

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第5494860号
(P5494860)

(45) 発行日 平成26年5月21日 (2014. 5. 21)

(24) 登録日 平成26年3月14日 (2014. 3. 14)

(51) Int. Cl.	F I
G 0 6 F 12/00 (2006.01)	G O 6 F 12/00 5 4 5 A
G 0 6 F 17/30 (2006.01)	G O 6 F 12/00 5 1 1 A
	G O 6 F 17/30 1 7 O H

請求項の数 3 (全 42 頁)

(21) 出願番号	特願2013-76465 (P2013-76465)	(73) 特許権者	000005223
(22) 出願日	平成25年4月1日 (2013. 4. 1)		富士通株式会社
(62) 分割の表示	特願2008-143527 (P2008-143527)		神奈川県川崎市中原区上小田中4丁目1番1号
原出願日	平成20年5月30日 (2008. 5. 30)	(74) 代理人	100104190
(65) 公開番号	特開2013-168163 (P2013-168163A)		弁理士 酒井 昭徳
(43) 公開日	平成25年8月29日 (2013. 8. 29)	(72) 発明者	片岡 正弘
審査請求日	平成25年4月1日 (2013. 4. 1)		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	佐藤 達博
			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	坪倉 孝
			東京都世田谷区玉川台二丁目13番2号
			株式会社ライフメディア内
			最終頁に続く

(54) 【発明の名称】 情報管理プログラム、情報管理装置および情報管理方法

(57) 【特許請求の範囲】

【請求項 1】

文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバーにアクセス可能なマスターサーバーを、

前記各スレーブサーバーから前記アーカイブを受信する受信手段、

前記受信手段によって受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成する圧縮パラメータ生成手段、

前記圧縮パラメータ生成手段によって生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成する共通ハフマン木生成手段、

前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記圧縮ファイルを割り当てる割当手段、

前記割当手段によって前記各スレーブサーバーに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通ハフマン木生成手段によって生成された共通のハフマン木とを、割当先となるスレーブサーバーに送信する送信手段、

として機能させることを特徴とする情報管理プログラム。

【請求項 2】

文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバーにアクセス可能な情報管理装置であって、

10

20

前記各スレーブサーバーから前記アーカイブを受信する受信手段と、

前記受信手段によって受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成する圧縮パラメータ生成手段と、

前記圧縮パラメータ生成手段によって生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成する共通ハフマン木生成手段と、

前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記圧縮ファイルを割り当てる割当手段と、

前記割当手段によって前記各スレーブサーバーに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通ハフマン木生成手段によって生成された共通のハフマン木とを、割当先となるスレーブサーバーに送信する送信手段と、

を備えることを特徴とする情報管理装置。

【請求項3】

文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバーにアクセス可能なマスターサーバーが、

前記各スレーブサーバーから前記アーカイブを受信する受信工程と、

前記受信工程によって受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成する圧縮パラメータ生成工程と、

前記圧縮パラメータ生成工程によって生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成する共通ハフマン木生成工程と、

前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記圧縮ファイルを割り当てる割当工程と、

前記割当工程によって前記各スレーブサーバーに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通ハフマン木生成工程によって生成された共通のハフマン木とを、割当先となるスレーブサーバーに送信する送信工程と、

を実行することを特徴とする情報管理方法。

【発明の詳細な説明】

【技術分野】

【0001】

この発明は、文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブに関する情報管理プログラム、情報管理装置および情報管理方法に関する。

【背景技術】

【0002】

従来、下記特許文献1により、検索対象ファイルを圧縮形式のまま、検索条件に該当する可能性のある圧縮ファイルを絞り込むことで、ファイルオープン回数を低減化する技術が開示されている。また一般的にファイルの検索に関して、オープン処理の回数が多くなり、且つファイル管理がセクタ単位のため、格納領域の断片化（フラグメンテーション）が発生し、格納サイズが増大するという問題について、アーカイブの適用が有効とされている。

【先行技術文献】

【特許文献】

【0003】

【特許文献1】国際公開第2006/123448号パンフレット

【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、上述のアーカイブでは、各圧縮ファイルが異なる圧縮パラメータにより

10

20

30

40

50

圧縮されているため、圧縮パラメータの演算が必要になり、総合的にオープン処理が長くなる問題があった。また、上述した特許文献およびアーカイブでは、検索対象ファイル数の増大にともない、オープン処理の比率が急増するという問題があった。特に、大規模な辞典ではオープン処理がファイル処理の全体の20～30%を占め、全文検索の速度低下の要因となっているという問題があった。また、約5000ファイルから、指定されたファイル名を特定するため、2分探索では13回の照合が必要になるという問題があった。さらに、ファイル管理がセクタ単位のため、ディスク領域の断片化（フラグメンテーション）が発生し、格納サイズが増大するという問題があった。

【0005】

一方、大容量の主記憶メモリを装備するシングルサーバのシステムにおける辞書検索では、多数のファイルに分割され、ディスク領域に分散されているため、キャッシュへの常駐化をするためには、オープンやリード処理に多大な時間を必要とし、且つキャッシュでは、格納領域の断片化（フラグメンテーション）が発生し、格納サイズが増大する問題があった。

【0006】

また、グリッドコンピューティングを利用した多くの辞書検索では、あるグリッドコンピュータの検索処理が遅れると全体の検索処理に影響を及ぼし、検索効率を低下させる原因となっていた。

【0007】

この発明は、上述した従来技術による問題点を解消するため、圧縮ファイルに対して圧縮形式のままで全文検索を実現する際に、ファイルアクセス処理の高速化を図ることにより検索処理の効率化を実現することができる情報管理プログラム、情報管理装置および情報管理方法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】

【0008】

上述した課題を解決し、目的を達成するため、文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバにアクセス可能であり、前記各スレーブサーバから前記アーカイブを受信し、受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成し、生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成し、前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバに前記圧縮ファイルを割り当て、前記各スレーブサーバに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通のハフマン木とを、割当先となるスレーブサーバに送信することを要件とする。

【0009】

このプログラム、装置および方法によれば、各スレーブサーバに対し、アクセス頻度が平準化された圧縮ファイル群を含むアーカイブを配信することができ、グリッドコンピュータ全体の検索の効率化と高速化を図ることができる。

【発明の効果】

【0010】

このプログラム、装置および方法によれば、圧縮ファイルに対して圧縮形式のままで全文検索を実現する際に、ファイルアクセス処理の高速化を図ることにより検索処理の効率化を実現することができるという効果を奏する。

【図面の簡単な説明】

【0011】

【図1】実施の形態1にかかる情報検索装置のハードウェア構成を示すブロック図である。

【図2】アーカイブの記憶内容を示す説明図である。

10

20

30

40

50

- 【図 3】 圧縮ファイルと検索対象ファイルとの関係を示す説明図である。
- 【図 4】 単字出現マップを示す説明図である。
- 【図 5】 連字出現マップを示す説明図である。
- 【図 6】 圧縮パラメータを示す説明図である。
- 【図 7】 ファイルパステーブルを示す説明図である。
- 【図 8】 文字出現マップ連携テーブルを示す説明図である。
- 【図 9】 ファイルパス連携テーブルを示す説明図である。
- 【図 10】 仮想アーカイブ容量テーブルを示す説明図である。
- 【図 11】 情報検索装置の機能的構成を示すブロック図である。
- 【図 12】 文字出現マップを用いた圧縮ファイルの絞込み例を示す説明図である。 10
- 【図 13】 情報検索装置の仮想アーカイブ設定機能が実行する仮想アーカイブ設定処理手順を示すフローチャートである。
- 【図 14】 情報検索装置の情報検索機能が実行する情報検索処理手順を示すフローチャートである。
- 【図 15】 実施の形態 2 にかかる検索システムのシステム構成図である。
- 【図 16】 アーカイブの共通化を示す説明図である。
- 【図 17】 新アーカイブの分配処理を示す説明図である。
- 【図 18】 アーカイブ 200-1 の圧縮符号テーブルおよび圧縮パラメータを示す説明図である。
- 【図 19】 アーカイブ 200-1 の圧縮符号テーブルから生成されるハフマン木である。 20
- 【図 20】 アーカイブ 200-2 の圧縮符号テーブルおよび圧縮パラメータを示す説明図である。
- 【図 21】 アーカイブ 200-2 の圧縮符号テーブルから生成されるハフマン木を示す説明図である。
- 【図 22】 統合アーカイブの圧縮符号テーブルおよび圧縮パラメータを示す説明図である。
- 【図 23】 統合アーカイブの圧縮符号テーブルから生成される共通ハフマン木を示す説明図である。
- 【図 24】 アーカイブ 200-1 の記憶内容を示す説明図である。
- 【図 25】 アーカイブ 200-1 の単字出現マップを示す説明図である。 30
- 【図 26】 アーカイブ 200-1 の連字出現マップを示す説明図である。
- 【図 27】 アーカイブ 200-1 の圧縮パラメータを示す説明図である。
- 【図 28】 アーカイブ 200-1 のファイルパステーブルを示す説明図である。
- 【図 29】 アーカイブ 200-1 の文字出現マップ連携テーブルを示す説明図である。
- 【図 30】 アーカイブ 200-1 のファイルパス連携テーブルを示す説明図である。
- 【図 31】 アーカイブ 200-2 の記憶内容を示す説明図である。
- 【図 32】 アーカイブ 200-2 の単字出現マップ Mb1 を示す説明図である。
- 【図 33】 アーカイブ 200-2 の連字出現マップ Mb2 を示す説明図である。
- 【図 34】 アーカイブ 200-2 の圧縮パラメータ P2 を示す説明図である。
- 【図 35】 アーカイブ 200-2 のファイルパステーブル 222b を示す説明図である。 40
- 【図 36】 アーカイブ 200-2 の文字出現マップ連携テーブルを示す説明図である。
- 【図 37】 アーカイブ 200-2 のファイルパス連携テーブルを示す説明図である。
- 【図 38】 共通パラメータの生成例を示す説明図である。
- 【図 39】 文字出現マップ連携テーブルの再構成を示す説明図である。
- 【図 40】 ファイルパス連携テーブルの再構成を示す説明図である。
- 【図 41】 ファイルパステーブルの再構成を示す説明図である。
- 【図 42】 単字出現マップの再構成を示す説明図である。
- 【図 43】 連字出現マップの再構成を示す説明図である。
- 【図 44】 圧縮ファイル群の再構成を示す説明図である。
- 【図 45】 新アーカイブ A1 の記憶内容を示す説明図である。 50

【図４６】新アーカイブＡ２の記憶内容を示す説明図である。

【図４７】マスターサーバー（情報管理装置）の機能的構成を示すブロック図である。

【図４８】マスターサーバーによるアーカイブ再構成処理手順（前半）を示すフローチャートである。

【図４９】マスターサーバーによるアーカイブ再構成処理手順（後半）を示すフローチャートである。

【図５０】変換用ハフマン木の生成例（その１）を示す説明図である。

【図５１】変換用ハフマン木の生成例（その２）を示す説明図である。

【図５２】変換用ハフマン木を示す説明図（その１）である。

【図５３】変換用ハフマン木を示す説明図（その２）である。

10

【図５４】実施の形態３にかかるマスターサーバー１５０１（情報管理装置）の機能的構成を示すブロック図である。

【図５５】実施の形態３におけるマスターサーバーによるアーカイブ再構成処理手順（後半）を示すフローチャートである。

【図５６】ハフマン木への圧縮符号設定処理の詳細な処理手順を示すフローチャートである。

【発明を実施するための形態】

【００１２】

以下に添付図面を参照して、このプログラム、装置および方法の好適な実施の形態を詳細に説明する。

20

【００１３】

<実施の形態１>

狭義のアーカイブとは、一般に、複数のフォルダとその配下の多くのファイルを１ファイル化する技術である。電子メールなどに添付し送受信され、データ交換などの目的に利用されており、圧縮技術との組合せが多いことから、広義では、圧縮の付属技術として紹介されることもある。インターネットの普及とともに、アーカイブ技術が進歩し、その操作性や圧縮方式と組合せ、多種・多様なツールが開発された。パソコンなどのハードウェアの進歩は目ざましく、特にＣＰＵの高速化、メモリやハードディスクや光ディスクなどの媒体の大容量化は顕著である。

【００１４】

30

ハードウェアの進歩に伴い、データの多様化や、分析・解析や調査・研究への業務適用などの変化も著しい。一方、個人情報の流出や粉飾決算などに関して、個人情報保護法などの法整備などに対応したセキュリティ機能の強化が求められている。アーカイブ技術では、これまで、圧縮・伸長の性能が重視されてきたが、今後、検索やセキュリティなどとの連携機能が重視される。

【００１５】

従来のアーカイブは、主にデータの保存や情報伝達・交換などの分野で発達した技術であり、圧縮と１ファイル化が特徴である。また、ファイルを利用する場合は、解凍（あるいは、一時的に伸長）が行われる。ＺＩＰなどのアーカイブには、全文検索の機能がない。ファイルが多くなればなるほど、検索は重要機能である。

40

【００１６】

実施の形態１では、圧縮ファイルに対して圧縮形式のまま全文検索を実現する際に、アクセス頻度の高いファイルをキャッシュメモリに常駐化することで、リードなどのファイルアクセス処理の時間の短縮化を実現する。

【００１７】

（コンピュータのハードウェア構成）

図１は、実施の形態１にかかる情報検索装置のハードウェア構成を示すブロック図である。図１において、情報検索装置は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）１０１と、ＲＯＭ（Ｒｅａｄ－Ｏｎｌｙ　Ｍｅｍｏｒｙ）１０２と、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）１０３と、磁気ディスクドライブ１０４と

50

、磁気ディスク105と、光ディスクドライブ106と、着脱可能な記録媒体の一例としての光ディスク107と、ディスプレイ108と、I/F (Interface) 109と、キーボード110と、マウス111と、スキャナ112と、プリンタ113と、を備えている。また、各構成部はバス100によってそれぞれ接続されている。

【0018】

ここで、CPU101は、コンピュータの全体の制御を司る。ROM102は、ブートプログラムなどのプログラムを記憶している。RAM103は、CPU101のワークエリアとして使用される。磁気ディスクドライブ104は、CPU101の制御にしたがって磁気ディスク105に対するデータのリード／ライトを制御する。磁気ディスク105は、磁気ディスクドライブ104の制御で書き込まれたデータを記憶する。

10

【0019】

光ディスクドライブ106は、CPU101の制御にしたがって光ディスク107に対するデータのリード／ライトを制御する。光ディスク107は、光ディスクドライブ106の制御で書き込まれたデータを記憶したり、光ディスク107に記憶されたデータをコンピュータに読み取らせたりする。

【0020】

ディスプレイ108は、カーソル、アイコンあるいはツールボックスをはじめ、文書、画像、機能情報などのデータを表示する。このディスプレイ108は、たとえば、CRT、TFT液晶ディスプレイ、プラズマディスプレイなどを採用することができる。

20

【0021】

インターフェース（以下、「I/F」と略する。）109は、通信回線を通じてLAN (Local Area Network)、WAN (Wide Area Network)、インターネットなどのネットワーク114に接続され、このネットワーク114を介して他の装置に接続される。そして、I/F109は、ネットワーク114と内部のインターフェースを司り、外部装置からのデータの入出力を制御する。I/F109には、たとえばモデムやLANアダプタなどを採用することができる。

【0022】

キーボード110は、文字、数字、各種指示などの入力のためのキーを備え、データの入力をおこなう。また、タッチパネル式の入力パッドやテンキーなどであってもよい。マウス111は、カーソルの移動や範囲選択、あるいはウィンドウの移動やサイズの変更などをおこなう。ポインティングデバイスとして同様に機能を備えるものであれば、トラックボールやジョイスティックなどであってもよい。

30

【0023】

スキャナ112は、画像を光学的に読み取り、情報検索装置内に画像データを取り込む。なお、スキャナ112は、OCR (Optical Character Reader) 機能を持たせてもよい。また、プリンタ113は、画像データや文書データを印刷する。プリンタ113には、たとえば、レーザプリンタやインクジェットプリンタを採用することができる。

【0024】

（アーカイブの記憶内容）

40

図2は、アーカイブの記憶内容を示す説明図である。アーカイブは、図1に示したRAM103、磁気ディスク105などの記憶領域に記憶されている。アーカイブ200は、書庫領域201と管理領域202とデータ領域203とを有する。書庫領域201には、文字出現マップ連携テーブル211とファイルパス連携テーブル212と仮想アーカイブ容量テーブル213とが記憶されている。管理領域202には、圧縮パラメータ221とファイルパステーブル222と文字出現マップM（単字出現マップM1および連字出現マップM2）とが記憶されている。データ領域203には、圧縮ファイル群f（圧縮ファイルf1～fn）が記憶されている。

