

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2018年5月17日(17.05.2018)



(10) 国際公開番号

WO 2018/088027 A1

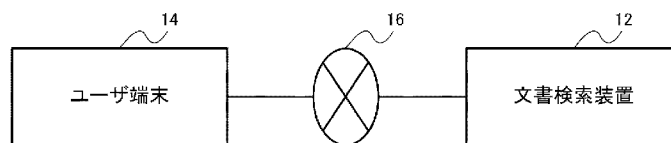
- (51) 国際特許分類:
G06F 17/30 (2006.01) *G06F 17/22* (2006.01)
- (21) 国際出願番号: PCT/JP2017/033316
- (22) 国際出願日: 2017年9月14日(14.09.2017)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2016-217884 2016年11月8日(08.11.2016) JP
- (71) 出願人: 株式会社野村総合研究所 (NOMURA RESEARCH INSTITUTE, LTD.)
[JP/JP]; 〒1000004 東京都千代田区大手町
1丁目9番2号 Tokyo (JP).
- (72) 発明者: 毛 羽翔(Mao Yuxiang); 〒1000004 東京都千代田区大手町二丁目2番1号 N R I
サイバーパテント内 Tokyo (JP).
- (74) 代理人: 森 下 賢 樹 (MORISHITA Sakaki);
〒1500021 東京都渋谷区恵比寿西2-
11-12 Tokyo (JP).
- (81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

(54) Title: DOCUMENT SEARCH DEVICE, DOCUMENT SEARCH METHOD, AND COMPUTER PROGRAM

(54) 発明の名称: 文書検索装置、文書検索方法およびコンピュータプログラム

【図1】

文書検索システム 10



10 Document search system
12 Document search device
14 User terminal

(57) **Abstract:** A document search device 12 stores a plurality of instances of document data in which a plurality of different words in original documents have been converted into codes of respectively different fixed lengths. A user terminal 14 transmits to the document search device 12 a search request which designates a plurality of words and which designates, with a quantity of words, a range in which the plurality of words are to be present. The document search device 12 acquires the fixed-length codes which correspond to each of the plurality of words which have been designated in the search request, and derives a character quantity reference range according to the word quantity reference range which has been designated in the search request and the fixed-length codes. The document search device 12 executes a neighborhood search with the codes of the plurality of words and the character quantity reference range as conditions, and extracts from the plurality of instances of document data the instances of document data which satisfy the conditions.

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

- 一 国際調査報告(条約第21条(3))

(57) 要約: 文書検索装置12は、オリジナル文書に記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換された複数の文書データを記憶する。ユーザ端末14は、複数の単語を指定し、かつ、それら複数の単語が存在すべき範囲を単語数で指定する検索要求を文書検索装置12へ送信する。文書検索装置12は、検索要求で指定された複数の単語のそれぞれに対応する固定長のコードを取得し、検索要求で指定された単語数基準の範囲と、固定のコード長とに応じて、文字数基準の範囲を導出する。文書検索装置12は、複数の単語のコードと文字数基準の範囲とを条件とする近傍検索を実行し、複数の文書データの中から上記条件を満たす文書データを抽出する。

明 細 書

発明の名称：

文書検索装置、文書検索方法およびコンピュータプログラム

技術分野

[0001] この発明は、データ処理技術に関し、特に文書を検索する技術に関する。

背景技術

[0002] N－g r a m方式の全文検索技術が広く普及している（例えば特許文献1参照）。また現在、文献の検索サービスの中には、ユーザが指定する複数のキーワードが、ユーザが指定する範囲（以下「近傍範囲」とも呼ぶ。）に記載された文献を検索する近傍検索を提供するものもある。

先行技術文献

特許文献

[0003] 特許文献1：特開2006－99427号広報

発明の概要

発明が解決しようとする課題

[0004] N－g r a m方式の検索エンジンでは、文字数基準の近傍検索は可能であるが、単語数基準の近傍検索はできない。言い換えれば、近傍検索の条件である近傍範囲を文字数では指定できるが、単語数では指定できない。単語ごとに文字数は異なりうるところ、近傍範囲を単語数で指定することへのニーズは高い。しかし、単語数基準の近傍検索が可能なように既存の検索エンジンを改修し、または、既存の検索エンジンを、単語数基準の近傍検索が可能な別の検索エンジンに置き換えるには多大なコストを要する。

[0005] 本発明は、こうした課題に鑑みてなされたものであり、主たる目的は、N－g r a m方式の検索エンジンを使用しつつ、単語数基準の近傍検索を実現する技術を提供することである。

課題を解決するための手段

[0006] 上記課題を解決するために、本発明のある態様の文書検索装置は、複数の文書データを記憶する第1文書記憶部であって、複数の文書データのそれぞれは、オリジナル文書に記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換されたものである第1文書記憶部と、複数の単語を指定する検索要求であって、かつ、それら複数の単語が存在すべき範囲を単語数で指定する検索要求を受け付ける受付部と、検索要求で指定された複数の単語のそれぞれに対応する固定長のコードを取得する取得部と、検索要求で指定された単語数基準の範囲と、固定のコード長とに応じて、文字数基準の範囲を導出する導出部と、取得部により取得された複数の単語のコードと、導出部により導出された文字数基準の範囲とを条件とする近傍検索を実行し、第1文書記憶部に記憶された複数の文書データの中から条件を満たす文書データを抽出する検索部と、を備える。

[0007] 本発明の別の態様は、文書検索方法である。この方法は、複数の文書データを記憶する文書記憶部であって、複数の文書データのそれぞれは、オリジナル文書に記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換されたものである文書記憶部にアクセス可能なコンピュータが、複数の単語を指定する検索要求であって、かつ、それら複数の単語が存在すべき範囲を単語数で指定する検索要求を受け付けるステップと、検索要求で指定された複数の単語のそれぞれに対応する固定長のコードを取得するステップと、検索要求で指定された単語数基準の範囲と、固定のコード長とに応じて、文字数基準の範囲を導出するステップと、取得するステップで取得された複数の単語のコードと、導出するステップで導出された文字数基準の範囲とを条件とする近傍検索を実行し、文書記憶部に記憶された複数の文書データの中から条件を満たす文書データを抽出するステップと、を実行する。

[0008] なお、以上の構成要素の任意の組合せ、本発明の表現をシステム、コンピュータプログラム、コンピュータプログラムを格納した記録媒体などの間で変換したものもまた、本発明の態様として有効である。

発明の効果

[0009] 本発明によれば、N－g r a m方式の検索エンジンを使用しつつ、単語数基準の近傍検索を実現することができる。

図面の簡単な説明

[0010] [図1]第1実施例の検索システムを示す図である。

[図2]図1の文書検索装置の機能構成を示すブロック図である。

[図3]辞書記憶部の構成例を模式的に示す図である。

[図4]第1実施例の文書検索装置の動作を示すフローチャートである。

[図5]第1実施例の文書検索装置の動作を示すフローチャートである。

[図6]第2実施例の文書検索装置の動作を示すフローチャートである。

[図7]第2実施例の文書検索装置の動作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

[0011] 本発明の実施例を詳細に説明する前にまず概要を説明する。

実施例の文書検索装置が備える検索エンジンは、N－g r a m方式の文書検索を実行する。例えば、「e a t」で検索すれば、「b e a t」「h e a t」「b e a t e n」は全てヒットする。N－g r a m方式の検索エンジンは、単語の区切りを認識しない。例えば、「e a t」と「b e a t e n」はどちらも「e a t」による検索にヒットする。

[0012] また、N－g r a m方式の検索エンジンは、近傍検索において単語数を考慮しない。例えば、「e a t p a t h o l o g i c a l l y y o u a r e d i s g u s t i n g」と「e a t i t m u c h f a s t e r y o u c a n h u r t y o u r s e l f」は、「e a t」と「y o u」の間の文字数は同じだが、間の単語数は1と3で相当違う。N－g r a m方式の検索エンジンでは、文字数に基づく近傍検索はできるが、単語数に基づく近傍検索はできない。

[0013] 近傍検索においてユーザは、複数のキーワードと、それら複数のキーワードが存在すべき範囲（すなわち近傍範囲）を条件として指定する。この範囲は、複数のキーワードの近さの程度を示すものと言え、以下では「距離」と呼ぶ。文字数で指定される距離であり、言い換えれば、文字数基準の近傍範

囲を「文字距離」と呼び、単語数で指定される距離であり、言い換えれば、単語数基準の近傍範囲を「単語距離」と呼ぶ。

[0014] ここで、文字数に基づく近傍検索の例を示す。例えば、キーワード「eat」、キーワード「you」、文字距離「15」を条件とする近傍検索が実行されると、15文字の範囲内に「eat」と「you」の両方を含む文書がヒットする。なお、実施例のN-gram方式の検索エンジンは、繋がっている文字列を完全一致で検索する。例えば、3-gramの場合に、キーワード「aboard」による検索では「aboveboard」はヒットしない。

