として、１２０．０～２２２．３の範囲で辞書データ３１を用意する。身長として２３１．２という値が出現した場合は、ＮＡやＮＵＬＬ、または、動的に符号を割当てて符号化する。

【００９１】

また、図示した各装置の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的状態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。例えば、符号化装置１０の特定部４１、符号化部４２、生成部４３、受付部５１、検索部５２、出力部５３、受付部６１および復号化部６２の各処理部が適宜統合されてもよい。符号化装置１０の上記各処理部の処理が適宜複数の処理部の処理に分離されてもよい。さらに、各処理部にて行なわれる各処理機能は、その全部または任意の一部が、ＣＰＵおよび当該ＣＰＵにて解析実行されるプログラムにて実現され、あるいは、ワイヤードロジックによるハードウェアとして実現され得る。

【００９２】

［符号化プログラム］

また、上記の実施例で説明した各種の処理は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーションなどのコンピュータシステムで実行することによって実現することもできる。そこで、以下では、上記の実施例と同様の機能を有するプログラムを実行するコンピュータシステムの一例を説明する。最初に、符号化処理を行う符号化プログラムについて説明する。図１６は、符号化プログラムを実行するコンピュータの一例を示す図である。

【００９３】

図１６に示すように、コンピュータ４００は、ＣＰＵ（Central Processing Unit）４１０、ＨＤＤ（Hard Disk Drive）４２０、ＲＡＭ（Random Access Memory）４４０を有する。これら４００～４４０の各部は、バス５００を介して接続される。

【００９４】

ＨＤＤ４２０には上記の符号化装置１０の特定部４１、符号化部４２および生成部４３と同様の機能を発揮する符号化プログラム４２０ａが予め記憶される。なお、符号化プログラム４２０ａについては、適宜分離してもよい。

【００９５】

また、ＨＤＤ４２０は、各種情報を記憶する。例えば、ＨＤＤ４２０は、ＯＳや符号化に用いる各種データを記憶する。

【００９６】

そして、ＣＰＵ４１０が、符号化プログラム４２０ａをＨＤＤ４２０から読み出して実行することで、実施例の各処理部と同様の動作を実行する。すなわち、符号化プログラム４２０ａは、特定部４１、符号化部４２および生成部４３と同様の動作を実行する。

【００９７】

なお、上記した符号化プログラム４２０ａについては、必ずしも最初からＨＤＤ４２０に記憶させることを要しない。

【００９８】

［検索プログラム］

次に、符号化データ３２を検索する検索プログラムについて説明する。図１７は、復号化プログラムを実行するコンピュータの一例を示す図である。なお、図１６と同一の部分については同一の符号を付して、説明を省略する。

【００９９】

図１７に示すように、ＨＤＤ４２０には上記の符号化装置１０の受付部５１、検索部５２および出力部５３と同様の機能を発揮する検索プログラム４２０ｂが予め記憶される。なお、検索プログラム４２０ｂについては、適宜分離してもよい。

【０１００】

また、ＨＤＤ４２０は、各種情報を記憶する。例えば、ＨＤＤ４２０は、ＯＳや検索に用いる各種データを記憶する。

【０１０１】

そして、ＣＰＵ４１０が、検索プログラム４２０ｂをＨＤＤ４２０から読み出して実行することで、実施例の各処理部と同様の動作を実行する。すなわち、検索プログラム４２０ｂは、受付部５１、検索部５２および出力部５３と同様の動作を実行する。

【０１０２】

なお、上記した検索プログラム４２０ｂについても、必ずしも最初からＨＤＤ４２０に記憶させることを要しない。

【０１０３】

［復号化プログラム］

次に、検索条件を満たすファイルを復号化する復号化プログラムについて説明する。図１８は、復号化プログラムを実行するコンピュータの一例を示す図である。なお、図１６および図１７と同一の部分については同一の符号を付して、説明を省略する。

【０１０４】

図１７に示すように、ＨＤＤ４２０には上記の符号化装置１０の受付部６１および復号化部６２と同様の機能を発揮する復号化プログラム４２０ｃが予め記憶される。なお、復号化プログラム４２０ｃについては、適宜分離してもよい。

