(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4261299号
(P4261299)
(45)発行日　平成21年4月30日(2009.4.30)　(24)登録日　平成21年2月20日(2009.2.20)

(51)Int.Cl.　F I
H03M 7/42 (2006.01)　H03M 7/42

請求項の数 8　(全 25 頁)

(21)出願番号　特願2003-328428(P2003-328428)
(22)出願日　平成15年9月19日(2003.9.19)
(65)公開番号　特開2005-94652(P2005-94652A)
(43)公開日　平成17年4月7日(2005.4.7)
審査請求日　平成18年4月11日(2006.4.11)

(73)特許権者　392026693
株式会社エヌ・ティ・ティ・ドコモ
東京都千代田区永田町二丁目１１番１号
(74)代理人　100088155
弁理士　長谷川　芳樹
(74)代理人　100092657
弁理士　寺崎　史朗
(74)代理人　100114270
弁理士　黒川　朋也
(74)代理人　100122507
弁理士　柏岡　潤二
(74)代理人　100123995
弁理士　野田　雅一

最終頁に続く

(54)【発明の名称】データ圧縮装置、データ復元装置およびデータ管理装置

(57)【特許請求の範囲】
【請求項１】
型と値をそれぞれ有する複数の頂点と、該頂点間の参照情報とを有する入力データを、前記頂点間の参照情報を有する相互参照関係データと、前記型と値を有する複数の頂点からなる頂点群とに分離し、その分離された前記頂点群のデータを出力する分離手段と、
特定のパターンを有する前記頂点間の参照情報を、前記頂点間の参照情報が共有可能なテンプレートとして蓄積するテンプレート蓄積手段と、
前記分離手段により分離された前記相互参照関係データから、前記テンプレート蓄積手段に蓄積されているテンプレートと一致する箇所を検出するテンプレート一致箇所検出手段と、
前記分離手段により分離された相互参照関係データのうち、前記テンプレート一致箇所検出手段により検出された一致箇所を前記テンプレートで前記頂点間の参照情報を参照可能な状態に置換し、その置換された相互参照関係データを出力するテンプレート置換手段とを有することを特徴とするデータ圧縮装置。
【請求項２】
それぞれの値を有し、該各値が属性情報として型を有することが可能な複数の頂点と、該頂点間の参照情報とを有する入力データを、前記頂点間の参照情報を有する相互参照関係データと、前記値を有する複数の頂点からなる頂点群とに分離し、その分離された前記頂点群のデータを出力する分離手段と、
特定のパターンを有する前記頂点間の参照情報を、前記頂点間の参照情報が共有可能な

テンプレートとして蓄積するテンプレート蓄積手段と、
前記分離手段により分離された前記相互参照関係データから、前記テンプレート蓄積手段に蓄積されているテンプレートと一致する箇所を検出するテンプレート一致箇所検出手段と、
前記分離手段により分離された相互参照関係データのうち、前記テンプレート一致箇所検出手段により検出された一致箇所を前記テンプレートで前記頂点間の参照情報を参照可能な状態に置換し、その置換された相互参照関係データを出力するテンプレート置換手段とを有することを特徴とするデータ圧縮装置。
【請求項３】
前記テンプレートが、前記頂点間の参照情報の一部または全部の参照方向を反転可能なことを特徴とする請求項１または２記載のデータ圧縮装置。
【請求項４】
第１から第ＮまでのＮ個の頂点を有し、前記第１の頂点と第２の頂点以外の連続番号を有する前記頂点は相互に参照し、前記第１の頂点が前記第２の頂点を参照し、かつ外部への参照を保持し、前記第Ｎの頂点が第Ｎ－１の頂点を参照し、前記第２から第Ｎまでの各頂点が、外部への参照を保持しないか、またはすべて同数の参照を保持する連続兄弟参照部を有する前記相互参照関係データに適用するための接続情報を有する連続兄弟参照用テンプレートが、前記テンプレート蓄積手段に蓄積されていることを特徴とする請求項１～３のいずれか一項記載のデータ圧縮装置。
【請求項５】
前記頂点間の参照情報に、前記テンプレートを適用可能な親テンプレートが前記テンプレート蓄積手段に蓄積されていることを特徴とする請求項１～４のいずれか一項記載のデータ圧縮装置。
【請求項６】
複数の前記入力データに共用可能な共用テンプレートが前記テンプレート蓄積手段に蓄積されていることを特徴とする請求項１～５のいずれか一項記載のデータ圧縮装置。
【請求項７】
特定のパターンを有する複数の頂点間の参照情報を、前記頂点間の参照情報が共有可能なテンプレートとして蓄積するテンプレート蓄積手段と、
前記テンプレートにより置換され、圧縮された相互参照関係データを第１の入力データとして入力し、前記相互参照関係データから、前記テンプレートを用いて圧縮前の元の相互参照関係データを復元する展開手段と、
型と値をそれぞれ有する複数の前記頂点からなる頂点群のデータを第２の入力データとして入力し、前記頂点群のデータを前記展開手段により復元された前記相互参照関係データと合成したデータを出力する合成手段とを有することを特徴とするデータ復元装置。
【請求項８】
圧縮可能なデータを蓄積する第１のデータ蓄積手段と、
請求項１～６のいずれか一項記載のデータ圧縮装置により圧縮されたデータを蓄積する第２のデータ蓄積手段と、
前記第１のデータ蓄積手段及び第２のデータ蓄積手段に蓄積されたそれぞれのデータの利用頻度を観測し、該観測された利用頻度に応じて移動要求を出力する利用頻度観測手段と、
該利用頻度観測手段からの移動要求にしたがい、前記利用頻度が高い高頻度データを請求項７記載のデータ復元装置により復元して前記第１のデータ蓄積手段へ格納し、前記利用頻度が前記高頻度データよりも低いデータを前記データ圧縮装置により圧縮して前記第２のデータ蓄積手段へ格納するようにして、データを移動させるための制御を行う制御手段と、
前記第１のデータ蓄積手段と第２のデータ蓄積手段のいずれかから、前記制御手段の指示に応じてデータを取得して出力する選択手段とを有することを特徴とするデータ管理装置。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、電子データの圧縮装置、復元装置およびデータ管理装置に関する。
【背景技術】
【０００２】
近年、ＷＷＷ(World Wide Web)の普及により、ＨＴＭＬ(Hyper Text Markup Language)やＸＭＬ(EXtensible Markup Language)等、構造化文書を用いたデータ交換が増加している。特に、ＸＭＬはＨＴＭＬを補う次世代の言語として注目を集めており、今後インターネットにおける情報交換の場において最も普及していくことが予想される。
【０００３】
ＸＭＬは要素の階層構造を表すデータ表現形式を伴った言語であり、ＸＭＬを用いた文書（ＸＭＬ文書）は、例えば図１８のように記述される。図１８は、ＸＭＬ文書１０を示す図である。図１８に示すとおり、ＸＭＬは大きくマークアップとテキスト情報に分けられる。図１８に示すＸＭＬ文書１０では、マークアップは、要素開始記号（開始タグ）Ｍａ、要素終了記号（終了タグ）Ｍｂ、空要素記号（空要素タグ）Ｍｃからなっている。図１８では、<book>、<title>、<authors>、<author>、<contents>および<chapter>が要素開始記号Ｍａを表している。また、</book>、</title>、</authors>、</author>、</contents>および</chapter>が要素終了記号Ｍｂを表し、<misc/>が空要素記号Ｍｃを表している。これらの要素開始記号Ｍａから対応する要素終了記号Ｍｂまでの領域、または空要素記号Ｍｃが要素（ＸＭＬの基本となる情報単位）を表している。
【０００４】
要素開始記号Ｍａと要素終了記号Ｍｂの間には、別な要素記号の他、テキスト情報を記述することができる。例えば、図１８に示すＸＭＬ文書１０では、要素<title>には、文字列“ＸＭＬの基礎”が、要素<authors>の中に現れる最初の要素<author>には、文字列“山田　太郎”がそれぞれテキスト情報として定義されている。
【０００５】
要素やテキスト情報の間には、親子関係、兄弟関係が定義されている。図１８に示すＸＭＬ文書１０の場合、要素開始記号Ｍａが<book>で始まり、要素終了記号Ｍｂが</book>で終了する要素（要素<book>）の中に、要素開始記号Ｍａが<title>で始まり、要素終了記号Ｍｂが</title>で終了する要素（要素<title>）が含まれている。このとき、要素<book>は要素<title>の親要素であるといい、要素<title>は要素<book>の子要素であるという。これが要素の親子関係である。
【０００６】
また、要素<title>と要素<authors>とは、同一の親要素<book>を持ち、かつ連続している。このとき、要素<title>と要素<authors>とは兄弟であるといい、要素<title>は要素<authors>の前兄弟、要素<authors>は要素<title>の次兄弟であるという。これが要素の兄弟関係である。
【０００７】
一般に、ＸＭＬは、コンピュータ間で通信を行う際や、ハードディスク装置やフラッシュメモリに蓄積する際には、図１８に示すＸＭＬ文書１０ようにテキスト形式で表現されている。一方、コンピュータ内部で検索や修正用に利用するときは、解析されてコンピュータ内部に適したデータ構造に変換されている。
【０００８】
図１９は、図１８に示すＸＭＬ文書１０を解析し、コンピュータの内部利用に適した形式に変換したデータ構造１１を示す図である。図１９では、各要素及びテキスト情報が型および値を有する頂点３０１～３１７として記述されている。型は各頂点３０１～３１７の左側に記述され、“Ｅ”であれば要素を表し、“Ｔ”であればテキスト情報を表している。例えば、頂点３０１では型３０１ａは“Ｅ”である。また、値は頂点の右側に記述され、例えば、頂点３０１では値３０１ｂは“ｂｏｏｋ”である。そして、頂点が要素を表

す場合は値に要素の名称（要素名）が記述され、テキスト情報を表す場合は文字列が記述される。例えば、頂点３０２では要素名<title>を表し、頂点３０６ではテキスト情報“ＸＭＬの基礎”を表している。

【０００９】

また、各頂点３０１～３１７は、もとの（変換前の）ＸＭＬ文書１０の親子関係および兄弟関係を表現するため、親参照、子参照、次兄弟参照および前兄弟参照の４つの参照を表す参照情報を有している。上述のＸＭＬ文書１０の場合、要素<title>は要素<book>の子要素であり、要素<book>は要素<title>の親要素であるから、図１９に示すデータ構造１１では、例えば、頂点３０１，３０２については、それぞれ、<book>から<title>への子参照Ｐ１と、<title>から<book>への親参照Ｐ２を有し、それらが矢印によって表現されている。また、要素<book>は<title>の次の子要素として要素<authors>も有している。この場合、頂点３０２，３０３については、要素<title>から要素<authors>への次兄弟参照Ｐ３、要素<authors>から要素<title>への前兄弟参照Ｐ４が保持されている。なお、兄弟関係にある要素では、先頭の子要素（例えば要素<title>）以外は親参照を直接に有しないものとされている。

