(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3945282号
(P3945282)
(45)発行日　平成19年7月18日(2007.7.18)　(24)登録日　平成19年4月20日(2007.4.20)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 230Z
G06F 17/30 210A
G06F 17/30 170A

請求項の数 5　(全 16 頁)

(21)出願番号　特願2002-76923(P2002-76923)
(22)出願日　平成14年3月19日(2002.3.19)
(65)公開番号　特開2003-281182(P2003-281182A)
(43)公開日　平成15年10月3日(2003.10.3)
審査請求日　平成16年10月20日(2004.10.20)

(73)特許権者　000002369
セイコーエプソン株式会社
東京都新宿区西新宿２丁目４番１号
(74)代理人　100098084
弁理士　川▲崎▼　研二
(72)発明者　田中　敬重
長野県諏訪市大和３丁目３番５号　セイコーエプソン株式会社内

審査官　辻本　泰隆

最終頁に続く

(54)【発明の名称】情報検索装置、情報検索方法、プログラムおよび記録媒体

(57)【特許請求の範囲】
【請求項１】
少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索する情報検索装置であって、
前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を記憶する第１の記憶手段と、
前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する関連情報取得手段と、
前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータに付加する重み付け単語付加手段と、
前記テキストデータからテキスト文を抽出する本文抽出手段と、
前記抽出されたテキスト文を複数の単語に分割して解析する形態素解析手段と、
前記複数の単語の各々が前記テキスト文に出現する回数を計数する出現頻度計数手段と、
前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて記憶する第２の記憶手段と、
前記項目指定情報によって指定された項目に則した検索条件を取得する検索条件取得手段と、

前記第２の記憶手段に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する検索手段と、を具備する
ことを特徴とする情報検索装置。
【請求項２】
前記重み付け単語は、前記検索条件取得手段によって取得される検索条件となる検索語である
ことを特徴とする請求項１に記載の情報検索装置。
【請求項３】
ＣＰＵと記憶装置とを有し、少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索する情報検索装置における情報検索方法であって、
前記ＣＰＵが、前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を前記記憶装置に記憶する第１の過程と、
前記ＣＰＵが、前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する第２の過程と、
前記ＣＰＵが、前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータに付加する第３の過程と、
前記ＣＰＵが、前記テキストデータからテキスト文を抽出する第４の過程と、
前記ＣＰＵが、前記抽出されたテキスト文を複数の単語に分割して解析する第５の過程と、
前記ＣＰＵが、前記複数の単語の各々が前記テキスト文に出現する回数を計数する第６の過程と、
前記ＣＰＵが、前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて前記記憶装置に記憶する第７の過程と、
前記ＣＰＵが、前記項目指定情報によって指定された項目に則した検索条件を取得する第８の過程と、
前記ＣＰＵが、前記記憶装置に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する第９の過程と、を備える
ことを特徴とする情報検索装置における情報検索方法。
【請求項４】
少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索するコンピュータを、
前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を記憶する第１の記憶手段、
前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する関連情報取得手段、
前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータに付加する重み付け単語付加手段、
前記テキストデータからテキスト文を抽出する本文抽出手段、
前記抽出されたテキスト文を複数の単語に分割して解析する形態素解析手段、
前記複数の単語の各々が前記テキスト文に出現する回数を計数する出現頻度計数手段、
前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて記憶する第２の記憶手段、
前記項目指定情報によって指定された項目に則した検索条件を取得する検索条件取得手

段、および
前記第２の記憶手段に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する検索手段、
として機能させるためのプログラム。
【請求項５】
少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索するコンピュータを、
前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を記憶する第１の記憶手段、
前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する関連情報取得手段、
前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータに付加する重み付け単語付加手段、
前記テキストデータからテキスト文を抽出する本文抽出手段、
前記抽出されたテキスト文を複数の単語に分割して解析する形態素解析手段、
前記複数の単語の各々が前記テキスト文に出現する回数を計数する出現頻度計数手段、
前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて記憶する第２の記憶手段と、
前記項目指定情報によって指定された項目に則した検索条件を取得する検索条件取得手段、および
前記第２の記憶手段に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する検索手段
として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、データベースの情報を検索する情報検索装置、情報検索方法、プログラムおよび記録媒体に関する。
【０００２】
【従来の技術】
企業などでは、例えばＬＡＮ（Local Area Network）などのコンピュータネットワーク（以下、単に「ネットワーク」と称する）が構成され、このネットワーク内における各種データの共有により、作業効率の向上化が図られている。具体的には、ネットワークを形成するいずれかのコンピュータにグループウェアやコラボレートウェアなどと呼ばれるソフトウェア（以下、「グループウェア」と称する）が導入されることで、このコンピュータ（以下、「グループウェアサーバ」と称する）が保持する各種データ（例えば、共有文書や各ユーザのスケジュールなど）に対してネットワークに接続された各コンピュータ（以下、「クライアント端末」と称する）からアクセス可能になる。
【０００３】
また、グループウェアには、クライアント端末からの要求に応じて、蓄積された文書データから該当する文書データを検索する機能が備えられている。これにより、ユーザは、クライアント端末を用いてグループウェアサーバが管理する大量の文書データから所望の文書データを見つけることが容易となる。
【０００４】
【発明が解決しようとする課題】
しかしながら、グループウェアサーバが文書データを検索する時には、全ての文書データを対象に検索処理を実行するのが一般的であり、文書データの数や各文書データの容量に比例して検索時間も長くなるといった問題がある。特に、顧客からの問い合わせに対応す

るコールセンターでは、グループウェアサーバが顧客からの問い合わせに応じた文書データを素早く検索して取り出す必要があるため、この問題は、より深刻化する。

【０００５】

本発明は、上述した事情を鑑みてなされたものであり、データベースに蓄積されている情報のうち、検索条件に該当する情報を特定するに要する時間を短縮することが可能な情報検索装置、情報検索方法、プログラムおよび記録媒体を提供することを目的とする。