【０１０５】

また、ＨＤＤ４２０は、各種情報を記憶する。例えば、ＨＤＤ４２０は、ＯＳや検索に用いる各種データを記憶する。

【０１０６】

そして、ＣＰＵ４１０が、復号化プログラム４２０ｃをＨＤＤ４２０から読み出して実行することで、実施例の各処理部と同様の動作を実行する。すなわち、復号化プログラム４２０ｃは、受付部６１および復号化部６２と同様の動作を実行する。

【０１０７】

なお、上記した復号化プログラム４２０ｃについても、必ずしも最初からＨＤＤ４２０に記憶させることを要しない。

【０１０８】

また、例えば、符号化プログラム４２０ａ、検索プログラム４２０ｂおよび復号化プログラム４２０ｃは、コンピュータ４００に挿入されるフレキシブルディスク（ＦＤ）、ＣＤ－ＲＯＭ、ＤＶＤディスク、光磁気ディスク、ＩＣカードなどの「可搬用の物理媒体」に記憶させてもよい。そして、コンピュータ４００がこれらからプログラムを読み出して実行するようにしてもよい。

【０１０９】

さらには、公衆回線、インターネット、ＬＡＮ、ＷＡＮなどを介してコンピュータ４００に接続される「他のコンピュータ（又はサーバ）」などにプログラムを記憶させておく。そして、コンピュータ４００がこれらからプログラムを読み出して実行するようにして