【００１０】

データ構造は、各頂点間の参照情報と、要素名やテキスト情報とを分離して管理することができ、例えば、そのそれぞれを図２０（ａ）、図２０（ｂ）のように表現することができる。ここで、図２０（ａ）は各頂点間の参照情報を有する相互参照関係データ４００を示す図であり、図２０（ｂ）は要素とテキスト情報のいずれかに設定される型と値を有する複数の頂点の集合（頂点群ともいう）を示すテーブル４５０を示す図である。

【００１１】

しかしながら、メモリ等の記憶装置の容量は有限であるため、データ構造を蓄積するときは、そのデータ構造を効率的に圧縮して蓄積することが求められる。この点に関し、非特許文献１には、図２０（ｂ）に示すような要素名やテキスト情報を圧縮する方法が開示されている。非特許文献１では、各頂点が保持する要素名やテキスト情報を辞書として別途蓄積し、各頂点には辞書のインデックスを持たせ、同じ文字列を複数蓄積しないようにすることで、圧縮する方法が開示されている。

【００１２】

一方、非特許文献２には、ＸＭＬ文書中の部分的な構造を再利用することで、ＸＭＬ文書を圧縮する方法が開示されている。この方法は元のＸＭＬ文書を構造、要素名情報、テキスト情報の３つに分離したのち、そのそれぞれをＬＺ７７等の一般的な圧縮アルゴリズムで圧縮するというものである（ＬＺ７７について詳しくは、Jacob Ziv,Abraham Lempel :A Universal Algorithm for Sequential Data Compression.IEEE Transactions on Information Theory 23(3):337-343(1977)を参照）。

ここで、非特許文献２に開示されている圧縮方法について説明する。この圧縮方法ではまず、要素開始記号や空要素記号をそれぞれ「＃１」、「＃２」のような短い要素名で置換し、要素終了記号を「／」で置換する。また、テキスト情報は「Ｃ」で置換する。

以上の圧縮方法を分離したＸＭＬ文書１０に適用すると、分離後のデータ構造１２、要素名情報１３およびテキスト情報１４はそれぞれ図２１、図２２、図２３のように表現される。

【００１３】

また、非特許文献２に記載の圧縮方法では、ＬＺ７７等に代表される圧縮アルゴリズムを用いてそれぞれを独立に圧縮するが、ここではその圧縮アルゴリズムの概要について説明する。ＬＺ７７等の圧縮アルゴリズムは元の入力情報に含まれる部分的なパターンを発見し、それをテンプレートとして繰り返し再利用することにより、圧縮を行う。例えば、図２１に示すデータ構造１２の圧縮について説明すると、テンプレートとして、テンプレートＸ，Ｙ，Ｚ，Ｗ，Ｖを用いるとし、それぞれのテンプレートの割り当てを、

Ｘ＝“＃１　＃２　Ｃ　／　＃３”，Ｙ＝”＃４Ｃ　／”，

Ｚ＝“／　＃５”，Ｗ＝“＃６　Ｃ／”，Ｖ＝“／　＃７　／　／”のように設定すると、

図２１に示したデータ構造１２は“Ｘ Ｙ Ｙ Ｙ Ｚ Ｗ Ｗ Ｖ”のように表せる。これは一部の文書構造をあらわすテンプレートとして、Ｙ，Ｗを複数回利用している。このように、テンプレートが繰り返し利用でき、元の文書を少ないテンプレートで表現することができれば、元のＸＭＬ文書を表す情報量が少なくて済むから圧縮が可能になる。

【００１４】

【非特許文献１】Mathias Neumuller and John N. Wilson: “Compact In-Memory Representation of XML”Internal Report of University of strathclyde

【非特許文献２】Hartmut Liefke and Dan Suciu.: “XMill: An Efficient Compressor for XML Data”, In proceedings of ACM SIGMOD International Conference on Management of Data, 2000

【発明の開示】

【発明が解決しようとする課題】

【００１５】

しかし、上述した従来の技術では、テンプレートを用いてＸＭＬ文書のデータ構造を圧縮する際、テンプレート相互の情報が独立していたため、テンプレート数が増えると、その分だけ必要とされる記憶量が増えてしまい、メモリ等の記憶装置を圧迫するという問題があった。

【００１６】

そこで、本発明は上記課題を解決するためになされたもので、テンプレート数が増えても各テンプレートの格納に要するメモリ等の記憶装置を圧迫しないようにすることができる構成を備えたデータ圧縮装置、データ復元装置およびデータ管理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】

【００１７】

上記課題を解決するため、本発明は型と値をそれぞれ有する複数の頂点と、その頂点間の参照情報とを有する入力データを、頂点間の参照情報を有する相互参照関係データと、型と値を有する複数の頂点からなる頂点群とに分離し、その分離された頂点群のデータを出力する分離手段と、特定のパターンを有する頂点間の参照情報を、頂点間の参照情報が共有可能なテンプレートとして蓄積するテンプレート蓄積手段と、分離手段により分離された相互参照関係データから、テンプレート蓄積手段に蓄積されているテンプレートと一致する箇所を検出するテンプレート一致箇所検出手段と、分離手段により分離された相互参照関係データのうち、テンプレート一致箇所検出手段により検出された一致箇所をテンプレートで頂点間の参照情報を参照可能な状態に置換し、その置換された相互参照関係データを出力するテンプレート置換手段とを有するデータ圧縮装置を特徴とする。

このデータ圧縮装置は、テンプレート蓄積手段に蓄積されているテンプレートが頂点間の参照情報が共有可能な構成を有するので、テンプレート数を記憶するための記憶容量が少なくて済むようになる。

【００１８】

また、本発明は、それぞれの値を有し、その各値が属性情報として型を有することが可能な複数の頂点と、その頂点間の参照情報とを有する入力データを、頂点間の参照情報を有する相互参照関係データと、値を有する複数の頂点からなる頂点群とに分離し、その分離された頂点群のデータを出力する分離手段と、特定のパターンを有する頂点間の参照情報を、頂点間の参照情報が共有可能なテンプレートとして蓄積するテンプレート蓄積手段と、分離手段により分離された相互参照関係データから、テンプレート蓄積手段に蓄積されているテンプレートと一致する箇所を検出するテンプレート一致箇所検出手段と、分離手段により分離された相互参照関係データのうち、テンプレート一致箇所検出手段により検出された一致箇所をテンプレートで頂点間の参照情報を参照可能な状態に置換し、その置換された相互参照関係データを出力するテンプレート置換手段とを有するデータ圧縮装置を提供する。

このデータ圧縮装置も、テンプレート蓄積手段に蓄積されているテンプレートが頂点間

の参照情報が共有可能な構成を有するので、テンプレート数を記憶するための記憶容量が少なくて済むようになる。

【００１９】

上記いずれのデータ圧縮装置も、テンプレートが、頂点間の参照情報の一部または全部の参照方向を反転可能なことが好ましい。

このようなテンプレートを有すると、１つのテンプレートを反転させて適用することができるから、テンプレート数を少なくすることができる。

【００２０】

また、第１から第ＮまでのＮ個の頂点を有し、第１の頂点と第２の頂点以外の連続番号を有する頂点は相互に参照し、第１の頂点が第２の頂点を参照し、かつ外部への参照を保持し、第Ｎの頂点が第Ｎ－１の頂点を参照し、第２から第Ｎまでの各頂点が、外部への参照を保持しないか、またはすべて同数の参照を保持する連続兄弟参照部を有する相互参照関係データに適用するための接続情報を有する連続兄弟参照用テンプレートが、テンプレート蓄積手段に蓄積されているようにすることもできる。

このテンプレートは、各頂点間の参照情報を有しなくてもよいので、記憶容量が少なくて済む。

【００２１】

上記いずれのデータ圧縮装置も、頂点間の参照情報に、テンプレートを適用可能な親テンプレートがテンプレート蓄積手段に蓄積されているとよい。

このテンプレートは、テンプレートを定義するのに他のテンプレートの定義を利用できるため、テンプレートを表現するのに必要な記憶容量を削減することができる。

【００２２】

さらに、複数の入力データに共用可能な共用テンプレートがテンプレート蓄積手段に蓄積されていることが好ましい。

共用テンプレートは、複数の入力データに共用されるから、テンプレート数を少なくすることができるようになる。

【００２３】

そして、本発明は、特定のパターンを有する複数の頂点間の参照情報を、頂点間の参照情報が共有可能なテンプレートとして蓄積するテンプレート蓄積手段と、テンプレートにより置換され、圧縮された相互参照関係データを第１の入力データとして入力し、相互参照関係データから、テンプレートを用いて圧縮前の元の相互参照関係データを復元する展開手段と、型と値をそれぞれ有する複数の頂点からなる頂点群のデータを第２の入力データとして入力し、頂点群のデータを展開手段により復元された相互参照関係データと合成したデータを出力する合成手段とを有するデータ復元装置を提供する。

このようなデータ復元装置によれば、圧縮された相互参照関係データをテンプレート蓄積手段に蓄積されたテンプレートを用いて圧縮前の元の相互参照関係データを復元することができる。

【００２４】

さらに、本発明は、圧縮可能なデータを蓄積する第１のデータ蓄積手段と、請求項１～６のいずれか一項記載のデータ圧縮装置により圧縮されたデータを蓄積する第２のデータ蓄積手段と、第１のデータ蓄積手段及び第２のデータ蓄積手段に蓄積されたそれぞれのデータの利用頻度を観測し、その観測された利用頻度に応じて移動要求を出力する利用頻度観測手段と、その利用頻度観測手段からの移動要求にしたがい、利用頻度が高い高頻度データを請求項７記載のデータ復元装置により復元して第１のデータ蓄積手段へ格納し、利用頻度が高頻度データよりも低いデータを上記データ圧縮装置により圧縮して第２のデータ蓄積手段へ格納するようにして、データを移動させるための制御を行う制御手段と、第１のデータ蓄積手段と第２のデータ蓄積手段のいずれかから、制御手段の指示に応じてデータを取得して出力する選択手段とを有するデータ管理装置を提供する。

このデータ管理装置は、利用頻度に応じて、第１のデータ蓄積手段と第２のデータ蓄積手段との間でデータを移動させることができ、データを圧縮済みまたは非圧縮の状態に適

宜変えて蓄積することができる。したがって、利用頻度の高いデータの利用速度が高いまま維持されるため、高速化が可能となる。

【発明の効果】

【００２５】

本発明によれば、データ構造の圧縮に用いるテンプレート数が増えても、各テンプレートの格納に要するメモリ等の記憶装置を圧迫しないようにすることができる。

【発明を実施するための最良の形態】

【００２６】

以下、図面を参照して本発明に係るデータ圧縮装置、データ復元装置およびデータ管理装置の実施の形態について、添付図面を用いて詳細に説明する。

データ圧縮装置の実施の形態

（第１の実施の形態）

図１は本実施の形態に係るデータ圧縮装置１０１の構成を示すブロック図である。図１に示すように、データ圧縮装置１０１はテンプレート蓄積手段１０２、分離手段１０３、テンプレート一致箇所検出手段１０４およびテンプレート置換手段１０５を有している。このデータ圧縮装置１０１は、入力データ１０６から、第１の出力データ１０７と第２の出力データ１０８を出力するようになっている。

