(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
2021年3月4日(04.03.2021)

WIPO | PCT

(10) 国際公開番号
WO 2021/038887 A1

(51) 国際特許分類:
G06F 16/383 (2019.01) *G06F 16/90* (2019.01)
G06F 16/31 (2019.01) *G06F 16/93* (2019.01)

(21) 国際出願番号： PCT/JP2019/034306

(22) 国際出願日： 2019年8月30日(30.08.2019)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 富士通株式会社(**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者: 馬場 謙介(**BABA, Kensuke**); 〒2118588 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 Kanagawa (JP). 野呂 智哉(**NORO, Tomoya**); 〒2118588 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 Kanagawa (JP). 福田 茂紀(**FUKUTA, Shigeki**); 〒2118588 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 Kanagawa (JP). 大倉 清司(**OKURA, Seiji**); 〒2118588 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 Kanagawa (JP).

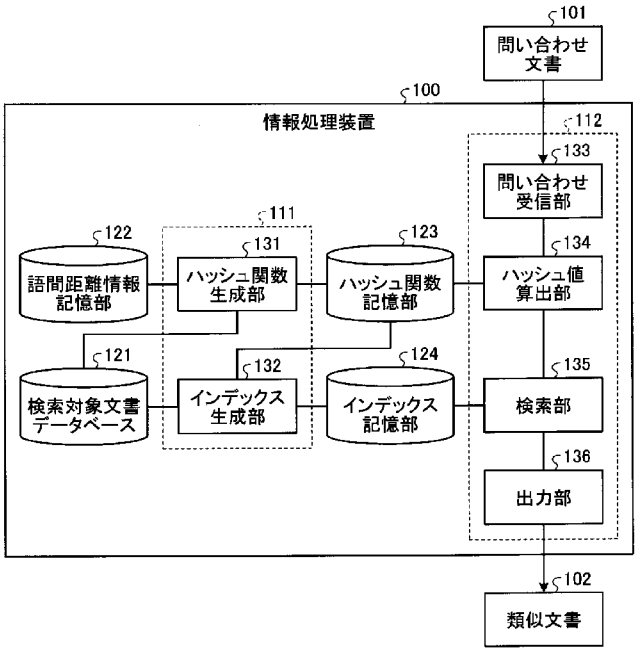
(74) 代理人: 特許業務法人酒井国際特許事務所 (**SAKAI INTERNATIONAL PATENT OFFICE**); 〒1000013 東京都千代田区霞が関３丁目８番１号 虎の門三井ビルディング Tokyo (JP).

(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,

(54) **Title:** SIMILAR DOCUMENT RETRIEVAL METHOD, SIMILAR DOCUMENT RETRIEVAL PROGRAM, SIMILAR DOCUMENT RETRIEVAL DEVICE, INDEX INFORMATION CREATION METHOD, INDEX INFORMATION CREATION PROGRAM, AND INDEX INFORMATION CREATION DEVICE

(54) 発明の名称： 類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置

[図1]



100... INFORMATION PROCESSING DEVICE
101... QUERY DOCUMENT
102... SIMILAR DOCUMENT
121... RETRIEVABLE DOCUMENT DATABASE
122... WORD INTERVAL DISTANCE INFORMATION STORAGE UNIT
123... HASH FUNCTION STORAGE UNIT
124... INDEX STORAGE UNIT
131... HASH FUNCTION GENERATION UNIT
132... INDEX GENERATION UNIT
133... QUERY RECEIVING UNIT
134... HASH VALUE COMPUTATION UNIT
135... RETRIEVAL UNIT
136... OUTPUT UNIT

(57) **Abstract:** In a similar document retrieval method according to an embodiment, a computer executes a generation process, a computation process, and a retrieval process. The generation process generates a hash function for allocating a value to each word included in a set of words in a retrievable document on the basis of the set of words and word

WO 2021/038887 A1

[続葉有]

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, JP, 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, 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)）

interval information indicating word meaning closeness, said hash function allocating a closer value with reference to a specific word in the order of closeness to said word. The computation process computes summary information for each of a plurality of retrievable documents on the basis of the generated hash function and computes summary information for an inputted document on the basis of the generated hash function. The retrieval process retrieves a document similar to the inputted document among the plurality of retrievable documents on the basis of a comparison between the computed summary information for the retrievable documents and the computed summary information for the inputted document.

(57) 要約：実施形態の類似文書検索方法は、生成する処理と、算出する処理と、検索する処理とをコンピュータが実行する。生成する処理は、検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成する。算出する処理は、生成したハッシュ関数に基づいて複数の検索対象文書それぞれの要約情報を算出し、生成したハッシュ関数に基づいて入力文書の要約情報を算出する。検索する処理は、算出した検索対象文書の要約情報と、入力文書の要約情報との間の比較に基づいて、複数の検索対象文書の中から入力文書に類似する文書を検索する。

明　細　書

発明の名称 ：

類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置

技術分野

[0001] 本発明の実施形態は、類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置に関する。

背景技術

[0002] 従来、コンピュータによる自然言語処理の一つとして、データベースに記憶された文書の中から入力文書に類似する文書を検索する検索処理がある。例えば、問い合わせ文書のサンプルと当該サンプルに対応する返答文書とをデータベースに登録しておく。そして、入力された問い合わせ文書に類似するサンプルを検索し、当該類似するサンプルに対応する返答文書を出力するチャットボット等の対話インタフェースを構築することが考えられる。

[0003] データベースに登録されたサンプルの中から入力文書に類似するサンプルを検索する方法としては、２つの文書の間で出現する単語の共通度を評価する方法がある。例えば、以下のような検索方法が考えられる。ある文書に含まれる単語集合から１つのハッシュ値を算出するハッシュ関数（Min-hash関数と言うことがある）を複数個定義しておく。各ハッシュ関数は、異なる単語に対して異なる値を対応付けた対応関係をもち、ある単語集合に含まれる単語に対応する値のうち最小の値をハッシュ値として出力する。データベースに登録されたサンプルに対して、この複数のハッシュ関数を用いて算出された複数のハッシュ値を列挙したベクトルを予め生成しておく。そして、入力文書に含まれる単語集合と上記の複数のハッシュ関数から同様にベクトルを算出し、データベースに登録されたサンプルのベクトルと近似するものを検索する。

先行技術文献

特許文献

[0004] 特許文献1：特開２０１８－１７３９０９号公報

発明の概要

発明が解決しようとする課題

[0005] しかしながら、上記の従来技術では、入力文書が短い、文書に用いられる表現が多様でサンプルと一致する表現が少ないなどの場合に、単語の共通度が低く評価され、検索精度が低下するという問題がある。例えば、チャットボットの入力文書は、主に話し言葉であり、１文が短く表現も多様であることから、入力文書と、サンプルとの間で共通の単語が出現する確率が全体的に低くなる。この結果、入力文書に対するサンプルの検索精度が低くなりやすい。

[0006] １つの側面では、類似する文書の検索精度を向上させる類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置を提供することを目的とする。

課題を解決するための手段

[0007] １つの案では、類似文書検索方法は、生成する処理と、算出する処理と、検索する処理とをコンピュータが実行する。生成する処理は、検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成する。算出する処理は、生成したハッシュ関数に基づいて複数の検索対象文書それぞれの要約情報を算出し、生成したハッシュ関数に基づいて入力文書の要約情報を算出する。検索する処理は、算出した検索対象文書の要約情報と、入力文書の要約情報との間の比較に基づいて、複数の検索対象文書の中から入力文書に類似する文書を検索する。

発明の効果

[0008] １つの側面では、類似する文書の検索精度が向上する。

図面の簡単な説明

[0009] [図1]図１は、実施形態にかかる情報処理装置の機能構成例を示すブロック図である。

[図2]図２は、実施形態にかかる情報処理装置の処理フローの一例を示す説明図である。

[図3]図３は、ハッシュ関数の生成処理の一例を示すフローチャートである。

[図4]図４は、ハッシュ関数を説明する説明図である。

[図5]図５は、語間の類似度を説明する説明図である。

[図6]図６は、Ｍｉｎ　ｈａｓｈによるハッシュ値を説明する説明図である。

[図7]図７は、実施形態にかかる情報処理装置の動作の概要を説明する説明図である。

[図8]図８は、検索対象文書の絞り込みを説明する説明図である。

[図9]図９は、操作画面の表示例を示す説明図である。

[図10]図１０は、プログラムを実行するコンピュータの一例を示す図である。

発明を実施するための形態

[0010] 以下、図面を参照して、実施形態にかかる類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置を説明する。実施形態において同一の機能を有する構成には同一の符号を付し、重複する説明は省略する。なお、以下の実施形態で説明する類似文書検索方法、類似文書検索プログラム、類似文書検索装置、索引情報作成方法、索引情報作成プログラムおよび索引情報作成装置は、一例を示すに過ぎず、実施形態を限定するものではない。また、以下の各実施形態は、矛盾しない範囲内で適宜組みあわせてもよい。

