【書類名】　特許願
【整理番号】　CP00744
【あて先】　特許庁長官殿
【国際特許分類】　G06F 3/06
【発明者】
　【住所又は居所】　大韓民国　７９０－７５１　ギョンサンブクド　ポハンシ　ナムグ　ジゴクドン　ギョス・アパート　９－１５０３
　【氏名】　パク　チャンイク
【発明者】
　【住所又は居所】　大韓民国　７０２－７８８　デグ　ブクグ　テジョンドン　ヒョップファ・アパート　１０３－１５０６
　【氏名】　パク　セジン
【特許出願人】
　【識別番号】　511258411
　【氏名又は名称】　ポハン工科大学校産学協力団
　【氏名又は名称原語表記】　ＰＯＳＴＥＣＨ　Ａｃａｄｅｍｙ－Ｉｎｄｕｓｔｒｙ　Ｆｏｕｎｄａｔｉｏｎ
【代理人】
　【識別番号】　100121728
　【弁理士】
　【氏名又は名称】　井関　勝守
　【電話番号】　06-6136-6951
　【ファクシミリ番号】　06-6136-6952
　【連絡先】　担当
【パリ条約による優先権等の主張】
　【国名】　大韓民国
　【出願日】　2013年 3月 7日
　【出願番号】　10-2013-0024356
【手数料の表示】
　【振替番号】　00016713
　【納付金額】　15,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　特許請求の範囲　1
　【物件名】　要約書　1
　【物件名】　図面　1
　【包括委任状番号】　1112740

【書類名】明細書
【発明の名称】データ重複除去方法及び装置
【技術分野】
【０００１】
本発明は、データ重複除去技術に関し、より詳細には、低い入出力レイテンシ(ｌａｔｅｎｃｙ)を提供するためのデータ重複除去方法及び装置に関する。
【背景技術】
【０００２】
データ重複除去技術は、データ保存装置内に重複されたデータを除去してより多い保存空間を確保するための技術を意味する。現在、多い企業、公共機関などでデータの安全な保管のためにデータのバックアップ(ｂａｃｋｕｐ)を周期的に実行している。バックアップデータは、その特性上多い重複的な要素を有し、これによって、バックアップデータの保存空間の効率を向上させるためにデータ重複除去技術が使われている。このようなデータ重複除去技術は、バックアップデータ保存装置の特性上、低い入出力レイテンシ(ｌａｔｅｎｃｙ)を必要としないので、重複除去率を高める技術を中心として発展している。
【０００３】
しかし、このようなデータ重複除去技術は、重複除去のための複雑なアルゴリズム(ａｌｇｏｒｉｔｈｍ)を基礎で実行されるため、ノート型ＰＣ(ｎｏｔｅｂｏｏｋ)、スマートホン(ｓｍａｒｔｐｈｏｎｅ)、タブレット型(ｔａｂｌｅｔ)ＰＣなどのような携帯用端末に適用しにくい問題点がある。すなわち、このような携帯用端末にデータ重複除去技術を適用する場合、順次的に保存されたデータの物理的手順が変わるようになるのでデータの入出力速度が深刻に遅くなる問題点がある。
【発明の概要】
【発明が解決しようとする課題】
【０００４】
したがって、前記のような従来の諸問題点を解消するために提案されたものであって、本発明の目的は、データの入出力特性に応じて適応的に決定された重複除去率を基盤としてデータの重複を除去するためのデータ重複除去方法を提供することにある。
【０００５】
本発明の他の目的は、データの入出力特性に応じて適応的に決定された重複除去率を基盤としてデータの重複を除去するためのデータ重複除去装置を提供することにある。
【課題を解決するための手段】
【０００６】
前記目的を達成するための本発明の一実施例によるデータの重複除去方法は、データの入力要請または出力要請を基盤として前記データに対する接近特性を獲得する段階と、前記接近特性を基盤として前記データの重複除去単位を決定する段階と、前記重複除去単位を基盤として前記データに対する重複除去を実行する段階と、を含む。
【０００７】
ここで、前記接近特性は、前記データに対する接近時間、前記データに対する変更時間、前記データに対する順次的接近回数及び前記データに対する任意的接近回数の中で少なくとも一つを含むことができる。
【０００８】
ここで、前記接近特性を獲得する段階は、前記データの入力要請を受信した場合、前記データの入力要請に対する時間情報を基盤として前記接近時間及び前記変更時間を獲得する段階と、前記データの入力要請の連続性を基盤として前記順次的接近回数または前記任意的接近回数を獲得する段階と、を含むことができる。
【０００９】
ここで、前記接近特性を獲得する段階は、前記データの出力要請を受信した場合、前記データの出力要請に対する時間情報を基盤として前記接近時間を獲得する段階と、前記データの出力要請の連続性を基盤として前記順次的接近回数または前記任意的接近回数を獲

得する段階と、を含むことができる。
【００１０】
ここで、前記重複除去単位を決定する段階は、前記データに対する現在接近時間と前記データに対する以前変更時間に対する第１の差を算出する段階と、前記第１の差が予め定義された第１の時間以下の場合、前記重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定する段階と、前記第１の差が予め定義された第１の時間を超過する場合、前記データに対する現在接近時間と前記データに対する以前接近時間に対する第２の差を算出する段階と、前記第２の差が予め定義された第２の時間を超過する場合、前記重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定する段階と、前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数以上であれば、前記重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定する段階と、前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数未満であれば、前記重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定する段階と、を含むことができる。
【００１１】
ここで、前記第４の重複除去単位は、前記データに対する重複除去を実行しないことを意味する。
【００１２】
ここで、前記データに対する重複除去を実行する段階は、前記重複除去単位を基盤として前記データに対する少なくても一つのデータブロックを生成する段階と、前記データブロックに対する固有の識別子を生成する段階と、前記固有の識別子がインデックステーブル内に存在するか否かを判断する段階と、前記固有の識別子が前記インデックステーブル内に存在する場合、前記固有の識別子に対応したデータブロックを除去する段階と、前記固有の識別子が前記インデックステーブル内に存在しない場合、前記固有の識別子と前記固有の識別子に対応したデータブロックを保存する段階と、を含むことができる。
【００１３】
ここで、前記固有の識別子を生成する段階は、ハッシュアルゴリズムを使用して前記データブロックに対する固有の識別子を生成することができる。
【００１４】
ここで、前記重複除去単位は、データに対する重複除去可能性を基盤として少なくとも一つの重複除去単位で分類されることができる。
【００１５】
本発明の他の目的を達成するための本発明の一実施例によるデータ重複除去装置は、データの入力要請または出力要請を基盤として前記データに対する接近特性を獲得し、前記接近特性を基盤として前記データの重複除去単位を決定し、前記重複除去単位を基盤として前記データに対する重複除去を実行する処理部と、前記処理部で処理される情報及び処理された情報を保存する保存部と、を含む。
【００１６】
ここで、前記接近特性は、前記データに対する接近時間、前記データに対する変更時間、前記データに対する順次的接近回数及び前記データに対する任意的接近回数の中で少なくとも一つを含むことができる。
【００１７】
ここで、前記処理部は、前記データの入力要請を受信した場合、前記データの入力要請に対する時間情報を基盤として前記接近時間及び前記変更時間を獲得し、前記データの入力要請の連続性を基盤として前記順次的接近回数または前記任意的接近回数を獲得することができる。
【００１８】
ここで、前記処理部は、前記データの出力要請を受信した場合、前記データの出力要請に対する時間情報を基盤として前記接近時間を獲得し、前記データの出力要請の連続性を

基盤として前記順次的接近回数または前記任意的接近回数を獲得することができる。
【００１９】
ここで、前記処理部は、前記重複除去単位を決定する場合、前記データに対する現在接近時間と前記データに対する以前変更時間に対する第１の差を算出し、前記第１の差が予め定義された第１の時間以下の場合、前記重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定し、前記第１の差が予め定義された第１の時間を超過する場合、前記データに対する現在接近時間と前記データに対する以前接近時間に対する第２の差を算出し、前記第２の差が予め定義された第２の時間を超過する場合、前記重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定し、前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数以上であれば、前記重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定し、前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数未満であれば、前記重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定することができる。
【００２０】
ここで、前記第４の重複除去単位は、前記データに対する重複除去を実行しないことを意味する。
【００２１】
ここで、前記処理部は、前記データに対する重複除去を実行する場合、前記重複除去単位を基盤として前記データに対する少なくても一つのデータブロックを生成し、前記データブロックに対する固有の識別子を生成し、前記固有の識別子がインデックステーブル内に存在するか否かを判断し、前記固有の識別子が前記インデックステーブル内に存在する場合、前記固有の識別子に対応したデータブロックを除去し、前記固有の識別子が前記インデックステーブル内に存在しない場合、前記固有の識別子と前記固有の識別子に対応したデータブロックを保存することができる。
【００２２】
ここで、前記処理部は、前記固有の識別子を生成する場合、ハッシュアルゴリズムを使用して前記データブロックに対する固有の識別子を生成することができる。
【００２３】
ここで、前記重複除去単位は、データに対する重複除去可能性を基盤として少なくとも一つの重複除去単位で分類されることができる。
【発明の効果】
【００２４】
本発明によれば、データの入出力特性を基盤として重複除去率(すなわち、チャンク(ｃｈｕｎｋ)のサイズ)を適応的に決定することができ、適応的に決定された重複除去率を基盤として重複を除去することができるので、低い入出力レイテンシ(ｌａｔｅｎｃｙ)を提供することができる。
【図面の簡単な説明】
【００２５】
【図１】図１は、本発明の一実施例によるデータ重複除去方法を示したフローチャートである。
【図２】図２は、データに対する接近特性を示した表である。
【図３】図３は、本発明の一実施例によるデータ重複除去方法において接近特性獲得段階を示した図である。
【図４】図４は、本発明の一実施例によるデータ重複除去方法において重複除去単位決定段階を示したフローチャートである。
【図５】図５は、重複除去単位に対する特性を示した表である。
【図６】図６は、本発明の一実施例によるデータ重複除去方法において重複除去実行段階を示したフローチャートである。