【０００６】

【課題を解決するための手段】

上記目的を達成するために、本発明は、少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索する情報検索装置であって、前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を記憶する第１の記憶手段と、前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する関連情報取得手段と、前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータの付加する重み付け単語付加手段と、前記テキストデータからテキスト文を抽出する本文抽出手段と、前記抽出されたテキスト文を複数の単語に分割して解析する形態素解析手段と、前記複数の単語の各々が前記テキスト文に出現する回数を計数する出現頻度計数手段と、前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて記憶する第２の記憶手段と、前記項目指定情報によって指定された項目に則した検索条件を取得する検索条件取得手段と、前記第２の記憶手段に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する検索手段と、を備える情報検索装置を提供する。

【０００７】

また、上記目的を達成するために、本発明は、ＣＰＵと記憶装置とを有し、少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索する情報検索装置における情報検索方法であって、前記ＣＰＵが、前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を前記記憶装置に記憶する第１の過程と、前記ＣＰＵが、前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する第２の過程と、前記ＣＰＵが、前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータに付加する第３の過程と、前記ＣＰＵが、前記テキストデータからテキスト文を抽出する第４の過程と、前記ＣＰＵが、前記抽出されたテキスト文を複数の単語に分割して解析する第５の過程と、前記ＣＰＵが、前記複数の単語の各々が前記テキスト文に出現する回数を計数する第６の過程と、前記ＣＰＵが、前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて前記記憶装置に記憶する第７の過程と、前記ＣＰＵが、前記項目指定情報によって指定された項目に則した検索条件を取得する第８の過程と、前記ＣＰＵが、前記記憶装置に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する第９の過程と、を備える情報検索装置における情報検索方法を提供する。

【０００８】

上述した情報検索装置および情報検索方法によれば、データベースに記憶されている複数の項目から検索の対象となり得る項目だけが予め抽出され、そして、その抽出された項目に対して検索が行われる。従って、本発明によれば、該当する文書データを特定するに要する時間が、データベースの全ての項目に対して検索が実行されるときに比べて早くなる

。また、利用者は、項目指定情報が指定する項目を変更するだけで、検索の対象とする項目を変更することができる。

【０００９】

ここで、上記情報検索装置において、前記テキストデータからテキスト文を抽出する本文抽出手段と、前記抽出されたテキスト文を複数の単語に分割して解析する形態素解析手段と、前記複数の単語の各々が前記テキスト文に出現する回数を計数する出現頻度計数手段とを備え、前記第２の記憶手段は、前記単語と当該単語の出現回数とを、前記テキスト文に対応するテキストデータの識別情報と対応付けて記憶する構成が望ましい。この構成によれば、検索条件として単語が取得された場合に、当該単語を多く含む順にテキストデータの識別情報を特定するといったことが行える。

【００１０】

また、上記目的を達成するために、本発明は、少なくともテキスト文を含むテキストデータと、当該テキストデータの識別情報とを対応付けるとともに、当該テキスト文に関連した複数の関連情報と、当該複数の関連情報を分類する項目と、当該テキスト文に対応するテキストデータの識別情報とを対応付けるデータベースを検索するコンピュータを、前記項目のうち、検索の対象となり得る重み付け単語を含む項目を指定する項目指定情報を記憶する第１の記憶手段、前記項目指定情報によって指定された項目に分類される関連情報を前記データベースから取得する関連情報取得手段、前記重み付け単語によって指定された単語を前記テキストデータから抽出して前記テキストデータの付加する重み付け単語付加手段、前記テキストデータからテキスト文を抽出する本文抽出手段、前記抽出されたテキスト文を複数の単語に分割して解析する形態素解析手段、前記複数の単語の各々が前記テキスト文に出現する回数を計数する出現頻度計数手段、前記関連情報取得手段によって取得された関連情報と、前記単語と当該単語の出現回数と、当該関連情報に対応する前記識別情報とを対応付けて記憶する第２の記憶手段と、前記項目指定情報によって指定された項目に則した検索条件を取得する検索条件取得手段、および前記第２の記憶手段に記憶された関連情報の中から、前記検索条件に該当する関連情報を特定し、当該関連情報に対応する前記識別情報を特定する検索手段として機能させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体に記憶されていても良いことは勿論である。

【００１１】

【発明の実施の形態】

以下、図面を参照して本発明の実施形態について説明する。

【００１２】

図１は、本発明の実施形態に係る情報検索システムの構成を示す図である。この図において、グループウェアサーバ２０は、例えば磁気ディスクなどの記憶装置に格納されたグループウェアデータベース２０ａを備えている。このグループウェアデータベース２０ａには、ネットワーク２を介して接続された多数のクライアント端末３０の間で共有される文書データが蓄積されている。ここで、文書データとは、テキスト文が含まれるデータのことである。また、グループウェアサーバ２０は、共有される文書データが蓄積されたデータベース（すなわち、上述したグループウェアデータベース２０ａ）の他にも、実際には、例えば利用者毎の電子メールデータが蓄積されたデータベースや、利用者毎のスケジュールデータが蓄積されたデータベースといった多種のデータベースを備えている。

【００１３】

さて、図１において、情報検索装置１０は、パーソナルコンピュータなどから構成されており、ネットワーク２を介してクライアント端末３０からの文書データの検索要求を取得し、この検索要求に該当する文書データの候補を当該クライアント端末３０に送信するものである。さらに説明すると、情報検索装置１０は、例えば磁気ディスクなどの記憶装置を備え、この記憶装置には、検索用データベース１０ａが格納されている。情報検索装置１０は、グループウェアデータベース２０ａに蓄積されている各文書データに関連する情報を検索用データベース１０ａに蓄積し、クライアント端末３０から検索要求を取得したときに、この検索用データベース１０ａに蓄積された情報を検索するようになっている。