[0011] ［概要］

本実施形態では、データベースに登録されたサンプル（以下、検索対象文書とも呼ぶ）の中から入力文書（以下、問い合わせ文書とも呼ぶ）に類似す

る文書を検索する情報処理装置を例示する。

[0012] この情報処理装置では、先ず事前処理を行い、次いで検索処理を行う。事前処理では、ハッシュ関数を用いて検索対象文書それぞれについてハッシュ値を算出し、算出したハッシュ値より検索対象文書を検索するための木構造などの索引構造（例えば探索木）を作る。

[0013] 検索処理では、ハッシュ関数を用いて問い合わせ文書のハッシュ値を算出する。次いで、情報処理装置は、問い合わせ文書のハッシュ値と、検索対象文書それぞれのハッシュ値とを比較し、索引構造で示された検索対象文書それぞれのハッシュ値の中から近いものを探す。次いで、情報処理装置は、問い合わせ文書のハッシュ値と近いハッシュ値の検索対象文書を類似する文書の検索結果とする。

[0014] ハッシュ値を算出するハッシュ関数については、データベースに登録された検索対象文書に含まれる単語を抽出して得られた単語の集合をもとに、複数個定義しておく。具体的には、情報処理装置は、Ｗを単語の集合、ｈをＷから｛０，１，…，｜Ｗ｜－１｝へのすべての単射の集合（複数のハッシュ関数）としてハッシュ関数を複数生成しておく。

[0015] 情報処理装置は、検索対象文書および問い合わせ文書について次の処理を行うことで、複数のハッシュ値によるベクトルを求めて検索対象文書および問い合わせ文書をハッシュ関数で要約した要約情報を得る。

・ｈからハッシュ関数をランダムに選択する。

・検索対象文書に含まれる単語を抽出して単語の集合を取得する。

・選択したｈにより各単語から得る整数の中で最小のものをハッシュ値とする。

・関数のランダムな選択を複数回行うことで複数のハッシュ値を得る。

[0016] 互いに類似する文書同士は、共通語の出現割合（ジャカール係数）が高くなる。ｈからのハッシュ関数の選択について、ハッシュ値が一致する確率はジャカール係数に一致する。よって、類似する文書を求める際の文書のハッシュ値同士の比較において、ベクトルの各要素の一致の割合からジャカール

係数を確率的に計算でき、ハッシュ値のハミング距離（例えば不一致の数）が文書間の近さ（類似度）を反映している。

[0017] また、ｎを検索対象文書の数、ｍを検索対象文書ごとの単語数（平均値）、ｋをハッシュ関数の数とする。すると、検索対象文書のベクトルを計算する計算量はＯ（ｋｍｎ）となる。１つの問い合わせ文書のベクトルを計算する計算量はＯ（ｋｍ）となり、このベクトルを用いた近傍探索の計算コストはＯ（ｌｏｇ（ｎ））となる。ハッシュ関数は予めランダムにｋ個生成される。ｋをＯ（ｌｏｇ（ｎ））とすると、異なる検索対象文書から同一のハッシュ値が算出される衝突確率を十分に小さくすることができる。

[0018] しかしながら、ハミング距離による類似度の検証では、ベクトルの各要素の差に意味はなく、要素が不一致の場合（共通語の出現がない場合）、類似度は０となる。このため、共通語の出現が少ない場合、類似検索の精度が低くなりやすい。

[0019] そこで、本実施形態では、情報処理装置は、単語間の語意の近さを示す語間距離情報をもとに、所定の単語を基準として語意の近い順に近い値を割り当てるハッシュ関数を生成する。具体的には、情報処理装置は、検索対象文書に含まれる単語を抽出して単語の集合し、ランダムに基準とする単語を選択する。次いで、情報処理装置は、語間距離情報をもとに、基準とする単語に対して語意の近い順に単語の集合に含まれる単語をソートし、ソート順に一意な値（例えばソート順に大きくなる整数値）の割り当てを行う。情報処理装置は、上記の処理を繰り返すことで、複数のハッシュ関数を生成する。

[0020] 例えば、単語の集合が｛猫、ご飯、…にゃんこ、えさ｝であり、基準とする単語を｛猫｝とする場合、情報処理装置は、｛猫｝を基準として語意の近い順にソートする。これにより、｛猫、にゃんこ、…ご飯、えさ｝が得られる。このようにソートした単語の集合について、情報処理装置は、｛猫＝０、にゃんこ＝１、…ご飯＝５、えさ＝６｝などのように、ソート順に整数値を割り当てる。

[0021] このようにして生成されたハッシュ関数によるハッシュ値は、所定の単語

に対する語意の距離（語間とも呼ぶ）に応じた値となる。このため、ハッシュ値間のハミング距離だけでなく、ユークリッド距離による類似度の検証が可能となる。したがって、共通語の出現がなく、ハッシュ値が不一致（要素が不一致の場合）であっても、ユークリッド距離により類似度を検証することができ、類似検索の精度を向上させることができる。

[0022] ［構成例］

図１は、実施形態にかかる情報処理装置の機能構成例を示すブロック図である。図１に示すように、情報処理装置１００は、入力された問い合わせ文書１０１について、検索対象文書データベース１２１に格納された検索対象文書の中から類似する類似文書１０２を求める処理を行う装置である。情報処理装置１００としては、例えばパーソナルコンピュータ等を適用できる。

[0023] 情報処理装置１００は、索引構造を作成する事前処理を行うインデックス生成モジュール１１１と、検索処理を行う検索モジュール１１２と、処理に関する各種データを格納する記憶部（検索対象文書データベース１２１、語間距離情報記憶部１２２、ハッシュ関数記憶部１２３およびインデックス記憶部１２４）とを有する。すなわち、情報処理装置１００は、類似文書検索装置および索引情報作成装置の一例である。

[0024] 検索対象文書データベース１２１は、問い合わせ文書１０１に対して検索対象となる検索対象文書が登録されたデータベースである。検索対象文書データベース１２１における検索対象文書は、事前に登録されていてもよいし、情報処理装置１００を使用するユーザとのチャットボット等の対話インタフェースにおける対話を通じて追加されてもよし、ネットワークから自動的に収集されてもよい。

[0025] 語間距離情報記憶部１２２は、単語それぞれについて、他の単語との語意の近さ（語間）を表す語間距離情報を格納する。具体的には、語間距離情報としては、単語それぞれ（ｖ）と、他の単語（ｗ）との語意の距離（語間）を表す関数ｄ（ｄ，ｗ）などがある。すなわち、語間距離情報は、単語の語意の近さを示す語間情報の一例である。

[0026] ハッシュ関数記憶部１２３は、インデックス生成モジュール１１１が生成した異なる複数のハッシュ関数を記憶する。各ハッシュ関数は、検索対象文書に出現し得る単語それぞれに対して一意な整数（所定の単語に対して語意の近い順）を対応付ける対応関係をもち、単語集合を受け付けて１つのハッシュ値を出力する。また、異なるハッシュ関数は異なる対応関係をもつ。

[0027] インデックス記憶部１２４は、問い合わせ文書１０１に類似する検索対象文書を検索するための索引構造を記憶する。索引構造は、複数のハッシュ関数を用いて検索対象文書から算出されたベクトル（要約情報）に基づいてインデックス生成部１３２が生成したものであり、検索対象文書それぞれの要約情報を探索する索引情報の一例である。

[0028] 索引構造としては、例えば探索木を適用できる。探索木は、木構造に接続された複数のノード（葉ノードと、葉ノードに至る各ノード）を含む。探索木の葉ノードは、検索対象文書を指し示す。例えば、葉ノードは、検索対象文書のベクトルと当該検索対象文書を識別する識別情報（例えば文書ＩＤ）とを含む。ただし、葉ノードがベクトルを含まなくてもよい。