【図７】図７は、データブロックに対する固有の識別子を生成する過程を示した概念図である。
【図８】図８は、本発明の一実施例によるデータ重複除去装置を示したブロック図である。
【発明を実施するための形態】
【００２６】
本発明は、多様に変更可能であり、さまざまな実施形態を有することができる。ここでは、特定の実施形態を図面に例示して詳細に説明する。
【００２７】
しかし、これは本発明の好ましい実施態様に過ぎず、本発明の実施の範囲を限定するものではなく、本発明の明細書及び図面内容に基づいてなされた均等な変更および付加は、いずれも本発明の特許請求の範囲内に含まれるものとする。
【００２８】
第１、第２などの用語は、多様な構成要素を説明するために使用することができるが、前記構成要素は前記用語により限定されものではない。前記用語は、一つの構成要素を他の構成要素から区別するための目的のみで使用される。例えば、本発明の権利範囲を脱しない範囲で、第１の構成要素は第２の構成要素と命名することができ、類似に第２の構成要素も第１の構成要素と命名することができる。及び/またはとの用語は、複数の関連された記載された項目の組み合わせまたは複数の関連された記載された項目の中のいずれか一項目を含む。
【００２９】
いかなる構成要素が他の構成要素に「連結されて」いるかあるいは「接続されて」いるとの用語は、ある構成要素が他の構成要素に直接的に連結されるかあるいは接続されることもできるが、中間に他の構成要素が介在することもできることを意味する。一方に、ある構成要素が他の構成要素に「直接連結されて」いるかあるいは「直接接続されて」いるとの用語は、中間に他の構成要素が存在しないことを意味する。
【００３０】
本明細書で使用した用語は、但し、特定の実施形態を説明するために使用されたもので、本発明はこれに限定されるものではない。単数の表現は、文脈上明白に相違に記載しない限り複数の表現を含む。本出願において、「含む」または「有する」などの用語は、明細書上に記載された特徴、数字、ステップ、動作、構成要素、部品またはこれらを組み合わせたものが存在することを指定するのであって、一つまたはその以上の他の特徴や数字、ステップ、動作、構成要素、部品またはこれらを組み合わせたものなどの存在または付加可能性をあらかじめ排除することではない。
【００３１】
特定しない限り、技術的や科学的な用語を含んでここで使用されるすべての用語は、本発明が属する技術分野で通常の知識を有する者により一般的に理解される意味と同一な意味を有する。一般的に使用される辞典に定義された用語は、関連技術の文脈上有する意味と一致することと解でき、本出願において明白に定義しない限り、理想的や過度に形式的な意味を有することで解釈できない。
【００３２】
以下、添付図面を参照して実施例を詳しく説明するが、図面符号に関係なく同一であるか対応する構成要素には同一な参照番号を付与し、その重複説明は省略する。
【００３３】
図１は、本発明の一実施例によるデータ重複除去方法を示したフローチャートである。
【００３４】
図１を参照すれば、本発明の一実施例によるデータ重複除去方法は、データの入力要請または出力要請を基盤としてデータに対する接近特性を獲得する段階(ステップＳ１００)と、接近特性を基盤としてデータの重複除去単位を決定する段階(ステップＳ２００)と、重複除去単位を基盤としてデータに対する重複除去を実行する段階(ステップＳ３００)と

、を含む。
【００３５】
ここで、図１に示した各段階は、図８に示したデータ重複除去装置で実行することができ、データ重複除去装置の具体的な構成とその機能に対しては後述する。
【００３６】
図２は、データに対する接近特性を示した表である。
【００３７】
図２を参照すれば、データに対する接近特性は、データに対する接近時間(ａ_ｔｉｍｅ)、データに対する変更時間(m_ｔｉｍｅ)、データに対する順次的接近回数(ｓｅｑＣｏｕｎｔ)、データに対する任意的接近回数(ｒａｎｄＣｏｕｎｔ)などを含むことができる。ここで、データに対する接近特性は、前記情報などに限定されないで、データの入力または出力に対する特性を示すことができる情報であれば、いずれも含むことができる。
【００３８】
データに対する接近時間(ａ_ｔｉｍｅ)は、データの入力要請(すなわち、データの書き込み要請)またはデータの出力要請(すなわち、データの読み取り要請)を受信した時間を意味する。データの入力要請を受けた場合、データ重複除去装置は、入力要請時間を該当データに対する接近時間で獲得することができ、獲得した接近時間を保存(すなわち、現在データに対するａ_ｔｉｍｅフィールドを現在時間で記録)することができる。一方、既に保存された接近時間がある場合、データ重複除去装置は、既に保存された接近時間を最近獲得した接近時間に更新することができる。
【００３９】
一方、データの出力要請を受けた場合、データ重複除去装置は、出力要請時間を該当データに対する接近時間で獲得することができ、獲得した接近時間を保存(すなわち、現在データに対するａ_ｔｉｍｅフィールドを現在時間で記録)することができる。一方、既に保存された接近時間がある場合、データ重複除去装置は、既に保存された接近時間を最近獲得した接近時間に更新することができる。
【００４０】
データに対する変更時間(m_ｔｉｍｅ)は、データの入力要請を受信した時間を意味する。データの入力要請を受けた場合、データ重複除去装置は、入力要請時間を該当データに対する変更時間で獲得することができ、獲得した変更時間を保存(すなわち、現在データに対するm_ｔｉｍｅフィールドを現在時間で記録)できる。一方、既に保存された変更時間がある場合、データ重複除去装置は、既に保存された変更時間を最近獲得した変更時間に更新することができる。
【００４１】
データに対する順次的接近回数(ｓｅｑＣｏｕｎｔ)は、現在のデータ要請と以前のデータ要請が連続された(すなわち、現在のデータ要請と以前のデータ要請の物理的または論理的ブロック番号が連続的な場合、またはその要請が連続的な形態を有する場合)回数を意味し、データに対する任意的接近回数(ｒａｎｄＣｏｕｎｔ)は、現在のデータ要請と以前データ要請が連続されない回数を意味する。ここで、順次的接近回数と任意的接近回数は、現在まで累積された回数を意味する。
【００４２】
現在のデータ要請と以前のデータ要請が連続される場合、データ重複除去装置は、現在データに対するｓｅｑＣｏｕｎｔフィールドの値を１増加させる。一方、現在のデータ要請と以前のデータ要請が連続されない場合、データ重複除去装置は、現在データに対するｒａｎｄＣｏｕｎｔフィールドの値を１増加させる。
【００４３】
図３は、本発明の一実施例によるデータ重複除去方法において接近特性獲得段階を示したフローチャートである。
【００４４】
図３を参照すれば、データに対する接近特性を獲得する段階(ステップＳ１００)は、受

信された要請がデータの入力要請に該当するかデータの出力要請に該当するかを判断する段階(ステップＳ１１０)と、データの入力要請を受信した場合、データの入力要請に対する時間情報を基盤として接近時間及び変更時間を獲得する段階(ステップＳ１２０)と、データの入力要請の連続性を基盤として順次的接近回数または任意的接近回数を獲得する段階(ステップＳ１３０)と、を含む。

【００４５】

また、データに対する接近特性を獲得する段階(ステップＳ１００)は、データの出力要請を受信した場合、データの出力要請に対する時間情報を基盤として接近時間を獲得する段階(ステップＳ１４０)と、データの出力要請の連続性を基盤として順次接近回数または任意的接近回数を獲得する段階(ステップＳ１５０)と、を含む。

【００４６】

ステップＳ１１０において、データ重複除去装置は、受信された要請がデータの入力要請に該当するかデータの出力要請に該当するかを判断する。受信された要請がデータの入力要請に該当する場合、データ重複除去装置は、次の段階として、ステップＳ１２０、ステップＳ１３０を実行する。一方、受信された要請がデータの出力要請に該当する場合、データ重複除去装置は、次の段階として、ステップＳ１４０、ステップＳ１５０を実行する。

【００４７】

ステップＳ１２０において、データ重複除去装置は、データの入力要請に対する時間情報を基盤として接近時間及び変更時間を獲得する。すなわち、データ重複除去装置は、データの入力要請を受信した時間を接近時間及び変更時間で獲得することができ、獲得した接近時間及び獲得した変更時間をデータベース(ｄａｔａｂａｓｅ)に保存する。この時、既に保存された接近時間が存在する場合、データ重複除去装置は、既に保存された接近時間を前記獲得した接近時間に更新することができ、既に保存された変更時間が存在する場合、データ重複除去装置は、既に保存された変更時間を前記獲得した変更時間に更新することができる。