【００１４】
図２は、本実施形態に係る情報検索装置１０の構成を示す機能ブロック図である。同図において、設定ファイル解析部１００は、設定ファイル２００に示される指示に従って、文書データに関連する情報のうち、検索用データベース１０ａに蓄積すべき情報（以下、「検索用情報」という）を特定し、データ収集部１０２に出力する。ここで、設定ファイル２００は、例えばグループウェアサーバ２０の管理者などによって作成されるデータファイルであり、その構成を図３に示す。同図に示すように、設定ファイル２００には、取得項目、重み付け単語、格納先アドレスおよび格納元アドレスの各々が指定されている。
【００１５】
取得項目は、グループウェアサーバ２０が管理するデータ項目のうち、どの項目を取得するかを指定するものである。詳述すると、グループウェアサーバ２０は、文書データに関連する関連情報をデータ項目ごとに分けて記録されたグループウェアファイル２２を、文書データごとに備えている。図４は、このグループウェアファイルの一例を示す図である。この図において、文字列「ITEM_NAME」は、データ項目を示すものであり、この文字列「ITEM_NAME」と等号（＝）にて結ばれた文字列がデータ項目名を示す。例えば、「ITEM_NAME=Classification」である場合、データ項目名は、「分類（Classification）」となる。また、データ項目名（すなわち、文字列「ITEM_NAME」）の次行がデータ項目名に対応する文書データの関連情報である。具体的には、例えば、文字列「ITEM_NAME=Classification」の次行に記載された文字列「TYPE_TEXT=テクニカルノート」は、データ項目名「分類」に対応する文書データの関連情報が「テクニカルノート」であることを示している。そこで、取得項目は、グループウェアファイル２２に含まれるデータ項目名（文字列「ITEM_NAME」によって示されるデータ項目名）のうち、取得すべきデータ項目名を指定する。なお、図示を省略するが、このグループウェアファイル２２には、当該グループウェアファイル２２が、どの文書データに対応しているかも示されている。
【００１６】
また、設定ファイル２００における重み付け単語は、検索語として頻繁に用いられる単語を指定するためのものである。格納元アドレスは、検索対象となるデータベースが格納されているアドレスを示すものである。詳述すると、グループウェアサーバ２０は、上述したように、多数のデータベースを備えるのが一般的であり、このため、どのデータベースを検索対象とするかが特定される必要がある。そこで、アドレスを指定することにより、検索対象となるデータベースを特定するのである。また、格納先アドレスは、上述した格納元アドレスによって特定されるデータベース内の各データから検索用情報に従って抽出した情報を検索用データベース１０ａに格納するときのアドレスを示すものである。このように、検索対象となるデータベースごとに、異なる格納先アドレスが指定されることで、検索対象となるデータベースごとに抽出した多数の情報を検索用データベース１０ａに格納することができるようになっている。
【００１７】
さて、図２において、データ収集部１０２は、設定ファイル解析部１００からの検索用情報によって示される取得項目をグループウェアサーバ２０からネットワーク２を介して受け取り、次の処理を行うものである。すなわち、データ収集部１０２は、文書データおよびグループウェアファイル２２から取得した各項目のうち、文書データにおける本文部分に対応するものから本文データファイル２０２を生成するとともに、本文部分以外のものから情報データファイル２０４を生成し、各々をインデキシング部１０４に出力する。図５に示すように、本文データファイル２０２には、重み付け単語によって指定された単語（図示例では、「インターフェースデバイスＹＹＹ」など）が本文データの末尾に付加される（詳細については、後述）。また、図６に示すように、情報データファイル２０４に含まれる情報は、例えば、文書データに付されたタイトル（TITLE）や、グループウェアデータベース２０ａにおける文書データの格納元アドレス（URL：Uniform Resource Locator）などである。なお、データ収集部１０２がグループウェアサーバ２０から文書データを取得する機能は、グループウェアの製造元が提供するＡＰＩ（Application Program

Interface）によって実現されている。

【００１８】

インデキシング部１０４は、データ収集部１０２から受け取った本文データファイル２０２に対して形態素解析を行った後に、インデキシング（目次化）を実行し、この実行結果を、インデックスファイル２０６に登録するものであり、コンピュータにおけるＣＰＵに相当する。インデックスファイル２０６は、検索用データベース１０ａに格納されているものであり、インデックスファイル２０６には、ページテーブル２０６ａ、キーワードテーブル２０６ｃおよび単語テーブル２０６ｂが含まれている（図７参照）。なお、各データテーブルについては、後述する。

【００１９】

ここで、インデキシング部１０４が実行する形態素解析とは、漢字仮名交じりで記載された日本語の文を単語（形態素）に分解し、各単語の読み仮名や品詞などを特定することである。形態素解析用辞書１０６は、インデキシング部１０４における形態素解析に用いられる辞書であり、様々な単語を収録している。さらに説明すると、インデキシング部１０４は、解析対象となる文の続きの部分と最も長く一致する単語を形態素解析用辞書１０６から抽出するといったことを繰り返して文を単語（形態素）に分解する。なお、単語同士が空白で区切られる言語（例えば英語）にて本文データファイルの本文が記載されている場合には、形態素解析が必要ないことは勿論である。

【００２０】

図８は、上述したページテーブルの一例を示す図である。このページテーブル２０６ａは、各文書データの概要を示す情報を管理するためのものである。このページテーブル２０６ａの１つのレコードには、文書識別情報と、サーバ識別情報と、格納元アドレスと、最終更新日時情報と、題名情報と、本文情報と、分類情報と、総単語数情報と、ソフト別文書識別情報と、参照レベル情報との各々が含まれている。