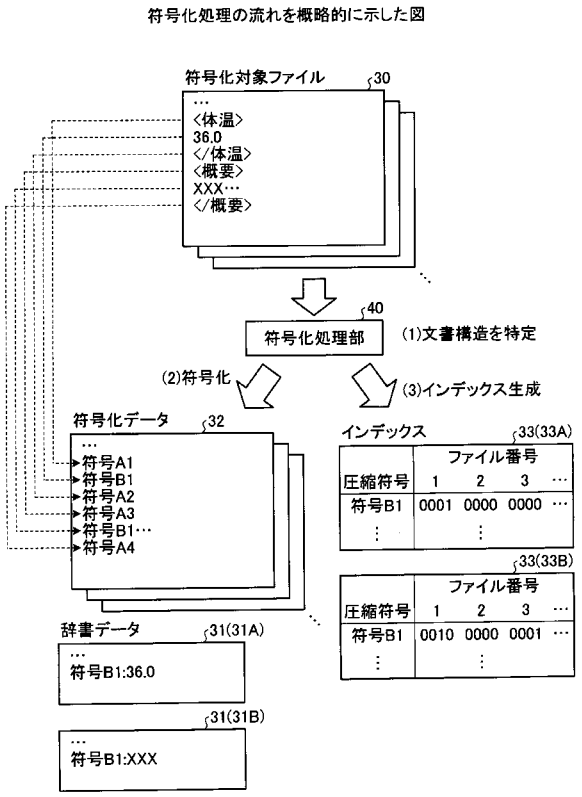
もよい。

【符号の説明】

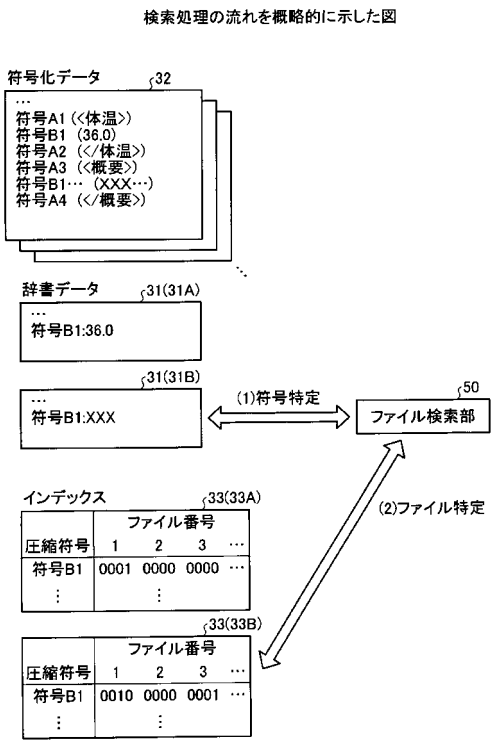
【０１１０】

１０　符号化装置
２０　記憶部
２１　制御部
３０　符号化対象ファイル
３１　辞書データ
３２　符号化データ
３３　インデックス
３４　静的辞書
３５　動的辞書
４０　符号化処理部
４１　特定部
４２　符号化部
４３　生成部
５０　ファイル検索部
５１　受付部
５２　検索部
５３　出力部
６０　復号化処理部
６１　受付部
６２　復号化部
７０　ＸＭＬスキーマ