【００２７】

図１におけるデータ圧縮装置１０１はテンプレート蓄積手段１０２、分離手段１０３、テンプレート一致箇所検出手段１０４、テンプレート置換手段１０５までが一体化された単一の装置として構成されているが、必ずしも単一の装置として実現される必要はなく、各手段を図示しない通信手段により接続して実現することも可能である。例えば、データ圧縮装置１０１からテンプレート蓄積手段１０２を取り除き、テンプレート蓄積手段１０２を別体の装置として実現し、データ圧縮装置１０１とテンプレート蓄積手段１０２とを通信手段によって接続する構成にしてもよい。そうすると、テンプレート蓄積手段１０２を含まない複数のデータ圧縮装置を複数設け、それらの間で１つのテンプレート蓄積手段１０２を共有するといったことも可能になる。

【００２８】

なお、本実施の形態においては、図２に示すＸＭＬ文書２０を圧縮する手順をもって発明の詳細を説明するが、本実施の形態におけるデータ圧縮装置１０１による圧縮の対象はＸＭＬ文書２０のようなＸＭＬ文書に限定されるものではなく、型と値を有する頂点、または値を有し、その各値が属性情報として型を有する頂点と、その頂点間の参照情報とを有するような一般的なグラフ構造によって表現されるデータにも適用することができる。値を有し、その各値が属性情報として型を有する頂点とは、例えば値として“１”、属性情報として“整数”といった情報を有するような頂点であり、属性情報から値の型が判定可能であることから、型と値を有する頂点を有するデータと同様に、本実施の形態におけるデータ圧縮装置１０１により、圧縮を行うことが可能である。

また、一般的なグラフ構造によって表現されるデータとは、ＸＭＬ文書のような根付順序木に限定されず、例えば閉路を有するような、より一般的な無向・有向グラフのことをいう。

【００２９】

図２（ａ）はＸＭＬ文書２０のテキスト表現の一例を示す図である。ＸＭＬ文書２０は既知の手法（例えばhttp://xml.apache.org/xerces2-j/ において示されるXercesなど）により、図２（ｂ）に示すようなコンピュータ内部に適したデータ構造２１に変換可能である。そこで、以下では、ＸＭＬ文書２０を図２（ｂ）に示すデータ構造２１に変換した後の圧縮工程について説明する。このデータ構造２１は、型と値をそれぞれ有する複数の頂点と、その頂点間の参照情報とを有している。

【００３０】

一方、図１に示した分離手段１０３は、図２（ｂ）に示すデータ構造２１を入力データ１０６として入力し、入力データ１０６を各頂点間の参照情報を有する相互参照関係デー

タと、型と値を有する複数の頂点からなる頂点群とに分離する。すなわち、分離手段１０３は、各頂点を一意に識別可能なＩＤ（頂点ＩＤ）を各頂点９０１～９２３に順に割り振ったものを相互参照関係データ９００とし、割り振った頂点ＩＤと、対応する頂点９０１～９２３がもともと有していた型および値との組を列記してテーブル９５０を生成し、そのテーブル９５０を型と値を持つ頂点９０１～９２３の集合（頂点群）とすることによって、データ構造２１を分離している。各頂点９０１～９２３への頂点ＩＤの割り振り方には、幅優先探索や、深さ優先探索等があるが、ここでは幅優先探索を用いている。また、分離手段１０３は、分離して得られる頂点群のデータを第２の出力データ１０８として出力する。

【００３１】

そして、ＸＭＬ文書２０から分離された相互参照関係データ９００、およびテーブル９５０はそれぞれ、図３（ａ），図３（ｂ）に示す通りである。ここで、図３（ｂ）は、頂点ＩＤ９５０ａ、型９５０ｂおよび値９５０ｃを一行とする形式で表現されている。

テンプレート蓄積手段１０２は、圧縮に先立ちあらかじめテンプレートとテンプレート実体を蓄積している。このとき、テンプレート蓄積手段１０２は、テンプレートやテンプレート実体として、例えば、あらかじめ高い頻度で適用されることが分かっている高頻度のものを蓄積している。このようなテンプレートとテンプレート実体としては、それぞれ例えば、図４（ａ），図４（ｂ）にそれぞれ示すテンプレート１００３とテンプレート実体１０２０とがある。

【００３２】

テンプレート１００３は、テンプレートＩＤ１００１、接続情報１００２およびパターン情報１００４を有している。テンプレートＩＤ１００１はテンプレート蓄積手段１０２に複数のテンプレートが蓄積された際に、その各テンプレートを一意に識別するために用いられる。パターン情報１００４はテンプレートによって表現される頂点間の参照情報のパターンを表し、複数の頂点とそれら相互の参照情報とを有している。パターン情報１００４に含まれる参照情報には、親参照、子参照、次兄弟参照、前兄弟参照の４つの参照が設けられている。なお、接続先の頂点がない参照のうち、テンプレートを適用し、後述するようにして、相互参照関係データ９００を圧縮する際に利用されない参照については、その旨がパターン情報１００４に記述されている。これは、例えば、無効な頂点を定義しておき、その頂点への参照とすることで実現可能である。接続情報１００２には、テンプレート１００３を適用して、相互参照関係データ９００を圧縮する際に他のテンプレートや頂点との接続を示す接続参照が列挙されている。

【００３３】

テンプレート１００３は、接続参照を有する接続情報１００２と、参照情報を有するパターン情報１００４とを区別して構成しているから、異なるテンプレート１００３同士でパターン情報１００４を共有することができる。つまり、接続情報１００２を異ならせることにより、接続され得る頂点や他のテンプレートを異ならせ、パターン情報１００４が同じでも、別テンプレートのようにして利用することができる。すると、テンプレート内に含まれる頂点間の参照情報が省略可能となり、テンプレート蓄積手段１０２のメモリ使用量（記憶領域）を効率よく利用することが可能となる。

【００３４】

図４（ｃ）は、テンプレート１００３の具体的な一例となる第１のテンプレート１０１１を示す図である。図４（ｃ）に示す第１のテンプレート１０１１におけるテンプレートＩＤ１０１２には、“１”が設定されている。第１のテンプレート１０１１におけるパターン情報１０１９は、５つの頂点１０１４～１０１８と、それらの間の参照とにより構成され、参照は矢印で記述されている。なお、参照の種類は矢印に対し、親参照はｐ、子参照はｃ、次兄弟参照はｎｓ、前兄弟参照はｐｓとして記述されている。例えば、頂点１０１４の子参照ｃは頂点１０１６を指定しており、頂点１０１６の次兄弟参照は頂点１０１７を指定している。

【００３５】

また、テンプレートを適用し相互参照関係データ９００を圧縮する際に利用されないことを示す参照は端点を「×」で記述し、テンプレートを適用し相互参照関係データ９００を圧縮する際に他のテンプレートや頂点と接続されることを示す参照は端点を「○」で記述している。後者に該当する４つの参照、すなわち、頂点１０１４の親参照、頂点１０１６、頂点１０１７、頂点１０１８の子参照については、接続情報１０１３に頂点ＩＤと参照の種類が列挙されている。

【００３６】

テンプレート実体１０２０は、入力データ１０６に対し、相互参照関係データ９００を圧縮する際に、テンプレートを適用したこと（テンプレート適用済み）を表すために用いられる。このテンプレート実体１０２０は、テンプレート実体ＩＤ１００５と、反転フラグ１００６と、テンプレート独自情報１００７とを有している。また、テンプレート独自情報１００７は適用するテンプレートを表す利用テンプレートＩＤ１００８と、実体接続情報１００９および実体情報１０１０とを有している。圧縮後の相互参照関係データにおいて、テンプレート実体１０２０を参照することにより、テンプレートの適用内容が把握できるようになる。

【００３７】

テンプレート実体ＩＤ１００５は、テンプレートを適用して相互参照関係データ９００を圧縮した際、そのテンプレートの適用箇所を一意に特定するために用いられる。反転フラグ１００６は、テンプレートを適用して相互参照関係データ９００を圧縮した際に兄弟関係の方向を反転して利用するか否かを示す。本実施の形態では、反転フラグ１００６が「偽」のときに次兄弟参照と、前兄弟参照は文字通りの意味を持ち、反転フラグ１００６が「真」のときに次兄弟参照が前兄弟参照の意味を、前兄弟参照が次兄弟参照の意味を持つものとしている。

【００３８】

テンプレート独自情報１００７に含まれる実体情報１０１０は、テンプレートを適用して相互参照関係データ９００を圧縮した際に接続する先の頂点を列挙するために設けられている。この実体接続情報１０１０には、テンプレートを適用して相互参照関係データ９００を圧縮した際にテンプレート内に内包される頂点のＩＤが蓄積されている。実体接続情報１００９については後述する。なお、反転フラグ１００６は、同じテンプレートを反転させて利用する場合には必要であるが、そのようなテンプレートの適用を行わない場合は設けなくてもよい。

【００３９】

次に、テンプレート一致箇所検出手段１０４は、入力データ１０６から分離手段１０３により分離された相互参照関係データ９００から、テンプレート蓄積手段１０２に蓄積されているテンプレートに一致する箇所を検出する。テンプレート蓄積手段１０２には、複数のテンプレートが蓄積されることが予想されるため、テンプレート一致箇所検出手段１０４による検出結果は複数通り存在すると考えられる。ただし、例えば後述する図８に示す一致箇所検出手順によれば、検出結果は一意に定まる。

【００４０】

本実施の形態では、テンプレート蓄積手段１０２に第１のテンプレート１０１１のみが蓄積されているときに、相互参照関係データ９００に対して、図８に示す手順により求めた一致箇所を示す。そのテンプレートの一致箇所１５０１は、例えば図９（ａ）に示すように、利用テンプレートＩＤ１５０２、反転フラグ１５０３およびテンプレートの頂点から元の相互参照関係データ９００の頂点への割り当てを表す頂点対応情報１５０４によって表すことができる。

【００４１】

そして、テンプレート一致箇所検出手段１０４により、相互参照関係データ９００からテンプレート蓄積手段１０２に蓄積されているテンプレートの一致箇所を検出した結果を表すテンプレート一致箇所情報１５０５は図９（ｂ）に示す通りである。このテンプレート一致箇所情報１５０５は、相互参照関係データ９００から、テンプレート蓄積手段１０

２に第１のテンプレート１０１１のみが蓄積されているとき、図８に示す手順にて検出した結果であり、ここでは、第１、第２、第３の一致箇所１５０６，１５０７，１５０８を有し、一致箇所が３箇所あったことを示している。

【００４２】

第１の一致箇所１５０６及び第２の一致箇所１５０７は第１のテンプレート１０１１を反転せず、テンプレートの頂点と元の相互参照関係データ９００の頂点をそれぞれ頂点対応情報１５０９、頂点対応情報１５１０に示すように対応させることで一致することを表している。また、第３の一致箇所１５０８は第１のテンプレート１０１１を反転し、テンプレートの頂点と元の相互参照関係データ９００の頂点を頂点対応情報１５１１に示すように対応させることで一致することを表している。

そして、テンプレート一致箇所検出手段１０４は、このようなテンプレート一致箇所情報１５０５をテンプレート置換手段１０５に伝達（入力）する。

【００４３】

テンプレート置換手段１０５は、テンプレート一致箇所検出手段１０４からテンプレート一致箇所情報１５０５を入力し、そのテンプレート一致箇所情報１５０５を用いて、元の相互参照関係データ９００に対してテンプレートを適用し、テンプレート実体１０２０を用いて、頂点間の参照情報を参照可能な状態のまま置換し、置換した結果を第１の出力データ１０７として出力する。