【００２１】

ここで、文書識別情報は、グループウェアデータベース２０ａから取得した文書データごとに、情報検索装置１０が固有に割り当てる識別情報である。サーバ識別情報は、その文書データの取得元であるグループウェアサーバ２０を特定する情報であり、本実施形態にあっては、図８に示すように、情報検索装置１０がサーバごとに固有に割り当てた番号によって示される。格納元アドレスは、グループウェアデータベース２０ａにおける文書データの格納アドレスを示すものであり、図８に示すように、ＵＲＬによって指定されている。最終更新日時情報は、情報検索装置１０が文書データの情報を更新した最終日時を示す情報である。題名情報は、その文書データの題名（TITLE）を示す情報であり、例えば２５６バイトといった所定バイト数の文字列によって示される。本文情報は、その文書データの本文の先頭から所定文字数（例えば２５６バイト）分の文を示すものである。

【００２２】

また、分類情報は、文書データの文書の分類を示す情報である。より具体的には、例えば、文書データがコールセンター内のネットワークで共有されるものである場合、分類情報には、その文書データが製品のテクニカルサポート用文書なのか、製品のマニュアルなのかといったことを示す情報が記録される。総単語情報は、文書データの本文における総単語数を示すものである。ソフト別文書識別情報は、グループウェアサーバ２０が文書データに割り当てた固有の識別情報を示すものである。参照レベル情報は、その文書データの閲覧がネットワークに接続された各クライアント端末に限定されているか、または、ネットワーク外の端末にも許可されているかといった情報を示すものである。ここで、サーバ識別情報と、ソフト別文書識別情報とがページテーブル２０６ａに含まれているのは、多数のサーバに同一のグループウェアが導入されている場合に、各々のサーバが同一の識別情報を文書データに割り当てたときでも、どのサーバのどの文書データなのかを一意に特定できるようにするためである。

【００２３】

次いで、図９は、上述した単語テーブルの一例を示す図である。この単語テーブル２０６

ｂは、各文書データの本文に含まれる単語を管理するためのものである。より具体的には、図９に示すように、単語テーブル２０６ｂの１つのレコードには、単語と、情報検索装置１０が単語ごとに固有に割り当てられる単語識別情報と、グループウェアデータベース２０ａに蓄積されている全文書データのうち、この単語を本文に含む文書データの数を示す単語使用文書数とが含まれている。ここで、単語使用文書数は、インデキシング部１０４が文書データの本文データファイル２０２に対して形態素解析を行った結果に従って算出されるものである。具体的には、インデキシング部１０４は、１つの本文データファイル２０２に形態素解析を行って本文を単語（形態素）に分解した後に、各々の単語ごとに固有の識別情報を割り当てて、単語テーブル２０６ｂに登録する。そして、インデキシング部１０４は、登録した単語識別情報に対応する単語使用文書数の値を「１」だけインクリメントする。係る処理がグループウェアデータベース２０ａに蓄積されている全ての文書データについて行われた結果、単語ごとの単語使用文書数が得られる。

【００２４】

また、図１０は、上述したキーワードテーブルの一例を示す図である。このキーワードテーブル２０６ｃは、各文書データの本文に含まれる単語ごとに、１つの単語が何回出現しているかなどを管理するためのものである。具体的には、図１０に示すように、キーワードテーブル２０６ｃの１つのレコードには、上述した単語テーブル２０６ｂに含まれる単語識別情報と、上述したページテーブル２０６ａに含まれる文書識別情報と、出現回数と、重要度とが含まれている。出現回数は、単語が、文書識別情報によって特定される文書データの本文内に何回出現するかを示すものであり、インデキシング部１０４が行う形態素解析により得られる。さらに説明すると、インデキシング部１０４は、文書データの本文データファイル２０２の本文を単語（形態素）に分解した後に、その本文内に、単語識別情報によって示される単語が幾つ含まれるかを計数することにより、出現頻度を算出する。重要度は、全文書データの本文における単語の頻出度を示すものであり、次の式を用いてインデキシング部１０４により算出される。

（重要度）＝Ｓ×ｌｏｇ(Ｎ／ｎ)

ここで、Ｓは、出現回数、Ｎは、グループウェアデータベース２０ａに蓄積されている文書データの数、ｎは、上述した単語使用文書数である。この式によって示されるように、本文に同じ単語が含まれる文書データが多くなる程、その単語の重要度が小さくなり、また、１つの文書データの本文に同じ単語が頻繁に出現する程、その単語の重要度が高くなる。ここで、上述したように、文書データの本文データファイル２０２の末尾には、データ収集部１０２により重み付け単語が付与されているため、この重み付け単語の重要度は、相対的に高くなるのである。特に、文書データの題目（TITLE）には、その文書データの本文の内容を顕著に反映した単語が含まれることが多いため、この題目を本文データファイル２０２に重み付けするようにしても良い。

【００２５】

図２において、検索要求取得応答部１０８は、ネットワーク２を介してクライアント端末３０から検索要求を受け取り、検索部１１０に出力する。この検索要求取得応答部１０８は、コンピュータにおけるネットワークインターフェースデバイスに相当する。また、検索部１１０は、検索要求取得応答部１０８からの検索要求に応じて検索用データベース１０ａに格納されているインデックスファイル２０６を検索し、検索結果を、検索要求取得応答部１０８に出力する。検索要求取得応答部１０８は、検索部１１０から検索結果を受け取ると、この検索結果をネットワーク２を介してクライアント端末３０に送信する。

【００２６】

次いで、本実施形態に係る情報検索装置１０の動作について説明する。

ここで、以下に説明する各処理手順を規定するプログラムは、情報検索装置１０が備えるＲＯＭや磁気ディスクなどの記録媒体に格納されている。なお、このプログラムは、例えば、光ディスクや光磁気ディスク、磁気ディスクなどの可搬型の記録媒体に記録されたものが情報検索装置１０にインストールされたものでも良く、また、ネットワーク２を介して当該情報検索装置１０にインストールされたものであっても良い。