[0015] 実施例では、N-gram方式の検索エンジンを使用して、単語数に基づく近傍検索を実現する。例えば、「車にコンピュータ」は「車」「に」「コンピュータ」の3単語で構成されるが、キーワードを「車」および「コンピュータ」とし、単語距離を「3」とする近傍検索において、「車にコンピュータ」を含む文書がヒットするようにする。なお、単語距離「2」は2つのキーワードが隣接することを意味する。単語距離「N（Nは3以上）」は複数のキーワードが、キーワードも含むN単語内に存在することを意味する。

[0016] 具体的には、実施例の文書検索装置は、検索対象の複数の文書として、オリジナル文書の各単語を固定長のコードに変換した文書を記憶する。文書検索装置は、検索要求を受け付けると、検索要求で指定されたキーワードを固定長のコードに変換する。それとともに、検索要求で指定された単語距離と固定のコード長とに基づいて文字距離を導出する。文書検索装置は、キーワードのコードと文字距離とをN-gram方式の検索エンジンに入力して、文字数に基づく近傍検索を実行させる。このように、検索対象文書内の各単語および検索キーワードを固定長のコードへ変換することで、ユーザが指定する単語距離を文字距離に変換可能になる。実施例の文書検索装置12によると、既存のN-gram方式の検索エンジンを使用しつつ、単語数に基づく近傍検索を実現することができる。言い換えれば、文字数に基づく近傍検索を実行する既存の検索エンジンを使用して、単語数に基づく近傍検索を実

現することができる。

[0017] (第1実施例)

図1は、第1実施例の文書検索システム10を示す。文書検索システム10は、文書検索装置12とユーザ端末14を備える。文書検索装置12は、キーワード検索や近傍検索等の文書検索サービスを提供する情報処理装置である。例えば、文書検索装置12は、インターネットを介して、特許文献の検索サービスを提供するサーバであってもよい。

[0018] ユーザ端末14は、文書検索装置12が提供する文書検索サービスの利用者（以下「ユーザ」と呼ぶ。）により操作される情報処理装置である。ユーザ端末14は、例えば、PC、スマートフォン、タブレット端末であってもよい。文書検索装置12とユーザ端末14は、LAN・WAN・インターネット等を含む通信網16を介して接続される。なお、図1には1つのユーザ端末14を描いているが、実際には、複数のユーザにより操作される複数のユーザ端末14が文書検索装置12に接続されてよい。

[0019] 図2は、図1の文書検索装置12の機能構成を示すブロック図である。文書検索装置12は、制御部20、記憶部22、通信部24を備える。制御部20は、文書検索に関する各種データ処理を実行する。記憶部22は、制御部20により参照または更新されるデータを記憶する記憶領域である。通信部24は、所定の通信プロトコルにしたがって外部装置と通信する。制御部20は、通信部24を介してユーザ端末14とデータを送受する。

[0020] 本明細書のブロック図において示される各ブロックは、ハードウェア的には、コンピュータのCPU・メモリをはじめとする素子や機械装置で実現でき、ソフトウェア的にはコンピュータプログラム等によって実現されるが、ここでは、それらの連携によって実現される機能ブロックを描いている。したがって、これらの機能ブロックはハードウェア、ソフトウェアの組合せによっていろいろなかたちで実現できることは、当業者には理解される場所である。

[0021] 例えば、制御部20の各ブロックの機能はコンピュータプログラムとして

実装され、そのコンピュータプログラムが文書検索装置 12 のストレージにインストールされてもよい。そして文書検索装置 12 の CPU が、そのコンピュータプログラムをメインメモリへ読み出して実行することにより制御部 20 の各ブロックの機能が発揮されてもよい。また、記憶部 22 は、文書検索装置 12 のメインメモリやストレージにより実現されてもよい。なお、文書検索装置 12 は、ウェブサーバ、アプリケーションサーバ、データベースサーバ等の複数台の装置が通信網を介して連携することにより実現されてもよい。

[0022] 記憶部 22 は、辞書記憶部 26、原本文書記憶部 28、コード化文書記憶部 30 を含む。原本文書記憶部 28 は、複数のオリジナル文書データを記憶する。オリジナル文書データは、文書内に記載された単語がコードに変換される前の文書データであり、言い換えれば、作成時点の内容の文書データである。オリジナル文書データは、日本語や英語等、様々な言語で記載された文書を含む。例えば、オリジナル文書データは、公開特許公報と管理データ（各種番号等）、特許掲載公報と管理データを含んでもよい。

[0023] 辞書記憶部 26 は、オリジナル文書データに含まれる複数の単語と、複数のコードとの対応関係を記憶する。この単語は、形態素とも言え、日本語や英語等、様々な言語において定められた単語を含む。図 3 は、辞書記憶部 26 の構成例を模式的に示す。辞書記憶部 26 は、様々な言語における互いに異なる複数の単語を、互いに異なる固定長のコードに対応付けた辞書データを記憶する。コードは、複数の単語それぞれのユニークな ID とも言える。実施例のコードは、11 バイトの固定長データであり、具体的には、11 文字固定の文字列とする。各コードの先頭（例えば 1 バイト目）には先頭以外では使用されない特殊値（言い換えれば区切り文字）が設定される。特殊値は、コードの 2 ～ 11 バイト目では使用されない値（ビットの並び）であってもよい。実施例ではコードの先頭の特殊値を「#」で示す。

[0024] 図 2 に戻り、コード化文書記憶部 30 は、オリジナル文書データに記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換された

文書データ（以下「コード化文書データ」とも呼ぶ。）を記憶し、複数のコード化文書データを記憶する。コード化文書データは、例えば、オリジナルの公開特許公報に記載された複数の単語がコードに変換された公開特許公報のコード化データと、オリジナルの特許掲載公報に記載された複数の単語がコードに変換された特許掲載公報のコード化データを含んでもよい。

[0025] コード化文書データの例を図3を参照しつつ説明する。例えば、オリジナル文書データに文字列「私の発明」（「私」「の」「発明」の3単語）が含まれる場合、コード化文書データには、上記文字列が「#0024F76DA7#0024F76DD8#0024F76DA6」に変換されて記録される。

[0026] なお、コード化文書記憶部30に記憶された複数のコード化文書データのそれぞれは、原本文書記憶部28に記憶された変換前のオリジナル文書データに対応付けられる。複数のコード化文書データのそれぞれは、対応するオリジナル文書データの識別子を含んでもよく、対応するオリジナル文書データのメモリ上のアドレス（すなわちポインタ）を含んでもよい。

[0027] 制御部20は、文書変換部32、検索要求受付部34、コード取得部36、文字距離導出部38、検索指示部40、検索実行部42、検索結果提供部44を含む。不図示だが、制御部20は、公知のウェブサーバの機能を含んでもよい。

[0028] 文書変換部32は、原本文書記憶部28に記憶されたオリジナル文書データからコード化文書データを生成する。例えば、文書変換部32は、公知の形態素解析処理を実行して、オリジナル文書データに記載された複数の単語を識別する。文書変換部32は、辞書記憶部26の辞書データを参照して、オリジナル文書データに記載された複数の単語を、各単語に対応する固定長コードに置換することによりコード化文書データを生成する。文書変換部32は、生成したコード化文書データをコード化文書記憶部30に格納する。

[0029] 検索要求受付部34は、文書検索装置12から送信された検索要求を受け付ける。実施例の検索要求は、近傍検索を要求するクエリ電文である。具体

的には、検索要求では、複数の単語が指定され、かつ、それら複数の単語が存在すべき範囲が単語数で指定される。言い換えれば、実施例の検索要求は、複数のキーワードと、それらキーワード間の単語距離とを指定するものである。

[0030] コード取得部 36 は、辞書記憶部 26 の辞書データを参照して、検索要求受付部 34 により受け付けられた検索要求で指定された複数の単語のそれぞれに対応するコードを取得する。

[0031] 文字距離導出部 38 は、検索要求受付部 34 により受け付けられた検索要求で指定された単語距離と、固定のコード長（実施例では 11 文字）とに応じて、近傍検索の条件となる文字距離を導出する。具体的には、文字距離導出部 38 は、（単語距離 × コード長）の結果を文字距離として導出する。

[0032] 検索指示部 40 と検索実行部 42 は互いに連携して、近傍検索を含む各種検索処理を実行する検索部として機能する。検索指示部 40 は、キーワードとしてコード取得部 36 により取得された複数のコードを指定する検索指示であり、文字距離導出部 38 により導出された文字距離を指定する検索指示を検索実行部 42 へ入力する。