【００４８】

ステップＳ１３０において、データ重複除去装置は、データの入力要請と該当データに対する以前要請の連続性を基盤として順次的接近回数または任意的接近回数を獲得することができる。すなわち、データ重複除去装置は、データの入力要請と該当データに対する以前要請が連続される場合、順次的接近回数を１増加させることができ、データの入力要請と該当データに対する以前要請が連続されない場合、任意的接近回数を１増加させることができる。

【００４９】

例えば、データベースに保存された接近特性の中で順次的接近回数が７であり、任意的接近回数が５である場合、データ重複除去装置は、データの入力要請と該当データに対する以前要請が連続される場合、順次接近回数を８に更新し、データの入力要請と該当データに対する以前要請が連続されない場合、任意的接近回数を６に更新する。

【００５０】

ここで、ステップＳ１２０を先に実行した後、ステップＳ１３０を実行することで説明したが、ステップＳ１２０とステップＳ１３０の実行順序は、これに限定されない。すなわち、ステップＳ１３０は、ステップＳ１２０と同時に実行するか、またはステップＳ１２０より先に実行してもよい。

【００５１】

ステップＳ１４０において、データ重複除去装置は、データの出力要請に対する時間情報を基盤として接近時間を獲得することができる。すなわち、データ重複除去装置は、データの出力要請を受信した時間を接近時間で獲得することができ、獲得した接近時間をデータベースに保存することができる。この時、既に保存された接近時間が存在する場合、データ重複除去装置は、既に保存された接近時間を前記獲得した接近時間に更新することができる。

【００５２】

ステップＳ１５０において、データ重複除去装置は、データの出力要請と該当データに対する以前要請の連続性を基盤として順次的接近回数または任意的接近回数を獲得することができる。すなわち、データ重複除去装置は、データの出力要請と該当データに対する以前要請が連続される場合、順次接近回数を１増加させることができ、データの入力要請と該当データに対する以前要請が連続されない場合、任意的接近回数を１増加させることができる。

【００５３】

例えば、データベースに保存された接近特性の中で順次的接近回数が７であり、任意的接近回数が５である場合、データ重複除去装置は、データの出力要請と該当データに対する以前要請が連続される場合、順次接近回数を８に更新し、データの出力要請と該当データに対する以前要請が連続されない場合、任意的接近回数を６に更新する。

【００５４】

ここで、ステップＳ１４０を先に実行した後、ステップＳ１５０を実行することで説明したが、ステップＳ１４０とステップＳ１５０の実行順序は、これに限定されない。すなわち、ステップＳ１５０は、ステップＳ１４０と同時に実行するか、またはステップＳ１４０より先に実行してもよい。

【００５５】

図４は、本発明の一実施例によるデータ重複除去方法において重複除去単位決定段階を示したフローチャートである。

【００５６】

図４を参照すれば、重複除去単位を決定する段階(ステップＳ２００)は、データに対する現在接近時間とデータに対する以前変更時間に対する第１の差を算出する段階(ステップＳ２１０)と、第１の差が予め定義された第１の時間を超過するか否かを判断する段階(ステップＳ２２０)と、第１の差が予め定義された第１の時間以下の場合、重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定する段階(ステップＳ２３０)と、第１の差が予め定義された第１の時間を超過する場合、データに対する現在接近時間とデータに対する以前接近時間に対する第２の差を算出する段階(ステップＳ２４０)と、第２の差が予め定義された第２の時間以下であるか否かを判断する段階(ステップＳ２５０)と、第２の差が予め定義された第２の時間を超過する場合、重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定する段階(ステップＳ２６０)と、第２の差が予め定義された第２の時間以下の場合、データに対する任意的接近回数がデータに対する順次的接近回数以下であるか否かを判断する段階(ステップＳ２７０)と、第２の差が予め定義された第２の時間以下の場合、データに対する任意的接近回数が順次的接近回数以上であれば、重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定する段階(ステップＳ２８０)と、第２の差が予め定義された第２の時間以下の場合、データに対する任意的接近回数が順次的接近回数未満であれば、重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定する段階(ステップＳ２９０)と、を含むことができる。

【００５７】

以下、ステップＳ２００で決定される重複除去単位に対して図５を参照して詳細に説明する。

【００５８】

図５は、重複除去単位に対する特性を示した表である。

【００５９】

図５を参照すれば、重複除去単位は、第１の重複除去単位、第２の重複除去単位、第３の重複除去単位、第４の重複除去単位で分類することができる。第１の重複除去単位は、よく使われないデータに適用することができ、よく使われないデータは、データに対する現在接近時間と以前接近時間の差が予め定義された臨界値より大きいデータを意味する。第１の重複除去単位は、全ての重複除去単位の中で一番小さいチャンク(ｃｈｕｎｋ)を使

用する。これによって、全ての重複除去単位の中で一番高いデータ重複除去率を提供することができる。すなわち、第１の重複除去単位は、低いレイテンシ(ｌａｔｅｎｃｙ)より高いデータ重複除去率を提供するために使われることができる。

【００６０】

第２の重複除去単位は、順次的接近より任意的接近がよく発生するデータに適用することができる。第２の重複除去単位は、第１の重複除去単位より大きくて第３の重複除去単位より小さいサイズのチャンクを使用する。これによって、全ての重複除去単位の中で相対的に高いデータ重複除去率(すなわち、第１の重複除去単位より低く第３の重複除去単位より高い重複除去率)を提供することができる。すなわち、任意的接近がよく発生するデータの場合、物理的な歪みが発生しても任意的接近性能に問題が発生しないので、第２の重複除去単位は、高い重複除去率を提供するために相対的に小さいサイズのチャンクを使用することができる。

【００６１】

第３の重複除去単位は、任意的接近より順次的接近がよく発生するデータに適用することができる。第３の重複除去単位は、第２の重複除去単位より大きいサイズのチャンクを使用する。これによって、全ての重複除去単位の中で相対的に小さい重複除去率(すなわち、第２の重複除去単位より低い重複除去率)を提供することができる。すなわち、順次的な接近がよく発生するデータに対して低い入出力レイテンシを提供するため、第３の重複除去単位は、相対的に大きいサイズのチャンクを使用することができる。

【００６２】

第４の重複除去単位は、入力がよく発生するデータに適用することができる。第４の重複除去単位は、第３の重複除去単位より大きいサイズのチャンクを使用する。これによって、全ての重複除去単位の中で一番小さい重複除去率(すなわち、第３の重複除去単位より低い重複除去率)を提供することができる。一方、第４の重複除去単位は、データに対する重複除去を実行しないことを意味する。すなわち、データの重複除去は、出力を主とするデータに有利なので、入力を主とするデータの場合、重複除去を実行しないこともある。

【００６３】

ここで、重複除去単位の分類は、上述の説明に限定されないで、多様に構成することができる。例えば、重複除去単位を３個分類または５個分類で構成することができる。重複除去単位が３個分類で構成される場合、第１の重複除去単位は、一番高い重複除去率を提供することができ、第２の重複除去単位は、第１の重複除去単位より低い重複除去率を提供することができ、第３の重複除去単位は、第２の重複除去単位より低い重複除去率(すなわち、一番低い重複除去率)を提供することができる。

【００６４】

さらに、図４を参照すれば、ステップＳ２１０において、データ重複除去装置は、データに対する現在接近時間とデータに対する以前変更時間に対する第１の差を算出することができる。すなわち、データ重複除去装置は、データの入力要請または出力要請から獲得した現在接近時間と、同一なデータに対する以前変更時間(すなわち、データの以前入力要請から獲得した変更時間)の差である第１の差を算出することができる。ここで、第１の差は、該当データがどれくらいの頻度で変更されるか(すなわち、データの入力要請がどれくらい頻度で発生するか)を意味する。

【００６５】

ステップＳ２２０において、データ重複除去装置は、第１の差が予め定義された第１の時間を超過するか否かを判断することができる。ここで、予め定義された第１の時間は、データ重複除去を実行するために基準になる時間を意味し、ユーザの設定によって異なる値を有することができる。例えば、予め定義された第１の時間は、１時間、２時間、３時間などで設定することができる。第１の差が予め定義された第１の時間以下の場合、データ重複除去装置は、次の段階として、ステップＳ２３０を実行することができ、第１の差が予め定義された第１の時間を超過する場合、データ重複除去装置は、次の段階として、