【００２７】
さて、情報検索装置１０は、グループウェアデータベース２０ａに蓄積されている各文書データの情報を示すインデックスファイル２０６に登録するための登録処理を実行する。具体的には、図１１に示すように、先ず、設定ファイル解析部１００が設定ファイル２００を読み出して、設定ファイル２００によって指示される取得項目、重み付け単語、格納元アドレスおよび格納先アドレスを特定し、これらの特定した情報を検索用情報としてデータ収集部１０２に出力する（ステップＳａ１）。
【００２８】
次に、データ収集部１０２は、設定ファイル解析部１００からの検索用情報によって示される取得項目をグループウェアサーバ２０からネットワーク２を介して受け取り、本文データファイル２０２（図５参照）および情報データファイル２０４（図６参照）を生成し、各々をインデキシング部１０４に出力する（ステップＳａ２）。
【００２９】
そして、インデキシング部１０４は、データ収集部１０２から受け取った本文データファイル２０２に対して形態素解析を行った後に、インデキシングを実行し、この実行結果を、３つのデータテーブルを含むインデックスファイル２０６に登録する。（ステップＳａ３）。これにより、１つの文書データに関する情報がインデックスファイル２０６に登録されることとなる。次いで、データ収集部１０２は、グループウェアデータベース２０ａ内に処理されてない文書データがあるかを判別し（ステップＳａ４）、この判別結果がＹＥＳであれば、残りの文書データの情報をインデックスファイル２０６に登録すべく、処理手順をステップＳａ２に戻す。一方、ステップＳａ４における判別結果がＮＯであれば、データ収集部１０２は、処理を終了する。これにより、グループウェアデータベース２０ａに蓄積されている全ての文書データの情報がインデックスファイル２０６に登録されることとなる。
【００３０】
ところで、グループウェアデータベース２０ａに蓄積されている文書データに対して、追加または削除が行われたり、また、１つの文書データに対して編集が行われたりといった編集処理が頻繁に行われる。そこで、情報検索装置１０は、インデックスファイル２０６に登録されている情報とグループウェアデータベース２０ａ内の各文書データの整合性が崩れないように、次のインデックスファイル修正処理を一定時間ごとに行っている。
【００３１】
すなわち、図１２に示すように、先ず、データ収集部１０２は、設定ファイル解析部１００からの検索用情報によって示される取得項目をグループウェアサーバ２０からネットワーク２を介して受け取り、本文データファイル２０２および情報データファイル２０４を生成し、各々をインデキシング部１０４に出力する（ステップＳｂ１）。インデキシング部１０４は、本文データファイル２０２、情報データファイル２０４およびインデックスファイル２０６に登録されている情報から、文書データが、▲１▼追加されたものであるか、▲２▼修正されたものであるか、▲３▼編集が加えられていないものか、を判別する（ステップＳｂ２）。
【００３２】
より具体的には、インデキシング部１０４は、情報データファイル２０４に含まれているサーバ識別情報およびソフト別文書識別情報に該当するものがインデックスファイル２０６のページテーブル２０６ａに登録されていなければ、この文書データが追加されたものであると判別する。一方、情報データファイル２０４に含まれているサーバ識別情報およびソフト別文書識別情報に該当するものが、インデックスファイル２０６のページテーブル２０６ａに既に登録されているものの、最終更新日時情報が情報データファイル２０４とインデックスファイル２０６との間で異なる場合には、インデキシング部１０４は、この文書データが修正されたと判別する。さらにまた、サーバ識別情報、ソフト別文書識別情報および最終更新日時情報の各々がいずれも情報データファイル２０４とインデックスファイル２０６との間で同じであれば、インデキシング部１０４は、この文書データに対

して何ら編集処理が成されていないと判別する。

【００３３】

さて、ステップＳｂ２における判別結果が、▲１▼追加されたものである、と判別された場合には、インデキシング部１０４は、上述した登録処理におけるステップＳａ３と同様の処理を実行し、この文書データの情報をインデックスファイル２０６に登録する（ステップＳｂ３）。次いで、データ収集部１０２は、グループウェアデータベース２０ａ内に処理されていない文書データがあるかを判別し（ステップＳｂ４）、この判別結果がＹＥＳであれば、残りの文書データを処理すべく、処理手順をステップＳｂ１に戻す。これにより、グループウェアデータベース２０ａに追加された文書データの情報がインデックスファイル２０６に新たに登録されることとなる。

【００３４】

一方、ステップＳｂ２の判別において、▲２▼修正されたものである、と判別された場合には、インデキシング部１０４は、この文書データに対応するインデックスファイル２０６の情報を一旦削除した後に、この文書データに対応する情報を新たに生成し、インデックスファイル２０６に登録する。より具体的には、インデキシング部１０４は、先ず、この文書データに対応する文書識別情報（図８参照）を特定し（ステップＳｂ５）、インデックスファイル２０６に含まれるページテーブル２０６ａ、単語テーブル２０６ｂ、キーワードテーブル２０６ｃの各々のテーブルから、特定した文書識別情報に関する情報を一括して削除する（ステップＳｂ６）。次いで、インデキシング部１０４は、この文書データに対応する情報を上述したインデキシング処理により生成し、インデックスファイル２０６に登録する（ステップＳｂ７）。次いで、データ収集部１０２は、グループウェアデータベース２０ａ内に処理されていない文書データがあるかを判別し（ステップＳｂ４）、この判別結果がＹＥＳであれば、残りの文書データを処理すべく、処理手順をステップＳｂ１に戻す。これにより、文書データに対して行われた修正がインデックスファイル２０６に反映されることとなる。また、ステップＳｂ２における判別結果が、▲３▼編集が加えられていないものであると判別された場合にも、インデキシング部１０４は、処理ステップをステップＳｂ４に進める。