[0033] 検索実行部 42 は、検索指示部 40 から入力された検索指示に応じて、N-gram 方式による文書検索処理を実行する検索エンジンである。検索実行部 42 は、ユーザ端末 14 から単語距離に基づく近傍検索が要求されたか否かに関わらず、文字距離に基づく近傍検索を実行する。

[0034] 具体的には、検索実行部 42 は、検索指示で指定された複数のキーワードと文字距離とを条件とする近傍検索を実行する。検索実行部 42 は、コード化文書記憶部 30 に記憶された複数のコード化文書データの中から上記条件を満たすコード化文書データを抽出する。また、検索実行部 42 は、コード化文書記憶部 30 から抽出した（すなわち近傍検索にヒットした）コード化文書データに対応するオリジナル文書データを原本文書記憶部 28 から抽出する。検索実行部 42 は、原本文書記憶部 28 から抽出したオリジナル文書

データ、および／または、コード化文書記憶部 30 から抽出したコード化文書データを検索結果提供部 44 へ出力する。

[0035] 検索結果提供部 44 は、検索実行部 42 により抽出されたオリジナル文書データを、検索要求元のユーザ端末 14 へ送信する。例えば、オリジナル文書データを示す識別情報（各種管理番号等）や、オリジナル文書データの本文の少なくとも一部をユーザ端末 14 へ送信してもよい。変形例として、検索結果提供部 44 は、オリジナル文書データとともに、またはオリジナル文書データに代えて、検索実行部 42 により抽出されたコード化文書データ（例えばオリジナル文書データと共通の管理番号等）をユーザ端末 14 へ送信してもよい。また、変形例として、ユーザに対してオリジナル文書データの全部又は一部を提示する場合に、オリジナル文書データ中検索指示で指示された複数のキーワードが出現する箇所をハイライト（顕示）してもよい。

[0036] 既述したように、コード化文書記憶部 30 に記憶されるコード化文書データは、オリジナル文書データに記録された単語に対して予め対応付けられたコードが記録されたものである。文書変換部 32 は、オリジナル文書データに記録された或る単語（「特定単語」と呼ぶ。）に対応するコードが辞書記憶部 26 の辞書データに存在しない場合、特定単語を複数の単文字単語に分解する。単文字単語は、1 文字で構成された単語であり、例えば「a」「b」「あ」「い」「日」「月」等を含む。辞書記憶部 26 の辞書データには、複数の単文字単語に対応する複数のコードが予め格納される。文書変換部 32 は、辞書記憶部 26 の辞書データを参照して、複数の単文字単語のそれぞれに予め対応付けられたコードをコード化文書データに記録する。

[0037] コード取得部 36 は、検索要求で指定された複数の単語のうち一部の単語に対応するコードが辞書データに存在しない場合、対応するコードが存在しない単語（すなわち特定単語）を複数の単文字単語に分解し、複数の単文字単語それぞれのコードを取得する。

[0038] 文字距離導出部 38 は、検索要求で指定された単語距離を、特定単語を分解した単文字単語の個数に応じて拡大する。具体的には、（検索要求で指定

された単語距離－特定単語数＋各特定単語の単文字単語の個数）を新たな単語距離として導出する。例えば、検索要求で指定された単語距離が５、特定単語数が１、この特定単語が３つの単文字単語に分解される場合（言い換えれば特定単語の文字数が３の場合）、 $(5 - 1 + 3 = 7)$ を新たな単語距離として導出する。文字距離導出部３８は、新たな単語距離に基づいて文字距離を導出する。例えば、新たな単語距離が「７」の場合、固定のコード長「１１」との積である「７７」を文字距離として導出する。このように、近傍検索における単語数の増加に応じて単語距離を拡大することで、特定単語を複数の単文字単語に分解した結果、ユーザが指定した単語距離が実質的に短くなってしまい検索精度が低下することを回避できる。

[0039] 検索指示部４０は、検索要求で指定された複数の単語のうち対応するコードが存在する１つ以上の単語のコードと、複数の単文字単語のコードと、文字距離導出部３８により導出された文字距離を条件とする近傍検索指示を検索実行部４２に入力する。検索実行部４２は、これらの条件に基づく近傍検索を実行し、これらの条件を満たすコード化文書データをコード化文書記憶部３０から抽出する。

[0040] 検索指示部４０はさらに、近傍検索により抽出された１つ以上のコード化文書データを母集団とした新たな近傍検索を検索実行部４２にさらに実行させる。具体的には、複数の単文字単語のコードが隣接することを条件とする新たな近傍検索をさらに実行させる。例えば、或る１つの特定単語を４つの単文字単語に分解した場合、文字距離「４」に基づく新たな近傍検索をさらに実行させる。検索実行部４２は、検索指示にしたがって新たな近傍検索を実行し、先の近傍検索の結果を絞り込む。

[0041] 以上の構成による文書検索装置１２の動作を説明する。

図４は、第１実施例の文書検索装置１２の動作を示すフローチャートである。同図は、コード化文書データ生成時の動作を示している。原本文書記憶部２８には、検索の母集団に含めるべきオリジナル文書データ（例えば新たに公開された公開特許公報等）が随時追加されていく。文書変換部３２は、

新たなオリジナル文書データが原本文書記憶部 28 に格納されるまで待機する (S 10 の N)。文書変換部 32 は、新たなオリジナル文書データが原本文書記憶部 28 に格納されたことを検出すると (S 10 の Y)、新たなオリジナル文書データに記載された文字列に対して形態素解析処理を実行し、新たなオリジナル文書データに記載された複数の単語を抽出する (S 11)。

[0042] 文書変換部 32 は、辞書データを参照して、S 11 で抽出した各単語に対応するコードを取得する (S 12)。オリジナル文書データに含まれる少なくとも 1 つの単語 (「特定単語」と呼ぶ。) に対応するコードが辞書データに未定義の場合 (S 14 の Y)、文書変換部 32 は、特定単語を複数の単文字単語に分解し、各単文字単語に対応するコードを取得する (S 16)。オリジナル文書データに特定単語が存在しなければ、すなわちオリジナル文書データの全ての単語が辞書データに定義されていれば (S 14 の N)、S 16 をスキップする。文書変換部 32 は、オリジナル文書データの各単語 (単文字単語を含みうる) を固定長のコードに変換したコード化文書データをコード化文書記憶部 30 に格納する (S 18)。

[0043] 図 5 も、第 1 実施例の文書検索装置 12 の動作を示すフローチャートである。同図は、検索時の動作を示している。検索要求受付部 34 は、ユーザ端末 14 から送信された検索要求を受信するまで待機する (S 20 の N)。検索要求 (具体的には複数の単語と、単語距離を指定した近傍検索要求) を検索要求受付部 34 が受信すると (S 20 の Y)、コード取得部 36 は、辞書データを参照して、検索要求で指定された複数の単語に対応する複数のコードを識別する (S 22)。

[0044] 少なくとも 1 つの単語 (「特定単語」と呼ぶ。) に対応するコードが辞書データに未定義であれば (S 24 の Y)、コード取得部 36 は、メモリの所定領域に設けた特殊フラグをオンに設定する (S 26)。コード取得部 36 は、特定単語を複数の単文字単語に分解し、各単文字単語に対応するコードを取得する (S 28)。文字距離導出部 38 は、検索要求で指定された単語距離を、単文字単語の個数に応じて拡大する (S 30)。具体的には、単文

字単語の個数と特定単語の個数との差分だけ単語距離を増加させる。検索要求で指定された単語に特定単語が含まれず、すなわち、指定された全ての単語が辞書データに定義されていれば（S 2 4 の N）、S 2 6～S 3 0 をスキップする。

[0045] 文字距離導出部 3 8 は、検索要求で指定された単語距離または S 3 0 で拡大した単語距離と、固定のコード長とに基づいて、近傍検索のための文字距離を導出する（S 3 2）。検索指示部 4 0 は、検索要求で指定された複数の単語のコードと、特定単語があればその特定単語を分解した複数の単文字単語のコードとをキーワードとして指定した検索指示であり、S 3 2 で導出された文字距離をさらに指定した近傍検索指示を検索実行部 4 2 へ入力する。検索実行部 4 2 は、コード化文書記憶部 3 0 に格納された複数のコード化文書データを母集団とする近傍検索処理であり、N - g r a m 方式による文字数基準の近傍検索処理を実行する（S 3 4）。検索実行部 4 2 は、近傍検索にヒットしたコード化文書データをコード化文書記憶部 3 0 から抽出し、抽出したコード化文書データに対応するオリジナル文書データを原本文書記憶部 2 8 から抽出する。