[0029] 葉ノード以外の各ノードには２つの子ノードが接続されている。葉ノード以外の各ノードは、ベクトルの中の特定の次元に対する閾値をもつ。入力されたベクトルの中の特定の次元のハッシュ値が閾値以上である場合は右子ノードに進み、特定の次元のハッシュ値が閾値未満である場合は左子ノードに進む。このようにして、探索木をルートノードから葉ノードに向かって辿ることで、所定のベクトルに近い検索対象文書を効率的に検索することができる。

[0030] インデックス生成モジュール１１１は、ハッシュ関数生成部１３１と、インデックス生成部１３２とを有する。

[0031] ハッシュ関数生成部１３１は、ハッシュ関数生成部１３１に記憶された検索対象文書と、問い合わせ受信部１３３に記憶された語間距離情報とに基づいて、複数のハッシュ関数を生成し、生成した複数のハッシュ関数をハッシュ関数記憶部１２３に格納する。

[0032] 具体的には、ハッシュ関数生成部１３１は、検索対象文書に含まれる単語を抽出して単語の集合し、ランダムに基準とする単語を選択する。次いで、ハッシュ関数生成部１３１は、語間距離情報記憶部１２２の語間距離情報を参照し、基準とする単語に対して語意の近い順に単語の集合に含まれる単語を整列（ソート）する。次いで、ハッシュ関数生成部１３１は、単語の集合に含まれる単語について、整列順に一意な値（例えばソート順に大きくなる整数値）を割り当ててハッシュ関数を生成する。ハッシュ関数生成部１３１は、上記のハッシュ関数を生成する処理を繰り返すことで、複数のハッシュ関数を生成する。

[0033] 例えば、語間距離情報として単語それぞれ（ｖ）と、他の単語（ｗ）との語意の距離（語間）を表す関数ｄ（ｄ，ｗ）が与えられているものとする。この関数ｄ（ｄ，ｗ）については、語間の類似度や単語のベクトル表現などを参照して予め作成することができる。また、関数ｄ（ｄ，ｗ）は、次のとおりである。

・任意のｖ，ｗ∈Ｗについて、０≦ｄ（ｖ，ｗ）である。

・任意のｗ∈Ｗについて、ｄ（ｗ，ｗ）＝０である。

[0034] また、ハッシュ関数については、任意のｕ、ｖ∈Ｗについて、ｈ（ｕ）＞ｈ（ｖ）⇔ｄ（ｕ，ｗ）＞ｄ（ｖ，ｗ）となるｗ∈Ｗがあるものとする。

[0035] ハッシュ関数生成部１３１は、Ｗからｗをランダムに選択し、すべてのｖ∈Ｗをｄ（ｖ，ｗ）の小さい順にソートする。次いで、ハッシュ関数生成部１３１は、ソートされた語ｗ，v_1，v_2…に整数０，１，２，…を割り当てる。なお、ハッシュ関数生成部１３１は、整数を割り当てる代わりにｄ（ｖ，ｗ）の値をそのまま使ってもよい（重複がないものとする）。

[0036] インデックス生成部１３２は、検索対象文書データベース１２１に記憶された検索対象文書とハッシュ関数記憶部１２３に記憶されたハッシュ関数に基づいて、索引構造を生成し、生成した索引構造をインデックス記憶部１２４に格納する。

[0037] 具体的には、インデックス生成部１３２は、検索対象文書ごとに単語集合

を抽出し、抽出した単語集合を複数のハッシュ関数それぞれに入力して、ハッシュ値のベクトル、すなわち検索対象文書の要約情報を算出する。次いで、インデックス生成部１３２は、複数の検索対象文書に対応する複数のベクトルを効率的に検索できるように、索引構造を生成する。例えば、インデックス生成部１３２は、ベクトルの中の１つの次元に着目し、ベクトルの集合が二分割されるように当該次元のハッシュ値の閾値を決定することを繰り返すことで、探索木を生成する。このとき、インデックス生成部１３２は、探索木の葉ノードにはできる限り単一のベクトルが対応付けられるように中間ノードを生成する。

[0038] 検索モジュール１１２は、問い合わせ受信部１３３と、ハッシュ値算出部１３４と、検索部１３５と、出力部１３６とを有する。

[0039] 問い合わせ受信部１３３は、問い合わせ文書１０１を受信する。問い合わせ受信部１３３は、ユーザから文字列として入力された問い合わせ文書１０１を受信してもよいし、ユーザが口頭で発した問い合わせ発話の音声信号を文字列に変換してもよい。また、問い合わせ受信部１３３は、他の情報処理装置から文字列または音声信号を受信してもよい。

[0040] ハッシュ値算出部１３４は、ハッシュ関数記憶部１２３に記憶された複数のハッシュ関数に基づいて、問い合わせ文書１０１に対応するベクトル、すなわち問い合わせ文書１０１の要約情報を生成する。具体的には、ハッシュ値算出部１３４は、問い合わせ文書１０１から単語集合を抽出し、抽出した単語集合を複数のハッシュ関数それぞれに入力して、ハッシュ値のベクトルを算出する。

[0041] 検索部１３５は、インデックス記憶部１２４に記憶された索引構造と問い合わせ文書１０１のベクトルに基づいて、近傍探索により問い合わせ文書１０１に最も類似する検索対象文書を検索する。具体的には、問い合わせ文書１０１に最も類似する検索対象文書は、ベクトル同士を比較したときにハッシュ値が一致する次元が最も多いものである。

[0042] 例えば、検索部１３５は、問い合わせ文書１０１のベクトルの中の特定の

次元のハッシュ値と閾値とを比較しながら、インデックス記憶部１２４に記憶された探索木をルートノードから葉ノードに向かって辿り、特定の葉ノードに到達する。検索部１３５は、到達した葉ノードに対応する検索対象文書を選択する。

[0043] なお、到達した葉ノードに２以上の検索対象文書が対応付けられている場合、すなわち、ハッシュ値が一致する次元数が同じであり、探索木では検索対象文書を１つに絞り込めない場合、検索部１３５は、ハッシュ値同士を比較してユークリッド距離を求める。次いで、検索部１３５は、ユークリッド距離がより近いものを最も類似する検索対象文書とする。

[0044] 出力部１３６は、検索された検索対象文書を類似文書１０２として出力する。例えば、出力部１３６は、類似文書１０２の文字列をディスプレイ等に表示してもよいし、類似文書１０２を音声信号に変換してスピーカにより音声を再生してもよい。また、出力部１３６は、他の情報処理装置に類似文書１０２の文字列または音声信号を送信してもよい。

[0045] また、出力部１３６は、類似文書１０２を出力する代わりに、検索された検索対象文書に対して検索対象文書データベース１２１において予め紐付けられた処理を実施してもよい。具体的には、検索対象文書データベース１２１には、検索対象文書ごとに、所定の処理（例えばスケジュール登録、メール送信）が登録されているものとする。出力部１３６は、検索された検索対象文書に紐付けられた処理を検索対象文書データベース１２１より読み出して実行することで、問い合わせ文書１０１に対応した処理を行うことが可能となる。

[0046] ［動作例］

図２は、実施形態にかかる情報処理装置の処理フローの一例を示す説明図である。図２に示すように、情報処理装置１００は、索引構造を作成する事前処理（Ｓ１）と、問い合わせ文書１０１に対する類似文書１０２を検索して出力する検索処理（Ｓ２）とを行う。

[0047] 先ず、事前処理（Ｓ１）について説明する。事前処理（Ｓ１）では、先ず

検索対象文書データベース１２１より検索対象文書が読み出され、ハッシュ関数生成部１３１に入力される（Ｓ１１）。

[0048] ハッシュ関数生成部１３１では、問い合わせ受信部１３３より語間の距離の入力を受け付け（Ｓ１３）、入力された検索対象文書と、語間の距離とに基づき、複数のハッシュ関数１２３ａを生成する（Ｓ１２）。

[0049] 図３は、ハッシュ関数の生成処理の一例を示すフローチャートである。図３に示すように、ハッシュ関数生成部１３１は、検索対象文書の入力を受け付け（Ｓ３１）、検索対象文書に含まれる単語（出現語）を抽出することで（Ｓ３２）、単語の集合（語集合）を取得する（Ｓ３３）。