【００３５】

次いで、ステップＳｂ４における判別結果がＮＯであれば、グループウェアデータベース２０ａ内の全ての文書データに対して処理が実行されたこととなる。従って、上述した一連の処理の間、インデックスファイル２０６（ページテーブル２０６ａ）において、一度も参照されなかった文書識別情報に対応する文書データは、グループウェアデータベース２０ａ内に存在しないこととなる。従って、インデキシング部１０４は、インデックスファイル２０６のページテーブル２０６ａから、参照されなかった文書識別情報を全て抽出し（ステップＳｂ８）、抽出した文書識別情報に対応する各情報を、インデックスファイル２０６に含まれる全てのテーブルから削除して（ステップＳｂ９）、処理を終了する。これにより、グループウェアデータベース２０ａから削除された文書データに対応する情報がインデックスファイル２０６から削除されることとなる。また、文書データが削除された場合、その文書識別情報に対応する情報をインデックスファイル２０６から削除するだけでよいため、インデックスファイル２０６の修正に要する時間が短縮される。

【００３６】

このように、インデックスファイル２０６には、グループウェアデータベース２０ａに蓄積されている各文書データの情報が登録され、文書データに対して、追加や削除、修正といった編集処理が行われたとしても、上述したインデックスファイル修正処理が一定時間ごとに繰り返し行われることで、その編集処理に応じて変更された情報がインデックスファイル２０６に即座に反映される。

【００３７】

さて、情報検索装置１０の検索要求取得応答部１０８は、クライアント端末３０からネットワーク２を介して検索要求を受け取ると、この検索要求を検索部１１０に出力する。検索部１１０は、受け取った検索要求に従ってインデックスファイル２０６を検索し、該当

する文書データの情報を抽出する。より具体的には、検索要求には、検索語として、検索用の単語、または、設定ファイル２００によって指定されたデータ項目が含まれている。例えば、検索要求に単語が検索語として含まれている場合、検索部１１０は、キーワードテーブル２０６ｃを参照し、その単語（詳細には、単語識別情報）の重要度が最も大きい順に文書識別情報を抽出する。そして、検索部１１０は、重要度の上位から所定の数（例えば２０など）だけの文書識別情報に対応する題名情報、本文情報および格納元アドレス（ＵＲＬ）などをページテーブル２０６ａから抽出し、検索要求取得応答部１０８を介してクライアント端末３０に送信する。これにより、クライアント端末３０に検索語に対応した文書データの候補が送信されることとなる。また、検索語として、例えば最終編集日時が検索要求に含まれていた場合には、検索部１１０は、ページテーブル２０６ａの各レコードを検索し、該当する文書識別情報に対応する題名情報、本文情報および格納元アドレス（ＵＲＬ）を検索要求取得応答部１０８を介してクライアント端末３０に送信する。なお、検索要求には、検索語として、単語およびデータ項目の各々が含まれていても良いことは勿論である。

【００３８】

このように、本実施形態によれば、グループウェアデータベース２０ａに蓄積されている文書データごとに、検索条件となり得る情報だけがインデックスファイル２０６に予め登録されている。情報検索装置１０は、検索要求を受けた場合には、このインデックスファイル２０６を検索すれば良く、インデックスファイル２０６のデータ量は、グループウェアデータベース２０ａに蓄積されている文書データのデータ量よりも小さいため、グループウェアデータベース２０ａの各文書データを対象として検索するよりも、速く検索が行える。さらに、利用者などが設定ファイル２００によって指定する取得項目を変更すれば、インデックスファイル２０６に登録されるデータ項目を変更することができるため、検索の用途に合わせてインデックスファイル２０６を構成しておくことができる。

また、本実施形態にて説明した情報検索装置１０は、複数のグループウェア間で汎用的に用いられ得るものである。さらに詳述すると、グループウェア毎に設定ファイル２００に記述する取得項目を変更するだけで、グループウェア毎にインデックスファイル２０６が構築されることになる。また、このような構成により、グループウェア毎にインデックスファイル２０６を構築すべく設定ファイル２００を変更したとしても、変更された設定ファイル２００に対応させて情報検索装置１０を動作させるべく、本実施形態に係る情報検索のためのプログラムを再度コンパイルする必要がない。

【００３９】

＜変形例＞

上述した実施形態は、あくまでも例示であって、本発明の一態様を示すものであり、本発明の範囲内で任意に変形可能である。そこで、以下に、各種の変形例について説明する。

【００４０】

例えば、上述した実施形態では、ネットワーク２にグループウェアサーバ２０が１つだけ接続される構成について例示したが、これに限らず、グループウェアサーバ２０が複数接続される構成であっても良い。さらに、夫々のグループウェアサーバ２０には、互いに異なるグループウェアが導入されていても良い。さらに詳述すると、互いに異なる複数のグループウェアサーバの各々のデータベースを統括的に検索することは、グループウェア毎にデータの管理形式（例えばデータ項目の数や名前など）が異なるため、一般的に困難である。これに対して、本変形例は、検索対象となり得るデータ項目の情報だけをインデックスファイル２０６のページテーブル２０６ａに登録する構成となっている。従って、情報検索装置１０がページテーブル２０６ａを検索することは、複数のグループウェアサーバの各々のデータベースを検索することと同等なことであり、これにより、複数のグループウェアサーバの各々のデータベースの検索が実現される。

【００４１】

また、例えば、インデキシング部１０４は、本文データファイル２０２に対して形態素解析を行う際に、例えば「ＰＣ」、「パーソナルコンピュータ」、「パソコン」といった、

互いに同一のものを指す単語を一つの単語として扱っても良い。これにより、例えば、検索語として「パソコン」が検索要求に含まれていた場合でも、「ＰＣ」や「パーソナルコンピュータ」といった単語を含む文書データも該当する文書データとして抽出され、検索の精度が向上する。

【００４２】

【発明の効果】

本発明によれば、データベースに蓄積されている情報のうち、検索条件に該当する情報を特定するに要する時間を短縮することが可能な情報検索装置、情報検索方法、プログラムおよび記録媒体が提供される。