テンプレートを用いて元の相互参照関係データ９００を置換する置換手順は、図１０に示すとおりで、また、置換した結果は図５に示す相互参照関係データ１１００のようになる。この相互参照関係データ１１００は圧縮済みの相互参照関係データ（以下「圧縮済み参照データ」ともいう）である。

【００４４】

図１０は、置換手順を示す図である。置換手順は、処理開始後、ステップ１に進みｉに０をセットして、ステップ２に進み、テンプレート一致箇所情報（テンプレート一致箇所情報１５０５）に含まれるすべての一致箇所（上述の場合は第１、第２、第３の一致箇所１５０６，１５０７，１５０８）について、それぞれ１つづつステップ３以下の処理を繰り返す。

選択した一致箇所はＭｉとする。

ステップ３では、テンプレート実体を１つ作成し、実体ＩＤ＝ｉとする。このテンプレート実体をＯｉとし，以下の処理を行う。

利用テンプレートＩＤ、反転フラグＭｉの利用、テンプレートＩＤ、反転フラグよりそれぞれ複製する。

実体情報をＭｉの頂点対応情報より複製する。

実体接続情報は実体情報に記述された対応関係より、元の参照をそのまま代入する。

次に、ステップ４に進んでｉ＝ｉ＋１を計算する、

続くステップ５では、作成済のテンプレート実体を１つずつ選択し、以下の処理を繰り返す。選択したテンプレート実体をＯｉとする。

次に、ステップ６に進み、実体接続情報に記述された参照の接続先頂点が他のテンプレート実体に含まれる場合はテンプレート実体ＩＤとテンプレートの頂点の組に置換する。

【００４５】

図５において、テンプレート実体として、３つのテンプレート実体１１２４、テンプレート実体１１３３、テンプレート実体１１４０が存在している。各テンプレート実体１１２４、テンプレート実体１１３３、テンプレート実体１１４０はそれぞれ実体ＩＤ１１２５、実体ＩＤ１１３１、実体ＩＤ１１３８を有し、そのそれぞれが、“１”，“２”，“３”の値を持っていることで識別可能になっている。各テンプレート実体は、すべて利用テンプレートＩＤ１１２８，１１３５，１１４２に“１”が設定され、反転フラグ１１２６，１１３２，１１３９は、前２者が偽、後者が真の値を有している。これにより、テンプレート実体１１２４，１１３３は図４（ｃ）に示す第１のテンプレート１０１１がそのまま適用され、テンプレート実体１１４０は第１のテンプレート１０１１が反転して適用

されたことを表す。

【００４６】

図５における各テンプレート実体１１２４，１１３３，１１４０の実体情報は、テンプレートが内包する頂点とテンプレート適用前の相互参照関係データの頂点との対応を示す情報が設定されている。そのため、置換された相互参照関係データ１１００では、頂点間の参照情報が残り、これらが参照可能になっている。例えば、テンプレート実体１１２４の場合、実体情報１１３０には、第１のテンプレート１０１１の頂点１０１４，１０１５，１０１６，１０１７および１０１８がそれぞれ図３（ａ）に示す圧縮前の相互参照データ９００の頂点９０２，９０３，９０４，９０５，９０６にそれぞれ一致することを表す情報が設定されている。

また、各テンプレート実体の実体接続情報には、他のテンプレート実体や頂点との接続関係を示す情報が設定されている。各テンプレート実体が適用しているテンプレートは第１のテンプレート１０１１であるが、第１のテンプレート１０１１は外部と接続できる参照を４つ保持していることがその接続情報１０１３に記述されている。

【００４７】

そこで、各テンプレート実体の実体接続情報には、これらの参照先がどの頂点となるのかを記述する。例えばテンプレート実体１１２４の場合、実体接続情報１１２９には、頂点１０１４の親参照は頂点９０７へ、頂点１０１６、頂点１０１７の子参照はそれぞれテンプレート実体ＩＤが“２”の頂点１０１４、テンプレート実体ＩＤが“３”の頂点１０１４へ、頂点１０１８の子参照はどの頂点にも接続しないことを示す情報が設定されている。

【００４８】

図３に示すように、テンプレート適用前の相互参照関係データ９００は、各頂点９０２～９１７がそれぞれ４つの参照を持っていたが、図５に示すテンプレート適用後の相互参照関係データ１１００は、各頂点間の参照情報を持たないテンプレート実体により置換されている。このような置換を行うことにより、テンプレートの圧縮が可能となっている。また、適用するテンプレートも、従来技術とは以下のような相違がある。つまり、テンプレート一致箇所１５０８は、第１のテンプレート１０１１と一致しなかったため、従来技術では、テンプレート一致箇所１５０８のための別なテンプレートが必要であったが、本実施の形態では、上述したように、第１のテンプレート１０１１を反転させてテンプレート一致箇所１５０８に一致させることができるから、テンプレート一致箇所１５０８のための別なテンプレートを設ける必要がない。そのため、テンプレート蓄積手段１０２のメモリ使用量（記憶領域）を効率よく利用することが可能となる。

【００４９】

なお、本実施の形態では、図３（ｂ）に示す入力データから分離された型と値を持つ頂点の集合である第２の出力データ１０８については、分離手段１０３により分離された後に出力されるだけで圧縮については何ら触れられていない。第２の出力データ１０８は、例えば、非特許文献１に示される方法等と組み合わせることにより圧縮することが可能である。

また、本実施の形態では、反転フラグは各テンプレート実体に１つずつ用意しているが相互参照関係データ全体で１つとしてもよいし、また両方を設定してもよい。

【００５０】

一方、上述した本実施の形態で示すデータ圧縮装置１０１を複数種類の入力データに適用する場合、テンプレートについてはその複数種類の入力データ間で共用する共用テンプレートとすることができる。その共用テンプレートは、複数種類の入力データに適用可能であるから、それぞれの入力データに対応してテンプレートを設けることを要しない。したがって、テンプレート蓄積手段１０２のメモリ使用量の効率化が可能である。

【００５１】

例えば、図１１（ａ）に示すＸＭＬ文書３０の場合、図８に示す一致箇所検出手順により、図２に示すＸＭＬ文書２０と同様、図４（ｃ）に示す第１のテンプレート１０１１を

適用することができる。そこで、テンプレート蓄積手段１０２は双方のＸＭＬ文書２０，３０に適用したテンプレートを区別せずに同一の共用テンプレートとして蓄積することにより、テンプレート蓄積手段１０２におけるメモリ使用量の効率化（メモリ利用効率の向上）を図ることができる。

【００５２】

なお、図８に示す一致箇所検出手順は以下のとおりである。

処理開始後ステップ１１で、パターン蓄積手段に蓄積されたパターンから、頂点の数が多い順に１つずつ選択し、以下の処理を繰り返す。

選択したパターンをＰｊとする。

次にステップ１２に進んで、反転フラグの値を偽、真のそれぞれに対し、以下を繰り返す。

続くステップ１３では、相互参照関係データに含まれる頂点から、選択したパターンＰの頂点の数と一致する頂点を選択する組み合わせをＸ１，Ｘ２，Ｘｍとし、その中から１つずつ選択して、以下を繰り返す。

選択した組み合わせをＸｋとする。

次に、ステップ１４に進み、Ｘｋに含まれる頂点はすべて置換済みマークが無いか否かを判断し、無ければステップ１５に進み、そうでなければ処理を終了する。ステップ１５に進むと、ＸｋがＰｊと同型か否かを判断し、同型であればステップ１６に進み、そうでなければ処理を終了する。ステップ１６に進むと、Ｘｋを一致箇所として登録し、Ｘｋに含まれる頂点は置換済みとしてマークする。

【００５３】

（第２の実施の形態）

図６（ａ）に示すような連続する複数の兄弟参照を有する相互参照関係データ１２０４を圧縮する場合について説明する。データ圧縮装置１０１によれば、この相互参照関係データ１２０４は、図６（ａ）に示す連続兄弟参照部１２００を図６（ｂ）に示す第２のテンプレート１２０１を用いて圧縮する。図６（ｂ）に示す第２のテンプレート１２０１はテンプレートＩＤ１２０２と、接続情報１２０３を有するが、第１のテンプレート１１０１とは異なり、パターン情報を有していない。この第２のテンプレート１２０１は、相互参照関係データ１２０４のような連続兄弟参照部を有する相互参照関係データを圧縮するために設けた連続兄弟参照用テンプレートである。

【００５４】

ここで、相互参照関係データ１２０４では、図６（ａ）に示すように、連続兄弟参照部１２００が、第１、第２、第３から第ＮまでのＮ個の頂点２ａ，２ｂ，２ｃ，・・・２ｎを有し、第１の頂点２ａと第２の頂点２ｂ以外の連続番号を有する各頂点は必ず相互に参照し、第１の頂点２ａが第２の頂点２ｂを参照し、かつ連続兄弟参照部１２００の外部にある頂点２ｐへの参照を保持している。また、第Ｎの頂点２ｎは、図示しない第Ｎ－１の頂点を参照し、さらに、第２から第Ｎまでの各頂点が、連続兄弟参照部１２００の外部への参照をまったく保持しないようになっている（または、例えば、図３における頂点９１２，９１３のように同数の参照を保持するようになっていてもよい）。

このような相互参照関係データ１２０４を圧縮するには、少なくとも、連続兄弟参照部１２００を構成する頂点の個数と、第１の頂点の外部への参照がわかればよいので、第２のテンプレート１２０１における接続情報１２０３には、連続兄弟参照部を構成する頂点の個数Ｎと、テンプレートの親参照ｐが設定されている。なお、テンプレートＩＤ１２０２は“２”を有している。

【００５５】

図７は、図６（ｂ）に示す第２のテンプレート１２０１を用いて、相互参照関係データ１２０４を圧縮した後の相互参照関係データ１３００を示す図である。この場合の圧縮では、テンプレートにおける一致箇所の検出やテンプレートの適用は第１の実施の形態と同様にすることで可能である。

図７において、相互参照関係データ１３００には、テンプレート実体として、テンプレ

ート実体１３１８，１３２３，１３２９の３つが存在し、実体ＩＤ１３１７，実体ＩＤ１３２２，実体ＩＤ１３２８は、それぞれ“１”，“２”，“３”の値を有している。各テンプレート実体は、すべて利用テンプレートＩＤ１３２６，１３３４，１３３３を有していて、いずれも“２”が設定されている（これは、第２のテンプレート１２０１を用いて圧縮したことを意味している）。なお、本実施の形態では、反転フラグを用いないため各テンプレート実体１３１８，１３２３，１３２９には反転フラグが設けられていない。

【００５６】

各テンプレート実体１３１８，１３２３，１３２９の実体接続情報１３２０，１３２５，１３３１は連続兄弟参照部を構成する頂点の個数Ｎ（それぞれ、Ｎ＝４，３，２）と、各テンプレート実体の親参照ｐを記録している。例えばテンプレート実体１３１８の場合では、Ｎ＝４、ｐは４０１になっている。各テンプレート実体の実体情報１３２１，１３２７，１３３２には、テンプレートにより内包される頂点を示す情報が示されている。例えば、テンプレート実体１３１８の場合は、実体接続情報１３２０より示されている４つの頂点、すなわち、頂点４０２、４０３、４０４，４０５を示している。