【0025】

アーカイブ200はその記憶領域230に格納されており、先頭から圧縮ファイルf '

50

までがキャッシュ領域 240 に格納される。ここで、キャッシュ領域 240 とは、アーカイブ 200 の記憶領域 230 に対して相対的に決まる記憶領域 230 であり、アーカイブ 200 の記憶領域 230 よりも高速アクセスできる領域である。たとえば、アーカイブ 200 の記憶領域 230 が磁気ディスク 105 である場合、キャッシュ領域 240 はメインメモリヒープ領域などに設けられる。また、キャッシュ領域 240 には、一部または全部の文字出現マップやファイルパス、仮想アーカイブも記憶される。

【0026】

図 3 は、圧縮ファイル f_i と検索対象ファイルとの関係を示す説明図である。 n 個の圧縮ファイル $f_1 \sim f_n$ は共通のハフマン木で圧縮されており、そのハフマン木を用いて伸長される。伸長された検索対象ファイル群は、たとえば、辞書や用語辞典など文字列が記述されたファイル群である。このファイルは、たとえば、HTML (HyperText Markup Language) や XML (Extensible Markup Language) などコンピュータリダブルな言語で記述されたファイルである。日本語の辞書の場合、1 ファイルの文字数は 4000 文字以上であり、ファイル数 n は、 $n = 4000 \sim 6000$ 程度の膨大なファイル数となる。

【0027】

日本語の辞書を例に挙げると、(A) に示したように、ファイル番号 $i = 23$ の圧縮ファイル f_{23} を伸長すると、検索対象ファイル F_{23} が得られる。また、ファイル番号 $i = 158$ の圧縮ファイル f_{158} を伸長すると、(B) に示したように、検索対象ファイル F_{158} が得られる。さらに、ファイル番号 $i = 4971$ の圧縮ファイル f_{4971} を伸長すると、(C) に示したように、検索対象ファイル F_{4971} が得られる。なお、各検索対象ファイル F_{23} , F_{158} , F_{4971} 中、墨付き括弧内の文字列は見出し語である。

【0028】

図 4 は、単字出現マップ M_1 を示す説明図である。単字出現マップ M_1 では、文字ごとにビット列を有する。ビット列はビット番号順に並んでいる。ビット番号 i は、圧縮ファイルのファイル番号 i に対応する。ビット列中、“1” は、その文字が存在していることを示しており、“0” は存在しないことを示している。たとえば、かな文字「あ」のビット番号 $i = 1$ のビットは“1”であるため、圧縮ファイル f_1 を伸長した検索対象ファイルにはかな文字「あ」が存在することとなる。一方、漢字文字「足」のビット番号 $i = 1$ のビットは“0”であるため、圧縮ファイル f_1 を伸長した検索対象ファイルには漢字文字「足」が存在しないこととなる。

【0029】

図 5 は、連字出現マップ M_2 を示す説明図である。連字とは、連続する文字列である。ここでは、2 連字について例示しているが、3 連字以上でもよい。マップ形式は、図 4 に示した単字出現マップ M_1 とおなじである。たとえば、数字連字「99」のビット番号 $i = 1$ のビットは“0”であるため、圧縮ファイル f_1 を伸長した検索対象ファイルには数字連字「99」が存在しないこととなる。一方、数字連字「99」のビット番号 $i = 2$ のビットは“1”であるため、圧縮ファイル f_2 を伸長した検索対象ファイルには数字連字「99」が存在することとなる。

【0030】

図 6 は、圧縮パラメータ 221 を示す説明図である。圧縮パラメータ 221 とは、圧縮ファイル群 f を伸長した検索対象ファイル群に記述されている文字／連字とその出現回数に対応付けたテーブルである。この圧縮パラメータ 221 にしたがって、検索対象ファイル群を圧縮ファイル群 f に圧縮するためのハフマン木が生成される。

【0031】

図 7 は、ファイルパステーブル 222 を示す説明図である。ファイルパステーブル 222 とは、圧縮ファイル f_i へのパス（ファイルパス）が記述されたテーブルである。具体的には、たとえば、ファイル ID ごとにその圧縮ファイル f_i へのファイルパスとその圧縮ファイル f_i を伸長した検索対象ファイルに記述されている見出し語とそのアドレスと

10

20

30

40

50

サイズとを対応付けている。ファイルIDとは、圧縮ファイル*f i*を一意に特定する情報である。

【0032】

ここでは、説明上、圧縮ファイルに割り当てられた符号をファイルIDとしている。たとえば、ファイルID：*f 2 3*の圧縮ファイル*f 2 3*のファイルパスは、“*honmon ¥ f i l e 2 3 . h t m l*”であり、見出し語は「川端康成」である（図3を参照。）。なお、後述する常駐フラグが“1”に設定された場合、キャッシュ領域240へのファイルパス（たとえば、“*cash ¥ f i l e 2 3 . h t m l*”）が書き込まれ、“0”に戻った場合、キャッシュ領域240へのファイルパスは消去される。

【0033】

図8は、文字出現マップ連携テーブル211を示す説明図である。文字出現マップ連携テーブル211は、文字出現マップMの文字ごとのビット列を構成するビットごとに、アドレスとサイズとアクセス頻度と常駐フラグとが記憶されている。アドレスとは、文字出現マップMの文字ごとのビット列を構成するビット番号に対応する圧縮ファイルが記憶されているアドレスである。サイズとは、圧縮ファイル*f i*のサイズである。たとえば、ビット番号*i*に対応する圧縮ファイルは圧縮ファイル*f i*である。この圧縮ファイル*f i*が記憶されているアドレスは*adr i*であり、圧縮ファイル*f i*のサイズは*s i*である。

【0034】

また、アクセス頻度とは、そのビット番号*i*に対応する圧縮ファイル*f i*へのアクセスする度合いであり、ここでは、アクセス回数としている。アクセス頻度はアクセス回数のほか確率（圧縮ファイル*f i*へのアクセス回数／全圧縮ファイルへの総アクセス回数）であらわしてもよい。常駐フラグとは、そのビット番号*i*に対応する圧縮ファイル*f i*を、アーカイブ200の記憶領域230から当該記憶領域230に対するキャッシュ領域240に移動してキャッシュ領域240に常駐されているか否かを示すフラグである。

【0035】

キャッシュ領域240に常駐されている場合、常駐フラグは“1”である。一方、キャッシュ領域240に常駐されず、記憶領域230に格納されている場合、常駐フラグは“0”である。なお、常駐フラグが“1”に設定された場合、アドレスにはキャッシュ領域240のアドレスが書き込まれ、“0”に戻った場合、キャッシュ領域240のアドレスは消去される。

【0036】

図9は、ファイルパス連携テーブル212を示す説明図である。ファイルパス連携テーブル212は、ファイルパスとビット番号*i*を連携するテーブルである。常駐フラグが“1”に設定された場合、キャッシュ領域240へのファイルパスが書き込まれ、“0”に戻った場合、キャッシュ領域240へのファイルパスは消去される。

【0037】

図10は、仮想アーカイブ容量テーブル213を示す説明図である。仮想アーカイブとは、記憶領域230に記憶されているアーカイブ200のうちキャッシュ領域240に記憶された圧縮ファイル群*f'*である。なお、書庫領域201や管理領域202内のテーブル等も仮想アーカイブに含めてもよい。

【0038】

（情報検索装置の機能的構成）

図11は、情報検索装置の機能的構成を示すブロック図である。情報検索装置1100は、並び替え処理部1101と、積算部1102と、書込み部1103と、設定部1104と、入力部1105と、特定部1106と、読出し部1107と、伸長部1108と、検索部1109と、出力部1110と、更新部1111と、を含む構成である。

【0039】

並び替え処理部1101～入力部1105、更新部1111が仮想アーカイブ設定機能であり、入力部1105～出力部1110が情報検索機能である。情報検索機能については、国際公開第WO2006/123448号パンフレットに詳細に記載されているため

10

20

30

40

50

、ここでは、本実施の形態の特徴部分に言及しつつ簡略化して説明する。

【0040】

なお、並び替え処理部1101～更新部1111は、具体的には、たとえば、図1に示したROM、RAM103、磁気ディスク105などの記憶領域230に記憶されたプログラムをCPU101に実行させることによって、または、I/F109によってその機能を実現する。

【0041】

まず、並び替え処理部1101は、文字出現マップ連携テーブル211を、その各圧縮ファイルへのアクセス頻度の高い順（降順）に並び替える機能を有する。この並び替えは、常駐フラグの設定の前処理となる。

10

【0042】

積算部1102は、並び替え処理部1101による並び替え後のアクセス頻度の高い順から圧縮ファイルf iのサイズを積算する機能を有する。具体的には、圧縮ファイル群fが記憶されている記憶領域230に対するキャッシュ領域240の記憶容量を超えないように積算する。たとえば、キャッシュ領域240の記憶容量を超えない最大の積算値となるように、並び替え処理部1101による並び替え後のアクセス頻度の高い順から圧縮ファイルのサイズを積算する。このように、最大の積算値を計算することで、キャッシュ領域240の記憶容量を最大限に活用することができる。

【0043】

書込み部1103は、積算部1102によって積算された圧縮ファイル群を、当該ファイル群の検索に先立って、記憶領域230からキャッシュ領域240に書き込む機能を有する。書込み対象となる圧縮ファイル群fは、記憶領域230から消去されてもよく、また、記憶領域230に残存させておいてもよい。このように、キャッシュ領域240にアクセス頻度が高い圧縮ファイル群を書き込むことで、ファイルアクセスの高速化を図ることができる。

20

【0044】

設定部1104は、書込み部1103によってキャッシュ領域240に書き込まれた圧縮ファイル群について常駐フラグを設定する機能を有する。具体的には、たとえば、キャッシュ領域240に書き込まれた圧縮ファイル群の文字出現マップ連携テーブル211における常駐フラグを、“0”から“1”に変更する。常駐フラグがすでに“1”の場合は、そのままである。なお、前回キャッシュ領域240に書き込まれていたが今回消去された圧縮ファイルについては、常駐フラグが“1”に設定されているため、“0”に変更する。これにより、アクセス頻度が高い圧縮ファイルがキャッシュ領域240に常駐されていることが特定できる。

30

【0045】

入力部1105は、検索文字列の入力を受け付ける機能を有する。具体的には、図1に示したキーボードなどの入力装置を用いて入力された検索文字列を受け付ける。たとえば、「川端康成」といった検索文字列の入力を受け付ける。なお、検索文字列のほかに、前方一致、後方一致などの検索条件の入力を受け付けることとしてもよい。

【0046】

特定部1106は、入力部1105によって入力された検索文字列を構成するすべての文字が存在する圧縮ファイルを特定する機能を有する。具体的には、文字出現マップMを参照して、圧縮ファイル群を、検索文字列を構成するすべての文字が存在する圧縮ファイルに絞り込む。たとえば、検索文字列が「川端康成」の場合、「川」、「端」、「康」、「成」という単字に分解される。そして、単字出現マップM1から単字「川」、「端」、「康」、「成」のビット列を論理積演算し、論理積演算結果が“1”となったビット番号に対応する圧縮ファイルに絞り込まれる。

40

【0047】

図12は、文字出現マップMを用いた圧縮ファイルf iの絞込み例を示す説明図である。図12に示したように、各単字「川」、「端」、「康」、「成」のビット列を同一ビッ

50

ト番号で論理積（AND）演算をおこなう。論理積が“1”であるビット番号 $i = 1, 2, 3, 158, 4971$ に対応する圧縮ファイル $f_1, f_{23}, f_{158}, f_{4971}$ の伸長後の検索対象ファイルに、単字「川」、「端」、「康」、「成」がすべて含まれていることがわかる。

【0048】

ただし、この段階では、単字「川」、「端」、「康」、「成」がすべて含まれていることがわかるだけで、単字「川」、「端」、「康」、「成」がこの順で並ぶ文字列「川端康成」が含まれていることまでは特定できない。これにより、圧縮形式のまま、検索文字列を含んでいる可能性がある圧縮ファイルを特定することができる。

【0049】

読出し部1107は、特定部1106によって特定された圧縮ファイルを、設定部1104によって設定された常駐フラグに基づく領域から読み出す機能を有する。具体的には、たとえば、特定部1106によって特定された圧縮ファイルについて、文字出現マップ連携テーブル211の常駐フラグを参照して、常駐フラグの値により、特定された圧縮ファイルが記憶されている領域を特定する。

【0050】

常駐フラグが“0”である場合、圧縮ファイルはキャッシュ領域240に記憶されておらず、アーカイブ200の記憶領域から読み出すこととなる。一般的なファイル名に基づくオープン処理では、ファイルパステーブル222が参照され、2分探索により一致したファイルIDの先頭アドレスとサイズにより該当の圧縮ファイル f_i がオープンされる。多数の圧縮ファイルが格納されるアーカイブでは、オープン処理が長くなる。

【0051】

一方、文字出現マップのビット番号に基づくオープン処理では、文字出現マップ211が参照され、ビット番号に対応した先頭アドレスとサイズを直接得ることができ、該当する圧縮ファイル f_i を高速にアクセスすることができる。常駐フラグが“1”である場合、キャッシュ領域240に記憶されていることがわかる。この場合、キャッシュ領域240から圧縮ファイル f_i をアクセスすることができ、さらに高速化を図ることができる。

【0052】

伸長部1108は、読出し部1107によって読み出された圧縮ファイルを伸長する機能を有する。具体的には、たとえば、圧縮パラメータ221に基づいて生成されたハフマン木により、読み出された圧縮ファイル f_i を伸長する。これにより、読み出された圧縮ファイル f_i のみ伸長すればよいため、ファイルアクセスの高速化を図ることができる。

【0053】

検索部1109は、伸長部1108によって伸長された検索対象ファイルの中から検索文字列に一致または関連する文字列を検索する機能を有する。具体的には、たとえば、伸長された検索対象ファイル群の中から検索文字列と一致する文字列が存在する検索対象ファイルを抽出する。検索文字列と前方一致または後方一致する文字列が存在する検索対象ファイルを、関連する検索対象ファイルとして抽出する。その他、検索文字列に共起する文字列が設定されている場合には、共起する文字列が存在する検索対象ファイルを、関連する検索対象ファイルとして抽出する。

【0054】

出力部1110は、伸長部1108によって伸長された検索対象ファイルを出力する機能を有する。出力部1110の出力形式は、具体的には、たとえば、ディスプレイへの表示やプリンタへの印刷出力、他のコンピュータへの送信、情報検索装置1100内の記憶領域230への格納などがある。ディスプレイに表示する場合、伸長された検索対象ファイルをそのまま表示することとしてもよい。また、伸長された検索対象ファイル名を一覧表示して、ユーザが検索対象ファイル名をポイントすることで、リンク先となる検索対象ファイルを読み出して画面表示することとしてもよい。

【0055】

また、検索部1109により検索された場合、検索された検索対象ファイルをそのまま

10

20

30

40

50

表示することとしてもよい。また、検索された検索対象ファイル名を一覧表示して、ユーザが検索対象ファイル名をポイントすることで、リンク先となる検索対象ファイルを読み出して画面表示することとしてもよい。

【0056】

更新部1111は、伸長部1108によって圧縮ファイルが伸長された場合、当該圧縮ファイルへのアクセス頻度を更新する機能を有する。具体的には、たとえば、アクセス頻度がアクセス回数で表されている場合、伸長された圧縮ファイルf_iのアクセス回数を1加算する。また、アクセス頻度が確率で表されている場合、伸長された圧縮ファイルに対しては、アクセス回数を1加算するとともに全圧縮ファイルf₁～f_nへの総アクセス回数も1加算する。

10

【0057】

一方、伸長されなかった圧縮ファイルf_iに対しては全圧縮ファイルf₁～f_nへの総アクセス回数のみ1加算する。したがって、並び替え処理部1101は、更新後の圧縮ファイル群fのアクセス頻度にしたがってソート処理を実行することとなる。これにより、文字出現マップMにより絞り込まれやすい圧縮ファイルf_iのアクセス頻度が高くなるため、次回以降の伸長速度の高速化を実現することができる。