【図面の簡単な説明】

【図１】　本発明の実施形態に係る情報検索システムの構成を示すブロック図である。

【図２】　情報検索装置の機能的構成を示すブロック図である。

【図３】　同設定ファイルの一例を示す図である。

【図４】　同グループウェアファイルの一例を示す図である。

【図５】　同本文データファイルの一例を示す図である。

【図６】　同情報データファイルの一例を示す図である。

【図７】　同インデックスファイルのデータ構成を示す概念図である。

【図８】　同ページテーブルの一例を示す図である。

【図９】　同単語テーブルの一例を示す図である。

【図１０】　同キーワードテーブルの一例を示す図である。

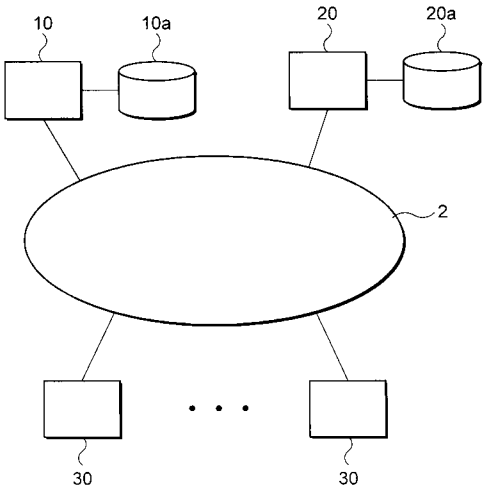
【図１１】　情報検索装置によって実行される登録処理の手順を示すフローチャートである。

【図１２】　情報検索装置によって実行されるインデックスファイル修正処理の手順を示すフローチャートである。

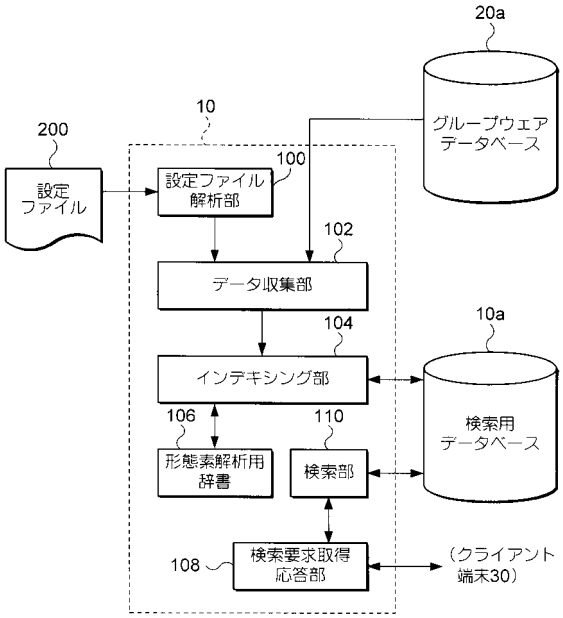
【符号の説明】

１０・・・情報検索装置、１０ａ・・・検索用データベース、２０・・・・グループウェアサーバ、２０ａ・・・グループウェアデータベース、３０・・・クライアント端末、１００・・・設定ファイル解析部、１０２・・・データ収集部、１０４・・・インデキシング部、１０６・・・形態素解析用辞書、１０８・・・検索要求取得応答部、１１０・・・検索部、２００・・・設定ファイル、２０６・・・インデックスファイル。

【図１】



【図２】



【図３】

200

取得項目	“Classification”, “Tifle”, …
重み付け単語	“動作”, “インストール”, …
格納元アドレス	http://abcde/xxxx
格納先アドレス	http://vwx/zzzz

【図４】

22

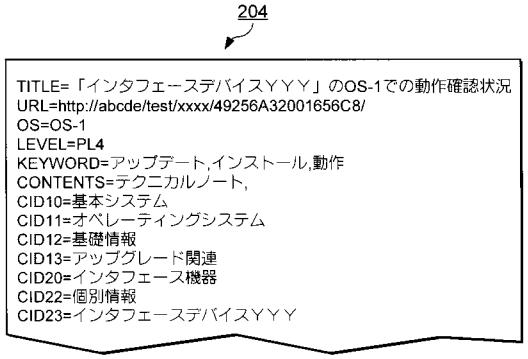
```
ITEM_NAME=Classification_m
TYPE_TEXT=テクニカルノート
ITEM_NAME=Subject
TYPE_TEXT「=インターフェースデバイスＹＹＹ」のOS-1での動作確認状況
ITEM_NAME=PC_m
TYPE_PC=PC-1
TYPE_PC=PC-2
ITEM_NAME=OS_m
TYPE_OS=OS-1
TYPE_OS=OS-2
ITEM_NAME=Category1_0_m
TYPE_TEXT=インターフェースデバイス
ITEM_NAME=Category1_2_m
TYPE_TEXT=トラブル
ITEM_NAME=Keyword_m
TYPE_KEYWORD=インターフェースデバイス
TYPE_KEYWORD=動作確認
TYPE_KEYWORD=OS
```