[0050] 次いで、ハッシュ関数生成部１３１は、Ｓ３４～Ｓ３９の処理を生成するハッシュ関数の数であるｋ回繰り返すことで、複数のハッシュ関数１２３ａを生成する。

[0051] 具体的には、ハッシュ関数生成部１３１は、語集合の中からランダムに１つの単語を選択し（Ｓ３５）、語間の距離を示す語間距離情報の入力を受け付け（Ｓ３６）、選択した単語と他の単語との距離を語間距離情報より参照する（Ｓ３７）。次いで、ハッシュ関数生成部１３１は、選択した単語との距離の近い順に語集合の単語をソートし（Ｓ３８）、整列した各単語に整列順に大きくなるような整数値をハッシュ値として割り当てる（Ｓ３９）。

[0052] 次いで、ハッシュ関数生成部１３１は、Ｓ３４～Ｓ３９の処理をｋ回繰り返して得られた複数のハッシュ関数１２３ａを出力し、インデックス生成部１３２に格納する（Ｓ４０）。

[0053] 図４は、ハッシュ関数１２３ａを説明する説明図である。図４に示すように、ハッシュ関数１２３ａにおけるh_1，h_2，…ｈ２，…が１つのハッシュ関数である。

[0054] 例えば、h_1は、（猫）を基準の単語としており、語集合における各単語について（猫）に対する語間の距離に応じた整数値が割り当てられている。また、h_2は、（ごはん）を基準の単語としており、語集合における各単語について（ごはん）に対する語間の距離に応じた整数値が割り当てられている。

また、h_3は、（にゃんこ）を基準の単語としており、語集合における各単語について（にゃんこ）に対する語間の距離に応じた整数値が割り当てられている。また、h_4は、（えさ）を基準の単語としており、語集合における各単語について（えさ）に対する語間の距離に応じた整数値が割り当てられている。また、h_5は、（花）を基準の単語としており、語集合における各単語について（花）に対する語間の距離に応じた整数値が割り当てられている。また、h_6は、（水）を基準の単語としており、語集合における各単語について（水）に対する語間の距離に応じた整数値が割り当てられている。

[0055] ここで、図４に例示したハッシュ関数１２３ａにより次の文書Ａ～文書Ｃのハッシュ値を求める場合を例示する。

文書Ａ：「猫にごはんをやる」

文書Ｂ：「にゃんこにえさをやる」

文書Ｃ：「花に水をやる」

[0056] 文書Ａの語集合は｛猫、ご飯｝、文書Ｂの語集合は｛にゃんこ、えさ｝、文書Ｃの語集合は｛花、水｝である。よって、図４に例示したハッシュ関数１２３ａによる文書Ａのハッシュ値H_Aは、H_A＝００１１３３となる。同様に、文書Ｂのハッシュ値H_Bは、H_B＝１１００２２となる。また、文書Ｃのハッシュ値H_Cは、H_C＝４２４２００となる。

[0057] ここで、文書Ａ～Ｃのハッシュ値を比較してハミング距離を計算すると、H_A，H_B：６、H_A，H_C：６、H_B，H_C：６となる。また、文書Ａ～Ｃのハッシュ値を比較してユークリッド距離を計算すると、H_A，H_B：１、H_A，H_C：６．９、H_B，H_C：６．２となる。このように、ハミング距離では類似度の検証が困難な場合（共通語が含まれていない場合）でも、ユークリッド距離により互いに類似する文書（図示例では文書Ａ、Ｂ）を検証することができる。

[0058] 図２に戻り、Ｓ１２に次いで、インデックス生成部１３２は、入力された検索対象文書と、生成したハッシュ関数１２３ａに基づいて、索引構造１２４ａを生成し（Ｓ１４）、生成した索引構造１２４ａをインデックス記憶部

１２４に格納する。

[0059] 検索処理（Ｓ２）では、問い合わせ受信部１３３が受信した問い合わせ文書１０１がハッシュ値算出部１３４に入力される（Ｓ２１）。ハッシュ値算出部１３４は、ハッシュ関数記憶部１２３に記憶された複数のハッシュ関数に基づいて、入力された問い合わせ文書１０１のハッシュ値を複数生成し（Ｓ２２）、問い合わせ文書１０１に対応するベクトルを得る。

[0060] 次いで、検索部１３５は、インデックス記憶部１２４に記憶された索引構造における検索対象文書それぞれのベクトルと、問い合わせ文書１０１のベクトルのハッシュ値を照合し（Ｓ２３）、問い合わせ文書１０１に最も類似する検索対象文書を検索する。出力部１３６は、検索部１３５が検索した類似文書１０２を出力する（Ｓ２４）。

[0061] 図５は、語間の類似度を説明する説明図である。具体的には、図５は、語間の類似度を示す高次元の空間における語Ｗ１～Ｗ６それぞれの配置を俯瞰した図である。図５の語Ｗ１～Ｗ６それぞれは、文書に含まれる単語を示す。ここで、語Ｗ１～Ｗ３は、例えば「猫」などの単語について類似するものであり、点線で示すクラスターを形成している。同様に、語Ｗ４～Ｗ６は、例えば「犬」などの単語について類似するものであり、語Ｗ１～Ｗ３とは別のクラスターを形成している。

[0062] 図５に示すように、語間の類似度を示す高次元の空間において、単純な射影では、類似度をうまく評価することは困難である。例えば、軸Ａ１への直交射影では、類似する単語（語Ｗ１～Ｗ３または語Ｗ４～Ｗ６）の値が近くなる。しかしながら、軸Ａ１とは異なる軸Ａ２への直交射影では、互いに類似しない単語（例えば語Ｗ１と語Ｗ４）が互いに類似する単語（例えば語Ｗ１と語Ｗ３）よりも近くなる場合がある。

[0063] 本実施形態では、ランダムに基準とする単語を選択し、Ｍｉｎ　ｈａｓｈによるハッシュ値を用いることから、基準とする単語（基準点）からの距離による射影を用いている。

[0064] 図６は、Ｍｉｎ　ｈａｓｈによるハッシュ値を説明する説明図である。こ

こで、語Ｗ１は「猫」、語Ｗ２は「にゃんこ」、語Ｗ３は「キャット」、語Ｗ４は「犬」、語Ｗ５は「鼠」、語Ｗ６は「ドッグ」、語Ｗ７は「わんこ」であるものとする。また、基準点は語Ｗ１であるものとする。

[0065] 図６に示すように、基準点である語Ｗ１をもとに、ハッシュ関数１２３ａでは、１：「猫」、２：「にゃんこ」、３：「キャット」、４：「犬」、５：「鼠」、６：「ドッグ」、７：「わんこ」の値が割り当てられる。ここで、基準点からの距離ならば、基準点に近いほど類似度の大小関係が保たれている（例えば１：「猫」、２：「にゃんこ」、３：「キャット」など）。しかしながら、基準点からの距離が遠いところでは（例えば５：「鼠」、６：「ドッグ」）、類似しない語の値が近くなる場合がある。

[0066] 本実施形態では、語集合のハッシュ値の中の最小値を用いる（Ｍｉｎ ｈａｓｈ）ので、類似度の大小関係が保たれた、語の類似度を表現できる射影として適切なものとなっている。

[0067] 図７は、実施形態にかかる情報処理装置１００の動作の概要を説明する説明図である。図７に示すように、事前処理（Ｓ１）において、情報処理装置１００は、語間距離情報による語間の類似度をもとに、検索対象文書データベース１２１の検索対象文書１２１ａそれぞれに含まれる単語の集合について、所定の単語を基準として語意の近い順に近い値を割り当てるハッシュ関数を複数生成する。次いで、情報処理装置１００は、生成した複数のハッシュ関数を用いて検索対象文書１２１ａそれぞれをＭｉｎ Ｈａｓｈによる変換を行い、算出されたベクトルを検索するための索引構造を生成してインデックス記憶部１２４に格納する。

[0068] 検索処理（Ｓ２）において、情報処理装置１００は、入力された問い合わせ文書１０１について、Ｓ１で生成した同じ複数のハッシュ関数を用いてＭｉｎ Ｈａｓｈによる変換を行う。次いで、情報処理装置１００は、問い合わせ文書１０１より算出したハッシュ値のベクトルと、インデックス記憶部１２４に格納されたハッシュ値のベクトルとを比較することで、問い合わせ文書１０１に最も類似する検索対象文書１２１ａを検索する。

[0069] 図示例では、「会議を調整したい」とする問い合わせ文書１０１に対し、検索対象文書１２１ａの中で「スケジュールの調整」と、「打ち合わせの調整」とが「調整」という単語においてハッシュ値が一致することから、ハミング距離が近いものとなる。ここで、問い合わせ文書１０１に含まれる「会議」に対し、「スケジュール」よりも「会議」の方がより語意が近く、ユークリッド距離が近いものとなる。したがって、「会議を調整したい」とする問い合わせ文書１０１に対しては、「打ち合わせを調整したい」とする類似文書１０２が得られることとなる。