ステップＳ２４０を実行することができる。
【００６６】
ステップＳ２３０において、データ重複除去装置は、重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定することができる。すなわち、第１の差が予め定義された第１の時間以下の場合、これは入力がよく発生するデータを意味するので、データ重複除去装置は、重複除去単位の中でデータ重複除去率が一番低い(または、データ重複除去を実行しない)第４の重複除去単位を選択することができる。
【００６７】
ステップＳ２４０において、データ重複除去装置は、データに対する現在接近時間とデータに対する以前接近時間に対する第２の差を算出することができる。すなわち、データ重複除去装置は、データの入力要請または出力要請から獲得した現在接近時間と、同一なデータに対する以前接近時間(すなわち、データの以前入力要請または以前出力要請から獲得した接近時間)の差である第２の差を算出することができる。ここで、第２の差は、該当データに対する接近がどれくらいの頻度で発生するかを意味する。
【００６８】
ステップＳ２５０において、データ重複除去装置は、第２の差が予め定義された第２の時間以下であるか否かを判断することができる。ここで、予め定義された第２の時間は、データに対する接近が発生する可能性が低いデータを区別するために基準になる時間を意味し、ユーザの設定によって異なる値を有することができる。例えば、予め定義された第２の時間は、１時間、２時間、３時間などで設定することができる。第２の差が予め定義された第２の時間を超過する場合、データ重複除去装置は、次の段階として、ステップＳ２６０を実行することができ、第２の差が予め定義された第２の時間以下の場合、データ重複除去装置は、次の段階として、ステップＳ２７０を実行することができる。
【００６９】
ステップＳ２６０において、データ重複除去装置は、重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定することができる。すなわち、第２の差が予め定義された第２の時間を超過する場合、これはよく使われないデータ(すなわち、接近可能性が低いデータ)を意味するので、データ重複除去装置は、重複除去単位の中でデータ重複除去率が一番高い第１の重複除去単位を選択することができる。
【００７０】
ステップＳ２７０において、データ重複除去装置は、データに対する任意的接近回数がデータに対する順次的接近回数未満であるか否かを判断する。任意的接近回数が順次的接近回数以上の場合、データ重複除去装置は、次の段階として、ステップＳ２８０を実行することができ、任意的接近回数が順次的接近回数未満の場合、データ重複除去装置は、次の段階として、ステップＳ２９０を実行することができる。
【００７１】
ステップＳ２８０において、データ重複除去装置は、重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定することができる。すなわち、データに対する任意的接近回数が順次的接近回数以上の場合、これは任意的接近がよく発生するデータを意味するので、データ重複除去装置は、重複除去単位の中でデータ重複除去率が相対的に高い第２の重複除去単位を選択することができる。
【００７２】
ステップＳ２９０において、データ重複除去装置は、重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定することができる。すなわち、データに対する任意的接近回数が順次的接近回数未満の場合は、順次的接近がよく発生するデータを意味するので、データ重複除去装置は、重複除去単位の中でデータ重複除去率が相対的に低い第３の重複除去単位を選択することができる。
【００７３】
図６は、本発明の一実施例によるデータ重複除去方法において重複除去実行段階を示したフローチャートであり、図７は、データブロックに対する固有の識別子を生成する過程

を示した概念図である。
【００７４】
以下、図６及び図７を参照して、重複除去を実行する段階(ステップＳ３００)に対して詳細に説明する。
【００７５】
本発明の一実施例によるデータ重複除去方法において重複除去を実行する段階(ステップＳ３００)は、重複除去単位を基盤としてデータに対する少なくても一つのデータブロックを生成する段階(ステップＳ３１０)と、データブロックに対する固有の識別子(ｉｄｅｎｔｉｆｉｅｒ)を生成する段階(ステップＳ３２０)と、固有の識別子がインデックステーブル(ｉｎｄｅｘ　ｔａｂｌｅ)内に存在するか否かを判断する段階(ステップＳ３３０)と、固有の識別子がインデックステーブル内に存在しない場合、固有の識別子と固有の識別子に対応したデータブロックを保存する段階(ステップＳ３４０)と、固有の識別子がインデックステーブル内に存在する場合、固有の識別子に対応したデータブロックを除去する段階(ステップＳ３５０)と、を含む。
【００７６】
ステップＳ３１０において、データ重複除去装置は、重複除去単位を基盤としてデータに対する少なくても一つのデータブロックを生成することができ、重複除去単位は、チャンクのサイズを意味する。すなわち、データ重複除去装置は、第１の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができ、データ重複除去装置は、第２の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができ、データ重複除去装置は、第３の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができ、データ重複除去装置は、第４の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができる。一方、第４の重複除去単位がデータ重複除去を実行しないことを意味する場合、データ重複除去装置は、データに対するデータブロックを生成しないこともある。
【００７７】
上述したステップＳ３１０を基礎として、データ重複除去装置は、データ３０(図７)から複数のデータブロック３１(図７)を生成することができる。
【００７８】
ステップＳ３２０において、データ重複除去装置は、データブロックに対する固有の識別子を生成することができる。ここで、データ重複除去装置は、ハッシュアルゴリズム(例えば、ＳＨＡ－１、ＳＨＡ－２、ＳＨＡ－３など)を使用してデータブロックに対する固有の識別子を生成することができる。データブロックに対する固有の識別子を生成する方法は、上述した方法に限定されなるものではなくて、公知の多様な方法を使用してデータブロックに対する固有の識別子を生成することができる。
【００７９】
上述のステップＳ３２０を基礎として、データ重複除去装置は、各々のデータブロック３１(図７)から固有の識別子３２(図７)を生成することができる。
【００８０】
ステップＳ３３０において、データ重複除去装置は、固有の識別子がインデックステーブル内に存在するか否かを判断する。インデックステーブルは、固有の識別子と固有の識別子に対応したデータブロックを含むことができる。ここで、固有の識別子がインデックステーブル内に存在する場合、これは固有の識別子に対応したデータブロックが既に保存されていることを示し、固有の識別子がインデックステーブル内に存在しない場合、これは固有の識別子に対応したデータブロックが保存されていないことを示す。固有の識別子がインデックステーブル内に存在しない場合、データ重複除去装置は、次の段階として、ステップＳ３４０を実行することができ、固有の識別子がインデックステーブル内に存在する場合、データ重複除去装置は、次の段階として、ステップＳ３５０を実行することができる。
【００８１】

ステップＳ３４０において、データ重複除去装置は、固有の識別子と固有の識別子に対応したデータブロックを保存することができる。すなわち、固有の識別子に対応したデータブロックが保存されていない状態であるので、データ重複除去装置は、重複除去を実行しないで固有の識別子とデータブロックをデータベース(または、インデックステーブル)に保存する。

【００８２】

ステップＳ３５０において、データ重複除去装置は、固有の識別子に対応したデータブロックを除去することができる。すなわち、固有の識別子に対応したデータブロックが既に保存されている状態であるので、データ重複除去装置は、重複除去(すなわち、固有の識別子に対応したデータブロック削除)を実行することができる。

【００８３】

本発明の一実施例によるデータ重複除去方法は、多様なコンピュータ手段を通じて実行できるプログラム命令の形態で具現してコンピュータ判読可能媒体に記録することができる。コンピュータ判読可能媒体は、プログラム命令、データファイル、データ構造などを単独でまたは組み合わせて含むことができる。コンピュータ判読可能媒体に記録されるプログラム命令は、本発明のために特別に設計されて構成されたものであるか、あるいはコンピュータソフトウェア分野の当業者に公知されて使用可能なものであってもよい。

【００８４】

コンピュータ判読可能媒体の例には、ＲＯＭ(ＲｅａｄＯｎｌｙＭｅｍｏｒｙ)、ＲＡＭ、フラッシュメモリー(ｆｌａｓｈ　ｍｅｍｏｒｙ)などのようにプログラム命令を保存して実行するように特別に構成されたハードウェア装置が含まれる。プログラム命令の例には、コンパイラー(ｃｏｍｐｉｌｅｒ)により作られるもののような機械語コードだけではなくインタープリター(ｉｎｔｅｒｐｒｅｔｅｒ)などを使用してコンピュータにより実行できる高級言語コードを含む。上述したハードウェア装置は、本発明の動作を実行するために少なくとも一つのソフトウェアモジュールで作動するように構成されることができ、その逆も同様である。

【００８５】

図８は、本発明の一実施例によるデータ重複除去装置を示したブロック図である。

【００８６】

図８を参照すれば、本発明の一実施例によるデータ重複除去装置は、処理部１０及び保存部２０を含む。

【００８７】

処理部１０は、データの入力要請または出力要請を基盤としてデータに対する接近特性を獲得することができ、接近特性を基盤としてデータの重複除去単位を決定することができ、重複除去単位を基盤としてデータに対する重複除去を実行することができる。

【００８８】

ここで、処理部１０は、論理的構成である接近特性獲得部１１と、重複除去単位決定部１２と、重複除去実行部１３と、インデックステーブル管理部１４と、を含むことができる。一方、処理部１０は、物理的構成であるプロセッサ(ｐｒｏｃｅｓｓｏｒ)及びメモリー(ｍｅｍｏｒｙ)を含む。プロセッサは、汎用のプロセッサ(例えば、ＣＰＵ(Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ)及び/またはＧＰＵ(Ｇｒａｐｈｉｃｓ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ)など)またはデータ重複除去方法の実行のための専用のプロセッサを意味する。メモリーには、データ重複除去方法の実行のためのプログラムコード(ｐｒｏｇｒａｍ　ｃｏｄｅ)が保存される。すなわち、プロセッサは、メモリーに保存されたプログラムコードを読み出すことができ、読み出されたプログラムコードを基盤としてデータ重複除去方法の各段階を実行することができる。

【００８９】

ここで、データに対する接近特性は、データに対する接近時間(ａ_ｔｉｍｅ、図２参照)、データに対する変更時間(m_ｔｉｍｅ、図２参照)、データに対する順次的接近回数(ｓｅｑＣｏｕｎｔ、図２参照)、データに対する任意的接近回数(ｒａｎｄＣｏｕｎｔ、図

２参照)などを含むことができる。ここで、データに対する接近特性は、前記情報などに限定されないで、データの入力または出力に対する特性を示すことができる情報であれば、いずれも含むことができる。