以上のように相互参照関係データ１２０４は第２のテンプレート１２０１を用いて圧縮可能である。その圧縮に用いる第２のテンプレート１２０１はパターン情報を有していないため、第２のテンプレート１２０１を記憶するのに必要な記憶容量が少なくて済む。そのため、各テンプレート蓄積手段１０２のメモリ使用量を削減することが可能である。

【００５７】

データ復元装置の実施の形態

次に、データ復元装置１８０１について、図１２を用いて説明する。図１２は本実施の形態に係るデータ復元装置１８０１の構成を示すブロック図である。このデータ復元装置１８０１は、図５に示すような圧縮後の相互参照関係データ（圧縮済み参照データ）１１００と、図３（ｂ）に示す型と値を有する複数の頂点からなる頂点群のデータとから、それぞれが型と値を有する複数の頂点と、頂点間の参照情報とを有する元の入力データを復元する。データ復元装置１８０１は、テンプレート蓄積手段１８０２と、合成手段１８０３と、テンプレート展開手段１８０４とを有している。

【００５８】

なお、図１２において、データ復元装置１８０１は各手段が一体化された単一の装置として構成されているが、必ずしも単一の装置として実現される必要はなく、複数の装置を図示しない通信手段により接続して実現することも可能である。例えば、データ復元装置１８０１からテンプレート蓄積手段１８０２を分離した上で、テンプレート蓄積手段１８０２を別な単一装置として実現し、両装置間を図示しない通信手段により接続する構成にしてもよい。そうすると、テンプレート蓄積手段１８０２を有しない複数のデータ圧縮装置間でテンプレート蓄積手段１８０２を共有するといったことも可能になる。

テンプレート展開手段１８０４は第１の入力データ１８０６として与えられた圧縮後の相互参照関係データをテンプレート蓄積手段１８０２に蓄積されたテンプレートを用いて展開する。その展開は、例えば上述した図１３に示した復元手順で行うことができる。テンプレート展開手段１８０４により復元された相互参照関係データは、図３に示す相互参照関係データ９００のようになる。

【００５９】

合成手段１８０３は展開された相互参照関係データと、図３（ｂ）に示す第２の入力データとして与えられた複数の型と値を有する頂点群のデータとを合成し、合成されたデータを出力データ１８０５として出力する。その合成は、図３（ｂ）に示すテーブル９５０における型と値を持つ頂点群において、各頂点に頂点ＩＤが割り振ってあるため、頂点ＩＤが一致する相互参照関係データの頂点に、型と値をあてはめていくことによって行う。

以上のような手順により、それぞれが型と値を有する元の複数の頂点と、頂点間の参照情報からなる入力データを復元することが可能である。

復元手順は以下のとおりである。

図１３において、開始後のステップ２１で、圧縮済みの相互参照関係データに含まれる

すべてのテンプレート実体をＸ１，Ｘ２，Ｘｎとし、すべてについて以下を行う。
選択したテンプレート実体をＸｉとする。
次にステップ２２に進み、テンプレート実体Ｘｉが利用するテンプレートが持つ頂点間の参照情報を複製し、テンプレート実体Ｘｉの実体情報に記述される頂点のＩＤを割り振る。
次いでステップ２３に進み、テンプレート実体Ｘｉの実体接続情報に記述された頂点が他のテンプレート実体Ｘｍに含まれる頂点の場合、テンプレート実体Ｘｍに記述される頂点ＩＤで置換する。

【００６０】

データ管理装置の実施の形態

本発明によるデータ圧縮装置により、相互参照関係データを圧縮すると、その圧縮後のデータへのアクセス速度の若干の低下が見込まれる。そのため、データの利用頻度を観測しておいて、その時々で利用頻度の高いものは非圧縮とし、いったん圧縮した相互参照関係データについても、利用頻度が高くなれば非圧縮の状態に戻し、逆に利用頻度が低くなれば再度圧縮する、といった方法でデータ管理を行うことが好ましい。このようなデータ管理を行うデータ管理装置を設ければ、装置全体の高速化と省メモリ化を両立させることも可能である。

【００６１】

図１４は、このようなデータ管理を行えるデータ管理装置２０００の構成を示すブロック図である。データ管理装置２０００は、第１のデータ蓄積手段２００１と、第２のデータ蓄積手段２００４と、データ圧縮装置２００２と、データ復元装置２００３とを有している。また、データ管理装置２０００は、利用頻度観測手段２００６と、制御手段２００５と、選択手段２００８とを有している。

【００６２】

なお、本実施の形態におけるデータ管理装置２０００は、各装置が一体化された単一の装置とされているが、本発明によるデータ管理装置は、必ずしも単一の装置として実現される必要はなく、各装置を図示しない通信手段により接続して実現することもできる。例えば、データ管理装置２０００より、第１のデータ蓄積手段２００１を分離し、データ圧縮装置２００２から、後述のテンプレート蓄積手段１０２を取り除いた上で、第１のデータ蓄積手段２００１を別な単一装置として実現し、両装置間を通信手段により接続する構成をとることができる。そうすると、第１のデータ蓄積手段２００１を有しない複数のデータ管理装置間で第１のデータ蓄積手段２００１を共有するといったことも可能になる。また、その他の構成手段についても同様である。

【００６３】

第１のデータ蓄積手段２００１は、圧縮可能なデータとして、圧縮前のコンピュータに適した形式のデータ（例えば図２０に示した相互参照関係データ４００等）を蓄積している。第２のデータ蓄積手段２００４は、圧縮されたデータ（例えば、図５に示す相互参照関係データ１１００等）を蓄積している。ここで、データ圧縮装置２００２は上述した本発明によるデータ圧縮装置１０１と同じ構成を有し、データ復元装置２００３は上述したデータ復元装置１８０１と同じ構成を有している。

【００６４】

制御手段２００５は第１のデータ蓄積手段２００１、第２のデータ蓄積手段２００４、データ圧縮装置２００２、データ復元装置２００３、選択手段２００８をシステム外部から入力されるデータ指定２００９に基づいて制御する。この制御手段２００５は、利用頻度観測手段２００６からの移動要求にしたがい、データ指定２００９の指定に対応するデータを移動させるための制御を行う。選択手段２００８は、制御手段２００５の指示にしたがい、第１のデータ蓄積手段２００１と第２のデータ蓄積手段２００３のいずれかから蓄積されているデータを取得して出力する。

【００６５】

利用頻度観測手段２００６は、第１のデータ蓄積手段２００１または第２のデータ蓄積

手段２００４に蓄積されているデータ（相互参照関係データ）の利用頻度を観測し、観測した利用頻度に応じて後述する移動要求を出力する。この利用頻度観測手段２００６は、利用履歴リスト２００７を内部に保持している。この利用履歴リスト２００７には、例えば、利用要求のあったデータの識別ＩＤを利用要求のあった順に複数個（Ｎ個）線形リスト（図示せず）として保存している。

本実施の形態におけるデータ管理装置２０００は、内部に蓄積するデータを一意に識別するための識別ＩＤを各データに割り振っている。データ指定２００９は、そのための識別ＩＤを外部から入力する手段である。

【００６６】

以降、実際の動作内容について説明する。

データ管理装置２０００の場合、データは第１、第２いずれかのデータ蓄積手段２００１、２００４に保存されているが、初期状態では、第１データ蓄積手段２００１、第２のデータ蓄積手段２００４のいずれにデータを蓄積しておいてもよい。以下の説明では、すべて第２のデータ蓄積手段２００４に蓄積しておくことを想定している。

制御手段２００５は、データ指定２００９により、外部から識別ＩＤが入力されると、それを受けて利用頻度観測手段２００６に指示を入力する。利用頻度観測手段２００６は制御手段２００５の指示を受けて、該当するデータが第１のデータ蓄積手段２００１と第２のデータ蓄積手段２００４のいずれに蓄積されているか、および、両手段の間でのデータの移動があるか否かを通知する情報を制御手段２００５に入力する。

【００６７】

ここで、利用頻度観測手段２００６は、第１、第２のデータ蓄積手段２００１，２００４の指定について、利用履歴リスト２００７を参照し、データ指定２００９により指定されるデータがその利用履歴リスト２００７に有るか否かを判断する。そして、例えばそのデータが有れば第１のデータ蓄積手段２００１に蓄積されているとし、無ければ第２のデータ蓄積手段２００４から蓄積されている、というようにして返答する。

【００６８】

さらに、利用頻度観測手段２００６は、第１、第２のデータ蓄積手段２００１，２００４の間でデータ移動の有無については、次のようにして制御手段２００５に返答する。例えば、利用履歴リスト２００７が更新された時、その利用履歴リスト２００７に新規に載った（記録された）データは第２のデータ蓄積手段２００４から第１のデータ蓄積手段２００１へ移動したとし、利用履歴リスト２００７から外れたデータは第１のデータ蓄積手段２００１から第２のデータ蓄積手段２００４へ移動したとして返答する。

そして、制御手段２００５は利用頻度観測手段２００６からの上述した返答に基づき、第１のデータ蓄積手段２００１または第２のデータ蓄積手段２００４を制御して、記憶しているデータを出力させ、選択手段２００８により、いずれかから得たデータをデータ管理装置２０００の外部に出力する。

【００６９】

また、制御手段２００５は利用頻度観測手段２００６からデータの移動要求があったときに、データ蓄積手段２００１，２００４の間でデータを移動させるための制御を行う。例えば、第１のデータ蓄積手段２００１から第２のデータ蓄積手段２００４に移動するような移動要求があったときは、そのデータの利用頻度が低いため、第１のデータ蓄積手段２００１からデータを取り出し、そのデータをデータ圧縮装置２００２を用いて圧縮した上で、第２のデータ蓄積手段２００４に格納するように、データ移動の制御を行う。第１のデータ蓄積装置２００１からは取り出したデータを削除するように制御する。

【００７０】

逆に、第２のデータ蓄積手段２００４から第１のデータ蓄積手段２００１に移動するように、移動要求があったときは、そのデータの利用頻度が高いので、第２のデータ蓄積手段２００４からデータを取り出し、その圧縮されているデータをデータ復元装置２００３を用いて復元し、その復元されたデータを第１のデータ蓄積手段２００１に格納するように制御する。また、第２のデータ圧縮装置２００４からは取り出したデータを削除するよ

うに制御する。

【００７１】

以上のように、データ管理装置２０００によると、利用頻度観測手段２００６からの移動要求に応じて制御手段２００５がデータの移動を制御することにより、第１のデータ蓄積手段２００１と第２のデータ蓄積手段２００４との間でデータを移動させることができるから、利用頻度に応じて、相互参照関係データを圧縮済みか、非圧縮の状態に適宜変えて蓄積することができる。すると、利用頻度の高いデータに関しては、非圧縮の状態で蓄積することにより、利用速度が高いまま維持されるため、動作速度を高速にしつつ全体としてのメモリ使用量を抑えることができる。