[0070] なお、出力部１３６は、検索された検索対象文書について、最も類似する１つの類似文書１０２を出力してもよいし、ハミング距離およびユークリッド距離により得られた類似度が所定の閾値以上である複数の類似文書１０２を出力してもよい。

[0071] 図８は、検索対象文書の絞り込みを説明する説明図である。具体的には、図８は、類似度の閾値と検索対象文書１２１ａのヒット数との関係例を示すグラフである。

[0072] グラフ１０は、問い合わせ文書１０１と検索対象文書１２１ａの類似度の閾値と、類似度が閾値より大きい検索対象文書１２１ａの数（ヒット数）との間の関係を示す。類似度は、問い合わせ発話のベクトルと検索対象発話のベクトルの間でハッシュ値が一致する次元の数に、ハッシュ値同士のユークリッド距離を合わせたものである。

[0073] （Ａ）関連語を考慮せずにベクトルを算出する方法では、問い合わせ文書１０１と各検索対象文書１２１ａの類似度が全体として低く算出される。よって、類似度の閾値とヒット数との間の関係は曲線１１のようになる。すなわち、類似度の閾値を低く設定してもヒット数が少なくなり、類似する検索対象文書１２１ａの検索漏れが多くなる。

[0074] （Ｂ）関連語を同一視してベクトルを算出する方法では、問い合わせ文書１０１と各検索対象文書１２１ａの類似度が全体として高く算出される。よって、類似度の閾値とヒット数との間の関係は曲線１２のようになる。すな

わち、類似度の閾値を高く設定してもヒット数が多くなり、類似する検索対象文書１２１ａを効率的に絞り込むことが難しい。

[0075] （Ｃ）本実施形態の方法では、問い合わせ文書１０１と各検索対象文書１２１ａの類似度がユークリッド距離を加味したものとなる。よって、類似度の閾値とヒット数との間の関係は曲線１３のように連続的なものとなる。その結果、問い合わせ文書１０１に類似する検索対象文書１２１ａを効率的に絞り込むことができる。

[0076] 図９は、操作画面の表示例を示す説明図である。操作画面２０は、例えばチャットボット等の対話インタフェースにおいて情報処理装置１００のユーザに提示することでユーザからの各種操作を受け付ける画面である。操作画面２０において、表示領域２１は処理結果などを表示する領域であり、入力領域２２は文書などの入力を行う領域である。例えば、情報処理装置１００は、入力領域２２に入力された問い合わせ文書１０１に対して類似文書１０２の検索を行い、検索結果に応じた出力を表示領域２１に行う。

[0077] 例えば、情報処理装置１００は、「スケジュール調整したい」とする入力文書２１ａを問い合わせ文書１０１として類似文書１０２の検索を行い、検索結果２１ｂを表示領域２１に表示する。具体的には、「スケジュール調整したい」とする問い合わせ文書１０１に対し、検索対象文書１２１ａの中から類似文書１０２として得られた「スケジュールの調整」に対応するスケジュール登録の処理が実行される。

[0078] 情報処理装置１００では、図９の例におけるチャットボット等の対話インタフェースのように、問い合わせ文書１０１が主に話し言葉であり、１文が短く表現も多様である場合であっても、適切に類似文書１０２の検索を行うことができる。

[0079] ［効果］

以上のように、情報処理装置１００は、ハッシュ関数生成部１３１と、インデックス生成部１３２と、ハッシュ値算出部１３４と、検索部１３５とを有する。ハッシュ関数生成部１３１は、検索対象文書１２１ａに含まれる単

語の集合と、単語の語意の近さを示す語間距離情報とに基づき、単語の集合に含まれる単語それぞれに対して、所定の単語を基準として語意の近い順に近い値を割り当てるハッシュ関数を生成する。インデックス生成部１３２は、生成したハッシュ関数に基づいて複数の検索対象文書１２１ａそれぞれの要約情報を算出する。ハッシュ値算出部１３４は、生成したハッシュ関数に基づいて入力文書（問い合わせ文書１０１）の要約情報を算出する。検索部１３５は、算出した検索対象文書１２１ａの要約情報と、問い合わせ文書１０１の要約情報との間の比較に基づいて、複数の検索対象文書１２１ａの中から入力文書に類似する文書を検索する。

[0080] このため、情報処理装置１００において、生成したハッシュ関数に基づいて算出した検索対象文書１２１ａの要約情報と、問い合わせ文書１０１の要約情報とに含まれるハッシュ値は、所定の単語に対する語意の距離に応じた値となる。したがって、情報処理装置１００では、検索対象文書１２１ａの要約情報と、問い合わせ文書１０１の要約情報との間の比較において、例えば互いの要約情報の間におけるハミング距離だけでなく、ユークリッド距離を用いた類似度合いの検証を行うことができ、類似する文書の検索精度を向上させることができる。

[0081] また、インデックス生成部１３２およびハッシュ値算出部１３４のそれぞれは、生成したハッシュ関数において検索対象文書１２１ａまたは問い合わせ文書１０１に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する。このように、情報処理装置１００では、Min-hash関数により検索対象文書１２１ａまたは問い合わせ文書１０１の要約情報を算出できる。

[0082] また、ハッシュ関数生成部１３１は、所定の単語を選び直してハッシュ関数を生成する処理を繰り返すことでハッシュ関数を複数生成する。そして、インデックス生成部１３２およびハッシュ値算出部１３４のそれぞれは、生成した複数のハッシュ関数によって検索対象文書１２１ａまたは問い合わせ文書１０１に含まれる単語の集合から算出される複数のハッシュ値を含むベ

クトルを要約情報として算出する。これにより、情報処理装置１００では、検索対象文書１２１ａの要約情報と、問い合わせ文書１０１の要約情報との間の比較において、複数のハッシュ値を列挙したベクトルの比較により、例えばハミング距離、ユークリッド距離などを求めて類似度合いを検証することができる。

[0083] また、情報処理装置１００のインデックス生成部１３２は、算出した複数の検索対象文書１２１ａそれぞれの要約情報と、検索対象文書１２１ａとを対応付けた索引情報を生成する。検索部１３５は、問い合わせ文書１０１の要約情報と、索引情報において検索対象文書１２１ａと対応付けられた要約情報との間の比較を行う。情報処理装置１００では、索引情報を生成しておくことで、複数の検索対象文書１２１ａの中から問い合わせ文書１０１に類似する文書の検索を高速に行うことができる。

[0084] ［その他］

なお、図示した各装置の各構成要素は、必ずしも物理的に図示の如く構成されていることを要しない。すなわち、各装置の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷や使用状況などに応じて、任意の単位で機能的または物理的に分散・統合して構成することができる。

[0085] 例えば、本実施形態では、インデックス生成モジュール１１１と、検索モジュール１１２とを有する情報処理装置１００を例示したが、インデックス生成モジュール１１１と、検索モジュール１１２とはそれぞれ異なる情報処理装置が有していてもよい。すなわち、事前処理（Ｓ１）と、検索処理（Ｓ２）とは、それぞれ異なる情報処理装置で実施してもよい。

[0086] また、情報処理装置１００で行われる各種処理機能は、ＣＰＵ（またはＭＰＵ、ＭＣＵ（Micro Controller Unit）等のマイクロ・コンピュータ）上で、その全部または任意の一部を実行するようにしてもよい。また、各種処理機能は、ＣＰＵ（またはＭＰＵ、ＭＣＵ等のマイクロ・コンピュータ）で解析実行されるプログラム上、またはワイヤードロジックによるハードウエ

ア上で、その全部または任意の一部を実行するようにしてもよいことは言うまでもない。また、情報処理装置１００で行われる各種処理機能は、クラウドコンピューティングにより、複数のコンピュータが協働して実行してもよい。

[0087] ところで、上記の実施形態で説明した各種の処理は、予め用意されたプログラムをコンピュータで実行することで実現できる。そこで、以下では、上記の実施例と同様の機能を有するプログラムを実行するコンピュータ（ハードウエア）の一例を説明する。図１０は、プログラムを実行するコンピュータの一例を示す図である。