[0046] 特殊フラグがオンであれば（S 3 6 の Y）、検索指示部 4 0 は、S 3 4 の近傍検索処理で抽出された 1 つ以上のコード化文書データを母集団とした絞込検索であり、複数の単文字単語のコードが隣接することを条件とする新たな近傍検索を検索指示部 4 0 にさらに実行させる（S 3 8）。検索指示部 4 0 は、特殊フラグをオフに戻す（S 4 0）。特殊フラグがオフであれば（S 3 6 の N）、S 3 8、S 4 0 をスキップする。検索結果提供部 4 4 は、検索実行部 4 2 により抽出されたコード化文書データに関連する情報、および／または、オリジナル文書データに関連する情報を近傍検索結果としてユーザ端末 1 4 へ送信する（S 4 2）。検索結果提供部 4 4 は、S 3 8 の絞込検索が実行された場合、その絞込検索の結果を近傍検索結果としてユーザ端末 1 4 へ送信する。以上の処理により、ユーザから指定された単語距離に基づく近傍検索の結果がユーザへ提示される。

[0047] 具体例として、3つの単語「都知事」「小池氏」「外遊」と単語距離10とを指定する検索要求を受け付けた場合の動作を示す。この例では、「小池氏」が辞書データに存在しない特定単語である。オリジナル文書データが「小池氏」を含む場合、対応するコード化文書データには、3つの単文字単語「小」「池」「氏」それぞれのコードが記録される。文書検索装置12は、「小池氏」を3つの単文字単語「小」「池」「氏」に分解する。それとともに文書検索装置12は、単語距離を12 ($\div 10 - 1 + 3$) に拡大し、文字距離132 ($= 12 \times 11$ ※コード長は11とする) を導出する。文書検索装置12は、5つの単語「都知事」「小」「池」「氏」「外遊」と、文字距離132とを条件とする第1の近傍検索を実行して、この条件を満たすコード化文書データを抽出する。

[0048] 文書検索装置12は、第1の近傍検索の結果として1つ以上のコード化文書データを抽出した場合、それらのコード化文書データを母集団とした近傍検索であり、3つの単文字単語「小」「池」「氏」と、単語距離3 (文字距離は33) を条件とする第2の近傍検索を実行する。単語距離3の第2の近傍検索は、「小」「池」「氏」が隣接することを条件とするものである。第2の近傍検索の結果、「都知事」「小」「池」「氏」「外遊」が単語距離12 (実質的には単語距離10) 以内に記載され、かつ、「小」「池」「氏」が隣接するコード化文書データ (およびオリジナル文書データ) が抽出される。

[0049] なお、検索実行部42において順重視 (順番重視) の指定が可能であれば、第2の近傍検索では、3つの単文字単語「小」「池」「氏」を順重視で検索するようにさらに指定することが望ましい。これにより、ユーザが指定した検索条件に一層適合する検索結果を得やすくなる。なお、第2の近傍検索では、第1の近傍検索の結果としての1つ以上のコード化文書データに対応する1つ以上のオリジナル文書データを母集団として、それら1つ以上のオリジナル文書データの中からキーワード「小池氏」を含むものを通常検索により抽出してもよい。

[0050] また、キーワード間にワイルドカード（ここでは「*」とする）を指定可能な検索エンジンであれば、検索フレーズとして「#（都知事）*#（小）#（池）#（氏）*#（外遊）」（ただし#（単語）は単語のコードを示す。）を指定し、かつ、文字距離１３２（単語距離１２）を指定する検索指示を検索実行部４２へ入力してもよい。この場合も、ユーザが指定した検索条件に一層適合する検索結果を得やすくなる。

[0051] 第１実施例の文書検索装置１２によると、文字数指定の近傍検索が可能であるが、単語数指定の近傍検索に未対応の既存のN-gram方式の検索エンジンを使用して、単語数指定の近傍検索を実現することができる。したがって、単語数指定の近傍検索に対応するために、既存の検索エンジンを改修し、または、新規の検索エンジンを購入することが不要になる。特に、英語等、単語毎に区切って記述される言語の文書検索では、文字数よりも単語数の方がユーザにとって直観的であり、検索条件としての利便性が高い。文書検索装置１２によると、英語圏等のユーザにとって利便性が高い単語数指定の近傍検索を、N-gram方式の検索エンジンを使用して実現できる。

[0052] また、文書検索装置１２では、複数の単語に対応する複数のコードの先頭に、先頭のみで使用される区切り文字を設定することで、コードの検索時に必ずコードの先頭から一致有無が判定される。これにより、コードの途中（すなわち単語の途中）から一致判定されることを防止でき、また、或るコードから次のコードに跨って一致判定されることを防止できる。また、文書検索装置１２によると、対応するコードが存在しない単語をユーザが指定した場合も近傍検索を実現できる。

[0053] （第２実施例）

第２実施例の文書検索装置１２は、単語に対応するコードが辞書データに未定義の場合の処理が第１実施例とは異なる。第２実施例における文書検索システム１０の構成と、文書検索装置１２の機能ブロックは第１実施例と同様（図１および図２）である。以下、第１実施例と重複する構成の説明は適宜省略し、第１実施例と異なる点を主に説明する。

[0054] 文書変換部 32 は、オリジナル文書データに記録された或る単語（特定単語と呼ぶ。）に対応するコードが辞書記憶部 26 の辞書データに未定義の場合、文書変換部 32 は、特定単語のコード変換をスキップして、次の単語のコード変換を実行する。すなわち、文書変換部 32 は、特定単語に関するコードをコード化文書データに記録せず、言い換えれば、オリジナル文書データに記録された単語のうち辞書データにおいてコードに対応付けられた単語のみをコード化文書データに記録する。

[0055] コード取得部 36 は、検索要求で指定された複数の単語のうち一部の単語（特定単語と呼ぶ。）に対応するコードが辞書データで未定義の場合、特定単語のコード変換をスキップして、次の単語のコード変換を実行する。検索指示部 40 は、検索要求で指定された複数の単語のうち対応するコードが定義済みの単語のコードと、検索要求で指定された単語距離を変換した文字距離とを条件とする近傍検索を検索実行部 42 に実行させる。検索実行部 42 は、近傍検索の実行結果として、上記条件を満たす 1 つ以上のコード化文書データに対応する 1 つ以上のオリジナル文書データを原本文書記憶部 28 から抽出する。

[0056] 検索指示部 40 はさらに、近傍検索により抽出された 1 つ以上のオリジナル文書データを母集団として指定し、かつ、特定単語をキーワードとして指定する絞込検索の指示を検索実行部 42 へ入力する。検索実行部 42 は、絞込検索（ここでは一般的なキーワード検索）を実行し、近傍検索により抽出した 1 つ以上のオリジナル文書データの中から、特定単語を含むオリジナル文書データを抽出する。検索結果提供部 44 は、検索実行部 42 による絞込検索の結果を、近傍検索の要求元のユーザ端末 14 へ送信する。

[0057] 以上の構成による文書検索装置 12 の動作を説明する。

図 6 は、第 2 実施例の文書検索装置 12 の動作を示すフローチャートである。同図は、図 4 に対応し、コード化文書データ生成時の動作を示している。同図の S50～S52 は、図 4 の S10～S12 と同じであるため説明を省略する。

[0058] オリジナル文書データに含まれる少なくとも1つの単語（「特定単語」と呼ぶ。）に対応するコードが辞書データに未定義の場合（S 5 4のY）、文書変換部32は、特定単語の処理をスキップし、特定単語に関するコードをコード化文書データに格納しない（S 5 6）。オリジナル文書データに特定単語が存在しなければ、すなわちオリジナル文書データの全ての単語がコードに対応付けられていれば（S 5 4のN）、S 5 6をスキップする。文書変換部32は、オリジナル文書データの各単語（ただし特定単語を除く）を固定長のコードに変換したコード化文書データをコード化文書記憶部30に格納する（S 5 8）。

[0059] 図7も、第2実施例の文書検索装置12の動作を示すフローチャートである。同図は、図5に対応し、検索時の動作を示している。同図のS 6 0、S 6 2は、図5のS 2 0、S 2 2と同じであるため説明を省略する。

[0060] 少なくとも1つの単語（「特定単語」と呼ぶ。）に対応するコードが辞書データに未定義であれば（S 6 4のY）、コード取得部36は、メモリの所定領域に設けた特殊フラグをオンに設定する（S 6 6）。検索指示部40は、特定単語を近傍検索の対象から除外する（S 6 8）。検索要求で指定された単語に特定単語が含まれず、すなわち、指定された全ての単語に対応するコードが辞書データで定義されていれば（S 6 4のN）、S 6 6、S 6 8をスキップする。文字距離導出部38は、検索要求で指定された単語距離と、固定のコード長に基づいて、近傍検索のための文字距離を導出する（S 7 0）。

[0061] 検索指示部40は、検索要求で指定された複数の単語のうち特定単語を除外した残りの単語のコードをキーワードとして指定した検索指示であり、S 7 0で導出された文字距離をさらに指定した近傍検索指示を検索実行部42へ入力する。検索実行部42は、コード化文書記憶部30に格納された複数のコード化文書データを母集団とする近傍検索処理であり、N - g r a m方式による文字数基準の近傍検索処理を実行する。検索実行部42は、近傍検索処理により抽出したコード化文書データに対応するオリジナル文書データ