【0058】

また、更新部1111は、伸長された圧縮ファイルf_iではなく、検索部1109によって検索された検索対象ファイルの圧縮ファイルf_iについて、当該圧縮ファイルf_iへのアクセス頻度を更新することとしてもよい。具体的には、たとえば、アクセス頻度がアクセス回数で表されている場合、検索された検索対象ファイルの圧縮ファイルf_iへのアクセス回数を1加算する。また、アクセス頻度が確率で表されている場合、検索された検索対象ファイルの圧縮ファイルf_iに対しては、アクセス回数を1加算するとともに全圧縮ファイルf₁～f_nへの総アクセス回数も1加算する。

20

【0059】

一方、検索されなかった検索対象ファイルの圧縮ファイルf_iに対しては全圧縮ファイルf₁～f_nへの総アクセス回数のみ1加算する。したがって、並び替え処理部1101は、更新後の圧縮ファイル群のアクセス頻度にしたがってソート処理を実行することとなる。これにより、実際に検索された圧縮ファイルf_iのアクセス頻度が高くなるため、次回以降の検索速度の高速化を実現することができる。

30

【0060】

(仮想アーカイブ設定処理手順)

図13は、情報検索装置1100の仮想アーカイブ設定機能が実行する仮想アーカイブ設定処理手順を示すフローチャートである。まず、並び替え処理部1101により、文字出現マップ連携テーブル211を、アクセス頻度の降順にソートする(ステップS1301)。ここで、並び替え後の順位kをk=1とし(ステップS1302)、積算部1102により、順位1～k+1までの圧縮ファイルの合計サイズを計算する(ステップS1303)。そして、合計サイズs(1～k+1)が、s(1～k+1) > T_sであるか否かを判断する(ステップS1304)。ここで、T_sはキャッシュ領域240に記憶可能な最大の記憶容量である。

40

【0061】

s(1～k+1) > T_sでない場合(ステップS1304: No)、kをインクリメントして(ステップS1305)、ステップS1303に戻る。一方、s(1～k+1) > T_sである場合(ステップS1304: Yes)、キャッシュ領域240にはこれ以上圧縮ファイルを格納できないため、順位1～k+1までの圧縮ファイルのビット番号、アクセス頻度およびサイズとなるように仮想アーカイブ容量テーブル213を更新する(ステップS1306)。

【0062】

そして、書込み部1103により、順位1～kまでの圧縮ファイルをキャッシュ領域240に書き込む(ステップS1307)。この場合、順位k+1以降の圧縮ファイルはキ

50

キャッシュ領域240から消去される。このあと、設定部1104により、文字出現マップ連携テーブル211における順位1～kまでの圧縮ファイルの常駐フラグをON（“0”→“1”）にする（ステップS1308）。

【0063】

なお、順位k+1以降の圧縮ファイルについては、常駐フラグをOFF（“1”→“0”）にする。これにより一連の処理が終了する。この仮想アーカイブ設定処理手順によれば、検索に先立って、アクセス頻度が高い圧縮ファイルを優先的に仮想アーカイブとして設定することができるため、ファイルアクセスの高速化を実現することができる。

【0064】

（情報検索処理手順）

10

図14は、情報検索装置1100の情報検索機能が実行する情報検索処理手順を示すフローチャートである。入力部1105により、検索文字列の入力を受け付ける（ステップS1401）。そして、検索文字列を単字または連字（以下、単に「文字」）に分解する（ステップS1402）。つぎに、分解された文字ごとのビット列を文字出現マップMから抽出して（ステップS1403）、抽出されたビット列を同一ビット番号で論理積演算する（ステップS1404）。

【0065】

この論理積演算が“1”である圧縮ファイルfiを、分解された文字を含む圧縮ファイルとして特定する（ステップS1405）。このあと、特定された圧縮ファイルfiのうち未処理の圧縮ファイルfiがあるか否かを判断する（ステップS1406）。未処理の圧縮ファイルfiがある場合（ステップS1406：Yes）、特定された未処理の圧縮ファイルfiを1つ選択し（ステップS1407）、その選択された圧縮ファイルfiに対して常駐フラグがONであるか否かを判断する（ステップS1408）。

20

【0066】

ONである場合（ステップS1408：Yes）、読出し部1107により、キャッシュ領域240から直接CPU101のレジスタに転送して（ステップS1409）、ステップS1411に移行する。一方、OFFである場合（ステップS1408：No）、読出し部1107により、アーカイブ200の記憶領域230からキャッシュ領域240にリードして、CPU101に読み込ませて、ステップS1411に移行する（ステップS1410）。ステップS1411では、伸長部1108により、圧縮パラメータ221に基づくハフマン木を用いて伸長処理を実行して（ステップS1411）、ステップS1406に戻る。

30

【0067】

ステップS1406において、未処理の圧縮ファイルfiがない場合（ステップS1406：No）、検索部1109により、伸長された検索対象ファイルを検索文字列により検索する（ステップS1412）。そして、出力部1110により、検索結果を出力する（ステップS1413）。このあと、更新部1111により、文字出現マップ連携テーブル211の該当する圧縮ファイルfiのアクセス頻度を1加算する（ステップS1414）。これにより、一連の処理を終了する。

【0068】

40

この情報検索処理手順によれば、常駐フラグがON（“1”）に設定されている圧縮ファイルfiをキャッシュ領域240から読み出して伸長処理をおこなうため、ファイルアクセスの高速化を実現することができる。また、検索する都度、圧縮ファイルfiへのアクセス頻度が更新されるため、キャッシュ領域240に書き込まれる圧縮ファイルfiを逐次更新することができる。したがって、次回以降のファイルアクセスの高速化も実現することができる。

【0069】

このように、実施の形態1によれば、アーカイブの圧縮ファイルfiのアクセスにおいて、文字出現マップのビット番号をもとに文字出現マップ連携テーブル211を利用することで高速化を図ることができる。さらに、アクセス頻度の高いファイルをキャッシュメ

50

メモリに常駐化することで、ファイルのアクセス処理の時間の短縮化を図ることができる。したがって、伸長するまでの時間が大幅に短縮化され、検索速度の高速化を図ることができる。また、キャッシュ領域を有効活用することで、省メモリ化を図ることができる。なお、この実施の形態1の情報検索装置1100は、パーソナル・コンピュータや検索サーバーのほか、携帯電話機、携帯型ゲーム機、電子辞書などの携帯型端末に適用することができる。

【0070】

＜実施の形態2＞

つぎに、実施の形態2について説明する。インターネットでのサイト検索などは、定期的に各サイトの更新を監視し、形態素解析を行った要約データをもとに、大規模なインデックスを作成し、全文検索が行われている。サイトのデータ量の増大に対し、従来では、各サイトの監視の処理の高速化とスループットの増加、複数のコンピュータによる検索のスケラビリティが課題となっている。

10

【0071】

これらの課題に対し、実施の形態2では、アーカイブ200の追加やマージ、削除などの高速化を実現する。また、グリッドコンピュータなどに関するスケラビリティに対して、スレーブサーバーに検索を分割し並行処理したり、各スレーブサーバーの稼働率を平準化することで、検索速度の効率化を実現する。

【0072】

図15は、実施の形態2にかかる検索システムのシステム構成図である。検索システム1500は、マスターサーバー1501と複数のスレーブサーバー1502-1～1502-Nを含む構成である。マスターサーバー1501と複数のスレーブサーバー1502-1～1502-N、または、スレーブサーバー1502-1～1502-Nどうしは、ネットワーク114を介して相互に通信可能である。マスターサーバー1501は、スレーブサーバー1502-1～1502-Nを統括管理する。各スレーブサーバー1502-Iは、上述した実施の形態1の情報検索装置1100に相当し、上述した仮想アーカイブ設定機能と情報検索機能を有する。

20

【0073】

なお、各スレーブサーバー1502-I ($I=1\sim N$) が備えるアーカイブ200の種類は異なる。たとえば、あるスレーブサーバー1502-Iが保持するアーカイブ200-Iは国語辞典のアーカイブであり、他のスレーブサーバー1502-J ($J\neq I$) が保持するアーカイブ200-Jは用語辞典のアーカイブであり、別のスレーブサーバー1502-K ($K\neq I, J$) が保持するアーカイブ200-Kは英和辞典のアーカイブであるなど、それぞれ種類や出版社が異なるアーカイブである。

30

【0074】

なお、各スレーブサーバー1502-Iはそれぞれ異なるアーカイブ200-Iを有するが、各アーカイブ200-I内の圧縮パラメータ221-Iも異なる。したがって、各スレーブサーバー1502-Iが保持するハフマン木h-Iも異なる構造となる。

【0075】

このようなスレーブサーバー群1502に対して、共通の検索キーワードをマスターサーバー1501から与えることで、串刺し検索と呼ばれる多書籍検索を実行する。各スレーブサーバー1502-Iは、検索結果をマスターサーバー1501に返すことで、マスターサーバー1501は複数辞書の検索結果を得ることができる。以降、説明の簡略化のため、スレーブサーバー1502-Iの数を2台 ($N=2$) として説明する。

40

【0076】

図16は、アーカイブの共通化を示す説明図である。マスターサーバー1501は、各スレーブサーバー1502-1, 1502-2における検索処理の平準化のため、スレーブサーバー1502-1, 1502-2のアーカイブ200-1, 200-2とハフマン木h-1, h-2をネットワーク114経由で収集する。そして、各アーカイブ200-1, 200-2を集約した統合アーカイブAと、各アーカイブ200-1, 200-2の

50

ハフマン木 $h-1$ 、 $h-2$ を共通化した共通ハフマン木 H を生成する。

【0077】

図17は、新アーカイブの分配処理を示す説明図である。マスターサーバー1501は、スレーブサーバー1502-1、1502-2の検索処理が平準化するように統合アーカイブAを分割し、各スレーブサーバー1502-1、1502-2に固有な新アーカイブA1、A2としてスレーブサーバー1502-1、1502-2に配信する。また、新共通ハフマン木H1、H2も各スレーブサーバー1502-1、1502-2に配信される。本例では、一台のスレーブサーバーに対して1つのアーカイブが割り当てられているため、各スレーブサーバー1502-1、1502-2に対して共通ハフマン木Hを配信しているが、スレーブサーバーごとに複数のアーカイブとそのハフマン木がある場合、スレーブサーバーごとに固有の共通ハフマン木を配信することとなる。たとえば、スレーブサーバー1501-1にアーカイブ200-1、200-2とハフマン木 $h-1$ 、 $h-2$ がある場合、スレーブサーバー1502-1に共通ハフマン木Hを配信することとなる。

10

【0078】

(ハフマン木の再構成)

図18は、アーカイブ200-1の圧縮符号テーブルおよび圧縮パラメータ221を示す説明図である。ここでは、説明を単純化するため、アーカイブ200-1に含まれる圧縮された検索対象ファイル群には、文字a~fが記述されているものとする。図18中、(A)はアーカイブ200-1の圧縮符号テーブル1800であり、(B)はアーカイブ200-1の圧縮パラメータP1である。圧縮符号テーブル1800において、出現回数

20

【0079】

図19は、アーカイブ200-1の圧縮符号テーブル1800から生成されるハフマン木 $h-1$ である。図19中、丸図形(○)が節点である。特に、最上位の節点を根R、それ以外の節点を内部節点と呼ぶ。また、四角図形(□)が葉、節点間、または、節点と葉とを結ぶ線分が枝である。葉の中に示した文字は伸長後の文字であり、葉の下に示した文字列は、当該葉となる伸長後の文字に割り当てられた圧縮符号である。

【0080】

図20は、アーカイブ200-2の圧縮符号テーブルおよび圧縮パラメータ221を示す説明図である。ここでは、説明を単純化するため、アーカイブ200-2に含まれる圧縮された検索対象ファイル群には、文字a~fが記述されているものとする。図20中、(A)はアーカイブ200-2の圧縮符号テーブル2000であり、(B)はアーカイブ200-2の圧縮パラメータP2である。圧縮符号テーブル2000において、出現回数

30

【0081】

図22は、統合アーカイブAの圧縮符号テーブルおよび圧縮パラメータ221を示す説明図である。統合アーカイブAは、アーカイブ200-1、200-2を統合しているため、統合アーカイブAに含まれている圧縮された検索対象ファイルには、文字a~fが記述されていることとなる。したがって、(B)に示した共通圧縮パラメータPの出現回数は、アーカイブ200-1の圧縮パラメータP1とアーカイブ200-2の圧縮パラメータP2の出現回数を同一文字で集計した値となる。図23は、統合アーカイブAの圧縮符号テーブルから生成される共通ハフマン木Hを示す説明図である。

40

【0082】

(アーカイブの再構成)

つぎに、実施の形態2におけるアーカイブの再構成について説明する。ここでは、上述したアーカイブ200-1とアーカイブ200-2とを統合して、統合アーカイブAを生成する。まず、アーカイブ200-1の記憶内容について説明する。

【0083】

図24は、アーカイブ200-1の記憶内容を示す説明図である。アーカイブ200-

50

1は、図1に示したRAM103、磁気ディスク105などの記憶領域230に記憶されている。アーカイブ200-1は、書庫領域201と管理領域202とデータ領域203とを有する。書庫領域201には、文字出現マップ連携テーブル211aとファイルパス連携テーブル212aと仮想アーカイブ容量テーブル213aとが記憶されている。管理領域202には、圧縮パラメータP1とファイルパステーブル222aと文字出現マップMa（単字出現マップMa1および連字出現マップMa2）とが記憶されている。データ領域203には、図3に示したような圧縮ファイル群fa（圧縮ファイルfa_1～fa_n）が記憶されている。なお、これらは実施の形態1で説明した内容に準じている。

【0084】

図25は、アーカイブ200-1の単字出現マップMa1を示す説明図である。図26は、アーカイブ200-1の連字出現マップMa2を示す説明図である。アーカイブ200-1の文字出現マップMaでは、アーカイブ200-2と区別するため、便宜上、ビット番号をa_1～a_nとする。図27は、アーカイブ200-1の圧縮パラメータP1を示す説明図である。図28は、アーカイブ200-1のファイルパステーブル222aを示す説明図である。図29は、アーカイブ200-1の文字出現マップ連携テーブル211aを示す説明図である。図30は、アーカイブ200-1のファイルパス連携テーブル212aを示す説明図である。

【0085】

つぎに、アーカイブ200-2の記憶内容について説明する。図31は、アーカイブ200-2の記憶内容を示す説明図である。アーカイブ200-2は、図1に示したRAM103、磁気ディスク105などの記憶領域230に記憶されている。アーカイブ200-2は、書庫領域201と管理領域202とデータ領域203とを有する。書庫領域201には、文字出現マップ連携テーブル211bとファイルパス連携テーブル212bと仮想アーカイブ容量テーブル213bとが記憶されている。管理領域202には、圧縮パラメータP2とファイルパステーブル222bと文字出現マップMb（単字出現マップMb1および連字出現マップMb2）とが記憶されている。データ領域203には、図3に示したような圧縮ファイル群fb（圧縮ファイルfb_1～fb_m）が記憶されている。なお、これらは実施の形態1で説明した内容に準じている。

【0086】

図32は、アーカイブ200-2の単字出現マップMb1を示す説明図である。図33は、アーカイブ200-2の連字出現マップMb2を示す説明図である。アーカイブ200-2の文字出現マップMbでは、アーカイブ200-1と区別するため、便宜上、ビット番号をb_1～b_mとする。図34は、アーカイブ200-2の圧縮パラメータP2を示す説明図である。図35は、アーカイブ200-2のファイルパステーブル222bを示す説明図である。図36は、アーカイブ200-2の文字出現マップ連携テーブル211bを示す説明図である。図37は、アーカイブ200-2のファイルパス連携テーブル212bを示す説明図である。

【0087】

まず、圧縮パラメータP1、P2の再構成について説明する。図38は、共通パラメータの生成例を示す説明図である。アーカイブ200-1の圧縮パラメータP1とアーカイブ200-2の圧縮パラメータP2を、同一文字（単字／連字）で出現頻度を集計することで、共通圧縮パラメータPが生成される。

【0088】

つぎに、文字出現マップ連携テーブル211a、211bの再構成について説明する。図39は、文字出現マップ連携テーブル211a、211bの再構成を示す説明図である。アーカイブ200-1の文字出現マップ連携テーブル211aとアーカイブ200-2の文字出現マップ連携テーブル211bとを統合して、アクセス頻度の高い順にソートする。統合後の文字出現マップ連携テーブル3900は、合計でn+m個のビット番号のアクセス頻度を有することとなる。このあと、アクセス頻度が各スレーブサーバー1502-1、1502-2において平準化するように、スレーブサーバー1502-1、150