[0088] 図１０に示すように、コンピュータ１は、各種演算処理を実行するＣＰＵ２０１と、データ入力を受け付ける入力装置２０２と、モニタ２０３と、スピーカ２０４とを有する。また、コンピュータ１は、記憶媒体からプログラム等を読み取る媒体読取装置２０５と、各種装置と接続するためのインタフェース装置２０６と、有線または無線により外部機器と通信接続するための通信装置２０７とを有する。また、コンピュータ１は、各種情報を一時記憶するＲＡＭ２０８と、ハードディスク装置２０９とを有する。また、コンピュータ１内の各部（２０１～２０９）は、バス２１０に接続される。

[0089] ハードディスク装置２０９には、上記の実施形態で説明した各種の処理を実行するためのプログラム２１１が記憶される。また、ハードディスク装置２０９には、プログラム２１１が参照する各種データ２１２（例えば語間距離情報記憶部１２２、検索対象文書データベース１２１、ハッシュ関数記憶部１２３およびインデックス記憶部１２４の情報）が記憶される。入力装置２０２は、例えば、コンピュータ１の操作者から操作情報の入力を受け付ける。モニタ２０３は、例えば、操作者が操作する各種画面を表示する。インタフェース装置２０６は、例えば印刷装置等が接続される。通信装置２０７は、ＬＡＮ（Local Area Network）等の通信ネットワークと接続され、通信ネットワークを介した外部機器との間で各種情報をやりとりする。

[0090] ＣＰＵ２０１は、ハードディスク装置２０９に記憶されたプログラム２１

１を読み出して、ＲＡＭ２０８に展開して実行することで、ハッシュ関数生成部１３１、インデックス生成部１３２、問い合わせ受信部１３３、ハッシュ値算出部１３４、検索部１３５および出力部１３６に関する各種の処理を行う。なお、プログラム２１１は、ハードディスク装置２０９に記憶されていなくてもよい。例えば、コンピュータ１が読み取り可能な記憶媒体に記憶されたプログラム２１１を、コンピュータ１が読み出して実行するようにしてもよい。コンピュータ１が読み取り可能な記憶媒体は、例えば、ＣＤ－ＲＯＭやＤＶＤディスク、ＵＳＢ（Universal Serial Bus）メモリ等の可搬型記録媒体、フラッシュメモリ等の半導体メモリ、ハードディスクドライブ等が対応する。また、公衆回線、インターネット、ＬＡＮ等に接続された装置にこのプログラム２１１を記憶させておき、コンピュータ１がこれらからプログラムを読み出して実行するようにしてもよい。

符号の説明

[0091] １…コンピュータ

１０…グラフ

１１～１３…曲線

２０…操作画面

２１…表示領域

２１ａ…入力文書

２１ｂ…検索結果

２２…入力領域

１００…情報処理装置

１０１…問い合わせ文書

１０２…類似文書

１１１…インデックス生成モジュール

１１２…検索モジュール

１２１…検索対象文書データベース

１２１ａ…検索対象文書

１２２…語間距離情報記憶部

１２３…ハッシュ関数記憶部

１２３ａ…ハッシュ関数

１２４…インデックス記憶部

１２４ａ…索引構造

１３１…ハッシュ関数生成部

１３２…インデックス生成部

１３３…問い合わせ受信部

１３４…ハッシュ値算出部

１３５…検索部

１３６…出力部

２０１…ＣＰＵ

２０２…入力装置

２０３…モニタ

２０４…スピーカ

２０５…媒体読取装置

２０６…インタフェース装置

２０７…通信装置

２０８…ＲＡＭ

２０９…ハードディスク装置

２１０…バス

２１１…プログラム

２１２…各種データ

Ａ１～Ａ２…軸

Ｗ１～Ｗ７…語

請求の範囲

[請求項1] 検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、前記単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成し、

生成した前記ハッシュ関数に基づいて複数の前記検索対象文書それぞれの要約情報を算出し、

生成した前記ハッシュ関数に基づいて入力文書の要約情報を算出し、

算出した前記検索対象文書の要約情報と、前記入力文書の要約情報との間の比較に基づいて、複数の前記検索対象文書の中から前記入力文書に類似する文書を検索する、

処理をコンピュータが実行することを特徴とする類似文書検索方法。

[請求項2] 前記算出する処理のそれぞれは、生成した前記ハッシュ関数において前記検索対象文書または前記入力文書に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する、

ことを特徴とする請求項１に記載の類似文書検索方法。

[請求項3] 前記生成する処理は、前記所定の単語を選び直して前記ハッシュ関数を生成する処理を繰り返すことで前記ハッシュ関数を複数生成し、

前記算出する処理のそれぞれは、生成した複数の前記ハッシュ関数によって前記検索対象文書または前記入力文書に含まれる単語の集合から算出される複数のハッシュ値を含むベクトルを要約情報として算出する、

ことを特徴とする請求項２に記載の類似文書検索方法。

[請求項4] 算出した複数の前記検索対象文書それぞれの要約情報と、前記検索対象文書とを対応付けた索引情報を生成する処理をさらにコンピュー

タが実行し、

前記検索する処理は、前記入力文書の要約情報と、前記索引情報において前記検索対象文書と対応付けられた要約情報との間の比較を行う、

ことを特徴とする請求項１に記載の類似文書検索方法。

[請求項5] 検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、前記単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成し、

生成した前記ハッシュ関数に基づいて複数の前記検索対象文書それぞれの要約情報を算出し、

生成した前記ハッシュ関数に基づいて入力文書の要約情報を算出し、

算出した前記検索対象文書の要約情報と、前記入力文書の要約情報との間の比較に基づいて、複数の前記検索対象文書の中から前記入力文書に類似する文書を検索する、

処理をコンピュータに実行させることを特徴とする類似文書検索プログラム。

[請求項6] 前記算出する処理のそれぞれは、生成した前記ハッシュ関数において前記検索対象文書または前記入力文書に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する、

ことを特徴とする請求項５に記載の類似文書検索プログラム。

[請求項7] 前記生成する処理は、前記所定の単語を選び直して前記ハッシュ関数を生成する処理を繰り返すことで前記ハッシュ関数を複数生成し、

前記算出する処理のそれぞれは、生成した複数の前記ハッシュ関数によって前記検索対象文書または前記入力文書に含まれる単語の集合から算出される複数のハッシュ値を含むベクトルを要約情報として算

出する、

ことを特徴とする請求項６に記載の類似文書検索プログラム。

[請求項8] 算出した複数の前記検索対象文書それぞれの要約情報と、前記検索対象文書とを対応付けた索引情報を生成する処理をさらにコンピュータが実行し、

前記検索する処理は、前記入力文書の要約情報と、前記索引情報において前記検索対象文書と対応付けられた要約情報との間の比較を行う、

ことを特徴とする請求項５に記載の類似文書検索プログラム。

[請求項9] 検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、前記単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成するハッシュ関数生成部と、

生成した前記ハッシュ関数に基づいて複数の前記検索対象文書それぞれの要約情報を算出する第１の算出部と、

生成した前記ハッシュ関数に基づいて入力文書の要約情報を算出する第２の算出部と、

算出した前記検索対象文書の要約情報と、前記入力文書の要約情報との間の比較に基づいて、複数の前記検索対象文書の中から前記入力文書に類似する文書を検索する検索部と、

を有することを特徴とする類似文書検索装置。

[請求項10] 前記第１の算出部および前記第２の算出部のそれぞれは、生成した前記ハッシュ関数において前記検索対象文書または前記入力文書に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する、

ことを特徴とする請求項９に記載の類似文書検索装置。

[請求項11] 前記ハッシュ関数生成部は、前記所定の単語を選び直して前記ハッシュ関数を生成する処理を繰り返すことで前記ハッシュ関数を複数生

成し、

前記第１の算出部および前記第２の算出部のそれぞれは、生成した複数の前記ハッシュ関数によって前記検索対象文書または前記入力文書に含まれる単語の集合から算出される複数のハッシュ値を含むベクトルを要約情報として算出する、

ことを特徴とする請求項１０に記載の類似文書検索装置。

[請求項12] 算出した複数の前記検索対象文書それぞれの要約情報と、前記検索対象文書とを対応付けた索引情報を生成する索引情報生成部をさらに有し、

前記検索部は、前記入力文書の要約情報と、前記索引情報において前記検索対象文書と対応付けられた要約情報との間の比較を行う、

ことを特徴とする請求項９に記載の類似文書検索装置。

[請求項13] 検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、前記単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成し、