を原本文書記憶部 28 から抽出する (S 7 2)。

[0062] 特殊フラグがオンであれば (S 7 4 の Y)、検索指示部 40 は、近傍検索処理で抽出された 1 つ以上のオリジナル文書データを母集団とする絞込検索の指示であり、特定単語をキーワードとして指定する絞込検索指示を検索実行部 42 へ入力する。検索実行部 42 は、近傍検索処理で抽出した 1 つ以上のオリジナル文書データの中から、特定単語を含むオリジナル文書データを抽出する (S 7 6)。検索指示部 40 は、特殊フラグをオフに戻す (S 7 8)。特殊フラグが当初からオフであれば (S 7 4 の N)、S 7 6 と S 7 8 をスキップする。検索結果提供部 44 は、検索実行部 42 により抽出されたコード化文書データに関連する情報、および／または、オリジナル文書データに関連する情報を近傍検索結果として、ユーザ端末 14 へ送信する (S 80)。検索結果提供部 44 は、S 7 6 の絞込検索が実行された場合、その絞込検索の結果を近傍検索結果としてユーザ端末 14 へ送信する。以上の処理により、ユーザから指定された単語距離に基づく近傍検索の結果がユーザへ提示される。

[0063] 具体例として、3 つの単語「都知事」「小池氏」「外遊」と単語距離 10 とを指定する検索要求を受け付けた場合の動作を示す。この例では、「小池氏」が辞書データに存在しない特定単語である。オリジナル文書データが「小池氏」を含む場合、対応するコード化文書データには、「小池氏」に関するコードは記録されない。文書検索装置 12 は、2 つの単語「都知事」「外遊」と、文字距離 110 (= 単語距離 10 × コード長 11) を条件とする近傍検索を実行し、この条件を満たすコード化文書データを抽出し、さらに対応するオリジナル文書データを抽出する。文書検索装置 12 は、近傍検索の結果として 1 つ以上のオリジナル文書データを抽出した場合、それらのオリジナル文書データを母集団とした絞込検索であり、キーワード「小池氏」を条件とするキーワード検索を実行する。これにより、「都知事」と「外遊」が単語距離 10 以内に記載され、かつ、「小池氏」を含むオリジナル文書データが抽出される。

- [0064] 第2実施例の文書検索装置12においても第1実施例の文書検索装置12と同様の効果を奏する。例えば、第2実施例の文書検索装置12によると、対応するコードが存在しない単語をユーザが指定した場合も、ノイズは生じうるが、ユーザが指定した条件に可及的に適合する検索結果をユーザへ提示することができる。
- [0065] 以上、本発明を第1実施例と第2実施例をもとに説明した。これら実施例は例示であり、各構成要素や各処理プロセスの組合せにいろいろな変形例が可能なこと、またそうした変形例も本発明の範囲にあることは当業者に理解されるところである。
- [0066] 変形例を説明する。辞書記憶部26の辞書データで定義される固定長のコードの先頭には、先頭以外では使用されない複数の特殊値のいずれかが設定されてもよい。言い換えれば、ある範囲内の値を特殊値（区切り文字）として設定してもよい。具体的には、UnicodeにおけるU+9000～U+9FFFの範囲を特殊値として定め、コード（ここでは10文字とする）の先頭1文字目に上記範囲のいずれかの特殊値を設定してもよい。また、U+1000～U+8FFFの範囲を、コードの2文字目～10文字目に設定してもよい。
- [0067] 本変形例の態様でも、先頭を示す区切り文字がコードの途中では出現しないため、実施例の区切り文字と同様の効果を奏する。例えば、必ずコードの先頭から一致有無が判定されることになり、コードの途中から一致判定されることを防止できる。また、第1変形例の態様では、コードの先頭（例えば1文字目）もコードの識別に用いることができる。すなわち、コードは単語毎にユニークとなる必要があるが、コードの先頭にも所定範囲の特殊値群の中から異なる値を設定できるため、コード長を実施例より短くすることができる。これにより、辞書データおよびコード化文書データのサイズを低減できる。
- [0068] 別の変形例を説明する。コード化文書記憶部30に格納されるコード化文書データは、オリジナル文書データに記載された互いに関連性を有する複数

の単語を、共通のコードへ変換したものであってもよい。具体的には、辞書記憶部 26 の辞書データでは、互いに関連する複数の単語に対して共通のコードを割り当ててもよい。例えば、基本形が同じで活用形が異なる複数の単語に対して同じコードを割り当ててもよい。また、英語等において、或る単語の原形、過去形、過去分詞、複数形に対して同じコードを割り当ててもよい。さらにまた、語幹が同じだが、接頭語または接尾語が異なる複数の単語に対して同じコードを割り当ててもよい。文書変換部 32 は、辞書データを参照して、綴りが異なる複数の単語であるが、互いに関連を有する複数の単語を同じコードへ変換してもよい。

[0069] さらに別の変形例を説明する。第 1 実施例では、検索要求で指定された単語のコードが辞書データに未定義の場合、未定義の単語（特定単語と呼ぶ。）を複数の単文字単語に分解するため、特定単語の文字数が多いほど近傍検索の対象となる単語数が増加する。検索要求に特定単語が存在する場合、ユーザは単語距離をより長めに指定することが望ましい。そこで、第 1 実施例の文書検索装置 12 は、検索条件としての単語距離を現在の値（言い換えれば検索要求時の指定値）より長くすることを促す内容のメッセージをユーザ端末 14 へ送信して表示させる通知部をさらに備えてもよい。このメッセージの送信タイミングは、近傍検索結果の送信時でもよく、コード取得部 36 により特定単語が検出された時点でもよい。さらなる変形例として、第 1 実施例の文書検索装置 12 は、ユーザへメッセージを提示することなく、検索条件としての単語距離をユーザの指定値より自動で大きくしてもよい。ユーザの指定値に加算する値は、検索条件として指定されたキーワードの個数、特定単語の個数、単文字単語の個数、ユーザの指定値のうち少なくとも 1 つに基づいて決定されてもよい。また、開発者の知見や文書検索装置 12 を用いた実験により適切な加算値（言い換えれば加算値決定アルゴリズム）が定められてもよい。

[0070] また、第 2 実施例では、検索要求で指定された単語のコードが辞書データに未定義の場合、未定義の単語を無視して近傍検索を実行する。したがって

、ユーザが当初指定した単語距離のままでは、検索結果にノイズが混入しやすくなる。そこで、第2実施例の文書検索装置12は、検索条件としての単語距離を現在の値（言い換えれば先の検索要求時の指定値）より短くすることを促す内容のメッセージをユーザ端末14へ送信して表示させる通知部をさらに備えてもよい。このメッセージの送信タイミングは、近傍検索結果の送信時でもよく、コード取得部36により特定単語が検出された時点でもよい。さらなる変形例として、第2実施例の文書検索装置12は、ユーザへメッセージを提示することなく、検索条件としての単語距離をユーザの指定値より自動で小さくしてもよい。ユーザの指定値から減算する値は、検索条件として指定されたキーワードの個数、特定単語の個数、ユーザの指定値のうち少なくとも1つに基づいて決定されてもよい。また、開発者の知見や文書検索装置12を用いた実験により適切な減算値（言い換えれば減算値決定アルゴリズム）が定められてもよい。

[0071] さらに別の変形例を説明する。第1実施例および第2実施例の文書検索装置12は、検索要求で指定された単語のコードが辞書データに未定義の場合、未定義の単語（特定単語と呼ぶ。）を別の単語に変更するか否かを確認するメッセージをユーザ端末14へ送信して表示させる通知部をさらに備えてもよい。この確認の結果、特定単語を変更しない旨がユーザ端末14から返信された場合、文書検索装置12は、第1実施例および第2実施例に記載の特定単語に関連する処理を実行してもよい。さらなる変形例として、文書検索装置12は、ユーザへメッセージを提示することなく、特定単語を、辞書データでコードが定義された別の単語（ここでは「定義済単語」と呼ぶ。）へ自動で変換してもよい。例えば、文書検索装置12は、特定単語と定義済単語との対応関係を定めたテーブル保持し、特定単語が検出された場合、当該特定単語に対応する定義済単語を識別し、当該特定単語に対応する定義済単語のコードへ変換してもよい。上記テーブルは、意味が類似し、および／または、綴りが類似する特定単語と定義済単語とを対応付けたものでもよい。文書検索装置12は、特定単語がこのテーブルにも未定義である場合、第