【００９０】

データに対する接近時間(ａ_ｔ ｉｍｅ)は、データの入力要請(すなわち、データの書き込み要請)またはデータの出力要請(すなわち、データの読み取り要請)を受信した時間を意味する。データに対する変更時間(m_ｔ ｉｍｅ)は、データの入力要請を受信した時間を意味する。データに対する順次的接近回数(ｓｅｑＣｏｕｎｔ)は現在のデータ要請と以前のデータ要請が連続される(すなわち、現在のデータ要請と以前のデータ要請が同一である)回数を意味し、データに対する任意的接近回数(ｒａｎｄＣｏｕｎｔ)は、現在のデータ要請と以前のデータ要請が連続されない(すなわち、現在のデータ要請と以前のデータ要請が同一ではない)回数を意味する。

【００９１】

接近特性を獲得する場合、処理部１０は、データの入力要請に対する時間情報を基盤として接近時間及び変更時間を獲得することができ、データの入力要請の連続性を基盤として順次的接近回数または任意的接近回数を獲得することができる。また、処理部１０は、データの出力要請に対する時間情報を基盤として接近時間を獲得することができ、データの出力要請の連続性を基盤として順次的接近回数または任意的接近回数を獲得することができる。ここで、接近特性を獲得する過程は、処理部１０内の接近特性獲得部１１で実行することができる。

【００９２】

具体的に、処理部１０は、受信された要請がデータの入力要請に該当するかデータの出力要請に該当するかを判断することができる。受信された要請がデータの入力要請に該当する場合、処理部１０は、データの入力要請を受信した時間を接近時間及び変更時間で獲得することができ、獲得した接近時間及び獲得した変更時間を保存部２０に保存することができる。さらに、処理部１０は、データの入力要請と該当データに対する以前要請が連続される場合、順次的接近回数を１増加させることができ、データの入力要請と該当データに対する以前要請が連続されない場合、任意的接近回数を１増加させることができる。

【００９３】

一方、受信された要請がデータの出力要請に該当する場合、処理部１０は、データの出力要請を受信した時間を接近時間で獲得することができ、獲得した接近時間をデータベースに保存することができる。さらに、処理部１０は、データの出力要請と該当データに対する以前要請が連続される場合、順次的接近回数を１増加させることができ、データの入力要請と該当データに対する以前要請が連続されない場合、任意的接近回数を１増加させることができる。

【００９４】

重複除去単位を決定する場合、処理部１０は、データに対する現在接近時間とデータに対する以前変更時間に対する第１の差を算出することができ、第１の差が予め定義された第１の時間以下の場合、重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定することができる。一方、第１の差が予め定義された第１の時間を超過する場合、処理部１０は、データに対する現在接近時間とデータに対する以前接近時間に対する第２の差を算出することができる。

【００９５】

ここで、処理部１０は、第２の差が予め定義された第２の時間を超過する場合、重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定することができ、第２の差が予め定義された第２の時間以下の場合、データに対する任意的接近回数が順次的接近回数以上であるか否かを判断することができる。

【００９６】

ここで、処理部１０は、データに対する任意的接近回数が順次的接近回数以上の場合、重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定

することができ、データに対する任意的接近回数が順次的接近回数未満の場合、重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定することができる。

【００９７】

上述した重複除去単位を決定する過程は、処理部１０内の重複除去単位決定部１２で実行することができる。

【００９８】

ここで、重複除去単位は、第１の重複除去単位、第２の重複除去単位、第３の重複除去単位、第４の重複除去単位で分類することができる。第１の重複除去単位は、よく使われないデータに適用することができ、第２の重複除去単位は、順次的接近より任意的接近がよく発生するデータに適用することができ、第３の重複除去単位は、任意的接近より順次的接近がよく発生するデータに適用することができ、第４の重複除去単位は、入力がよく発生するデータに適用することができる。

【００９９】

具体的に、処理部１０は、データの入力要請または出力要請から獲得した現在接近時間と、同一なデータに対する以前変更時間(すなわち、データの以前入力要請から獲得した変更時間)の差である第１の差を算出することができる。ここで、第１の差は、該当データがどれくらいの頻度で変更されるか(すなわち、データの入力要請がどれくらいの頻度で発生するか)を意味する。

【０１００】

処理部１０は、第１の差が予め定義された第１の時間を超過するか否かを判断することができる。ここで、予め定義された第１の時間は、データ重複除去を実行するために基準になる時間を意味し、ユーザの設定によって異なる値を有することができる。

【０１０１】

第１の差が予め定義された第１の時間以下の場合、これは入力がよく発生するデータを意味するので、処理部１０は、重複除去単位の中でデータ重複除去率が一番低い(または、データ重複除去を実行しない)第４の重複除去単位を選択することができる。

【０１０２】

一方、第１の差が予め定義された第１の時間を超過する場合、処理部１０は、データの入力要請または出力要請から獲得した現在接近時間と、同一なデータに対する以前接近時間(すなわち、データの以前入力要請または以前出力要請から獲得した接近時間)の差である第２の差を算出することができる。ここで、第２の差は、該当データに対する接近がどれいくらの頻度で発生するかを意味する。

【０１０３】

処理部１０は、第２の差が予め定義された第２の時間以下であるか否かを判断することができる。ここで、予め定義された第２の時間は、データに対する接近が発生する可能性が低いデータを区別するために基準になる時間を意味し、ユーザの設定によって異なる値を有することができる。

【０１０４】

第２の差が予め定義された第２の時間を超過する場合、これはよく使われないデータ(すなわち、接近可能性が低いデータ)を意味するので、処理部１０は、重複除去単位の中でデータ重複除去率が一番高い第１の重複除去単位を選択することができる。一方、第２の差が予め定義された第２の時間以下の場合、処理部１０は、データに対する任意的接近回数がデータに対する順次接近回数未満であるか否かを判断することができる。

【０１０５】

データに対する任意的接近回数が順次的接近回数以上の場合、これは任意的接近がよく発生するデータを意味するので、処理部１０は、重複除去単位の中でデータ重複除去率が相対的に高い第２の重複除去単位を選択することができる。一方、データに対する任意的接近回数が順次的接近回数未満の場合は、順次的接近がよく発生するデータを意味するので、処理部１０は、重複除去単位の中でデータ重複除去率が相対的に低い第３の重複除去

単位を選択することができる。
【０１０６】
データの重複除去を実行する場合、処理部１０は、重複除去単位を基盤としてデータに対する少なくても一つのデータブロックを生成することができ、データブロックに対する固有の識別子を生成することができ、固有の識別子がインデックステーブル内に存在するか否かを判断することができ、固有の識別子がインデックステーブル内に存在する場合、固有の識別子に対応したデータブロックを除去することができ、固有の識別子がインデックステーブル内に存在しない場合、固有の識別子と固有の識別子に対応したデータブロックを保存することができる。
【０１０７】
ここで、データの重複を除去する過程は、処理部１０内の重複除去実行部１３で実行することができ、インデックステーブル内に情報を保存、削除、更新する過程は、処理部１０内のインデックステーブル管理部１４で実行することができる。
【０１０８】
具体的に、処理部１０は、第１の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができ、第２の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができ、第３の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができ、第４の重複除去単位に対応したチャンクのサイズを基礎としてデータブロックを生成することができる。一方、第４の重複除去単位がデータ重複除去を実行しないことを意味する場合、処理部１０は、データに対するデータブロックを生成しないこともある。
【０１０９】
処理部１０は、ハッシュアルゴリズム(例えば、ＳＨＡ－１、ＳＨＡ－２、ＳＨＡ－３など)を使用してデータブロックに対する固有の識別子を生成することができる。データブロックに対する固有の識別子を生成する方法は、上述の説明に限定されるものではなくて、公知の多様な方法を使用してデータブロックに対する固有の識別子を生成することができる。
【０１１０】
処理部１０は、固有の識別子がインデックステーブル内に存在するか否かを判断することができる。固有の識別子がインデックステーブル内に存在しない場合、これは固有の識別子に対応したデータブロックが保存されていない状態であるので、処理部１０は、重複除去を実行しないで固有の識別子とデータブロックを保存部２０に保存することができる。一方、固有の識別子がインデックステーブル内に存在する場合、これは固有の識別子に対応したデータブロックが既に保存されている状態であるので、処理部１０は、重複除去(すなわち、固有の識別子に対応したデータブロック削除)を実行することができる。
【０１１１】
保存部２０は、処理部１０で処理される情報及び処理された情報を保存することができる。例えば、保存部２０は、データの入力要請、データの出力要請、データに対する接近特性、第１の差、予め定義された第１の時間、第２の差、予め定義された第２の時間、インデックステーブル、重複除去単位情報などを保存することができる。
【０１１２】
以上、添付した図面を参照して本発明の実施形態について説明したが、本発明が属する技術の分野における通常の知識を有する者であれば、本発明の技術的思想を逸脱しない範囲内で、様々な置換、変形及び変更が可能であるので、上述した実施例及び添付された図面に限定されるものではない。
【符号の説明】
【０１１３】
１０：処理部
１１：接近特性獲得部
１２：重複除去単位決定部