生成した前記ハッシュ関数に基づいて複数の前記検索対象文書それぞれの要約情報を算出し、

算出した複数の前記検索対象文書それぞれの要約情報を探索する索引情報を生成する、

処理をコンピュータが実行することを特徴とする索引情報作成方法。

[請求項14] 前記算出する処理は、生成した前記ハッシュ関数において前記検索対象文書に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する、

ことを特徴とする請求項１３に記載の索引情報作成方法。

[請求項15] 前記生成する処理は、前記所定の単語を選び直して前記ハッシュ関数を生成する処理を繰り返すことで前記ハッシュ関数を複数生成し、

前記算出する処理は、生成した複数の前記ハッシュ関数によって前記検索対象文書に含まれる単語の集合から算出される複数のハッシュ値を含むベクトルを要約情報として算出する、

ことを特徴とする請求項１４に記載の索引情報作成方法。

[請求項16] 検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、前記単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成し、

生成した前記ハッシュ関数に基づいて複数の前記検索対象文書それぞれの要約情報を算出し、

算出した複数の前記検索対象文書それぞれの要約情報を探索する索引情報を生成する、

処理をコンピュータに実行させることを特徴とする索引情報作成プログラム。

[請求項17] 前記算出する処理は、生成した前記ハッシュ関数において前記検索対象文書に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する、

ことを特徴とする請求項１６に記載の索引情報作成プログラム。

[請求項18] 前記生成する処理は、前記所定の単語を選び直して前記ハッシュ関数を生成する処理を繰り返すことで前記ハッシュ関数を複数生成し、

前記算出する処理は、生成した複数の前記ハッシュ関数によって前記検索対象文書に含まれる単語の集合から算出される複数のハッシュ値を含むベクトルを要約情報として算出する、

ことを特徴とする請求項１７に記載の索引情報作成プログラム。

[請求項19] 検索対象文書に含まれる単語の集合と、単語の語意の近さを示す語間情報とに基づき、前記単語の集合に含まれる単語それぞれに対して、所定の単語を基準として当該単語に対して語意の近い順に近い値を割り当てるハッシュ関数を生成するハッシュ関数生成部と、

生成した前記ハッシュ関数に基づいて複数の前記検索対象文書それぞれの要約情報を算出する算出部と、

算出した複数の前記検索対象文書それぞれの要約情報を探索する索引情報を生成する索引情報生成部と、

を有することを特徴とする索引情報作成装置。

[請求項20] 前記算出部は、生成した前記ハッシュ関数において前記検索対象文書に含まれる単語の集合の単語それぞれに割り当てられた値の中で最小の値をハッシュ値として算出する、