【００７２】

（その他の実施の形態）

本発明によれば、図１５に示すようなテンプレートをテンプレート蓄積手段に格納することもできる。図１５は第３のテンプレート２１００を示す図である。図１５に示す第３のテンプレート２１００のパターン情報２１０３において、エリア２１０４，２１０５における頂点間の参照情報は、図４（ｃ）に示す第１のテンプレート１０１１のパターン情報に一致することが分かる。したがって、テンプレートやテンプレート実体を拡張し、内部に適用されたテンプレートの数や、テンプレート内部の各テンプレートを一意に識別できるＩＤ等を記述する情報を付加することにより、テンプレート内部のパターン情報に対しても、前述までのテンプレートを適用することが可能になる。

【００７３】

ここでは説明のためにテンプレート内部のパターン情報に前述までのテンプレートを適用するとき、元のテンプレートを親テンプレート、テンプレート内部のパターン情報に適用されるテンプレートを子テンプレートと呼ぶことにする。

テンプレート内部のパターン情報にも、テンプレートの適用を可能とするための親テンプレート２３００と親テンプレート実体２３２０の構成例をそれぞれ図１７（ａ），（ｂ）に示す。親テンプレート２３００と親テンプレート実体２３２０は、図４に示す第１の実施の形態で用いたテンプレート１００３及びテンプレート実体１０２０に対し、前者に内部テンプレート情報２３０１を追加し、後者に内部テンプレート実体接続情報２３０２及び内部テンプレート実体情報２３０３を追加している。

【００７４】

内部テンプレート情報２３０１には、例えば、子テンプレートの数や、子テンプレートのＩＤを記述する。内部テンプレート実体情報２３０３には、子テンプレートに内包される頂点が親テンプレートを実際に適用するとき、親テンプレート適用前の頂点とどのように対応するのかの対応関係を各頂点毎に記述する。内部テンプレート実体接続情報２３０２には、親テンプレートを実際に適用した際、子テンプレートが外部の親テンプレートと接続するときの接続情報を記述する。

【００７５】

そして、図１６に、図１７（ａ）に示すテンプレートの具体的な親テンプレート２２００を示す。この親テンプレート２２００は、内部テンプレート情報２２１６に２種類の子テンプレートを２箇所に適用していることから、ＩＤ＝１、ＩＤ＝２としている。パターン情報２２０３は、第１の実施の形態に示すように、相互参照関係データを圧縮する要領で圧縮されており、第１のテンプレート１０１１が適用されている。

このように構成した親テンプレート２２００を相互参照関係データに適用すれば、テンプレートを定義するのに他のテンプレートの定義を利用できるため、図１５に示すような第３のテンプレート２１００に比べて、テンプレートの情報量を削減することが可能である。

【図面の簡単な説明】

【００７６】

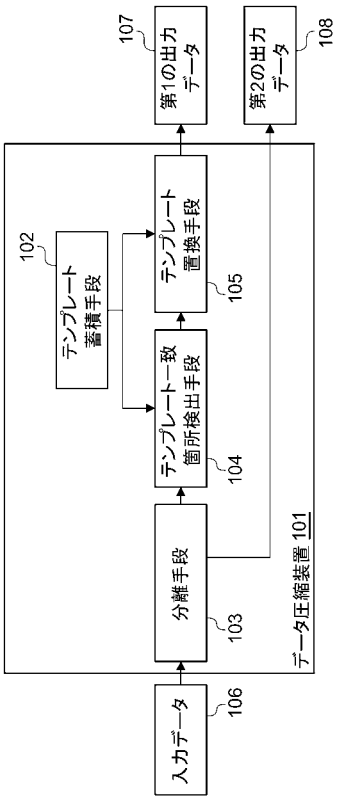
【図１】本発明の実施の形態に係るデータ圧縮装置の構成を示すブロック図である。

【図２】（ａ）はＸＭＬ文書の一例を示す図、（ｂ）は（ａ）のＸＭＬ文書のデータ構造

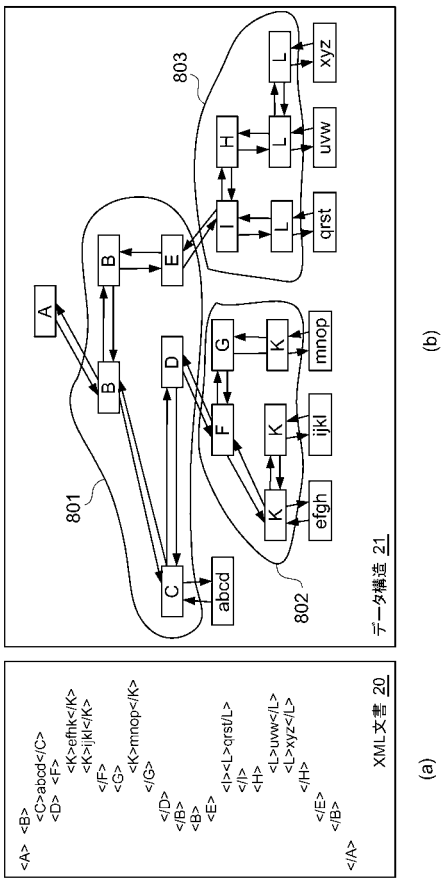
を示す図である。
【図３】（ａ）は図２のＸＭＬ文書から分離された相互参照関係データを示す図、（ｂ）は頂点の集合のテーブルを示す図である。
【図４】（ａ）はテンプレートの構成を示すブロック図、（ｂ）はテンプレート実体の構成を示すブロック図、（ｃ）は第１のテンプレートの構成を示すブロック図である。
【図５】圧縮後の相互参照関係データを示す図である。
【図６】（ａ）は別の相互参照関係データを示す図、（ｂ）は第１のテンプレートの構成を示すブロック図である。
【図７】図６における圧縮後の相互参照関係データを示す図である。
【図８】一致箇所の検出手順の一例を示す図である。
【図９】（ａ）はテンプレートの一致箇所を示すブロック図、（ｂ）はテンプレート一致箇所情報を示すブロック図である。
【図１０】置換手順の一例を示す図である。
【図１１】（ａ）は別のＸＭＬ文書を示す図、（ｂ）は（ａ）のＸＭＬ文書のデータ構造を示す図である。
【図１２】本実施の形態に係るデータ復元装置の構成を示すブロック図である。
【図１３】テンプレートから元の頂点の参照情報を復元する手順の一例を示す図である。
【図１４】データ管理装置の構成を示すブロック図である。
【図１５】第１のテンプレートの構成を示すブロック図である。
【図１６】内部にテンプレートを有する親テンプレートの一例を示すブロック図である。
【図１７】（ａ）は親テンプレートを示すブロック図、（ｂ）は親テンプレート実体を示すブロック図である。
【図１８】ＸＭＬ文書の別の一例を示す図である。
【図１９】図１８のＸＭＬ文書のデータ構造を示す図である。
【図２０】（ａ）は図１９のＸＭＬ文書から分離された相互参照関係データを示す図、（ｂ）は頂点の集合のテーブルを示す図である。
【図２１】図１８のＸＭＬ文書から分離されたデータ構造を示す図である。
【図２２】図１８のＸＭＬ文書から分離された要素名情報を示す図である。
【図２３】図１８のＸＭＬ文書から分離されたテキスト情報を示す図である。
【符号の説明】
【００７７】
２０，３０…ＸＭＬ文書
２１，３１…データ構造
１０１…データ圧縮装置
１０２，１８０２…テンプレート蓄積手段
１０３…分離手段
１０４…テンプレート一致箇所検出手段
１０５…テンプレート置換手段
１０６…入力データ、１０７…第１の出力データ
１０８…第２の出力データ
９００，１１００…相互参照関係データ
１２０４、１３００…相互参照関係データ
９０１…頂点、９５０…テーブル
１００３…テンプレート
１０２０，１１２４、１１３３…テンプレート実体
１１４０，１３２３，１３２９…テンプレート実体
１０１１…第１のテンプレート
１２０１…第２のテンプレート
１５０１…一致箇所
１５０５…テンプレート一致箇所情報

１８０１，２００３…データ復元装置
１８０３…合成手段
１８０４…テンプレート展開手段
２０００…データ管理装置
２００１…第１のデータ蓄積手段
２００２…データ圧縮装置
２００４…第２のデータ蓄積手段
２００５…制御手段、２００６…利用頻度観測手段
２００８…選択手段、２１００…第３のテンプレート
２２００、２３００…親テンプレート
２３２０…親テンプレート実体

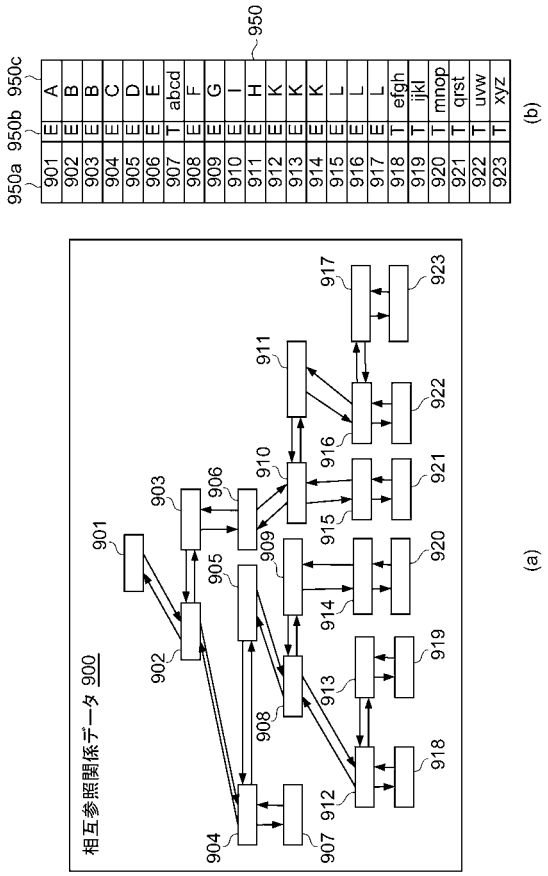
【図１】



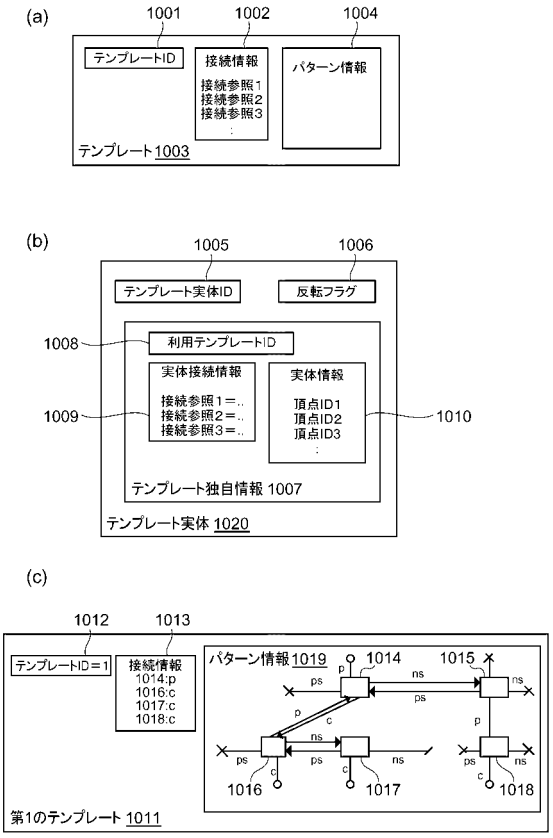
【図２】



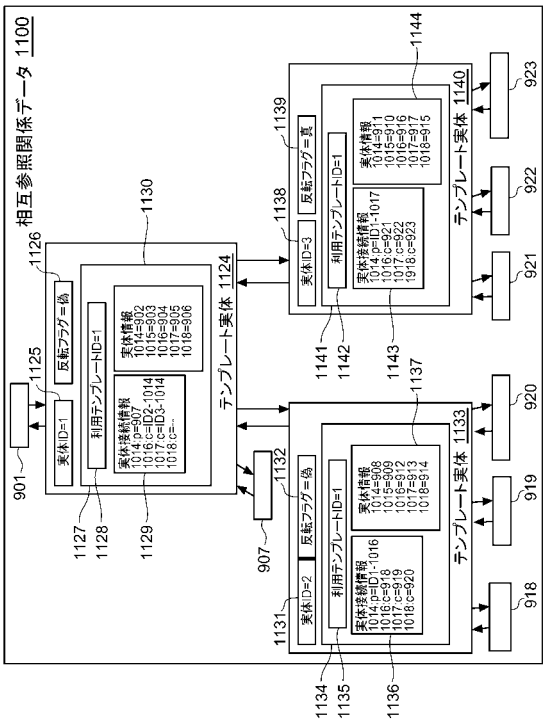
【図３】


950
950a
950b
950c
(b)
相互参照関係データ 900
(a)