10

20

30

40

50

2-2ごとに割り当てる。

【0089】

ここでは、アクセス頻度が高い順の順位が奇数番目となるビット番号について、スレーブサーバー1502-1に割り当てる新たな文字出現マップ連携テーブル3900aとする。一方、偶数番目となるビット番号について、スレーブサーバー1502-2に割り当てる新たな文字出現マップ連携テーブル3900bとする。割当方法としては、ここでは例として奇数番目と偶数番目としたが、一方のスレーブサーバー1502-1に対しては、1位、4位、5位、8位、9位…とし、他方のスレーブサーバー1502-2に対しては、2位、3位、6位、7位、10位…としてもよい。その他、割り当てた順位（またはアクセス頻度）の合計が均等になるように分配すればどのような割り当て方法でもよい。

10

【0090】

図40は、ファイルパス連携テーブル212a、212bの再構成を示す説明図である。まず、アーカイブ200-1のファイルパス連携テーブル212aとアーカイブ200-2のファイルパス連携テーブル212bとを統合する。統合後のファイルパス連携テーブル4000は、合計で $n+m$ 個のビット番号のファイルパスを有することとなる。このあと、上述した文字出現マップ連携テーブル3900で採用された割当方法により、ビット番号を割り当てる。

【0091】

これにより、一方のスレーブサーバー1502-1には、そのスレーブサーバー1502-1に割り当てられた文字出現マップ連携テーブル3900aのビット番号と同一ビット番号のファイルパス連携テーブル4000aが得られる。同様に、他方のスレーブサーバー1502-2には、そのスレーブサーバー1502-2に割り当てられた文字出現マップ連携テーブル3900bのビット番号と同一ビット番号のファイルパス連携テーブル4000bが得られる。

20

【0092】

図41は、ファイルパステーブル222a、222bの再構成を示す説明図である。まず、アーカイブ200-1のファイルパステーブル222aとアーカイブ200-2のファイルパステーブル222bとを統合する。統合後のファイルパステーブル4100は、合計で $n+m$ 個のファイルIDのファイルパスを有することとなる。このあと、上述した文字出現マップ連携テーブル3900で採用された割当方法により、ビット番号に対応するファイルIDを割り当てる。

30

【0093】

これにより、一方のスレーブサーバー1502-1には、そのスレーブサーバー1502-1に割り当てられた文字出現マップ連携テーブル3900aのビット番号と同一ビット番号に対応するファイルIDのファイルパステーブル4100aが得られる。同様に、他方のスレーブサーバー1502-2には、そのスレーブサーバー1502-2に割り当てられた文字出現マップ連携テーブル3900bのビット番号と同一ビット番号に対応するファイルIDのファイルパステーブル4100bが得られる。

【0094】

図42は、単字出現マップMa1、Mb1の再構成を示す説明図である。まず、アーカイブ200-1の単字出現マップMa1とアーカイブ200-2の単字出現マップMb1とを統合する。統合後の単字出現マップMab1は、文字ごとに合計で $n+m$ 個のビットからなるビット列を有することとなる。このあと、上述した文字出現マップ連携テーブル3900で採用された割当方法により、ビット番号を割り当てる。

40

【0095】

これにより、一方のスレーブサーバー1502-1には、そのスレーブサーバー1502-1に割り当てられた文字出現マップ連携テーブル3900aのビット番号と同一ビット番号となる単字出現マップMA1が得られる。同様に、他方のスレーブサーバー1502-2には、そのスレーブサーバー1502-2に割り当てられた文字出現マップ連携テーブル3900bのビット番号と同一ビット番号となる単字出現マップMB1が得られる

50

。

【0096】

図43は、連字出現マップM a 2, M b 2の再構成を示す説明図である。まず、アーカイブ200-1の連字出現マップM a 2とアーカイブ200-2の連字出現マップM b 2とを統合する。統合後の連字出現マップM a b 2は、文字ごとに合計でn+m個のビットからなるビット列を有することとなる。このあと、上述した文字出現マップ連携テーブル3900で採用された割当方法により、ビット番号を割り当てる。

【0097】

これにより、一方のスレーブサーバー1502-1には、そのスレーブサーバー1502-1に割り当てられた文字出現マップ連携テーブル3900 aのビット番号と同一ビット番号となる連字出現マップM B 1が得られる。同様に、他方のスレーブサーバー1502-2には、そのスレーブサーバー1502-2に割り当てられた文字出現マップ連携テーブル3900 bのビット番号と同一ビット番号となる連字出現マップM B 2が得られる。

10

。

【0098】

図44は、圧縮ファイル群f a, f bの再構成を示す説明図である。まず、アーカイブ200-1の圧縮ファイル群f aをそのハフマン木h-1を用いて伸長する。これにより、検索対象ファイル群F aが得られる。同様に、アーカイブ200-2の圧縮ファイル群f bをそのハフマン木h-2を用いて伸長する。これにより、検索対象ファイル群F bが得られる。

20

【0099】

つぎに、検索対象ファイル群F aを、共通ハフマン木Hを用いて再圧縮する。これにより、圧縮ファイル群g aが得られる。同様に、検索対象ファイル群F bを、共通ハフマン木Hを用いて再圧縮する。これにより、圧縮ファイル群g bが得られる。

【0100】

このあと、再圧縮された圧縮ファイル群g aおよび圧縮ファイル群g bを統合する。そして、上述した文字出現マップ連携テーブル3900で採用された割当方法により、アクセス頻度の高い順に並び替える。これにより、アクセス頻度順の統合圧縮ファイル群gが得られる。

【0101】

30

これにより、一方のスレーブサーバー1502-1には、そのスレーブサーバー1502-1に割り当てられた文字出現マップ連携テーブル3900 aのビット番号と同一ビット番号の圧縮ファイル群g 1が割り当てられる。同様に、他方のスレーブサーバー1502-2には、そのスレーブサーバー1502-2に割り当てられた文字出現マップ連携テーブル3900 bのビット番号と同一ビット番号の圧縮ファイル群g 2が割り当てられる。

。

【0102】

図45は、新アーカイブA 1の記憶内容を示す説明図である。新アーカイブA 1は、スレーブサーバー1502-1に配信される。新アーカイブA 1には、図38に示した共通圧縮パラメータP、図39に示した再構成後の文字出現マップ連携テーブル3900 a、図40に示した再構成後のファイルパス連携テーブル4000 a、図41に示した再構成後のファイルパステーブル4100 a、図42に示した再構成後の単字出現マップM A 1、図43に示した再構成後の連字出現マップM 2、図44に示した再構成後の圧縮ファイル群g 1が格納される。

40

【0103】

図46は、新アーカイブA 2の記憶内容を示す説明図である。新アーカイブA 2は、スレーブサーバー1502-2に配信される。新アーカイブA 2には、図38に示した共通圧縮パラメータP、図39に示した再構成後の文字出現マップ連携テーブル3900 b、図40に示した再構成後のファイルパス連携テーブル4000 b、図41に示した再構成後のファイルパステーブル4100 b、図42に示した再構成後の単字出現マップM B 1

50

、図４３に示した再構成後の連字出現マップMB２、図４４に示した再構成後の圧縮ファイル群g２が格納される。

【０１０４】

（マスターサーバー１５０１の機能的構成）

図４７は、マスターサーバー１５０１（情報管理装置）の機能的構成を示すブロック図である。マスターサーバー１５０１は、受信部４７０１と、共通圧縮パラメータ生成部４７０２と、共通ハフマン木生成部４７０３と、伸長部４７０４と、圧縮部４７０５と、再構成部４７０６と、送信部４７０７と、を含む構成である。受信部４７０１～送信部４７０７は、具体的には、たとえば、図１に示したROM１０２、RAM１０３、磁気ディスク１０５などの記憶領域に記憶されたプログラムをCPU１０１に実行させることによって、または、I/F１０９によってその機能を実現する。

10

【０１０５】

まず、受信部４７０１は、スレーブサーバー１５０２－Iから送信されてくるデータを受信する機能を有する。具体的には、たとえば、各スレーブサーバー１５０２－１～１５０２－Nからアーカイブ２００－１～２００－Nやハフマン木h－１～h－Nを受信する。

【０１０６】

共通圧縮パラメータ生成部４７０２は、全アーカイブ２００－１～２００－Nに対する共通圧縮パラメータPを生成する機能を有する。具体的には、たとえば、図３８に示したように、各スレーブサーバー１５０２－１，１５０２－２から受信されたアーカイブ２００－１，２００－２に含まれている圧縮パラメータP１，P２を抽出する。そして、抽出された各圧縮パラメータP１，P２の出現回数を同一文字ごとに集計することで、共通圧縮パラメータPを生成する。生成された共通圧縮パラメータPは、共通ハフマン木生成部４７０３とアーカイブ生成部４７１３に送られる。

20

【０１０７】

共通ハフマン木生成部４７０３は、全アーカイブ２００－１～２００－Nに対する共通ハフマン木Hを生成する機能を有する。具体的には、たとえば、共通圧縮パラメータPの出現回数の高い文字から２分探索で“０”と“１”を割り当てることにより、共通ハフマン木Hを生成する（図２２、図２３を参照。）。生成された共通ハフマン木は、伸長部４７０４とアーカイブ生成部４７１３に送られる。

30

【０１０８】

伸長部４７０４は、アーカイブ２００－Iごとに、アーカイブ２００－Iに含まれている圧縮ファイル群fを伸長する機能を有する。伸長処理に用いられるハフマン木は、アーカイブ２００－Iとともに送信されてきたハフマン木h－Iである。たとえば、図４４に示したように、圧縮ファイル群faをその圧縮に用いられたハフマン木h－１で伸長することで、検索対象ファイル群Faを得る。同様に、圧縮ファイル群fbをその圧縮に用いられたハフマン木h－２で伸長することで、検索対象ファイル群Fbを得る。

【０１０９】

圧縮部４７０５は、伸長部４７０４で伸長された検索対象ファイル群を再圧縮する機能を有する。再圧縮に用いられるハフマン木は共通ハフマン木Hである。たとえば、図４４に示したように、検索対象ファイル群Faを共通ハフマン木Hで再圧縮することで、圧縮ファイル群gaを得る。同様に、検索対象ファイル群Fbをその圧縮に用いられたハフマン木Hで再圧縮することで、圧縮ファイル群gbを得る。圧縮ファイル群ga，gbは、統合部４７１１で統合される。

40

【０１１０】

再構成部４７０６は、受信された各アーカイブ２００－Iと各ハフマン木h－Iを再構成する機能を有する。再構成部４７０６は、統合部４７１１と、割当部４７１２と、アーカイブ生成部４７１３と、を有する。統合部４７１１は、アーカイブ２００－I内のデータをそれぞれ統合する機能を有する。

【０１１１】

50

具体的には、たとえば、図 3 9～図 4 3 に示したように、各アーカイブ 2 0 0－1，2 0 0－2 の文字出現マップ連携テーブル 2 1 1 a，2 1 1 b、各ファイルパス連携テーブル 2 1 2 a，2 1 2 b、各ファイルパステーブル 2 2 2 a，2 2 2 b、各単字出現マップ M a 1，M b 1、各連字出現マップ M a 2，M b 2 をそれぞれ統合して、統合後の文字出現マップ連携テーブル 3 9 0 0、統合後のファイルパス連携テーブル 4 0 0 0、統合後のファイルパステーブル 4 1 0 0、統合後の単字出現マップ M a b 1、統合後の連字出現マップ M a b 2 を得る。

【0 1 1 2】

また、統合部 4 7 1 1 は、図 4 4 に示したように、圧縮部 4 7 0 5 によってアーカイブ 2 0 0－1，2 0 0－2 ごとに再圧縮された圧縮ファイル群 g a，g b を統合して、統合後の圧縮ファイル群 g を得る。

10

【0 1 1 3】

割当部 4 7 1 2 は、統合されたデータをアーカイブ別に、各スレーブサーバー 1 5 0 2－I において割当後のアクセス頻度が均等になるように、統合部 4 7 1 1 によって統合されたデータをスレーブサーバー 1 5 0 2－I ごとに割り当てる機能を有する。

【0 1 1 4】

具体的には、たとえば、図 3 9 に示したように、統合後の文字出現マップ連携テーブル 3 9 0 0 のレコードを、アクセス頻度またはその順位が均等になるように割り当てて、スレーブサーバー 1 5 0 2－1，1 5 0 2－2 ごとに再構成された文字出現マップ連携テーブル 3 9 0 0 a，3 9 0 0 b を得る。

20

【0 1 1 5】

また、図 4 0～図 4 3 に示した統合後のファイルパス連携テーブル 4 0 0 0、統合後のファイルパステーブル 4 1 0 0、統合後の単字出現マップ M a b 1、統合後の連字出現マップ M a b 2 についても同様である。また、割当部 4 7 1 2 は、図 4 4 に示したように、統合後の圧縮ファイル群 g についても、アクセス頻度またはその順位が均等になるように割り当てて、スレーブサーバー 1 5 0 2－1，1 5 0 2－2 ごとに再構成された圧縮ファイル群 g 1，g 2 を得る。

【0 1 1 6】

アーカイブ生成部 4 7 1 3 は、スレーブサーバー 1 5 0 2－I ごとに再構成された新アーカイブを生成する機能を有する。具体的には、たとえば、スレーブサーバー 1 5 0 2－1，1 5 0 2－2 ごとに割り当てられたデータを、スレーブサーバー 1 5 0 2－1，1 5 0 2－2 ごとにまとめて、図 4 5 および図 4 6 に示したような新アーカイブ A 1，A 2 とする。

30

【0 1 1 7】

送信部 4 7 0 7 は、各スレーブサーバー 1 5 0 2－I にデータを配信する機能を有する。具体的には、たとえば、アーカイブ 2 0 0－1～2 0 0－N やハフマン木 h－1～h－N の収集リクエストを配信する。また、再構成部 4 7 0 6 によって再構成された新アーカイブ A 1（A 2）と共通ハフマン木 H とを組み合わせて、割当先となるスレーブサーバー 1 5 0 2－1（1 5 0 2－2）にそれぞれ配信する。

【0 1 1 8】

40

（再構成処理手順）

つぎに、マスターサーバー 1 5 0 1 によるアーカイブ再構成処理手順について説明する。図 4 8 および図 4 9 は、マスターサーバー 1 5 0 1 によるアーカイブ再構成処理手順を示すフローチャートである。図 4 8 において、まず、受信部 4 7 0 1 により、各スレーブサーバー 1 5 0 2－1～1 5 0 2－N のアーカイブ 2 0 0－1～2 0 0－N とハフマン木 h－1～h－N とを収集する（ステップ S 4 8 0 1）。

【0 1 1 9】

つぎに、統合部 4 7 1 1 により、アーカイブ 2 0 0－1，2 0 0－2 内の文字出現マップ連携テーブル 2 1 1 a，2 1 1 b を抽出して統合する（ステップ S 4 8 0 2）。そして、割当部 4 7 1 2 により、統合後の文字出現マップ連携テーブル 3 9 0 0 をアクセス頻度

50

の高い順（降順）にソートして（ステップS4803）、スレーブサーバー1502-1、1502-2ごとに文字出現マップ連携テーブル3900a、3900bを割り当てる（ステップS4804）。

【0120】

つぎに、各ファイルパステーブル222a、222b、各ファイルパス連携テーブル212a、212b、および各文字出現マップMa、Mbを統合し、アクセス頻度の高いビット番号（または対応するファイルID）にしたがって、スレーブサーバー1502-1、1502-2ごとに割り当てることで再構成する（ステップS4805）。

【0121】

そして、共通圧縮パラメータ生成部4702により、共通圧縮パラメータPを生成し（ステップS4806）、この共通圧縮パラメータPを用いて、共通ハフマン木生成部4703により、共通ハフマン木Hを生成する（ステップS4807）。

10

【0122】

つぎに、図49において、伸長部4704により、アーカイブ200-1、200-2ごとの圧縮ファイル群fa、fbをその圧縮に用いたハフマン木h-1、h-2で伸長する（ステップS4901）。そして、圧縮部4705により、アーカイブ200-1、200-2ごとに伸長された検索対象ファイル群Fa、Fbを共通ハフマン木Hで再圧縮する（ステップS4902）。つぎに、統合部4711により、再圧縮された各圧縮ファイル群ga、gbを統合して（ステップS4903）、アクセス頻度の高いビット番号順（降順）にソートする（ステップS4904）。