1 実施例および第2 実施例に記載の特定単語に関連する処理を実行してもよい。

[0072] さらに別の変形例を説明する。文書検索装置1 2は、外部から入力された文書データ（以下「入力文書データ」とも呼ぶ。）をより検索に適した形式に変換することがある。文書検索装置1 2の原本文書記憶部2 8は、変換後の文書データ（以下「検索形文書データ」とも呼ぶ。）をオリジナル文書データとして保持してもよい。すなわち、オリジナル文書データは、入力文書データと検索形文書データの両方を含んでもよい。

[0073] 入力文書データと検索形文書データの例を示す。

入力文書データ1：

「I bought a book from that shopper, when I was a little girl.」

入力文書データ1を変換後の検索形文書データ1：

「I buy a book from that shop, when I be a little girl.」

入力文書データ2：

「The chef cooked a special food at a national event while we were devoted in eating.」

入力文書データ1を変換後の検索形文書データ2：

「The chef cook a special food at a nation event while we were devoted in eat.」

[0074] 上記の例で示すように、検索形文書データは、入力文書データに含まれる動詞の単語（過去形、過去分詞形、現在分詞形等）を基本形（言い換えれば原形）に変換したものでもよい。また、検索形文書データは、入力文書データに含まれる名詞の単語（複数形）を単数形に変換したものでもよく、ある名詞の単語をより一般的な名詞の単語に変換したものでもよい。文書検索装置1 2は、外部装置から入力された入力文書データを取得し、変換前の単語と変換後の単語とを予め対応付けたテーブルを参照して、入力文書データを検索形文書データに変換し、変換後の検索形文書データを原本文書記憶部2 8へ格納する検索形文書生成部をさらに備えてもよい。

- [0075] さらに別の変形例を説明する。第1実施例の文書検索装置12は、特定単語を分解した単文字単語を含むキーワードによる近傍検索実行後、複数の単文字単語のコードが隣接することを条件とする新たな近傍検索（絞込検索）を実行しなくてもよい。言い換えれば、図5のS38をスキップしてもよい。なお、複数の単文字単語のコードが隣接することを条件とする近傍検索を、単文字単語を含むキーワードによる近傍検索より先に実行してもよい。
- [0076] また、第2実施例の文書検索装置12は、特定単語を除外したキーワードによる近傍検索実行後、特定単語をキーワードとした絞込検索を実行しなくてもよい。言い換えれば、図7のS76をスキップしてもよい。なお、特定単語をキーワードとした検索を、特定単語を除外したキーワードによる近傍検索より先に実行してもよい。例えば、特定単語を含むオリジナル文書データを抽出後、抽出したオリジナル文書データに対応するコード化文書データを母集団として、特定単語を除外したキーワードによる近傍検索を実行してもよい。
- [0077] 上述した実施例および変形例の任意の組み合わせもまた本発明の実施の形態として有用である。組み合わせによって生じる新たな実施の形態は、組み合わせられる実施例および変形例それぞれの効果をあわせもつ。請求項に記載の各構成要件が果たすべき機能は、実施例および変形例において示された各構成要素の単体もしくはそれらの連携によって実現されることも当業者には理解されるところである。

符号の説明

- [0078] 10 文書検索システム、 12 文書検索装置、 26 辞書記憶部、
28 原本文書記憶部、 30 コード化文書記憶部、 34 検索要求受付部、
36 コード取得部、 38 文字距離導出部、 40 検索指示部、
42 検索実行部。

産業上の利用可能性

- [0079] この発明は、文書を検索する装置に適用できる。

請求の範囲

- [請求項1] 複数の文書データを記憶する第1文書記憶部であって、前記複数の文書データのそれぞれは、オリジナル文書に記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換されたものである第1文書記憶部と、
- 複数の単語を指定する検索要求であって、かつ、それら複数の単語が存在すべき範囲を単語数で指定する検索要求を受け付ける受付部と、
- 検索要求で指定された複数の単語のそれぞれに対応する固定長のコードを取得する取得部と、
- 検索要求で指定された単語数基準の範囲と、固定のコード長とに応じて、文字数基準の範囲を導出する導出部と、
- 前記取得部により取得された複数の単語のコードと、前記導出部により導出された文字数基準の範囲とを条件とする近傍検索を実行し、前記第1文書記憶部に記憶された複数の文書データの中から前記条件を満たす文書データを抽出する検索部と、
- を備えることを特徴とする文書検索装置。
- [請求項2] 前記固定長のコードの先頭には、先頭以外では使用されない特殊値が設定されることを特徴とする請求項1に記載の文書検索装置。
- [請求項3] 前記固定長のコードの先頭には、先頭以外では使用されない複数の特殊値のうちいずれかが設定されることを特徴とする請求項1に記載の文書検索装置。
- [請求項4] 前記第1文書記憶部に記憶された複数の文書データに対応する変換前の文書データである複数のオリジナル文書データを記憶する第2文書記憶部をさらに備え、
- 前記検索部は、前記第1文書記憶部から抽出した文書データに対応するオリジナル文書データを前記第2文書記憶部から抽出することを特徴とする請求項1から3のいずれかに記載の文書検索装置。

[請求項5] 前記第1文書記憶部に記憶される文書データは、オリジナル文書データに記載された単語に予め対応付けられたコードが記録されたものであり、オリジナル文書データに記載された或る単語に対応するコードが未定義の場合、その単語に関するコードは前記文書データに記録されず、

検索要求で指定された複数の単語のうち一部の単語に対応するコードが未定義の場合、前記検索部は、対応するコードが存在する単語のコードに基づく近傍検索を実行し、当該近傍検索で抽出した1つ以上のオリジナル文書データの中から、前記一部の単語を含むオリジナル文書データを抽出することを特徴とする請求項4に記載の文書検索装置。

[請求項6] 前記第1文書記憶部に記憶される文書データは、オリジナル文書データに記載された単語に予め対応付けられたコードが記録されたものであり、オリジナル文書データに記載された或る単語に対応するコードが未定義の場合、その単語は複数の単文字単語に分解され、複数の単文字単語に予め対応付けられたコードが前記文書データに記録され、

検索要求で指定された複数の単語のうち一部の単語に対応するコードが未定義の場合、前記検索部は、対応するコードが存在する単語のコードと、前記一部の単語を分解した複数の単文字単語のコードとに基づく近傍検索を実行することを特徴とする請求項1から4のいずれかに記載の文書検索装置。

[請求項7] 前記導出部は、検索要求で指定された単語数基準の範囲を、前記一部の単語を分解した単文字単語の個数に応じて拡大し、拡大した単語数基準の範囲に応じて文字数基準の範囲を導出することを特徴とする請求項6に記載の文書検索装置。

[請求項8] 前記検索部は、前記近傍検索により抽出した1つ以上の文書データを母集団とした新たな近傍検索であって、前記複数の単文字単語のコ

ードが隣接することを条件とする新たな近傍検索をさらに実行することを特徴とする請求項6または7に記載の文書検索装置。

[請求項9] 前記第1文書記憶部に記憶される文書データは、オリジナル文書データに記載された互いに関連性を有する複数の単語を、共通のコードへ変換したものである請求項1から8のいずれかに記載の文書検索装置。

[請求項10] 複数の文書データを記憶する文書記憶部であって、前記複数の文書データのそれぞれは、オリジナル文書に記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換されたものである文書記憶部にアクセス可能なコンピュータが、

複数の単語を指定する検索要求であって、かつ、それら複数の単語が存在すべき範囲を単語数で指定する検索要求を受け付けるステップと、

検索要求で指定された複数の単語のそれぞれに対応する固定長のコードを取得するステップと、

検索要求で指定された単語数基準の範囲と、固定のコード長とに応じて、文字数基準の範囲を導出するステップと、

前記取得するステップで取得された複数の単語のコードと、前記導出するステップで導出された文字数基準の範囲とを条件とする近傍検索を実行し、前記文書記憶部に記憶された複数の文書データの中から前記条件を満たす文書データを抽出するステップと、

を実行することを特徴とする文書検索方法。

[請求項11] 複数の文書データを記憶する文書記憶部であって、前記複数の文書データのそれぞれは、オリジナル文書に記載された互いに異なる複数の単語が、互いに異なる固定長のコードに変換されたものである文書記憶部にアクセス可能なコンピュータに、

複数の単語を指定する検索要求であって、かつ、それら複数の単語が存在すべき範囲を単語数で指定する検索要求を受け付ける機能と、

検索要求で指定された複数の単語のそれぞれに対応する固定長のコードを取得する機能と、

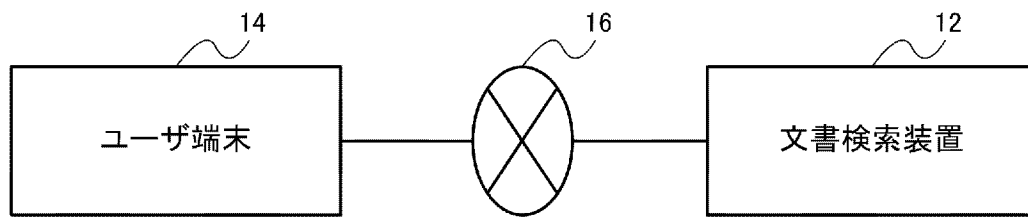
検索要求で指定された単語数基準の範囲と、固定のコード長とに応じて、文字数基準の範囲を導出する機能と、