１３：重複除去実行部
１４：インデックステーブル管理部
２０：保存部

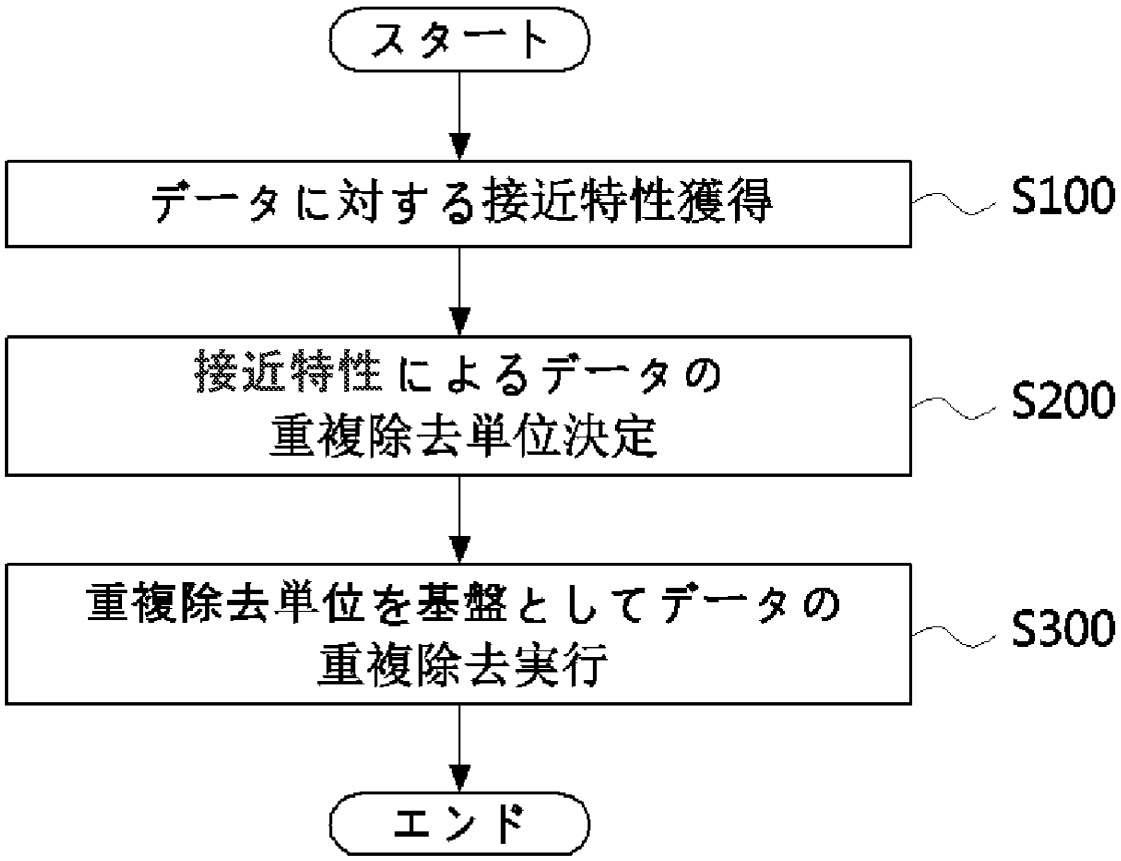
【書類名】特許請求の範囲
【請求項１】
データ重複除去装置で実行するデータ重複除去方法であって、
データの入力要請または出力要請を基盤として前記データに対する接近特性を獲得する段階と、
前記接近特性を基盤として前記データの重複除去単位を決定する段階と、
前記重複除去単位を基盤として前記データに対する重複除去を実行する段階と、
を含むことを特徴とするデータ重複除去方法。
【請求項２】
前記接近特性は、前記データに対する接近時間、前記データに対する変更時間、前記データに対する順次的接近回数及び前記データに対する任意的接近回数の中で少なくとも一つを含むことを特徴とする請求項１に記載のデータ重複除去方法。
【請求項３】
前記接近特性を獲得する段階は、
前記データの入力要請を受信した場合、前記データの入力要請に対する時間情報を基盤として前記接近時間及び前記変更時間を獲得する段階と、
前記データの入力要請の連続性を基盤として前記順次的接近回数または前記任意的接近回数を獲得する段階と、を含むことを特徴とする請求項２に記載のデータ重複除去方法。
【請求項４】
前記接近特性を獲得する段階は、
前記データの出力要請を受信した場合、前記データの出力要請に対する時間情報を基盤として前記接近時間を獲得する段階と、
前記データの出力要請の連続性を基盤として前記順次接近回数または前記任意的接近回数を獲得する段階と、を含むことを特徴とする請求項２に記載のデータ重複除去方法。
【請求項５】
前記重複除去単位を決定する段階は,
前記データに対する現在接近時間と前記データに対する以前変更時間に対する第１の差を算出する段階と、
前記第１の差が予め定義された第１の時間以下の場合、前記重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定する段階と、
前記第１の差が予め定義された第１の時間を超過する場合、前記データに対する現在接近時間と前記データに対する以前接近時間に対する第２の差を算出する段階と、
前記第２の差が予め定義された第２の時間を超過する場合、前記重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定する段階と、
前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数以上であれば、前記重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定する段階と、
前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数未満であれば、前記重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定する段階と、を含むことを特徴とする請求項２に記載のデータ重複除去方法。
【請求項６】
前記第４の重複除去単位は、前記データに対する重複除去を実行しないことを意味することを特徴とする請求項５に記載のデータ重複除去方法。
【請求項７】
前記データに対する重複除去を実行する段階は、
前記重複除去単位を基盤として前記データに対する少なくても一つのデータブロック(ｂｌｏｃｋ)を生成する段階と、
前記データブロックに対する固有の識別子を生成する段階と、
前記固有の識別子がインデックステーブル(ｉｎｄｅｘ　ｔａｂｌｅ)内に存在するか否

かを判断する段階と、
前記固有の識別子が前記インデックステーブル内に存在する場合、前記固有の識別子に対応したデータブロックを除去する段階と、
前記固有の識別子が前記インデックステーブル内に存在しない場合、前記固有の識別子と前記固有の識別子に対応したデータブロックを保存する段階と、を含むことを特徴とする請求項１に記載のデータ重複除去方法。
【請求項８】
前記固有の識別子を生成する段階は、ハッシュアルゴリズム(ｈａｓｈ　ａｌｇｏｒｉｔｈｍ)を使用して前記データブロックに対する固有の識別子を生成することを特徴とする請求項７に記載のデータ重複除去方法。
【請求項９】
前記重複除去単位は、データに対する重複除去可能性を基盤として少なくとも一つの重複除去単位で分類されることを特徴とする請求項１に記載のデータ重複除去方法。
【請求項１０】
データの入力要請または出力要請を基盤として前記データに対する接近特性を獲得し、前記接近特性を基盤として前記データの重複除去単位を決定し、前記重複除去単位を基盤として前記データに対する重複除去を実行する処理部と、
前記処理部で処理される情報及び処理された情報を保存する保存部と、
を含むことを特徴とするデータ重複除去装置。
【請求項１１】
前記接近特性は、前記データに対する接近時間、前記データに対する変更時間、前記データに対する順次的接近回数及び前記データに対する任意的接近回数の中で少なくとも一つを含むことを特徴とする請求項１０に記載のデータ重複除去装置。
【請求項１２】
前記処理部は、前記データの入力要請を受信した場合、前記データの入力要請に対する時間情報を基盤として前記接近時間及び前記変更時間を獲得し、前記データの入力要請の連続性を基盤として前記順次的接近回数または前記任意的接近回数を獲得することを特徴とする請求項１１に記載のデータ重複除去装置。
【請求項１３】
前記処理部は、前記データの出力要請を受信した場合、前記データの出力要請に対する時間情報を基盤として前記接近時間を獲得し、前記データの出力要請の連続性を基盤として前記順次的接近回数または前記任意的接近回数を獲得することを特徴とする請求項１１に記載のデータ重複除去装置。
【請求項１４】
前記処理部は、
前記重複除去単位を決定する場合、前記データに対する現在接近時間と前記データに対する以前変更時間に対する第１の差を算出し、前記第１の差が予め定義された第１の時間以下の場合、前記重複除去単位をデータに対する重複除去可能性が一番低い第４の重複除去単位で決定し、前記第１の差が予め定義された第１の時間を超過する場合、前記データに対する現在接近時間と前記データに対する以前接近時間に対する第２の差を算出し、前記第２の差が予め定義された第２の時間を超過する場合、前記重複除去単位をデータに対する重複除去可能性が一番高い第１の重複除去単位で決定し、
前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数以上であれば、前記重複除去単位を第１の重複除去単位より重複除去可能性が低い第２の重複除去単位で決定し、
前記第２の差が予め定義された第２の時間以下の場合、前記データに対する任意的接近回数が順次的接近回数未満であれば、前記重複除去単位を第２の重複除去単位より重複除去可能性が低い第３の重複除去単位で決定することを特徴とする請求項１１に記載のデータ重複除去装置。
【請求項１５】

前記第４の重複除去単位は、前記データに対する重複除去を実行しないことを意味することを特徴とする請求項１４に記載のデータ重複除去装置。
【請求項１６】
前記処理部は、前記データに対する重複除去を実行する場合、前記重複除去単位を基盤として前記データに対する少なくても一つのデータブロック(ｂｌｏｃｋ)を生成し、前記データブロックに対する固有の識別子を生成し、前記固有の識別子がインデックステーブル(ｉｎｄｅｘ　ｔａｂｌｅ)内に存在するか否かを判断し、前記固有の識別子が前記インデックステーブル内に存在する場合、前記固有の識別子に対応したデータブロックを除去し、前記固有の識別子が前記インデックステーブル内に存在しない場合、前記固有の識別子と前記固有の識別子に対応したデータブロックを保存することを特徴とする請求項１０に記載のデータ重複除去装置。
【請求項１７】
前記処理部は、前記固有の識別子を生成する場合、ハッシュアルゴリズム(ｈａｓｈ　ａｌｇｏｒｉｔｈｍ)を使用して前記データブロックに対する固有の識別子を生成することを特徴とする請求項１６に記載のデータ重複除去装置。
【請求項１８】
前記重複除去単位は、データに対する重複除去可能性を基盤として少なくとも一つの重複除去単位で分類されることを特徴とする請求項１０に記載のデータ重複除去装置。