20

【0123】

このあと、割当部4712により、アクセス頻度またはその順位の合計が均等になるように、各スレーブサーバー1502-1、1502-2に割り当てる（ステップS4905）。そして、アーカイブ生成部4713により、スレーブサーバー1502-1、1502-2ごとの新アーカイブA1、A2を生成して（ステップS4906）、新アーカイブA1（A2）および共通ハフマン木Hを各新アーカイブA1（A2）の割当先となるスレーブサーバー1502-1（1502-2）に配信する（ステップS4907）。これにより、一連の処理を終了する。

【0124】

このように、実施の形態2によれば、各スレーブサーバー1502-1、1502-2のアーカイブ200-1、200-2を再構成することで、スレーブサーバー1502-1、1502-2の検索速度の均一化を図ることができる。したがって、同一の検索文字列を各スレーブサーバー1502-1、1502-2に与えた場合、ほぼ同時に検索結果が各スレーブサーバー1502-1、1502-2から返ってくる。すなわち、最後の検索結果が返ってくるまでの待ち時間が短縮化されるため、検索速度の向上を図ることができる。

30

【0125】

<実施の形態3>

つぎに、実施の形態3について説明する。実施の形態3は、実施の形態2の一部を改良した構成である。実施の形態2では、各アーカイブ200-Iの圧縮ファイル群を、その圧縮に用いられたハフマン木h-Iでそれぞれ伸長する工程と、伸長された検索対象ファイル群を共通ハフマン木Hで再圧縮する工程とを実行する。この伸長および再圧縮という2パス処理をおこなっておくことにより、各スレーブサーバー1502-Iにおいて、共通ハフマン木Hを用いて圧縮／伸長することができる構成としている。

40

【0126】

これに対し、実施の形態3では、ハフマン木の伸長処理を利用して、スレーブサーバー1502-Iごとのハフマン木の葉に埋め込まれている伸長文字と同一文字の葉を、共通ハフマン木Hから特定する。そして、共通ハフマン木Hの特定された葉に割り当てられた圧縮符号を、特定元となるハフマン木h-Iの葉の伸長文字のかわりに設定する。設定後のハフマン木を変換用ハフマン木と称す。

50

【0127】

そして、変換用ハフマン木を用いて、設定前のハフマン木h-Iで圧縮した圧縮ファイル群の伸長処理を施すことにより、圧縮ファイル群が、共通ハフマン木Hの圧縮符号に対応する圧縮ファイル群に変換される。このように、1回の変換処理という1パス処理により、圧縮形式のまま共通ハフマン木Hの圧縮符号に対応する圧縮ファイル群をスレーブサーバー1502-Iごとに得ることができる。したがって、マスターサーバー1501における再構成処理の高速化を実現することができる。

【0128】

なお、上述した実施の形態1、2と同一構成には同一符号を付し、その説明を省略する。また、実施の形態3においても、説明の簡略化のため、マスターサーバー1502-Iの数を2台(N=2)として説明する。

10

【0129】

図50は、変換用ハフマン木の生成例(その1)を示す説明図である。図50では、アーカイブ200-1のハフマン木h-1と共通ハフマン木Hとを用いた生成例を示している。ここで、ハフマン木h-1の文字「b」が設定されている葉に注目する。まず、この注目葉の文字「b」と同一文字が設定されている共通ハフマン木の葉を特定する(ステップS5001)。つぎに、特定された共通ハフマン木の葉の文字「b」に割り当てられている圧縮符号「110」およびその圧縮符号長(この例では3)を読み出す(ステップS5002)。そして、注目葉に設定されている文字「b」に代えて、読み出された圧縮符号およびその圧縮符号長「110(3)」を書き込む(ステップS5003)。他の文字a, c~fについても同様に変換する。

20

【0130】

図51は、変換用ハフマン木の生成例(その2)を示す説明図である。図51では、アーカイブ200-2のハフマン木h-2と共通ハフマン木Hとを用いた生成例を示している。ここで、ハフマン木h-2の文字「f」が設定されている葉に注目する。まず、この注目葉の文字「f」と同一文字が設定されている共通ハフマン木の葉を特定する(ステップS5101)。つぎに、特定された共通ハフマン木の葉の文字「f」に割り当てられている圧縮符号「1110」およびその圧縮符号長(この例では4)を読み出す(ステップS5102)。そして、注目葉に設定されている文字「f」に代えて、読み出された圧縮符号およびその圧縮符号長「1110(4)」を書き込む(ステップS5103)。他の文字a~eについても同様に変換する。

30

【0131】

図52は、変換用ハフマン木を示す説明図(その1)である。この変換用ハフマン木H1は、図50に示したアーカイブ200-1のハフマン木h-1と共通ハフマン木Hとを用いた生成処理により生成されたハフマン木である。この変換用ハフマン木H1を用いて、たとえば、圧縮符号「1110」が割り当てられた伸長文字に伸長しようとする、図50に示した文字「b」に伸長されるのではなく、文字「b」の代わりに設定された圧縮符号「110」に変換される。したがって、圧縮ファイルの伸長および再圧縮という2パス処理が不要となり、圧縮形式のまま1パス処理により、共通ハフマン木Hで圧縮/伸長可能な圧縮ファイルを得ることができる。

40

【0132】

図53は、変換用ハフマン木を示す説明図(その2)である。この変換用ハフマン木H2は、図51に示したアーカイブ200-2のハフマン木h-2と共通ハフマン木Hとを用いた生成処理により生成されたハフマン木である。この変換用ハフマン木H2を用いて、たとえば、圧縮符号「1111」が割り当てられた伸長文字に伸長しようとする、図51に示した文字「f」に伸長されるのではなく、文字「f」の代わりに設定された圧縮符号「1110」に変換される。したがって、圧縮ファイルの伸長および再圧縮という2パス処理が不要となり、圧縮形式のまま1パス処理により、共通ハフマン木Hで圧縮/伸長可能な圧縮ファイルを得ることができる。

【0133】

50

(マスターサーバー 1501 の機能的構成)

図 54 は、実施の形態 3 にかかるマスターサーバー 1501 (情報管理装置) の機能的構成を示すブロック図である。マスターサーバー 1501 は、実施の形態 2 で説明した構成のほか、選択部 5401 と、特定部 5402 と、設定部 5403 と、変換部 5404 と、を含む構成である。選択部 5401 ~ 変換部 5404 は、具体的には、たとえば、図 1 に示した ROM 102、RAM 103、磁気ディスク 105 などの記憶領域に記憶されたプログラムを CPU 101 に実行させることによって、または、I/F 109 によってその機能を実現する。

【0134】

選択部 5401 は、アーカイブ 200-I ごとに、当該アーカイブ 200-I 内の圧縮ファイル群の圧縮に用いられたハフマン木 h-I から任意の葉を順次選択する機能を有する。具体的には、たとえば、図 50 および図 51 に示したハフマン木 h-I の葉を順次選択する。

10

【0135】

特定部 5402 は、選択部 5401 によって順次選択された葉で伸長される文字と同一文字の葉を、共通ハフマン木 H から特定する機能を有する。具体的には、図 50 および図 51 で示したように、同一文字が設定されている共通ハフマン木 H の葉を特定する。

【0136】

設定部 5403 は、ハフマン木 h-I において選択された葉に、当該選択された葉で伸長される文字に代えて、特定部 5402 によって特定された葉に割り当てられた圧縮符号を設定する機能を有する。具体的には、選択された葉の構造体中、伸長文字が書き込まれている領域に、共通ハフマン木 H から特定された圧縮符号を上書きする。また、その他の空き領域に圧縮符号長を書き込んでおく。なお、この設定対象となる葉の構造体では、上位節点へのポインタはそのままであるため、その選択された葉に割り当てられた圧縮符号を指定することで、葉の構造体書き込まれた圧縮符号に変換される。

20

【0137】

変換部 5404 は、変換用ハフマン木 H1、H2 を用いて、その変換前のハフマン木 h-1、h-2 で圧縮された圧縮ファイルを圧縮形式のまま変換する機能を有する。これにより、共通ハフマン木 H で圧縮/伸長可能な圧縮ファイル群 g a、g b となる。なお、変換後の圧縮ファイル群 g a、g b は、実施の形態 2 と同様、統合部 4711 により統合され、割当部 4712 によりアクセス頻度またはその順位の合計が均等になるように、スレーブサーバー 1502-1、1502-2 ごとに割り当てられる。

30

【0138】

(再構成処理手順)

つぎに、実施の形態 3 におけるマスターサーバー 1501 によるアーカイブ再構成処理手順について説明する。図 55 は、マスターサーバー 1501 によるアーカイブ再構成処理手順(後半)を示すフローチャートである。なお、再構成処理手順(前半)は、図 48 と同一であるため、説明を省略する。

【0139】

まず、ハフマン木 h-1、h-2 への圧縮符号設定処理を実行する(ステップ S5501)。つぎに、変換部 5404 により、1 パス変換処理を実行する(ステップ S5502)。1 パス変換処理は、アーカイブ 200-1、200-2 ごとに得られた変換用ハフマン木 H1、H2 を用いて、アーカイブ 200-1、200-2 内の圧縮ファイル群 g a、g b ごとに実行する。

40

【0140】

そして、統合部 4711 により、変換後の圧縮ファイル群 g a、g b を統合し(ステップ S5503)、アクセス頻度の高い順(降順)にソートする(ステップ S5504)。このあと、割当部 4712 により、アクセス頻度またはその順位の合計が均等になるように、各スレーブサーバー 1502-1、1502-2 に割り当てる(ステップ S5505)。

50

【0141】

そして、アーカイブ生成部4713により、スレーブサーバー1502-1, 1502-2ごとの新アーカイブA1, A2を生成して(ステップS5506)、新アーカイブA1(A2)および共通ハフマン木Hを各新アーカイブA1(A2)の割当先となるスレーブサーバー1502-1(1502-2)に配信する(ステップS5507)。これにより、一連の処理を終了する。

【0142】

図56は、ハフマン木への圧縮符号設定処理の詳細な処理手順を示すフローチャートである。まず、選択部5401により、各アーカイブ200-1, 200-2のハフマン木h-1, h-2の中から、未処理のハフマン木があるか否かを判断する(ステップS5601)。未処理のハフマン木がある場合(ステップS5601: Yes)、未処理のハフマン木を1つ選択する(ステップS5602)。そして、選択ハフマン木の中に未処理の葉があるか否かを判断する(ステップS5603)。

10

【0143】

未処理の葉がある場合(ステップS5603: Yes)、未処理の葉を1つ選択する(ステップS5604)。そして、特定部5402により、共通ハフマン木Hから選択葉に設定されている文字と同一文字が設定されている葉を特定する(ステップS5605)。このあと、設定部5403により、特定された葉に割り当てられた圧縮符号とその圧縮符号長を、選択葉の構造体に設定して(ステップS5606)、ステップS5603に戻る。

20

【0144】

ステップS5603において、未処理の葉がない場合(ステップS5603: No)、ステップS5601に戻る。ステップS5601において、未処理のハフマン木がない場合(ステップS5601: No)、1パス変換処理(ステップS5502)に移行する。

【0145】

このように、実施の形態3によれば、ハフマン木の伸長処理という既存の処理を利用することで、1回の変換処理という1パス処理により、圧縮形式のまま共通ハフマン木の圧縮符号に対応する圧縮ファイル群をスレーブサーバー1502-1, 1502-2ごとに得ることができる。したがって、共通ハフマン木Hに適用させるためのファイルオープン処理(伸長処理)が不要となり、マスターサーバー1501におけるアーカイブ再構成処理の高速化を実現することができる。また、ハフマン木の伸長処理という既存の処理を応用しているため、あらたなアルゴリズムを構築することなく、シンプルな構成でアーカイブ200再構成処理を実現することができる。

30

【0146】

以上のことから、上述した実施の形態1~3によれば、圧縮ファイルに対して圧縮形式のままで全文検索を実現する際に、ファイルアクセス処理の高速化を図ることにより検索処理の効率化を実現することができる。また、キャッシュ領域による常駐メモリ化や各サーバー資源の効率化により、全文検索の高速化を実現することができる。

【0147】

なお、本実施の形態で説明した方法は、予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等のコンピュータで実行することにより実現することができる。このプログラムは、ハードディスク、フレキシブルディスク、CD-ROM、MO、DVD等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、インターネット等のネットワークを介して配布することが可能な媒体であってもよい。

40

【0148】

上述した実施の形態に関し、さらに以下の付記を開示する。

【0149】

(付記1) 文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブにアクセス可能なコンピュータを、

50

前記各圧縮ファイルへのアクセス頻度の高い順に並び替える並び替え手段、

前記圧縮ファイル群が記憶されている記憶領域に対するキャッシュ領域の記憶容量を超えないように、前記並び替え手段による並び替え後のアクセス頻度の高い順から前記圧縮ファイルのサイズを積算する積算手段、

前記積算手段によって積算された圧縮ファイル群を、当該ファイル群の検索に先立って、前記記憶領域からキャッシュ領域に書き込む書込み手段、

として機能させることを特徴とする情報検索プログラム。

【0150】

(付記2) 前記積算手段は、

前記キャッシュ領域の記憶容量を超えない最大の積算値となるように、前記並び替え手段による並び替え後のアクセス頻度の高い順から前記圧縮ファイルのサイズを積算し、

前記書込み手段は、

前記最大の積算値として積算された圧縮ファイル群を、当該ファイル群の検索に先立って、前記記憶領域からキャッシュ領域に書き込む書込み手段、

として機能させることを特徴とする情報検索プログラム。

【0151】

(付記3) 前記コンピュータを、

前記書込み手段によって前記キャッシュ領域に書き込まれた圧縮ファイル群について常駐フラグを設定する設定手段、

検索文字列の入力を受け付ける入力手段、

前記圧縮ファイルごとの文字の存否を特定する文字出現マップを参照することにより、前記入力手段によって入力された検索文字列を構成するすべての文字が存在する圧縮ファイルを特定する特定手段、

前記特定手段によって特定された圧縮ファイルを、前記設定手段によって設定された常駐フラグに基づく領域から読み出す読出し手段、

前記読出し手段によって読み出された圧縮ファイルを伸長する伸長手段、

前記伸長手段によって伸長された検索対象ファイルを出力する出力手段、

として機能させることを特徴とする付記1または2に記載の情報検索プログラム。

【0152】

(付記4) 前記コンピュータを、

前記伸長手段によって伸長された検索対象ファイルの中から前記検索文字列に一致または関連する文字列を検索する検索手段として機能させ、

前記出力手段は、

前記検索手段によって検索された検索結果を出力することを特徴とする付記3に記載の情報検索プログラム。

【0153】

(付記5) 前記コンピュータを、

前記伸長手段によって前記圧縮ファイルが伸長された場合、当該圧縮ファイルへのアクセス頻度を更新する更新手段として機能させ、

前記並び替え手段は、

前記更新手段によって更新された更新後のアクセス頻度に基づいて、前記各圧縮ファイルへのアクセス頻度の高い順に並び替えることを特徴とする付記3または4に記載の情報検索プログラム。

【0154】

(付記6) 前記コンピュータを、

前記検索手段によって検索対象ファイルが検索された場合、当該検索対象ファイルが圧縮された圧縮ファイルへのアクセス頻度を更新する更新手段として機能させ、

前記並び替え手段は、

前記更新手段によって更新された更新後のアクセス頻度に基づいて、前記各圧縮ファイルへのアクセス頻度の高い順に並び替えることを特徴とする付記4に記載の情報検索プロ

10

20

30

40

50

グラム。

【0155】

(付記7) 文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバーにアクセス可能なマスターサーバーを、

前記各スレーブサーバーから前記アーカイブを受信する受信手段、

前記受信手段によって受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成する圧縮パラメータ生成手段、

前記圧縮パラメータ生成手段によって生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成する共通ハフマン木生成手段、

前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記圧縮ファイルを割り当てる割当手段、

前記割当手段によって前記各スレーブサーバーに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通ハフマン木生成手段によって生成された共通のハフマン木とを、割当先となるスレーブサーバーに送信する送信手段、