前記取得する機能により取得された複数の単語のコードと、前記導出する機能により導出された文字数基準の範囲とを条件とする近傍検索を実行し、前記文書記憶部に記憶された複数の文書データの中から前記条件を満たす文書データを抽出する機能と、

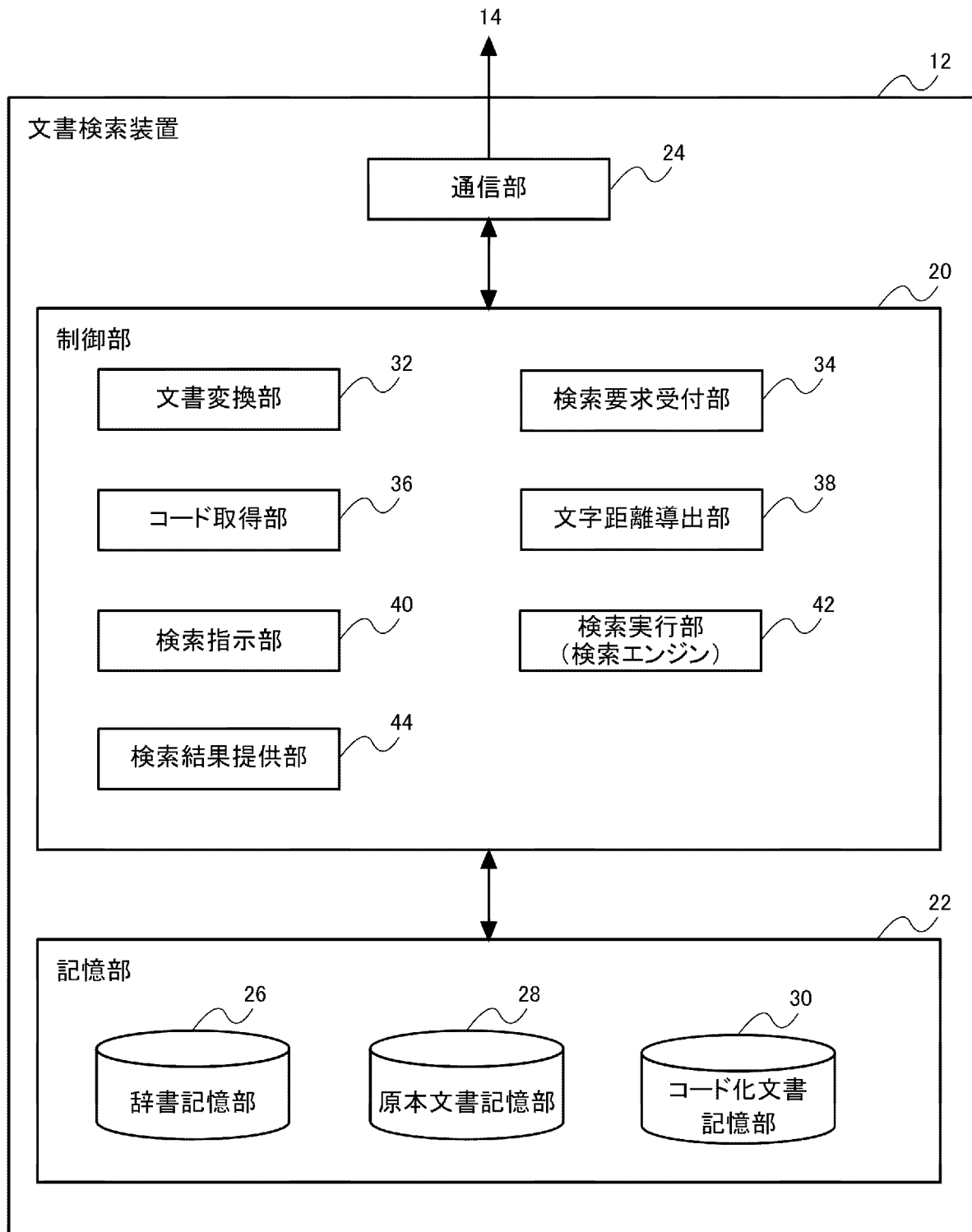
を実現させるためのコンピュータプログラム。

[図1]

文書検索システム 10



[図2]

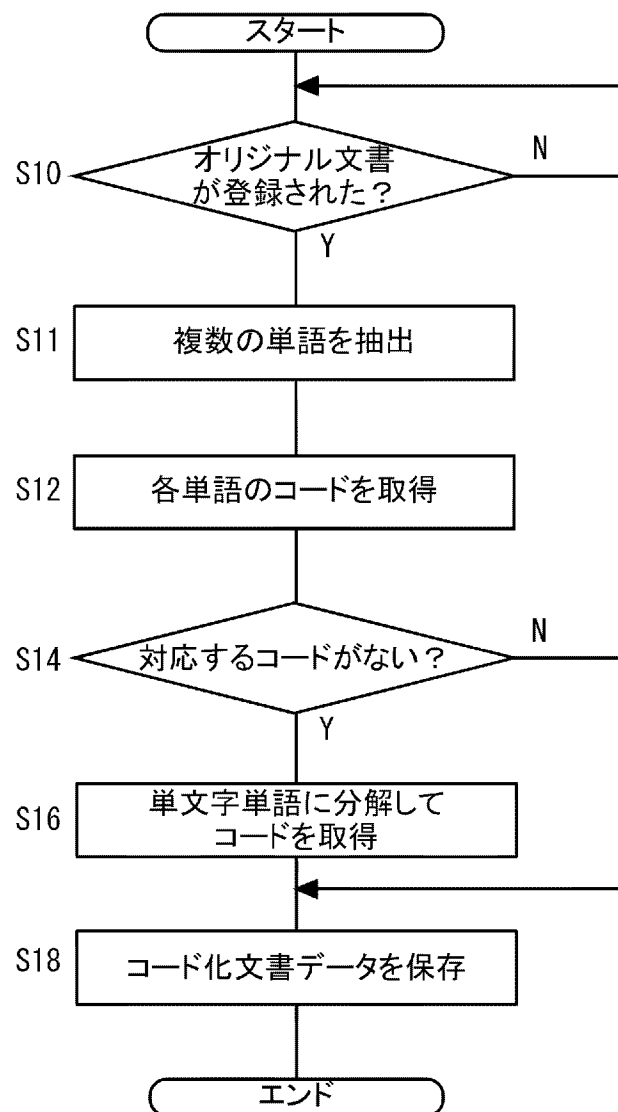


[図3]

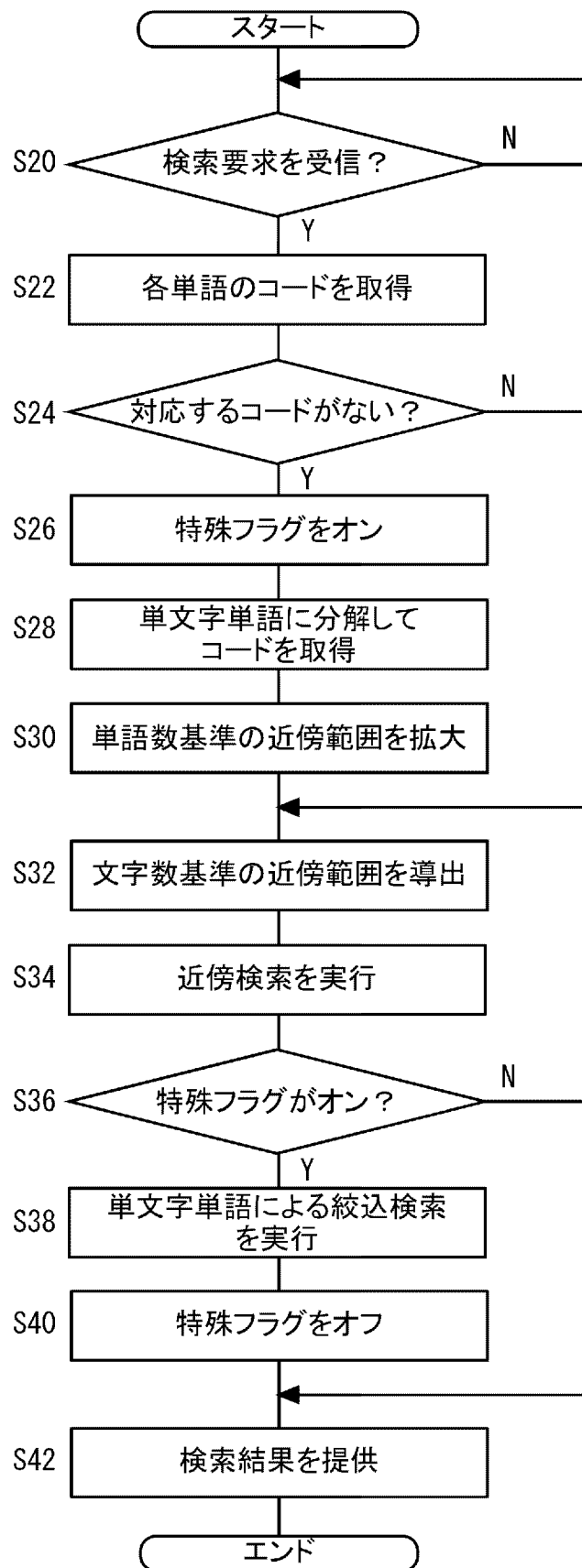
単語	コード
発明	#0024F76DA6
私	#0024F76DA7
の	#0024F76DD8
eat	#0100BB17CE
device	#0100BB17CF