【図１】

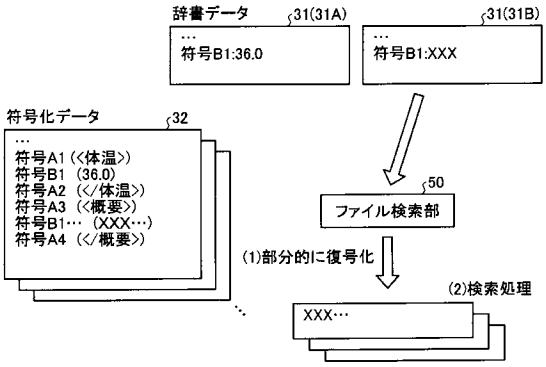


【図２Ａ】



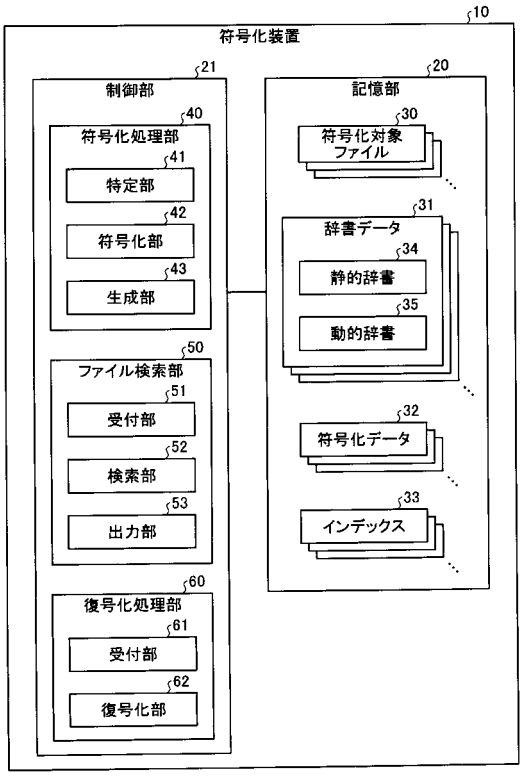
【図２Ｂ】

インデックスが生成されない場合の検索処理の流れを概略的に示した図



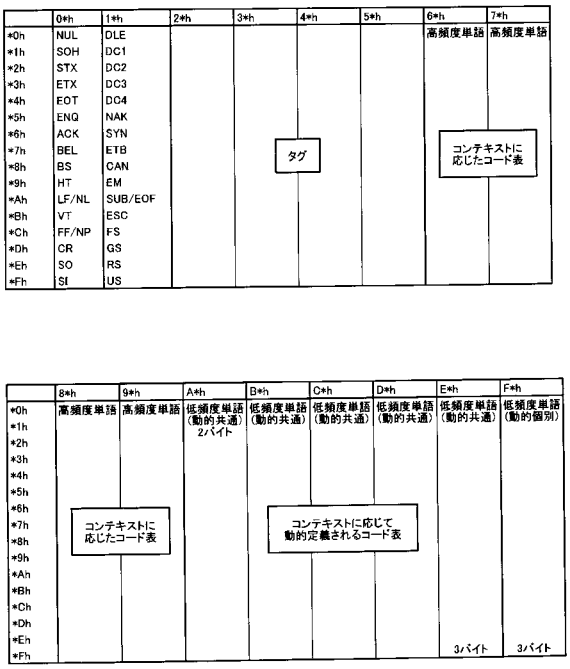
【図３】

符号化装置の構成の一例を示す図



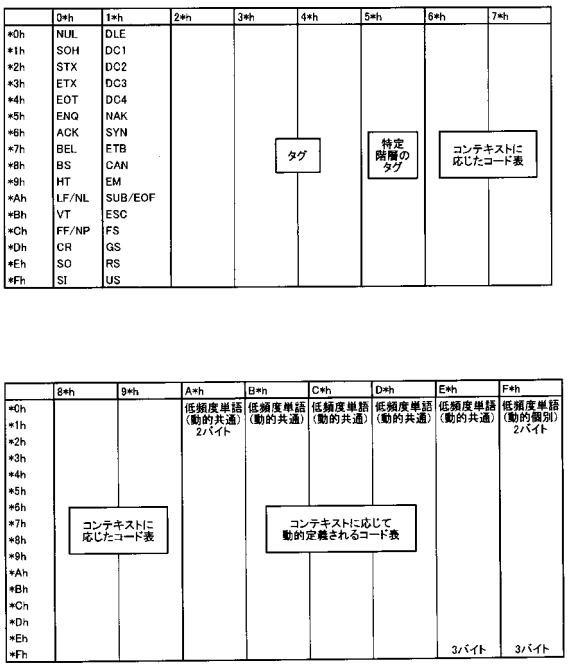
【図４】

符号の割当ての一例を示す図



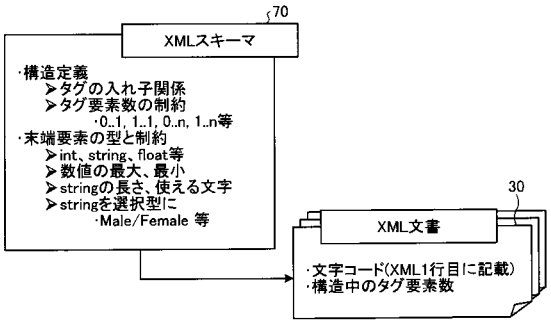
【図５】

符号の割当ての一例を示す図



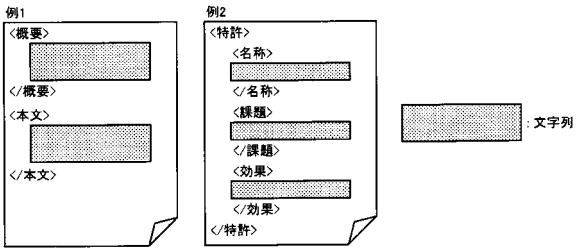
【図６】

スキーマの概略的構成を示した図



【図７Ａ】

タグにより文書構造を示した文書の一例を示す図



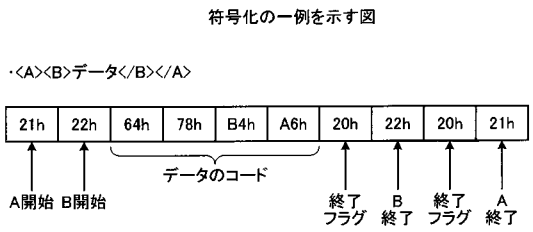
【図７Ｂ】

文書の一部にタグによりメタデータを付与した文書の一例を示す図

例3: AAAへのリンクはこちら
例4: <症状>BBB</症状>を訴えたため、<病名>CCC</病名>を疑い、<薬名>DDD </薬名>を投与した
例5: <日時>2015/3/6</日時>に、<地名>大阪</地名>で<人名>鈴木</人名>に会う