として機能させることを特徴とする情報管理プログラム。

【0156】

(付記8) 前記マスターサーバーを、

前記アーカイブごとに、当該アーカイブに含まれている圧縮ファイル群を、当該圧縮ファイル群の圧縮に用いられた圧縮パラメータに基づくハフマン木により伸長する伸長手段、

前記伸長手段によって伸長された検索対象ファイル群を、前記共通のハフマン木により圧縮する圧縮手段として機能させ、

前記割当手段は、

前記圧縮手段によって圧縮された再圧縮ファイルの元となる圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記再圧縮ファイルを割り当て、

前記送信手段は、

前記再圧縮ファイル群を含むあらたなアーカイブと前記共通のハフマン木とを、前記割当先となるスレーブサーバーに送信することを特徴とする付記7に記載の情報管理プログラム。

【0157】

(付記9) 前記マスターサーバーを、

前記アーカイブごとに、当該アーカイブ内の圧縮ファイル群の圧縮に用いられたハフマン木から任意の葉を順次選択する選択手段、

前記選択手段によって順次選択された葉で伸長される文字と同一文字の葉を、前記共通のハフマン木から特定する特定手段、

前記ハフマン木における前記選択された葉に、当該選択された葉で伸長される文字に代えて、前記特定手段によって特定された葉に割り当てられた圧縮符号を設定する設定手段、

前記設定手段による設定後のハフマン木を用いて、前記アーカイブ内の圧縮ファイル群を、前記設定手段によって設定された圧縮符号で圧縮された圧縮ファイル群に変換する変換手段として機能させ、

前記割当手段は、

前記変換手段によって変換された変換後の圧縮ファイルの変換元となる圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記変換後の圧縮ファイルを割り当て、

前記送信手段は、

前記変換手段による変換後の圧縮ファイル群を含むアーカイブと前記共通のハフマン木

10

20

30

40

50

とを、前記割当先となるスレーブサーバーに送信することを特徴とする付記 7 に記載の情報管理プログラム。

【0158】

(付記 10) 前記割当手段は、

前記アクセス頻度の順位の合計が互いに近似するように割り当てることを特徴とする付記 7～9 のいずれか一つに記載の情報管理プログラム。

【0159】

(付記 11) 文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブにアクセス可能な情報検索装置であって、

前記各圧縮ファイルへのアクセス頻度の高い順に並び替える並び替え手段と、

前記圧縮ファイル群が記憶されている記憶領域に対するキャッシュ領域の記憶容量を超えないように、前記並び替え手段による並び替え後のアクセス頻度の高い順から前記圧縮ファイルのサイズを積算する積算手段と、

前記積算手段によって積算された圧縮ファイル群を、当該ファイル群の検索に先立って、前記記憶領域からキャッシュ領域に書き込む書き込み手段と、

を備えることを特徴とする情報検索装置。

【0160】

(付記 12) 文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバーにアクセス可能な情報管理装置であって、

前記各スレーブサーバーから前記アーカイブを受信する受信手段と、

前記受信手段によって受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成する圧縮パラメータ生成手段と、

前記圧縮パラメータ生成手段によって生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成する共通ハフマン木生成手段と、

前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記圧縮ファイルを割り当てる割当手段と、

前記割当手段によって前記各スレーブサーバーに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通ハフマン木生成手段によって生成された共通のハフマン木とを、割当先となるスレーブサーバーに送信する送信手段と、

を備えることを特徴とする情報管理装置。

【0161】

(付記 13) 文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブにアクセス可能な情報検索装置が、

前記各圧縮ファイルへのアクセス頻度の高い順に並び替える並び替え工程と、

前記圧縮ファイル群が記憶されている記憶領域に対するキャッシュ領域の記憶容量を超えないように、前記並び替え工程による並び替え後のアクセス頻度の高い順から前記圧縮ファイルのサイズを積算する積算工程と、

前記積算工程によって積算された圧縮ファイル群を、当該ファイル群の検索に先立って、前記記憶領域からキャッシュ領域に書き込む書き込み工程と、

を実行することを特徴とする情報検索方法。

【0162】

(付記 14) 文字列が記述された検索対象ファイルが圧縮された圧縮ファイル群を含むアーカイブを記憶する複数のスレーブサーバーにアクセス可能なマスターサーバーが、

前記各スレーブサーバーから前記アーカイブを受信する受信工程と、

前記受信工程によって受信されたアーカイブごとの前記検索対象ファイル群に記述されている各文字および当該各文字の出現頻度に関する圧縮パラメータに基づいて、前記出現頻度を前記文字ごとに集計することにより、前記各圧縮ファイル群に共通の圧縮パラメータを生成する圧縮パラメータ生成工程と、

前記圧縮パラメータ生成工程によって生成された共通の圧縮パラメータに基づいて、前記各圧縮ファイル群に共通のハフマン木を生成する共通ハフマン木生成工程と、

前記圧縮ファイルへのアクセス頻度の合計が互いに近似するように、前記各スレーブサーバーに前記圧縮ファイルを割り当てる割当工程と、

前記割当工程によって前記各スレーブサーバーに割り当てられた圧縮ファイル群を含むあらたなアーカイブと前記共通ハフマン木生成工程によって生成された共通のハフマン木とを、割当先となるスレーブサーバーに送信する送信工程と、

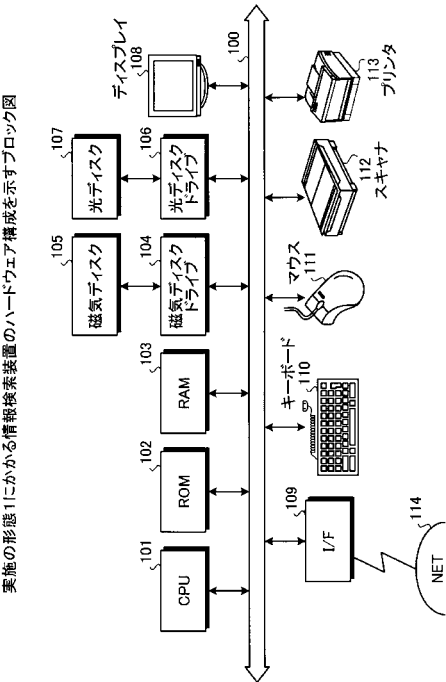
を実行することを特徴とする情報管理方法。

【符号の説明】

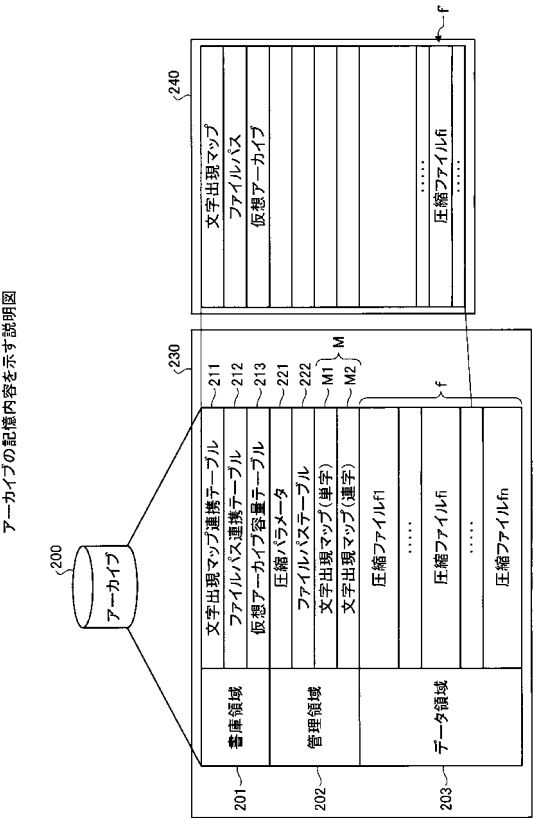
【0163】

1100	情報検索装置	
1101	並び替え処理部	
1102	積算部	
1103	書込み部	
1104	設定部	
1105	入力部	
1106	特定部	
1107	読出し部	
1108	伸長部	
1109	検索部	20
1110	出力部	
1111	更新部	
1500	検索システム	
1501	マスターサーバー	
1502	スレーブサーバー群	
4701	受信部	
4702	共通圧縮パラメータ生成部	
4703	共通ハフマン木生成部	
4704	伸長部	
4705	圧縮部	30
4706	再構成部	
4707	送信部	
4711	統合部	
4712	割当部	
4713	アーカイブ生成部	
5401	選択部	
5402	特定部	
5403	設定部	
5404	変換部	

【図 1】

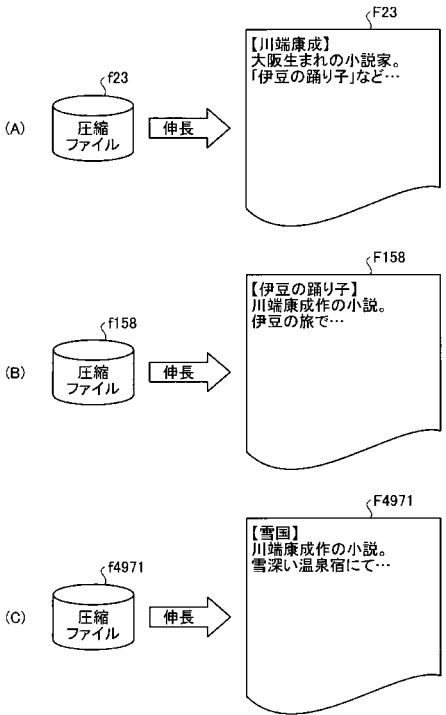


【図 2】



【図 3】

圧縮ファイルと検索対象ファイルとの関係を示す説明図



【図 4】

単字出現マップを示す説明図

文字出現マップ(単字)

文字種	文字	ビット番号									
		1	2	3	...	i	...	n			
数字	0	1	1	1	...	1	...	1			
	...	...	...	...	...	...	...	...			
アルファベット 小文字	9	1	1	1	...	1	...	1			
	a	1	1	1	...	1	...	1			
アルファベット 大文字	...	...	...	...	...	...	...	...			
	z	0	0	1	...	0	...	1			
かな	A	1	1	1	...	1	...	1			
	...	...	...	...	...	...	...	...			
カタカナ	Z	0	0	0	...	1	...	0			
	あ	1	1	1	...	1	...	1			
漢字	...	...	...	...	...	...	...	...			
	ん	1	1	1	...	1	...	1			
	ア	1	1	1	...	1	...	1			
	...	...	...	...	...	...	...	...			
	ン	1	1	1	...	1	...	1			
	足	0	0	0	...	0	...	0			
	...	...	...	...	...	...	...	...			
	吾	0	0	0	...	0	...	0			

【図 5】

連字出現マップを示す説明図

文字種	文字出現マップ(連字)									
	連字	ビット番号								
		1	2	3	...	i	...	n		
数字	00	0	0	0	...	0	...	0		
	...	...	...	...	...	...	...	...		
	99	0	1	0	...	0	...	1		
アルファベット 小文字	aa	0	0	0	...	0	...	0		
	...	...	...	...	...	...	...	...		
	zz	0	0	0	...	0	...	0		
アルファベット 大文字	AA	0	0	0	...	0	...	0		
	...	...	...	...	...	...	...	...		
	ZZ	0	0	0	...	0	...	0		
かな	ああ	0	0	0	...	0	...	0		
	...	...	...	...	...	...	...	...		
	んん	0	0	0	...	0	...	0		
カタカナ	アア	0	0	0	...	0	...	0		
	...	...	...	...	...	...	...	...		
	ンン	0	0	0	...	0	...	0		
漢字	足足	0	0	0	...	0	...	0		
	...	...	...	...	...	...	...	...		
	吾吾	0	0	0	...	0	...	0		

【図 6】

圧縮パラメータを示す説明図

圧縮パラメータ	
文字/連字	出現回数
0	98657
...	...
9	35726
a	45989
...	...
z	16572
A	5782
...	...
Z	6586
あ	85794
...	...
ん	28580
ア	65237
...	...
ン	29116
足	5871
...	...
吾	528
00	3895
...	...
99	369
...	...

【図 7】

ファイルバーステーブルを示す説明図

ファイルID	ファイルバーステーブル			
	ファイルパス	見出し語	アドレス	サイズ
f1	honmon\file1.html	あ	adr1	s1
...	...	...	...	...
f23	honmon\file23.html	川端康成	adr23	s23
...	...	...	...	...
f158	honmon\file158.html	伊豆の踊り子	adr158	s158
...	...	...	...	...
fi	honmon\filei.html	...	adri	si
...	...	...	...	...
f4971	honmon\file4971.html	雪国	adr4971	s4971
...	...	...	...	...
fn	honmon\filen.html	ん	adrn	sn

【図 8】

文字出現マップ連携テーブルを示す説明図

文字出現マップ連携テーブル				
ビット番号	アドレス	サイズ	アクセス頻度	常駐フラグ
1	adr1	s1	57	1
...	...	...	...	...
i	adri	si	103	1
...	...	...	...	...
n	adrn	sn	8	0

【図 9】

ファイルパス連携テーブルを示す説明図

212

ファイルパス連携テーブル	
ビット番号	ファイルパス
1	honmon¥file1.html
...	...
i	honmon¥filei.html
...	...
n	honmon¥filen.html

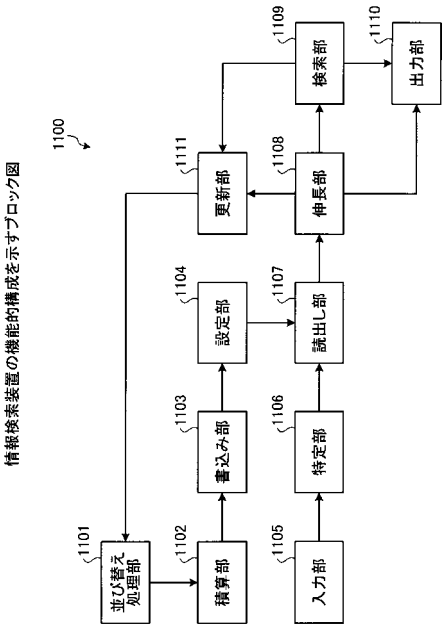
【図 10】

仮想アーカイブ容量テーブルを示す説明図

213

仮想アーカイブ容量テーブル		
ビット番号	アクセス頻度	サイズ
3015	241	s3015
867	239	s867
...	...	...
54	92	s54

【図 11】



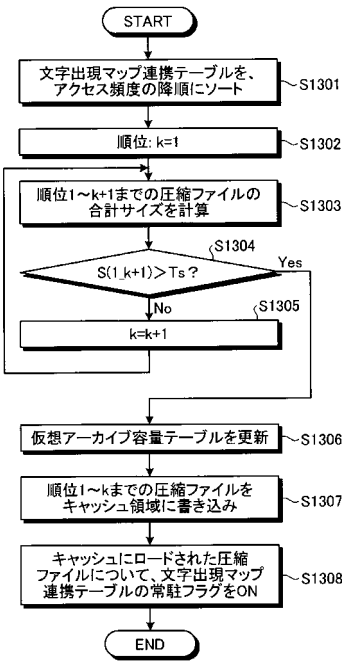
【図 12】

文字出現マップを用いた圧縮ファイルの絞込み例を示す説明図

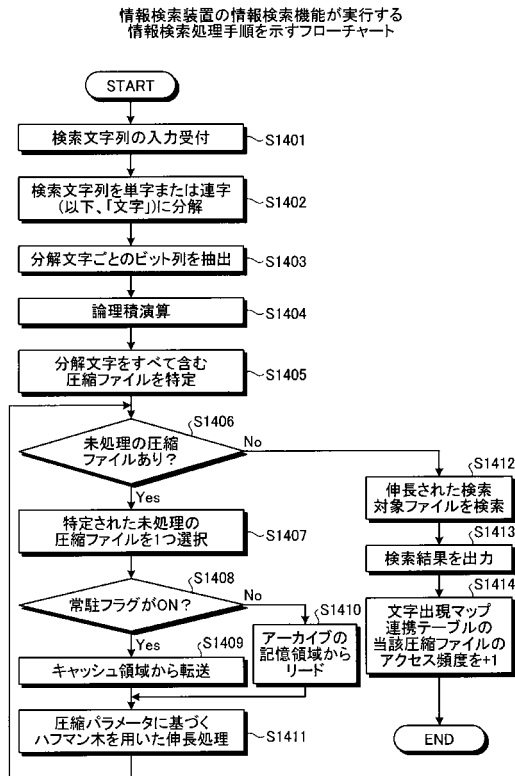
文字	ビット番号									
	1	...	23	...	158	...	4971	...	n	
川	1	...	1	...	1	...	1	...	0	
端	1	...	1	...	1	...	1	...	0	
康	1	...	1	...	1	...	1	...	0	
成	1	...	1	...	1	...	1	...	0	
論理積	1	...	1	...	1	...	1	...	0	