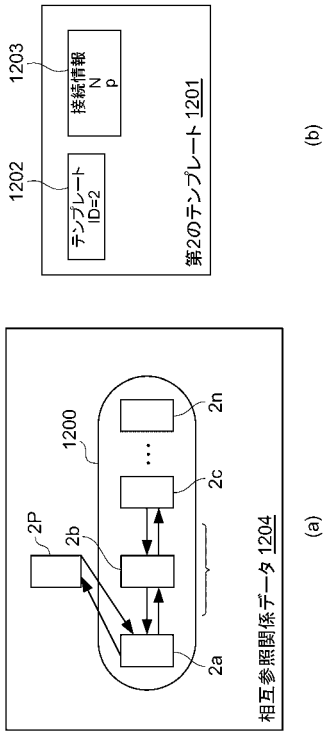
【図４】


(a)
テンプレートID
接続情報
パターン情報
テンプレート1003
(b)
テンプレート実体ID
反転フラグ
利用テンプレートID
実体接続情報
実体情報
テンプレート独自情報 1007
テンプレート実体 1020
(c)
テンプレートID＝1
接続情報
パターン情報1019
第1のテンプレート 1011

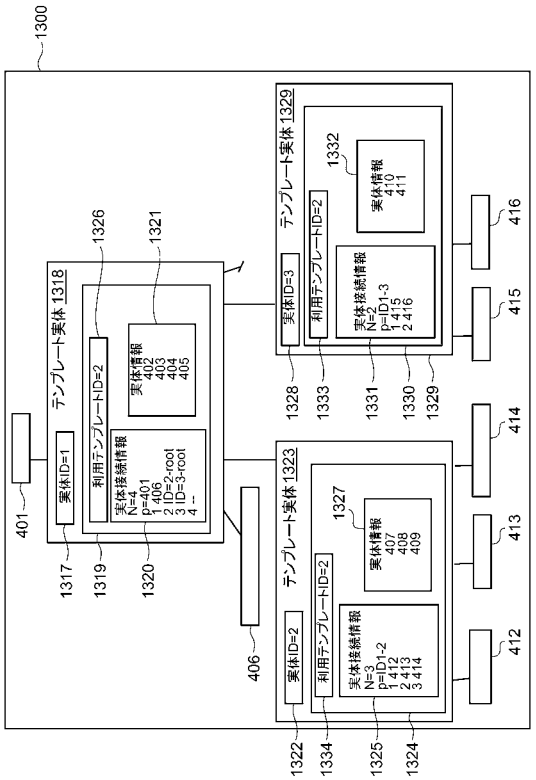
【図５】


相互参照関係データ 1100
テンプレート実体 1124
テンプレート実体 1133
テンプレート実体 1140

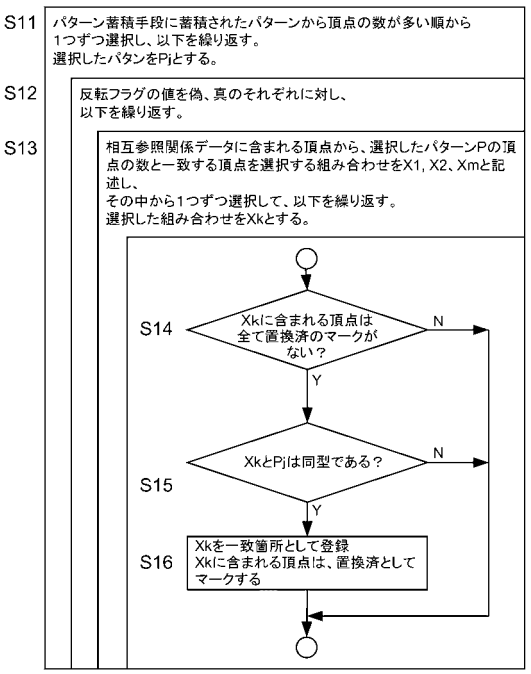
【図６】


(b)
第2のテンプレート 1201
(a)
相互参照関係データ 1204

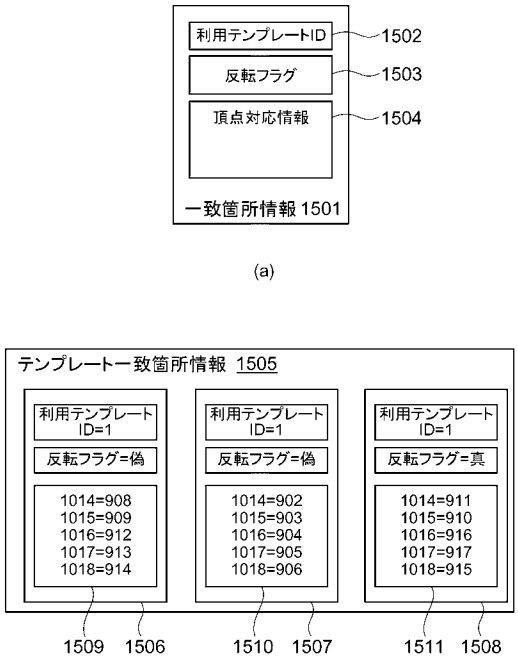
【図７】


1300
テンプレート実体 1318
実体ID=1
利用テンプレートID=2
実体接続情報
N=4
p=401
1 406
2 ID=2-root
3 ID=3-root
4 –
実体情報
402
403
404
405
テンプレート実体 1329
実体ID=3
利用テンプレートID=2
実体接続情報
N=2
p=ID1-3
1 415
2 416
実体情報
410
411
テンプレート実体 1323
実体ID=2
利用テンプレートID=2
実体接続情報
N=3
p=ID1-2
1 412
2 413
3 414
実体情報
407
408
409
401
406
412
413
414
415
416
1317
1319
1320
1321
1322
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334

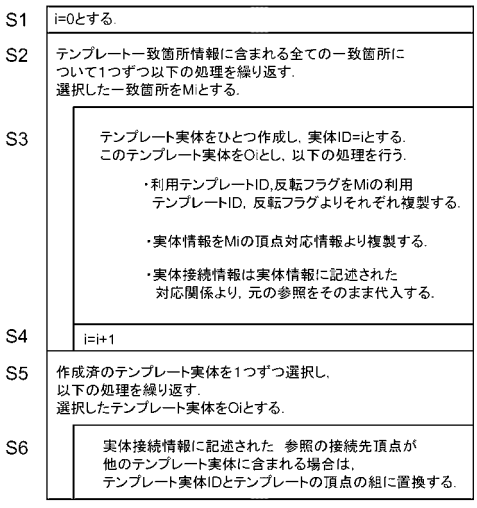
【図８】


S11
パターン蓄積手段に蓄積されたパターンから頂点の数が多い順から
1つずつ選択し、以下を繰り返す。
選択したパタンをPjとする。
S12
反転フラグの値を偽、真のそれぞれに対し、
以下を繰り返す。
S13
相互参照関係データに含まれる頂点から、選択したパターンPの頂
点の数と一致する頂点を選択する組み合わせをX1, X2、Xmと記
述し、
その中から1つずつ選択して、以下を繰り返す。
選択した組み合わせをXkとする。
S14
Xkに含まれる頂点は
全て置換済のマークが
ない？
N
Y
S15
XkとPjは同型である？
N
Y
S16
Xkを一致箇所として登録
Xkに含まれる頂点は、置換済として
マークする

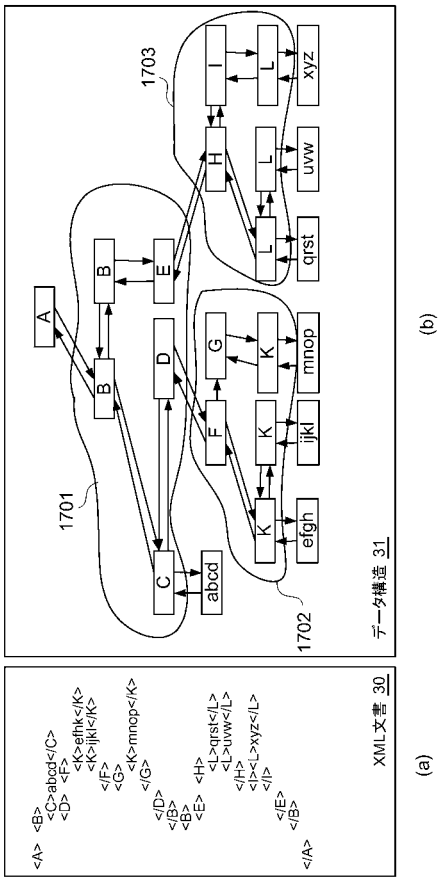
【図９】


利用テンプレートID
1502
反転フラグ
1503
頂点対応情報
1504
一致箇所情報 1501
(a)
テンプレート一致箇所情報 1505
利用テンプレート ID=1
反転フラグ=偽
1014=908
1015=909
1016=912
1017=913
1018=914
利用テンプレート ID=1
反転フラグ=偽
1014=902
1015=903
1016=904
1017=905
1018=906
利用テンプレート ID=1
反転フラグ=真
1014=911
1015=910
1016=916
1017=917
1018=915
1509
1506
1510
1507
1511
1508
(b)

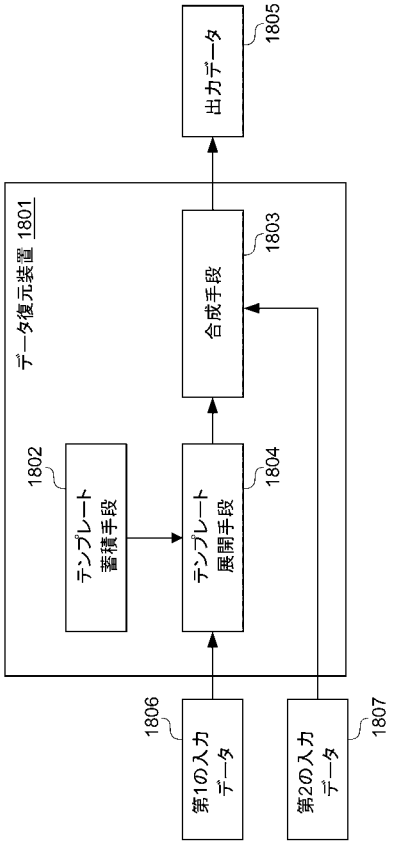
【図１０】


S1
i=0とする.
S2
テンプレート一致箇所情報に含まれる全ての一致箇所に
ついて1つずつ以下の処理を繰り返す.
選択した一致箇所をMiとする.
S3
テンプレート実体をひとつ作成し, 実体ID=iとする.
このテンプレート実体をOiとし, 以下の処理を行う.
・利用テンプレートID,反転フラグをMiの利用
テンプレートID, 反転フラグよりそれぞれ複製する.
・実体情報をMiの頂点対応情報より複製する.
・実体接続情報は実体情報に記述された
対応関係より, 元の参照をそのまま代入する.
S4
i=i+1
S5
作成済のテンプレート実体を1つずつ選択し,
以下の処理を繰り返す.
選択したテンプレート実体をOiとする.
S6
実体接続情報に記述された 参照の接続先頂点が
他のテンプレート実体に含まれる場合は,
テンプレート実体IDとテンプレートの頂点の組に置換する.