【書類名】要約書
【要約】
本発明は、データ重複除去方法及び装置に関する。データの重複除去方法は、データの入力要請または出力要請を基盤としてデータに対する接近特性を獲得する段階と、接近特性を基盤としてデータの重複除去単位を決定する段階と、重複除去単位を基盤としてデータに対する重複除去を実行する段階と、を含む。したがって、低い入出力レイテンシを提供することができる。
【選択図】図１

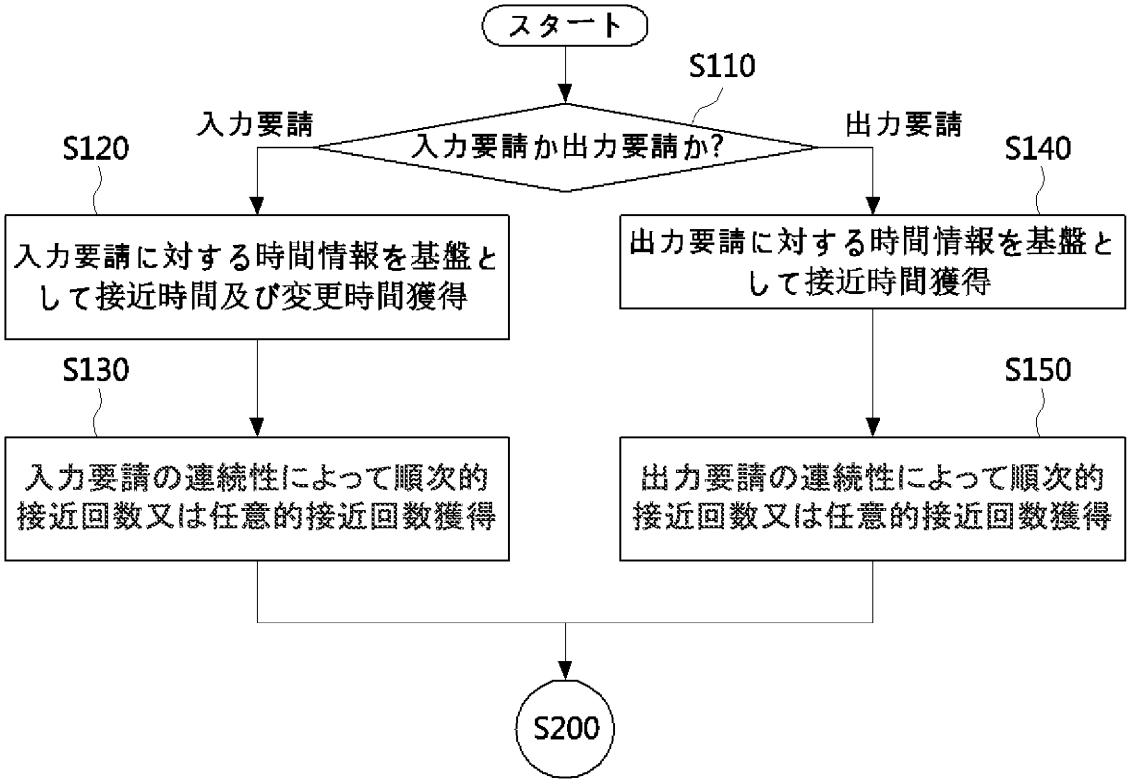
【書類名】図面
【図１】



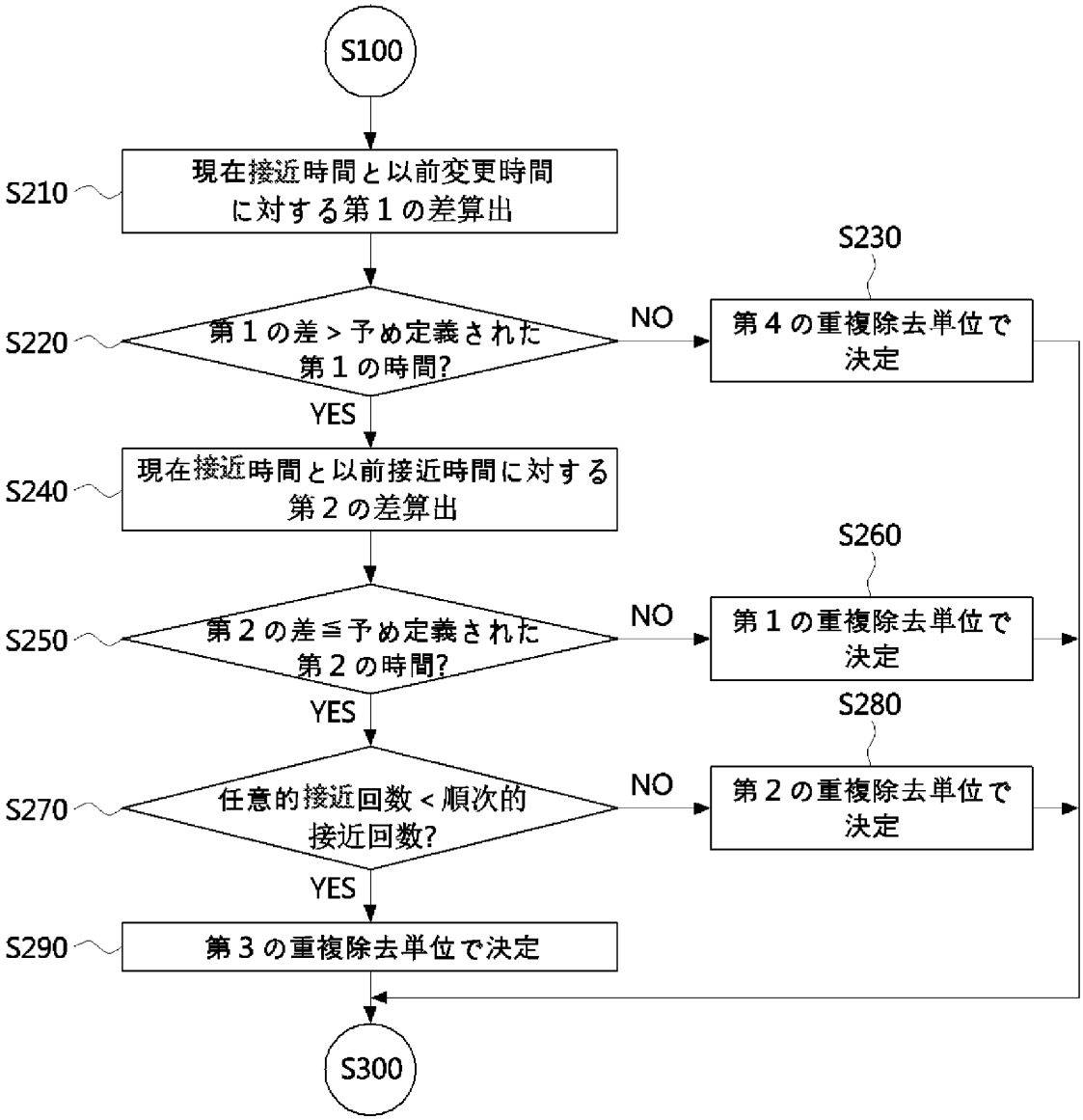

【図２】

| タイプ | フィールド | 備考 |
|---|---|---|
| 時間型 | m_time | 最近変更時間 |
| 時間型 | a_time | 最近接近時間 |
| 正数型 | seqCount | 順次的接近回数 |
| 正数型 | randCount | 任意的接近回数 |

【図３】


スタート
S110
入力要請か出力要請か?
入力要請
出力要請
S120
S140
入力要請に対する時間情報を基盤として接近時間及び変更時間獲得
出力要請に対する時間情報を基盤として接近時間獲得
S130
S150
入力要請の連続性によって順次的接近回数又は任意的接近回数獲得
出力要請の連続性によって順次的接近回数又は任意的接近回数獲得
S200