【図 13】

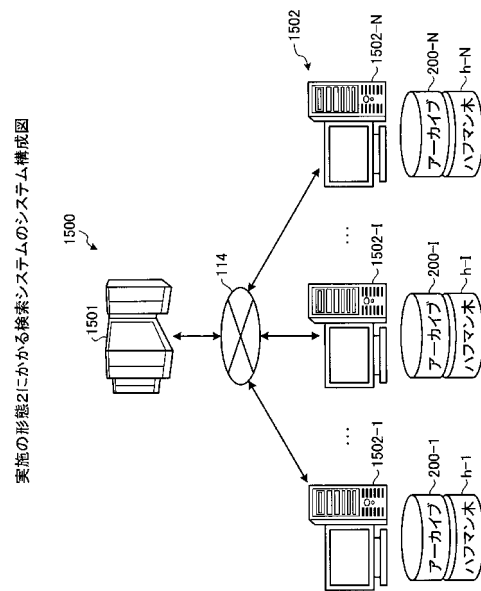
情報検索装置の仮想アーカイブ設定機能が実行する仮想アーカイブ設定処理手順を示すフローチャート



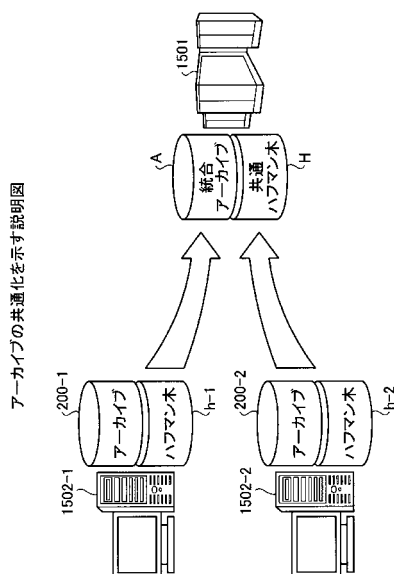
【図 14】



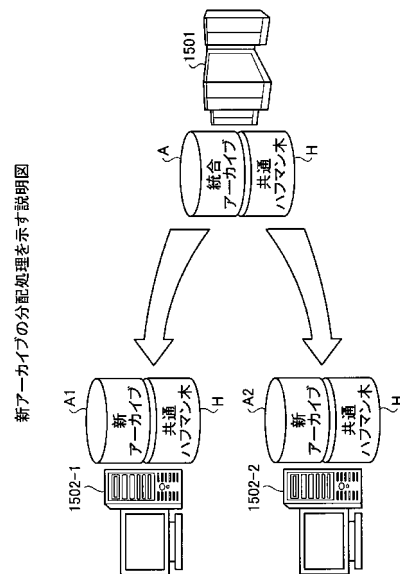
【図 15】



【図 16】



【図 17】



【図 18】

アーカイブ200-1の圧縮符号テーブルおよび圧縮パラメータを示す説明図

(A)

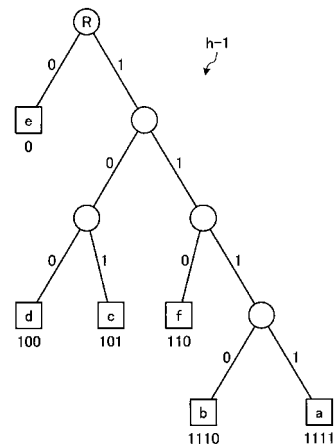
圧縮符号	区分	文字
0	8ビット	e
100	16ビット	d
101	16ビット	c
110	16ビット	f
1110	16ビット	b
1111	16ビット	a

(B)

圧縮パラメータ	
出現回数	文字
32	e
31	d
30	c
23	f
13	b
3	a

【図 19】

アーカイブ200-1の圧縮符号テーブルから生成されるハフマン木



【図 20】

アーカイブ200-2の圧縮符号テーブルおよび圧縮パラメータを示す説明図

(A)

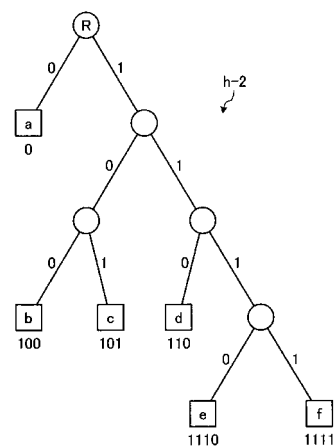
圧縮符号	区分	文字
0	16ビット	a
100	16ビット	b
101	16ビット	c
110	16ビット	d
1110	8ビット	e
1111	8ビット	f

(B)

圧縮パラメータ	
出現回数	文字
21	a
18	b
15	c
12	d
9	e
6	f

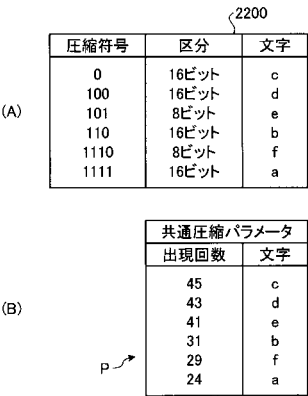
【図 21】

アーカイブ200-2の圧縮符号テーブルから生成されるハフマン木を示す説明図



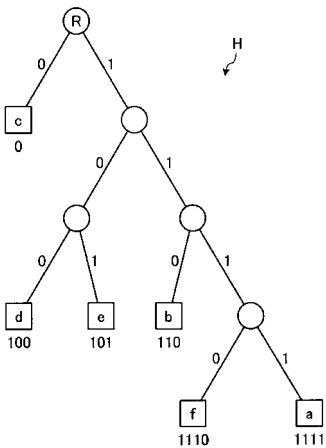
【図 2 2】

統合アーカイブの圧縮符号テーブルおよび圧縮パラメータを示す説明図



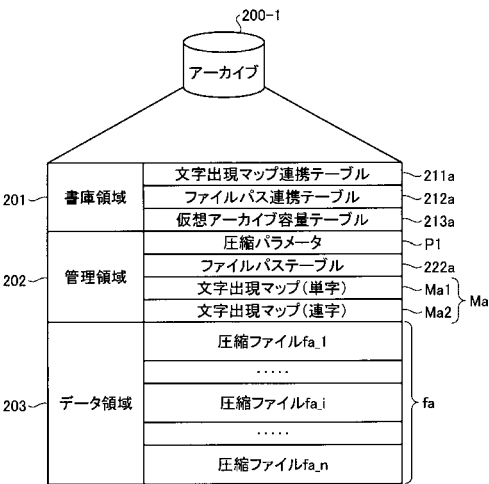
【図 2 3】

統合アーカイブの圧縮符号テーブルから生成される共通ハフマン木を示す説明図



【図 2 4】

アーカイブ200-1の記憶内容を示す説明図



【図 2 5】

アーカイブ200-1の単字出現マップを示す説明図

Ma1

文字出現マップ(単字)									
文字種	文字	ビット番号							
		a.1	a.2	a.3	...	a.i	...	a.n	
数字	0	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	9	1	1	1	...	1	...	1	
アルファベット 小文字	a	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	z	0	0	1	...	0	...	1	
アルファベット 大文字	A	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	Z	0	0	0	...	1	...	0	
かな	あ	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	ん	1	1	1	...	1	...	1	
カタカナ	ア	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	ン	1	1	1	...	1	...	1	
漢字	足	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	吾	0	0	0	...	0	...	0	

【図 26】

アーカイブ200-1の連字出現マップを示す説明図

文字出現マップ(連字)									
文字種	連字	ビット番号							
		a_1	a_2	a_3	...	a_i	...	a_n	
数字	00	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	99	0	1	0	...	0	...	1	
アルファベット 小文字	aa	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	zz	0	0	0	...	0	...	0	
アルファベット 大文字	AA	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	ZZ	0	0	0	...	0	...	0	
かな	ああ	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	んん	0	0	0	...	0	...	0	
カタカナ	アア	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	ンン	0	0	0	...	0	...	0	
漢字	足足	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	吾吾	0	0	0	...	0	...	0	

【図 27】

アーカイブ200-1の圧縮パラメータを示す説明図

圧縮パラメータ	
単字/連字	出現回数
0	98657
...	...
9	35726
a	45989
...	...
z	16572
A	5782
...	...
Z	6586
あ	85794
...	...
ん	28580
ア	65237
...	...
ン	29116
足	5871
...	...
吾	528
00	3895
...	...
99	369
...	...

【図 28】

アーカイブ200-1のファイルバスターテーブルを示す説明図

ファイルバスターテーブル				
ファイルID	ファイルパス	見出し語	アドレス	サイズ
fa_1	honmon\filea_1.html	あ	adra_1	sa_1
...	...	...	...	...
fa_23	honmon\filea_23.html	伊豆の踊り子	adra_23	sa_23
...	...	...	...	...
fa_158	honmon\filea_158.html	川端康成	adra_158	sa_158
...	...	...	...	...
fa_i	honmon\filea_i.html	...	adra_i	sa_i
...	...	...	...	...
fa_4971	honmon\filea_4971.html	雪国	adra_4971	sa_4971
...	...	...	...	...
fa_n	honmon\filea_n.html	ん	adra_n	sa_n

【図 29】

アーカイブ200-1の文字出現マップ連携テーブルを示す説明図

文字出現マップ連携テーブル				
ビット番号	アドレス	サイズ	アクセス頻度	常駐フラグ
a_i	adra_i	sa_i	57	1
...	...	...	...	...
a_j	adra_j	sa_j	103	1
...	...	...	...	...
a_n	adra_n	sa_n	8	0

【図 3 0】

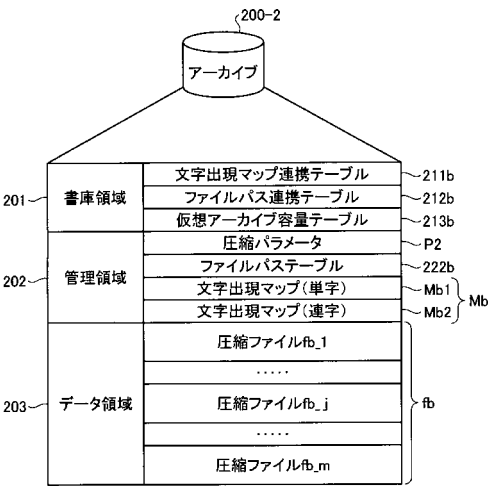
アーカイブ200-1のファイルパス連携テーブルを示す説明図

212a

ファイルパス連携テーブル	
ビット番号	ファイルパス
a.1	honmon¥filea.1.html
...	...
a.i	honmon¥filea.i.html
...	...
a.n	honmon¥filea.n.html

【図 3 1】

アーカイブ200-2の記憶内容を示す説明図



【図 3 3】

アーカイブ200-2の連字出現マップMb2を示す説明図

Mb2

文字出現マップ(連字)									
文字種	連字	ビット番号							
		b_1	b_2	b_3	...	b_j	...	b_m	
数字	00	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	99	0	1	0	...	0	...	1	
アルファベット 小文字	aa	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	zz	0	0	0	...	0	...	0	
アルファベット 大文字	AA	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	ZZ	0	0	0	...	0	...	0	
かな	ああ	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	んん	0	0	0	...	0	...	0	
カタカナ	アア	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	ンン	0	0	0	...	0	...	0	
漢字	足足	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	吾吾	0	0	0	...	0	...	0	

【図 3 2】

アーカイブ200-2の単字出現マップMb1を示す説明図

Mb1
↓

文字出現マップ(単字)									
文字種	文字	ビット番号							
		b_1	b_2	b_3	...	b_j	...	b_m	
数字	0	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	9	1	1	1	...	1	...	1	
アルファベット 小文字	a	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	z	0	0	1	...	0	...	1	
アルファベット 大文字	A	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	Z	0	0	0	...	1	...	0	
かな	あ	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	ん	1	1	1	...	1	...	1	
カタカナ	ア	1	1	1	...	1	...	1	
	...	...	...	...	...	...	...	...	
	ン	1	1	1	...	1	...	1	
漢字	足	0	0	0	...	0	...	0	
	...	...	...	...	...	...	...	...	
	吾	0	0	0	...	0	...	0	

【図 3 4】

アーカイブ200-2の圧縮パラメータP2を示す説明図

P2

圧縮パラメータ	
単字/連字	出現回数
0	57820
...	...
9	12879
a	226854
...	...
z	4780
A	6599
...	...
Z	3001
あ	77259
...	...
ん	41938
ア	52108
...	...
ン	10047
足	3180
...	...
吾	98
00	15983
...	...
99	74
...	...

【図 3 5】

アーカイブ200-2のファイルパステーブル222bを示す説明図

ファイルパステーブル				
ファイルID	ファイルパス	見出し語	アドレス	サイズ
fb_1	honmon%fileb_1.html	...	adfb_1	sb_1
...	...	...	...	...
fb_j	honmon%fileb_j.html	地球温暖化	adfb_j	sb_j
...	...	...	...	...
fb_m	honmon%fileb_m.html	...	adfb_m	sb_m

【図 3 6】

アーカイブ200-2の文字出現マップ連携テーブルを示す説明図

文字出現 マップ連携テーブル				
ビット番号	アドレス	サイズ	アクセス頻度	常駐フラグ
b_1	adfb_1	sb_1	954	1
...	...	...	...	...
b_j	adfb_j	sb_j	15	0
...	...	...	...	...
b_m	adfb_m	sb_m	1097	1

【図 3 7】

アーカイブ200-2のファイルパス連携テーブルを示す説明図

ファイルパス連携テーブル	
ビット番号	ファイルパス
b_1	honmon%fileb_1.html
...	...
b_j	honmon%fileb_j.html
...	...
b_m	honmon%fileb_m.html

【図 3 8】

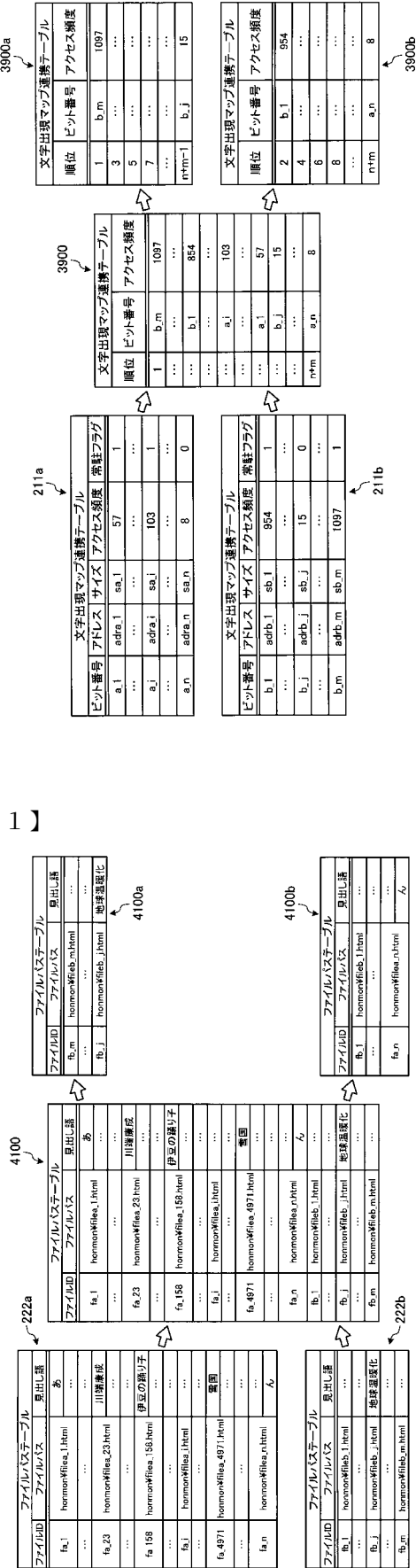
共通パラメータの生成例を示す説明図

圧縮パラメータ	
単字/連字	出現回数
0	98657
...	...
9	35726
a	45989
...	...
z	16572
A	5782
...	...
Z	6586
あ	85794
...	...
ん	28560
ア	65237
...	...
ン	29116
足	5871
...	...
書	528
00	3895
...	...
99	369
...	...