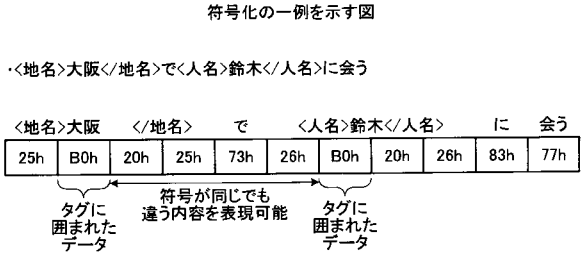
【図８Ａ】

符号化の一例を示す図

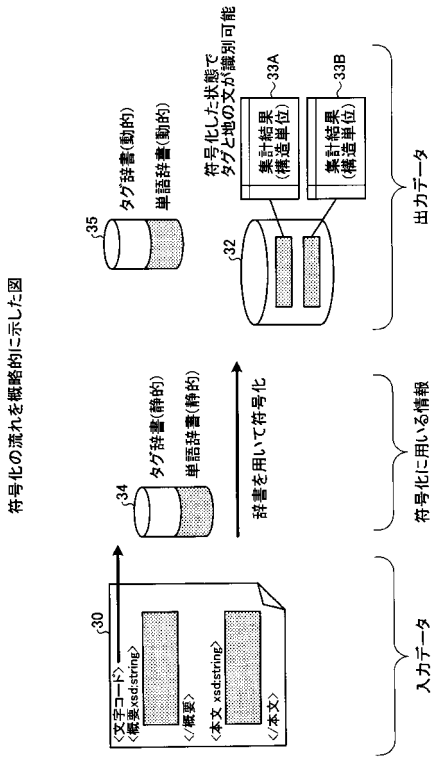


【図８Ｂ】

符号化の一例を示す図

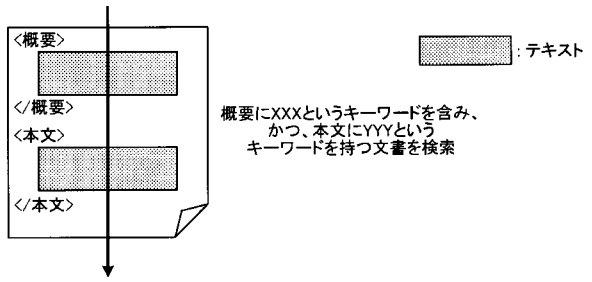


【図９】



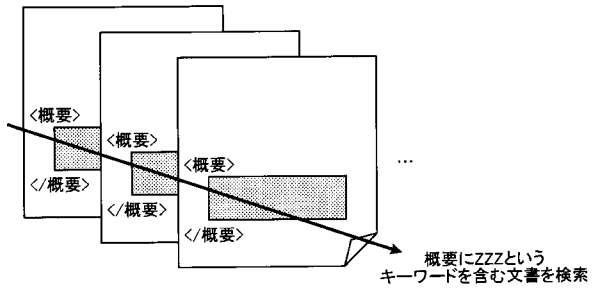
【図１０Ａ】

検索の一例を示した図



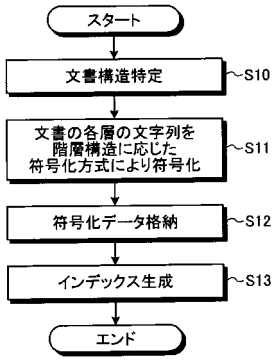
【図１０Ｂ】

検索の一例を示した図



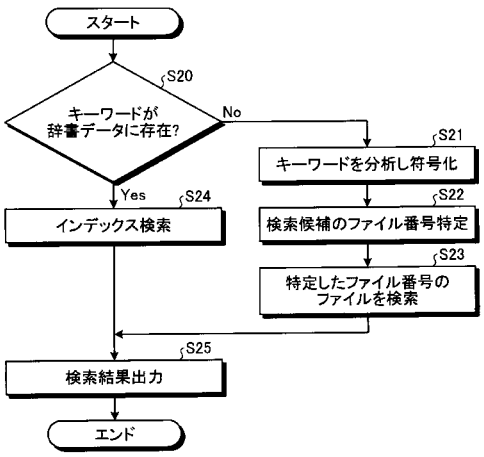
【図１１】

符号化処理の手順の一例を示すフローチャート



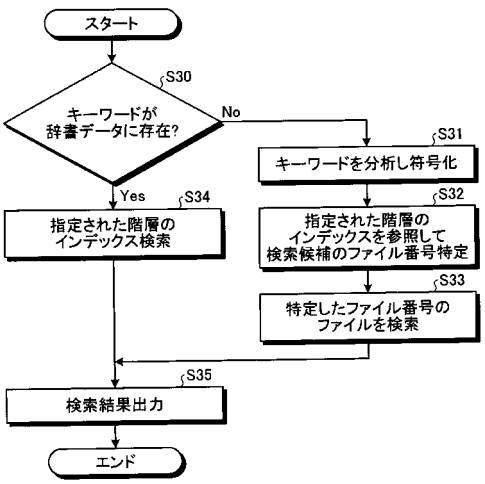