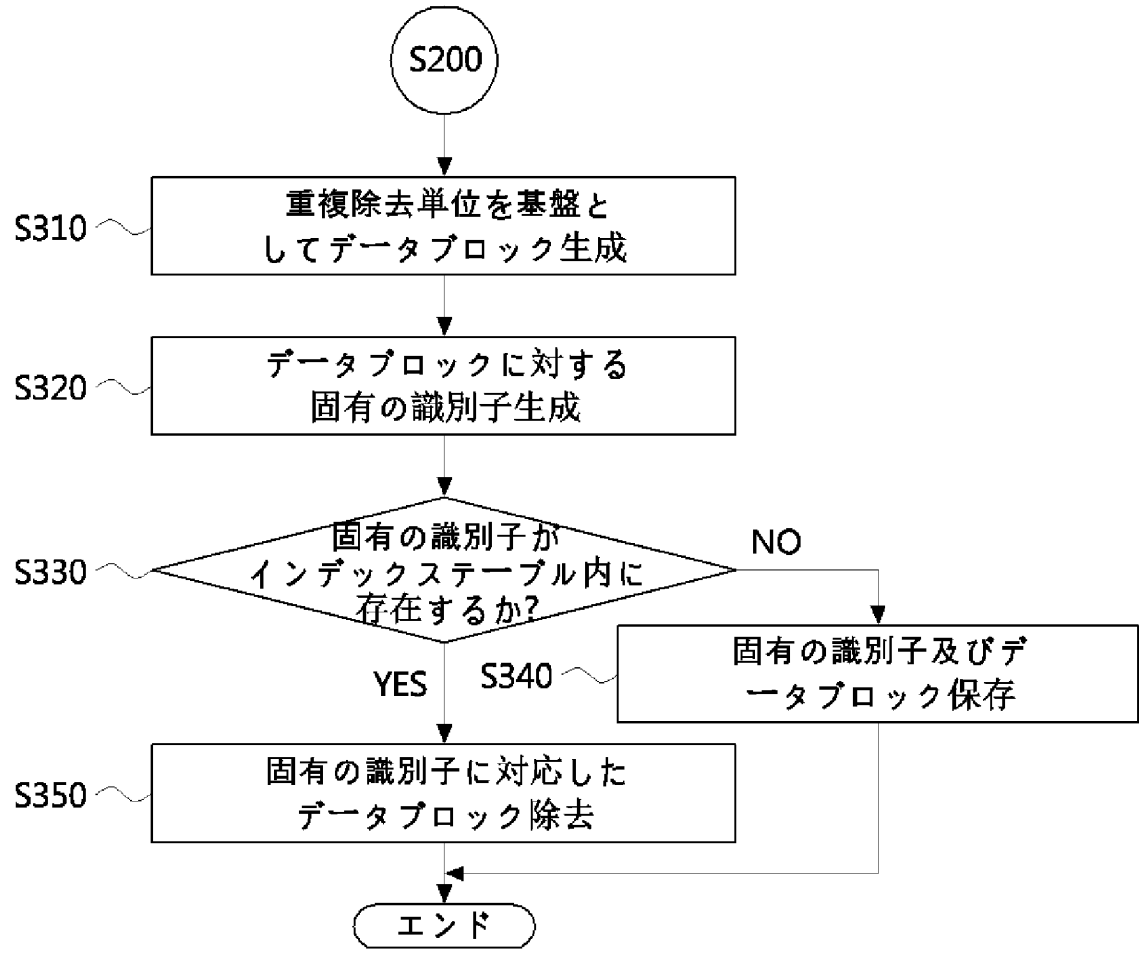
【図４】



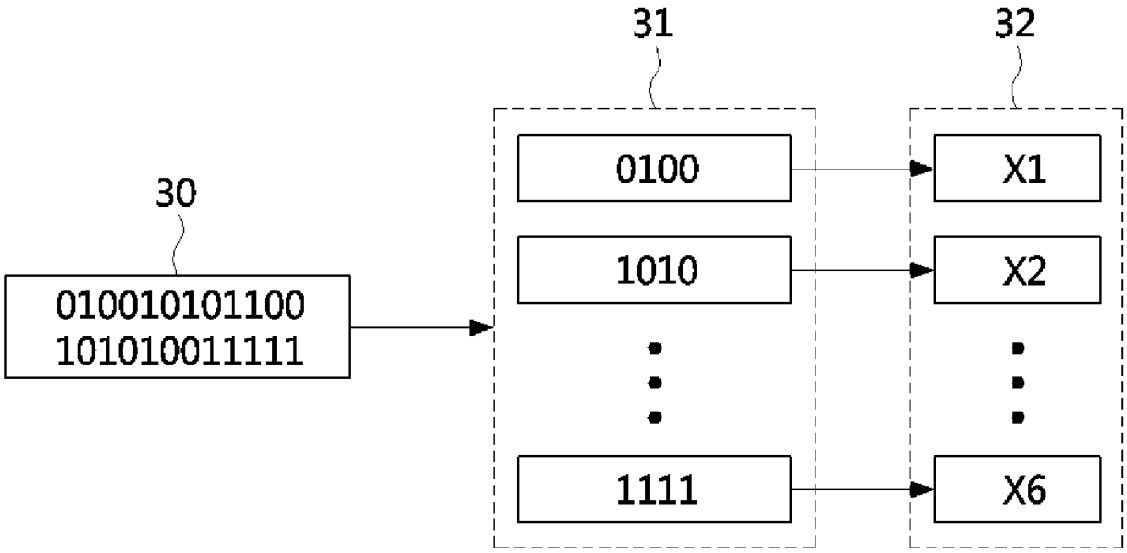

【図５】

| 分類 | 重複除去単位 | 備考 |
|---|---|---|
| よく使用しないデータ | 第１の重複除去単位 | －番小さいチャンク使用<br>－番大きい重複除去率 |
| 任意的接近がよく発生するデータ | 第２の重複除去単位 | - 相対的に小さいチャンク使用<br>- 相対的に大きい重複除去率 |
| 順次的接近がよく発生するデータ | 第３の重複除去単位 | - 相対的に大きいチャンク使用<br>- 相対的に小さい重複除去率 |
| 入力がよく発生するデータ | 第４の重複除去単位 | - 重複除去を実行しない |

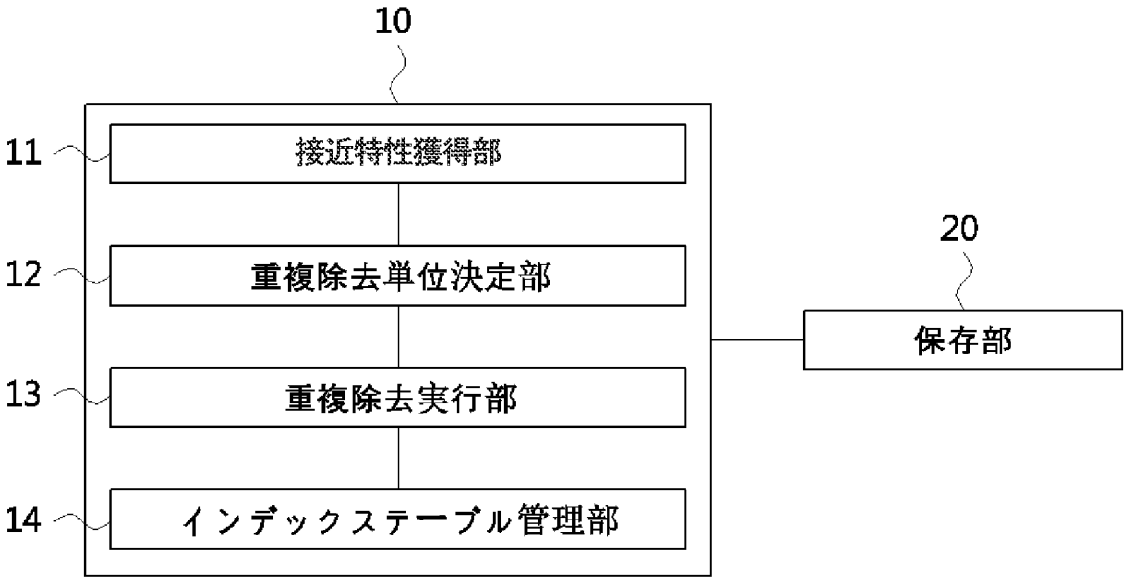
【図６】


S200
S310
重複除去単位を基盤としてデータブロック生成
S320
データブロックに対する固有の識別子生成
S330
固有の識別子がインデックステーブル内に存在するか?
NO
YES
S340
固有の識別子及びデータブロック保存
S350
固有の識別子に対応したデータブロック除去
エンド

【図７】


31
32
30
010010101100
101010011111
0100
1010
1111
X1
X2
X6

【図８】


10
11
接近特性獲得部
12
重複除去単位決定部
13
重複除去実行部
14
インデックステーブル管理部
20
保存部

【書類名】　出願審査請求書
【整理番号】　CP00744
【あて先】　特許庁長官　殿
【出願の表示】
　【出願番号】　特願2014- 45770
【請求項の数】　18
【請求人】
　【識別番号】　511258411
　【氏名又は名称】　ポハン工科大学校産学協力団
【代理人】
　【識別番号】　100121728
　【弁理士】
　【氏名又は名称】　井関　勝守
【手数料の表示】
　【振替番号】　00016713
　【納付金額】　190,000円

# 受領書

平成２６年　３月　７日
特　許　庁　長　官

識別番号　　１００１２１７２８
氏名（名称）　　井関　勝守　　様

以下の書類を受領しました。

| 項番 | 書類名 | 整理番号 | 受付番号 | 提出日 | 出願番号通知（事件の表示） |
|---|---|---|---|---|---|
| 1 | 特許願 | CP00744 | 51400497951 | 平26. 3. 7 | 特願2014- 45770 |

以　上

# 受領書

平成２６年　３月　７日
特　許　庁　長　官

識別番号　　１００１２１７２８
氏名（名称）　　井関　勝守　　様

以下の書類を受領しました。

| 項番 | 書類名 | 整理番号 | 受付番号 | 提出日 | 出願番号通知（事件の表示） |
|---|---|---|---|---|---|
| 1 | 出願審査請求 | CP00744 | 51400497964 | 平26. 3. 7 | 特願2014- 45770 |

以　上