26

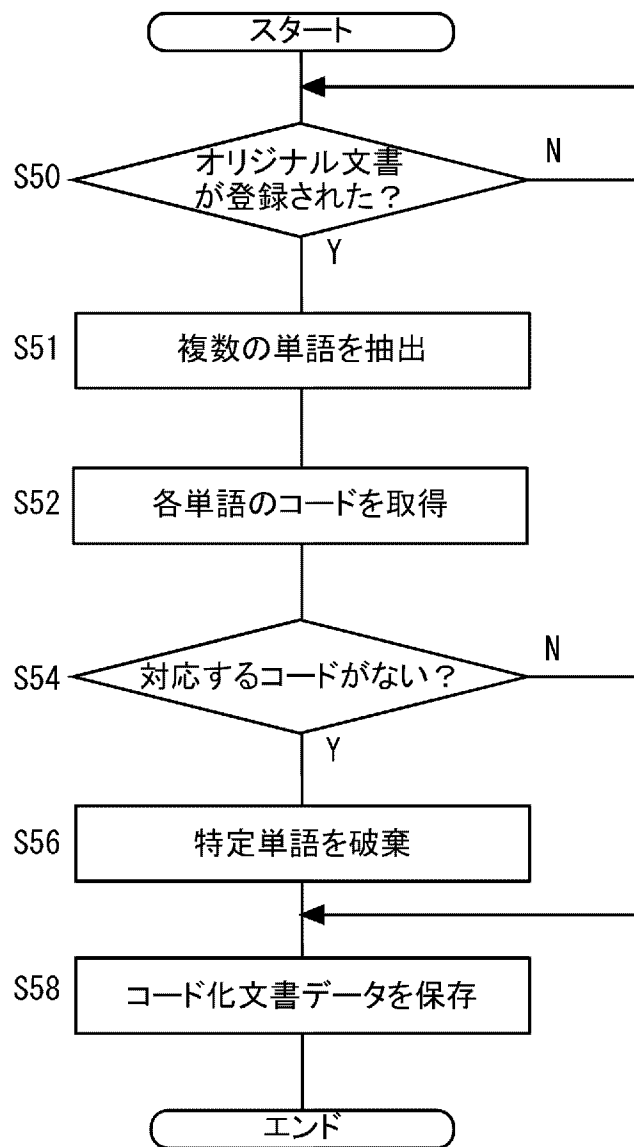
[図4]



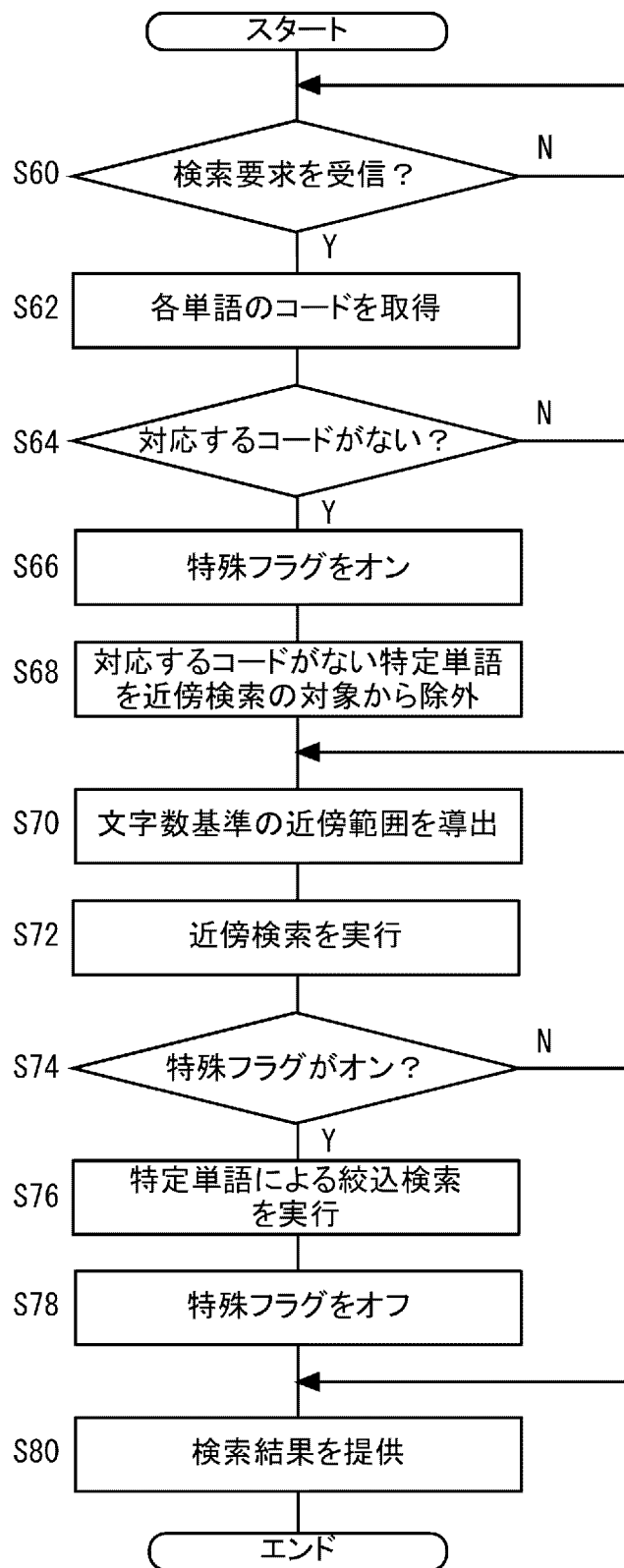
[図5]



[図6]



[図7]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/033316

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06F17/30 (2006.01) i, G06F17/22 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Int.Cl. G06F17/30, G06F17/22

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan	1922-1996
Published unexamined utility model applications of Japan	1971-2017
Registered utility model specifications of Japan	1996-2017
Published registered utility model applications of Japan	1994-2017

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	JP 2005-242416 A (SHOGAKUKAN INC.) 08 September 2005, paragraphs [0017], [0022], [0026]-[0034], [0037], [0062] & US 2005/0187964 A1, paragraphs [0033], [0038], [0042]-[0051], [0054], [0081]	1, 4, 9-11 2-3, 5-8
A	US 2002/0165707 A1 (CALL, Charles G.) 07 November 2002, paragraphs [0024]-[0032], [0093]-[0095], fig. 1 (Family: none)	1-11
A	US 5787386 A (XEROX CORPORATION) 28 July 1998, entire text, all drawings & US 5523946 A	1-11

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	"I" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search
20 November 2017 (20.11.2017)

Date of mailing of the international search report
28 November 2017 (28.11.2017)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G06F17/30(2006.01)i, G06F17/22(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int.Cl. G06F17/30, G06F17/22

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2017年
日本国実用新案登録公報	1996-2017年
日本国登録実用新案公報	1994-2017年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
X A	JP 2005-242416 A（株式会社小学館）2005.09.08, [0017], [0022], [0026]-[0034], [0037], [0062] & US 2005/0187964 A1, [0033], [0038], [0042]-[0051], [0054], [0081]	1, 4, 9-11 2-3, 5-8
A	US 2002/0165707 A1（CALL, Charles G.）2002.11.07, [0024]-[0032], [0093]-[0095], FIG.1 （ファミリーなし）	1-11

☒ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日

20.11.2017

国際調査報告の発送日

28.11.2017

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

田中 秀樹

5M

6291

電話番号 03-3581-1101 内線 3599

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	US 5787386 A (XEROX CORPORATION) 1998.07.28, 全文, 全図 & US 5523946 A	1-11