【図１２】

検索処理の手順の一例を示すフローチャート



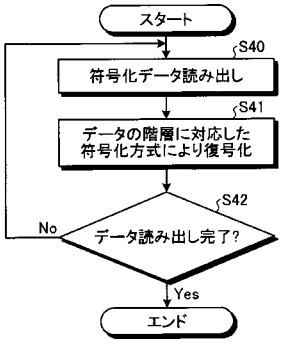
【図１３】

検索処理の手順の一例を示すフローチャート



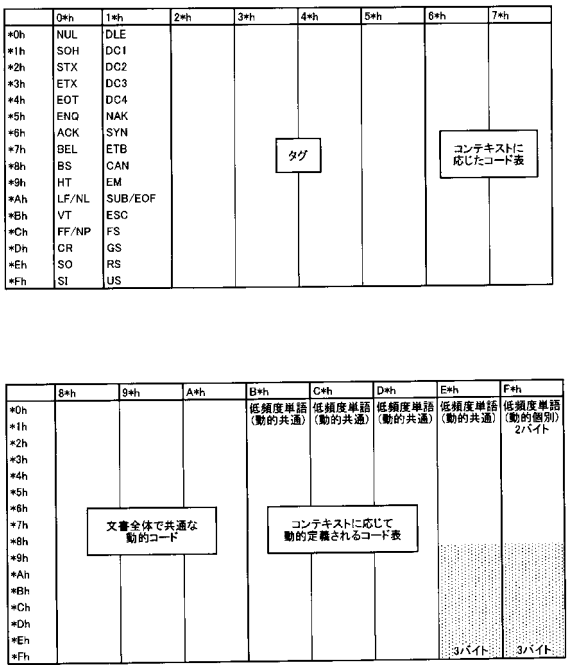
【図１４】

復号化処理の手順の一例を示すフローチャート



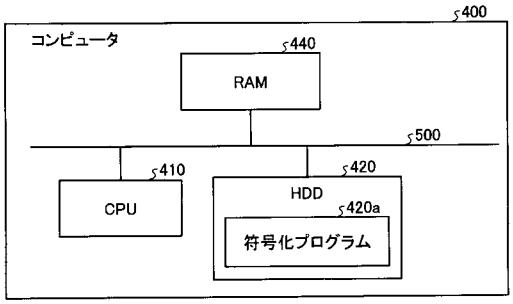
【図１５】

符号の割当ての一例を示す図



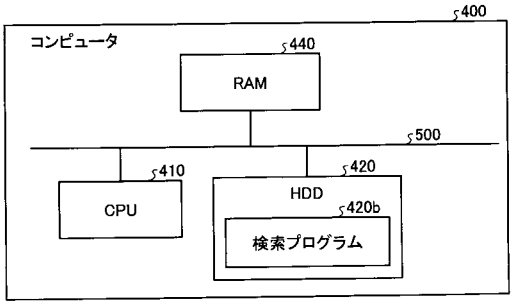
【図１６】

符号化プログラムを実行するコンピュータの一例を示す図



【図１７】

検索プログラムを実行するコンピュータを示す図



【図１８】

復号化プログラムを実行するコンピュータの一例を示す図