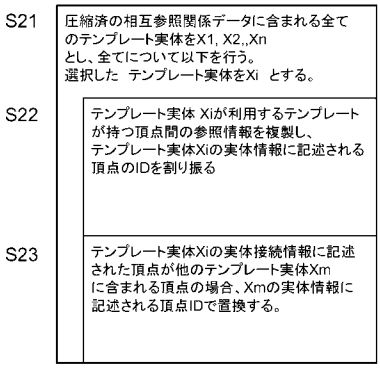
【図１１】



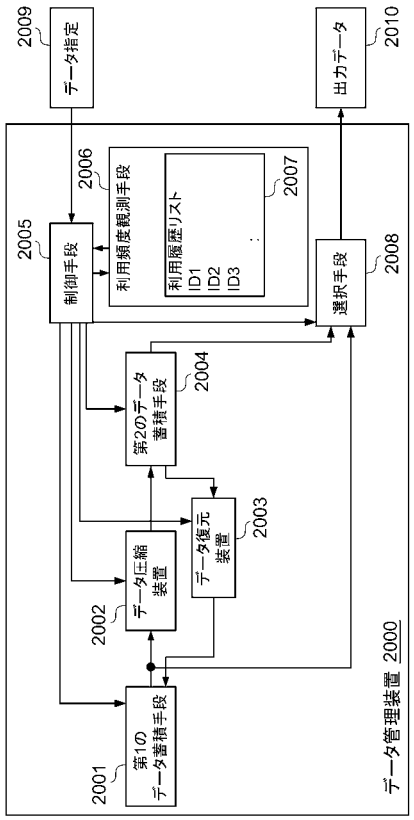
【図１２】



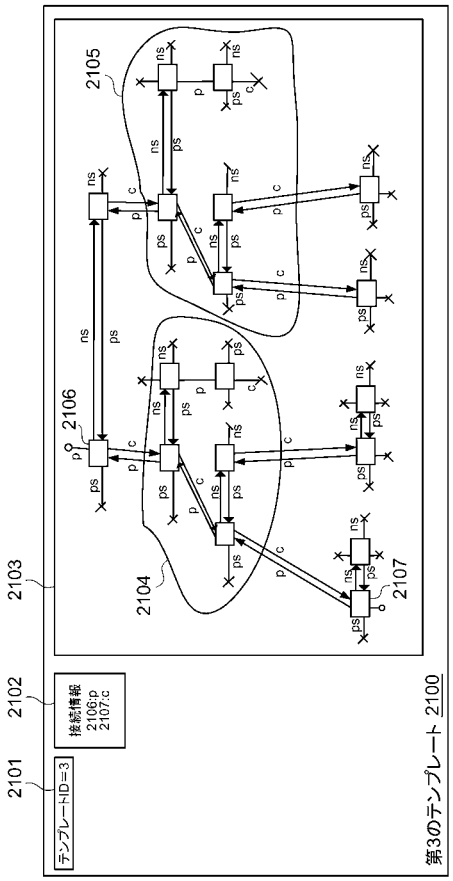
【図１３】



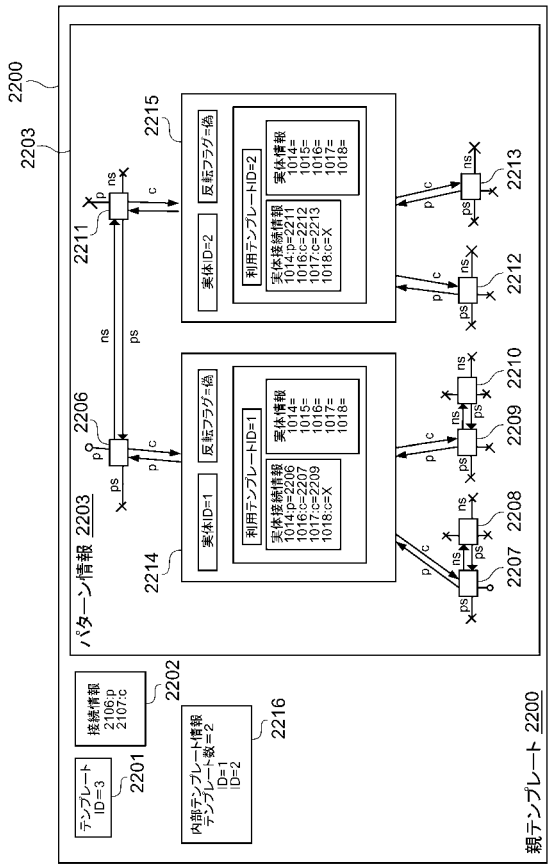
【図１４】



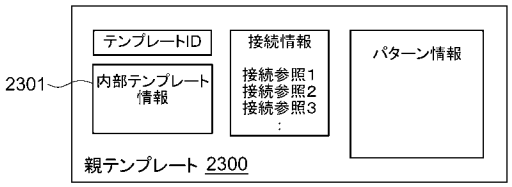
【図１５】



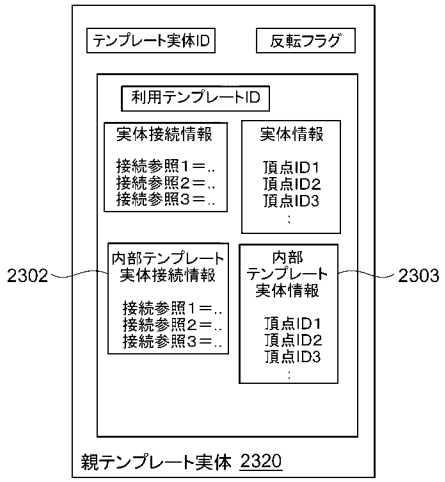
【図１６】



【図１７】



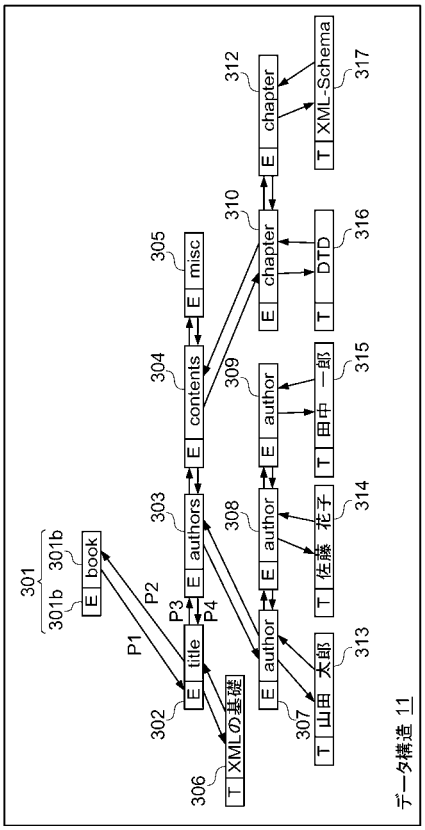
(a)



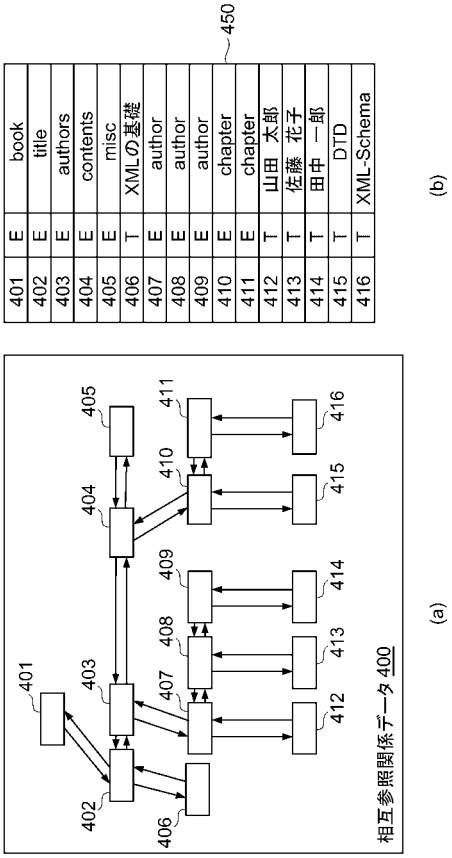
(b)

【図１８】

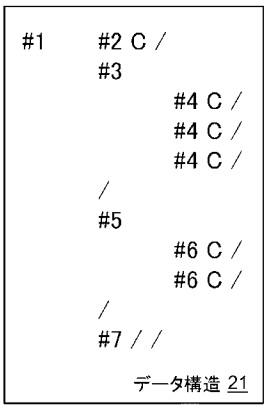
【図１９】



【図２０】



【図２１】



【図２２】

【図２３】

フロントページの続き

(72)発明者　行友　英記
東京都千代田区永田町二丁目１１番１号　株式会社エヌ・ティ・ティ・ドコモ内
(72)発明者　中山　雄大
東京都千代田区永田町二丁目１１番１号　株式会社エヌ・ティ・ティ・ドコモ内
(72)発明者　金野　晃
東京都千代田区永田町二丁目１１番１号　株式会社エヌ・ティ・ティ・ドコモ内
(72)発明者　竹下　敦
東京都千代田区永田町二丁目１１番１号　株式会社エヌ・ティ・ティ・ドコモ内

審査官　渡辺　未央子

(56)参考文献　特開２００１－２８２５１６（ＪＰ，Ａ）
特開平０９－１３０６１６（ＪＰ，Ａ）
特開２００１－２５１６１７（ＪＰ，Ａ）
特開２００２－１６３２４８（ＪＰ，Ａ）
特開２００３－０４４４５９（ＪＰ，Ａ）
特開２００４－０３２７７４（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０３Ｍ　　７／３０