共通圧縮パラメータ	
単字/連字	出現回数
0	156477
...	...
9	48605
a	272843
...	...
z	21352
A	12381
...	...
Z	9587
あ	163053
...	...
ん	70518
ア	117345
...	...
ン	39183
足	9051
...	...
書	626
00	19878
...	...
99	443
...	...

【図 39】

文字出現マップ再構成を示す説明図

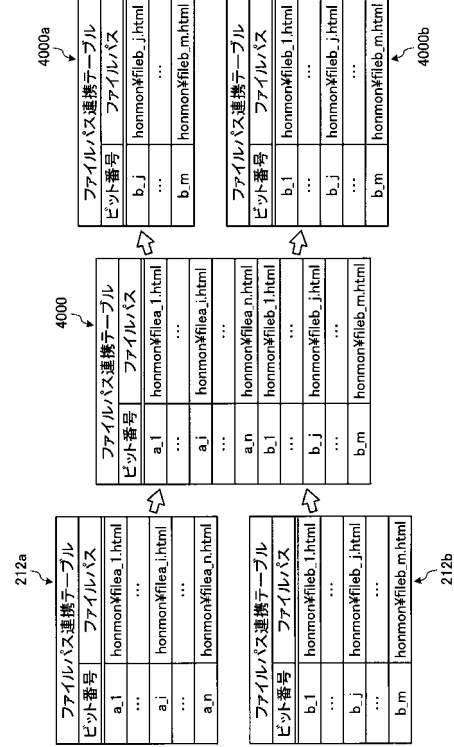


【図 41】

ファイルバスターテーブルの再構成を示す説明図

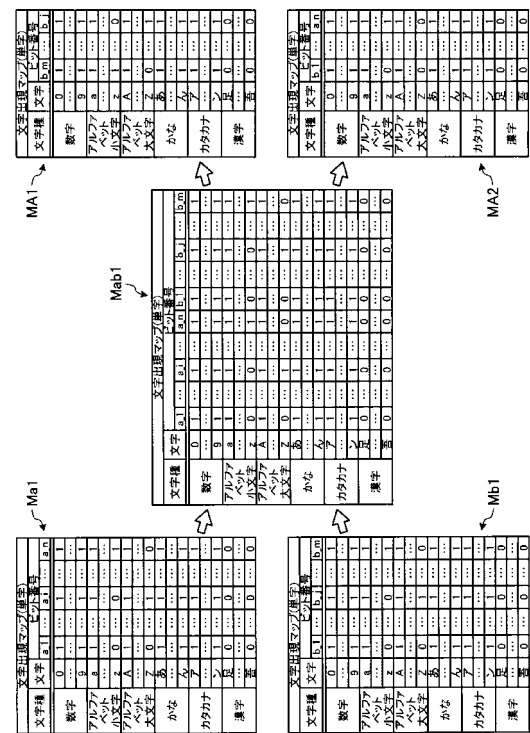
【図 40】

ファイルバスターテーブルの再構成を示す説明図

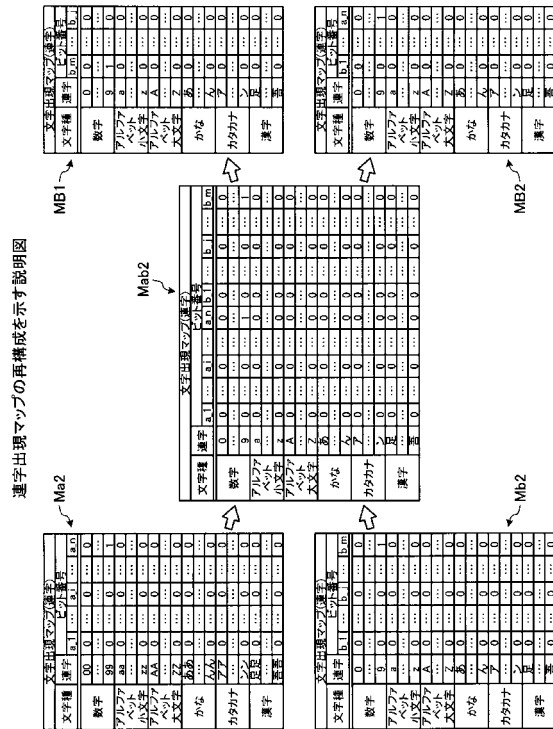


【図 42】

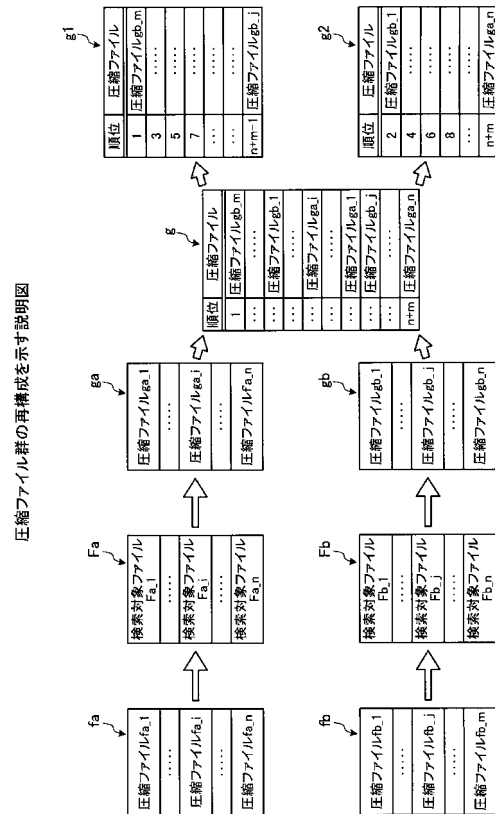
単語出現マップの再構成を示す説明図



【図 4 3】

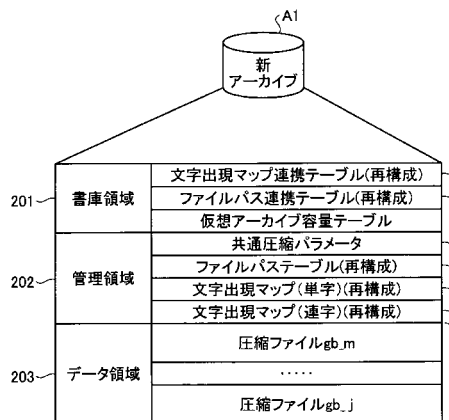


【图 4 4】



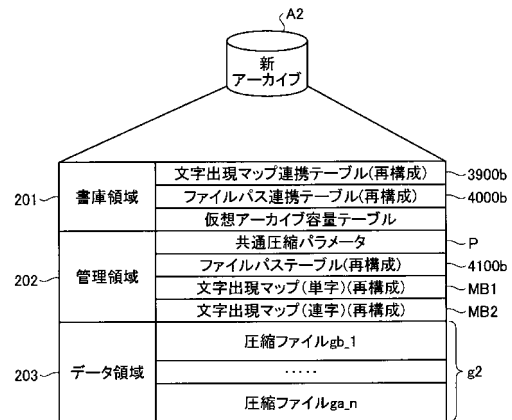
【図 4 5】

新アーカイブA1の記憶内容を示す説明図

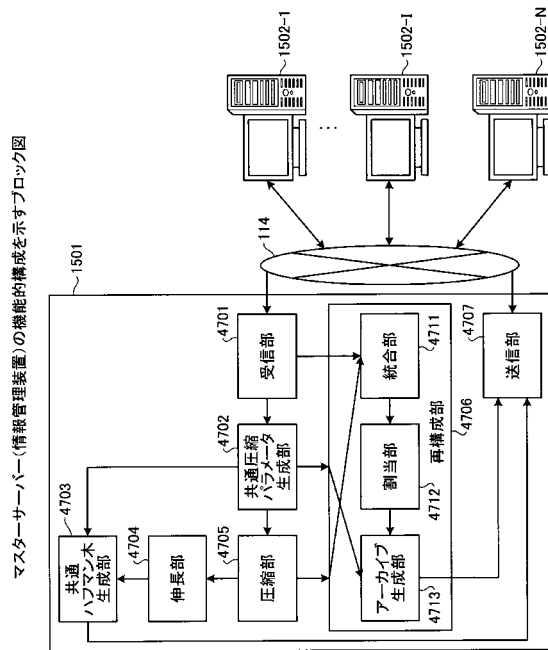


【図 4 6】

新アーカイブA2の記憶内容を示す説明図

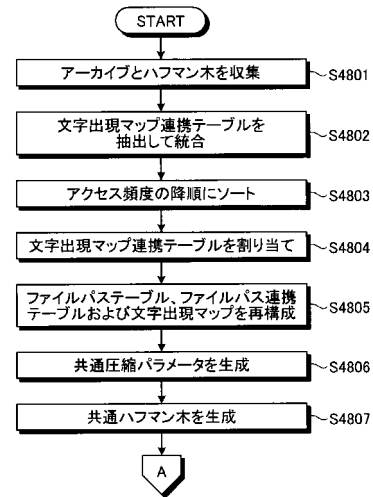


【図 47】



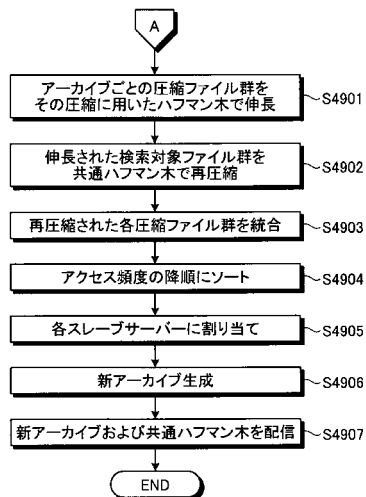
【図 48】

マスターサーバーによるアーカイブ再構成処理手順(前半)を示すフローチャート



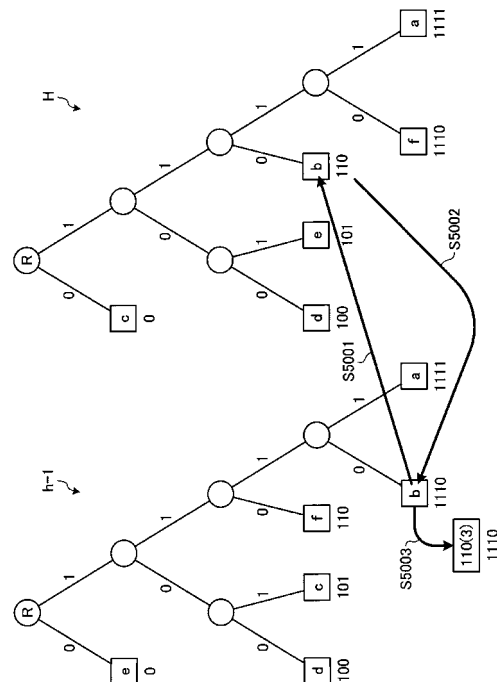
【図 49】

マスターサーバーによるアーカイブ再構成処理手順(後半)を示すフローチャート

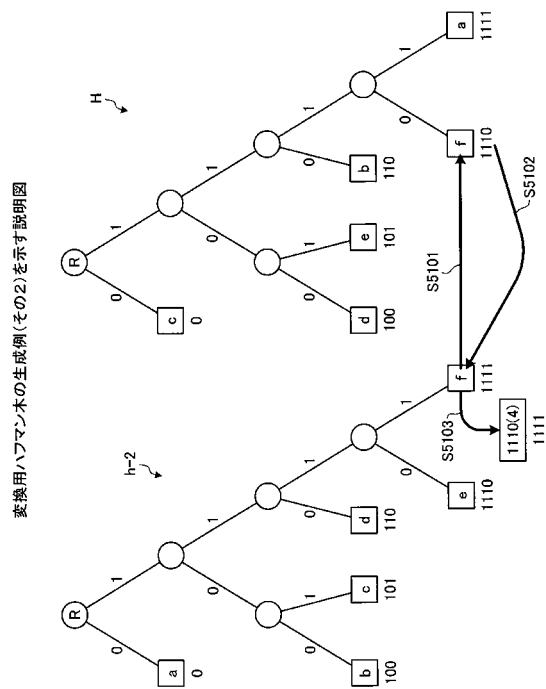


【図 50】

変換用ハフマン木の生成例(その1)を示す説明図

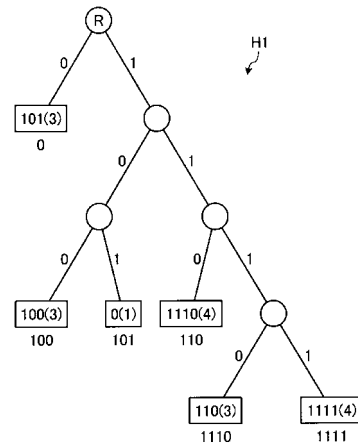


【図 5 1】



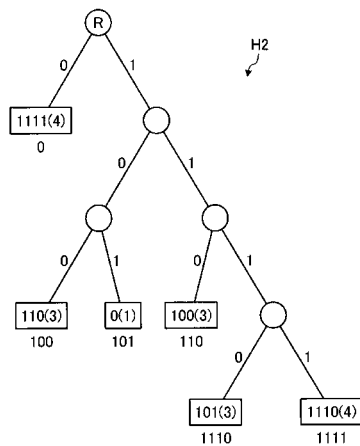
【図 5 2】

変換用ハフマン木を示す説明図(その1)



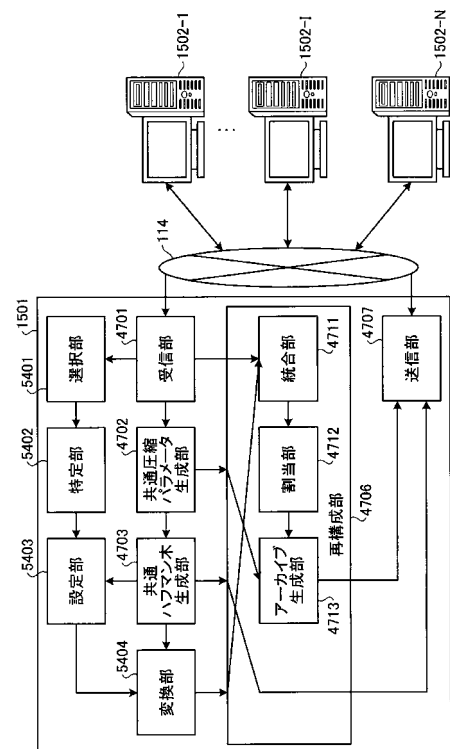
【図 5 3】

変換用ハフマン木を示す説明図(その2)



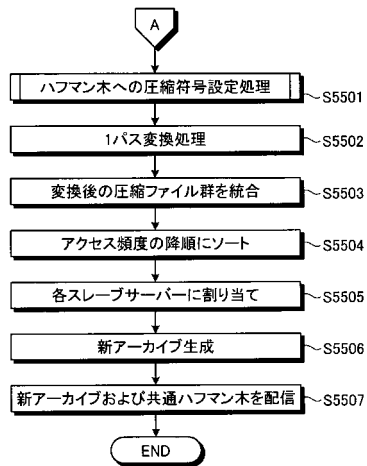
【図 5 4】

実施の形態3にかかるマスターサーバ1501(情報管理装置)の機能的構成を示すブロック図



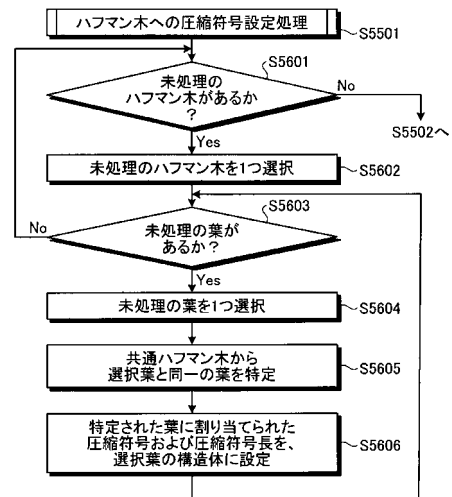
【図 55】

実施の形態3におけるマスターサーバーによる
アーカイブ再構成処理手順(後半)を示すフローチャート



【図 56】

ハフマン木への圧縮符号設定処理の詳細な処理手順を示すフローチャート



フロントページの続き

審査官 萩島 豪

(56)参考文献 国際公開第2006/123448 (WO, A1)
国際公開第2008/057815 (WO, A2)
特開平04-274557 (JP, A)

(58)調査した分野(Int.Cl., DB名)
G06F 12/00
G06F 17/30