【図５】

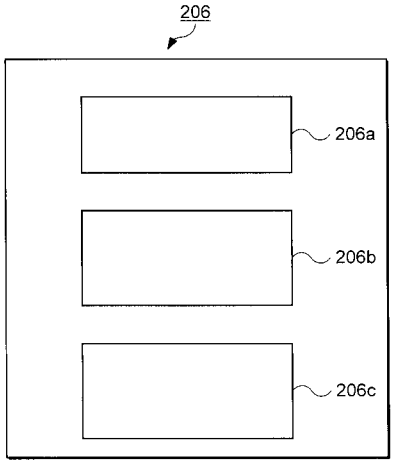
202

「インタフェースデバイスＹＹＹ」OS-1での動作確認状況
「使用ドライバ」
－－OS-1標準ドライバ＝
　OS-1をインストールすると同時にインストールされるドライバ
－－専用ドライバ＝
　インターフェースデバイスＹＹＹに最適なOS-1用のドライバ
－－OS-2用ドライバ
　OS-2用のドライバ
－－「OS-2標準ドライバ名」欄＝
　OS－2インストール時にインストールされるデバイスドライバ名を記載しています。 OS-2標準ドライバ名は次の方法で確認することができます。1.「コントロールパネル」の「システム」アイコンをダブルクリックします。
－－「注意事項/制限事項」欄＝
'00年9月20日現在、確認されている注意事項/制限事項を記載しています。
－－使用ドライバ－ OS-1標準ドライバ※OS-2用ドライバについては未評価
－－OS-1標準ドライバ名・インターフェースデバイスＹＹＹ ＺＺ
－－注意事項/制限事項---・OS-1標準ドライバを使用します。・SCSI HDDが接続されている場合、「SCSIのボードを低電力状態にできないため、コンピュータをスタンバイまたは休止状態にできません。」と表示され、スタンバイおよび休止状態に移行することができません。・CD-RWドライブで書き込みができない場合があります。
－－アップデート,インストール,動作,アップデート,インストール,動作,「インタフェースデバイスＹＹＹ」OS-1での動作確認状況,「インタフェースデバイスＹＹＹ」OS-1での動作確認状況,基本システム,オペレーティングシステム,基礎情報,アップグレード関連,インタフェース機器,SCSIボード,個別情報,インタフェースデバイスＹＹＹ,基本システム,オペレーティングシステム,基礎情報,アップグレード関連,インタフェース機器,SCSIボード,個別情報,インタフェースデバイスＹＹＹ

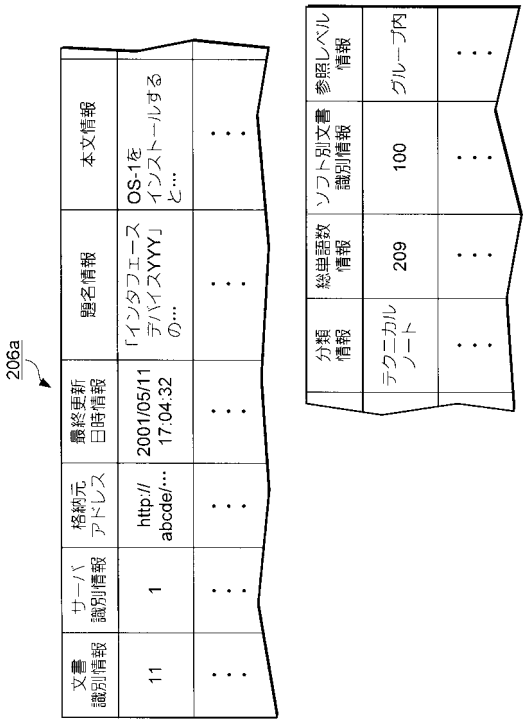
【図６】



【図７】



【図８】



【図９】

206b

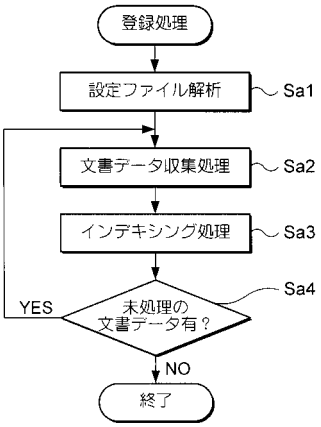
単語	単語識別情報	単語使用文書数
デバイス	1	2264
YYY	2	44
インストール	3	169
⋮	⋮	⋮

【図１０】

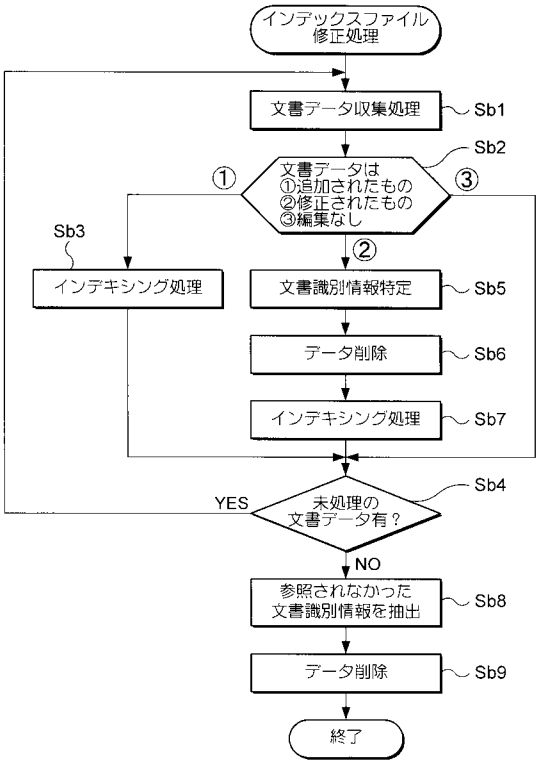
206c

単語識別情報	文書識別情報	出現回数	重要度
1	144	1	0.0000817
1	286	24	0.001961
⋮	⋮	⋮	⋮

【図１１】



【図１２】


インデックスファイル
修正処理
文書データ収集処理
Sb1
Sb2
文書データは
①追加されたもの
②修正されたもの
③編集なし
①
③
②
Sb3
インデキシング処理
文書識別情報特定
Sb5
データ削除
Sb6
インデキシング処理
Sb7
Sb4
YES
未処理の
文書データ有？
NO
参照されなかった
文書識別情報を抽出
Sb8
データ削除
Sb9
終了

フロントページの続き

(56)参考文献 特開２０００－３３０８３８（ＪＰ，Ａ）
特開２０００－２２２４３２（ＪＰ，Ａ）
特開２０００－２９８６６８（ＪＰ，Ａ）
データウェアハウス研究会，入門マネジメント＆ストラテジー　よくわかるデータウェアハウス，日本，株式会社　日本実業出版社，２０００年　８月１０日，第１版，76～77,86～87

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/30 ，
G06F 17/28