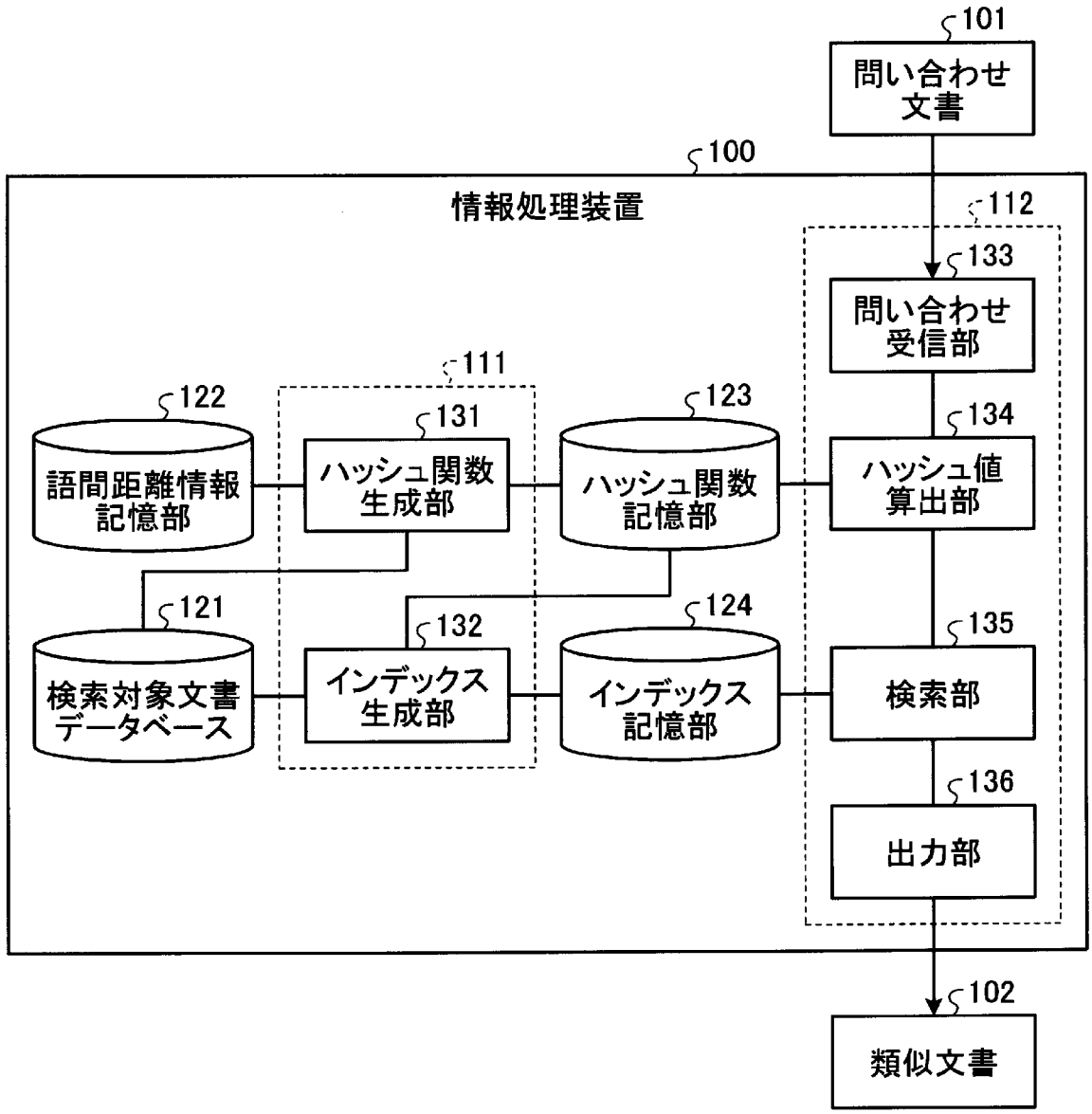
ことを特徴とする請求項１９に記載の索引情報作成装置。

[請求項21] 前記ハッシュ関数生成部は、前記所定の単語を選び直して前記ハッシュ関数を生成する処理を繰り返すことで前記ハッシュ関数を複数生成し、

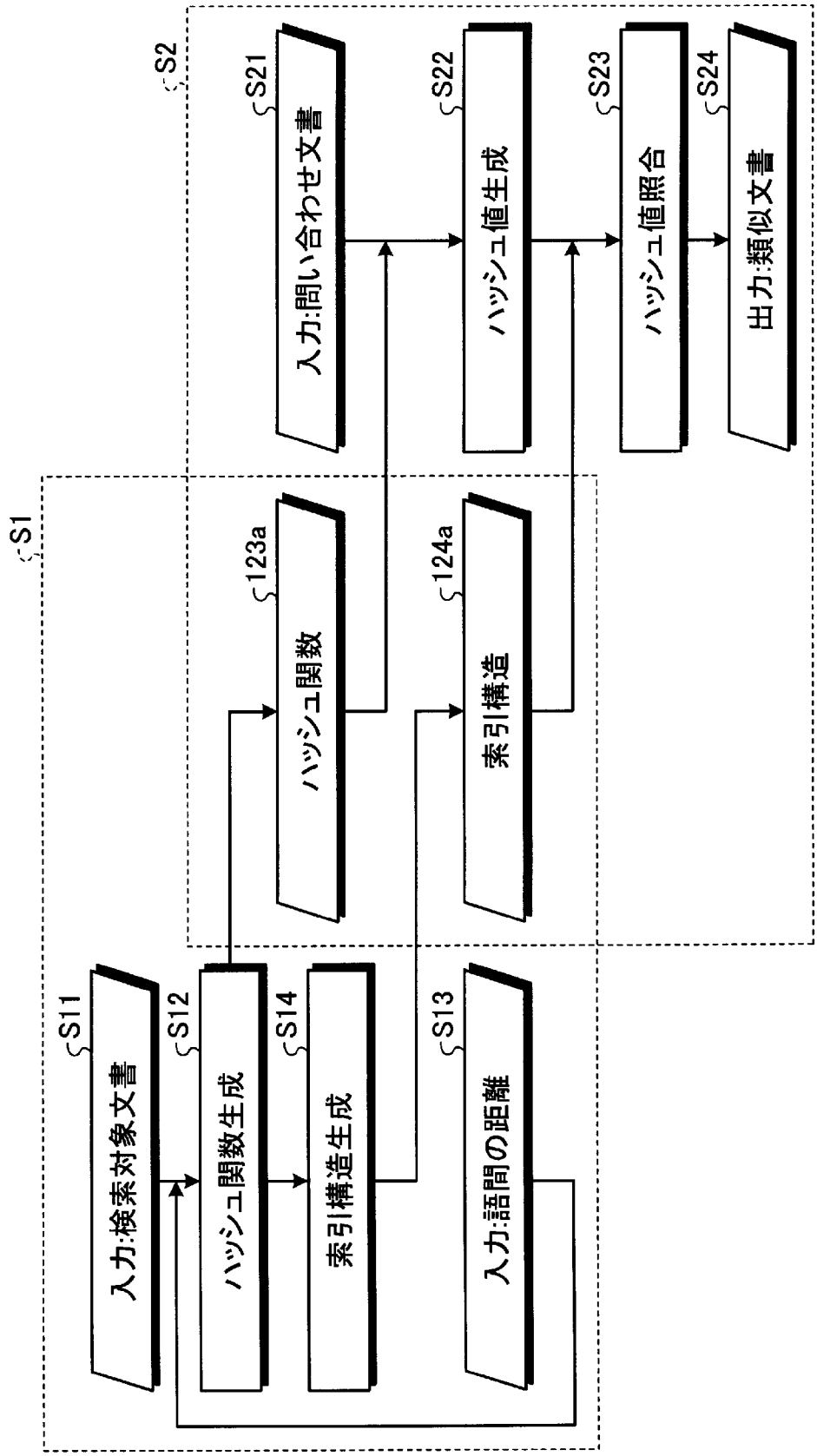
前記算出部は、生成した複数の前記ハッシュ関数によって前記検索対象文書に含まれる単語の集合から算出される複数のハッシュ値を含むベクトルを要約情報として算出する、

ことを特徴とする請求項２０に記載の索引情報作成装置。

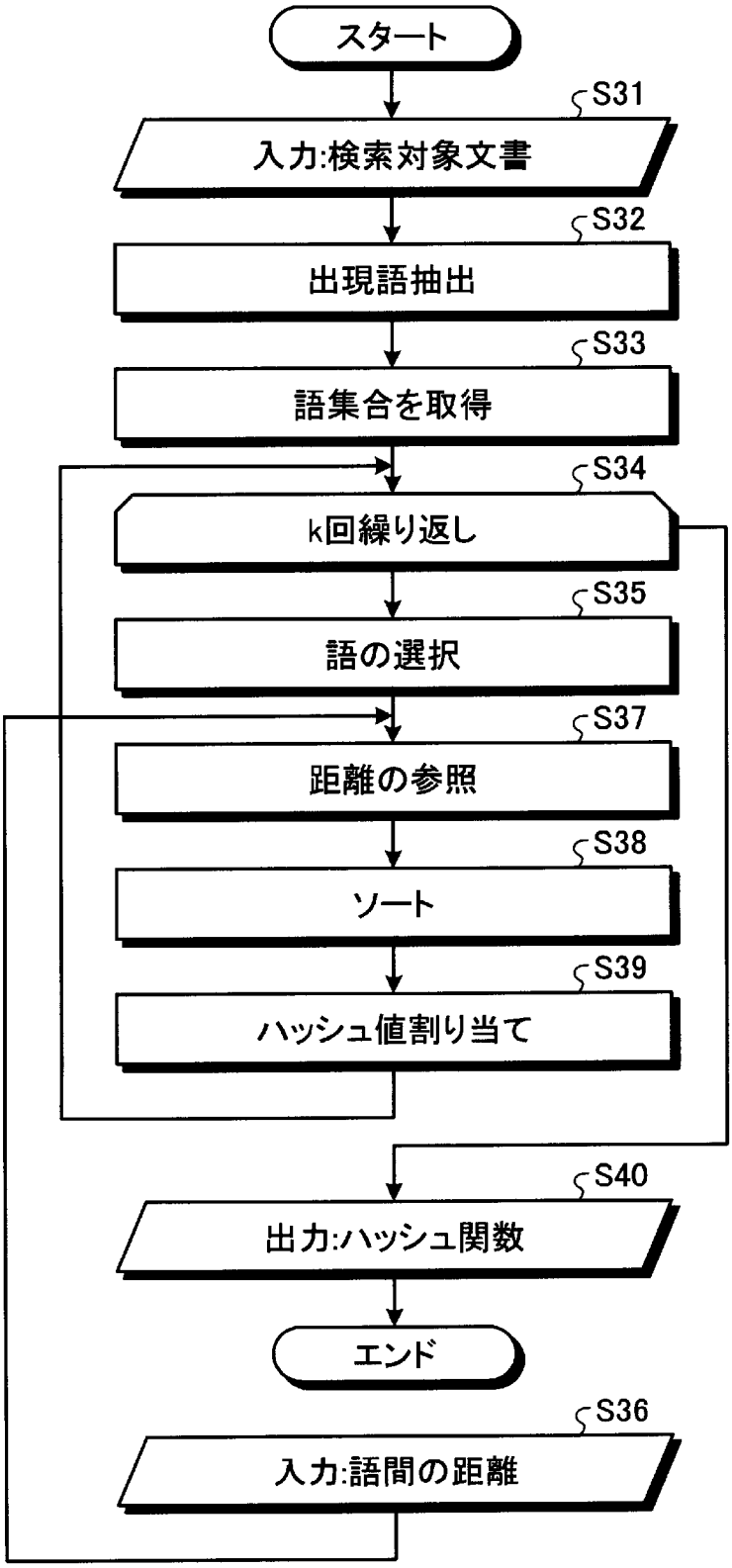
[図1]


101
問い合わせ文書
100
情報処理装置
112
133
問い合わせ受信部
111
122
131
123
134
語間距離情報記憶部
ハッシュ関数生成部
ハッシュ関数記憶部
ハッシュ値算出部
121
132
124
135
検索対象文書データベース
インデックス生成部
インデックス記憶部
検索部
136
出力部
102
類似文書

[図2]


S1
S2
S11
入力:検索対象文書
S12
ハッシュ関数生成
S13
入力:語間の距離
S14
索引構造生成
123a
ハッシュ関数
124a
索引構造
S21
入力:問い合わせ文書
S22
ハッシュ値生成
S23
ハッシュ値照合
S24
出力:類似文書

[図3]

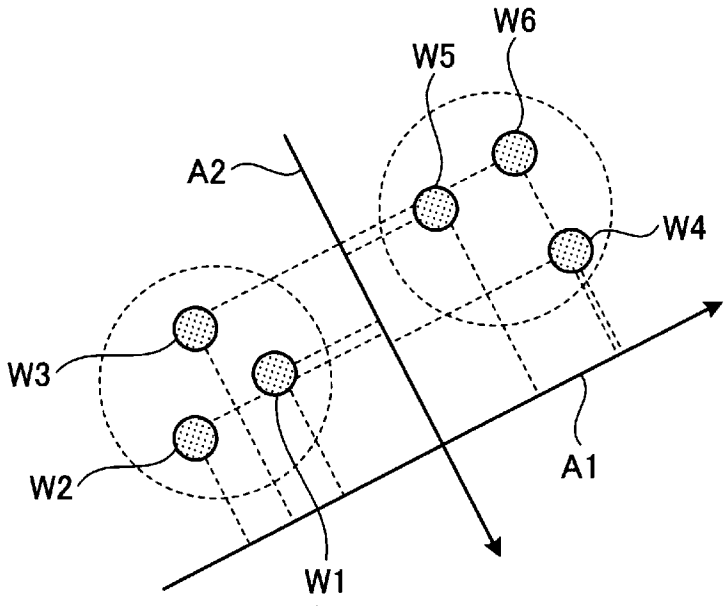

スタート
S31
入力:検索対象文書
S32
出現語抽出
S33
語集合を取得
S34
k回繰り返し
S35
語の選択
S37
距離の参照
S38
ソート
S39
ハッシュ値割り当て
S40
出力:ハッシュ関数
エンド
S36
入力:語間の距離

[図4]

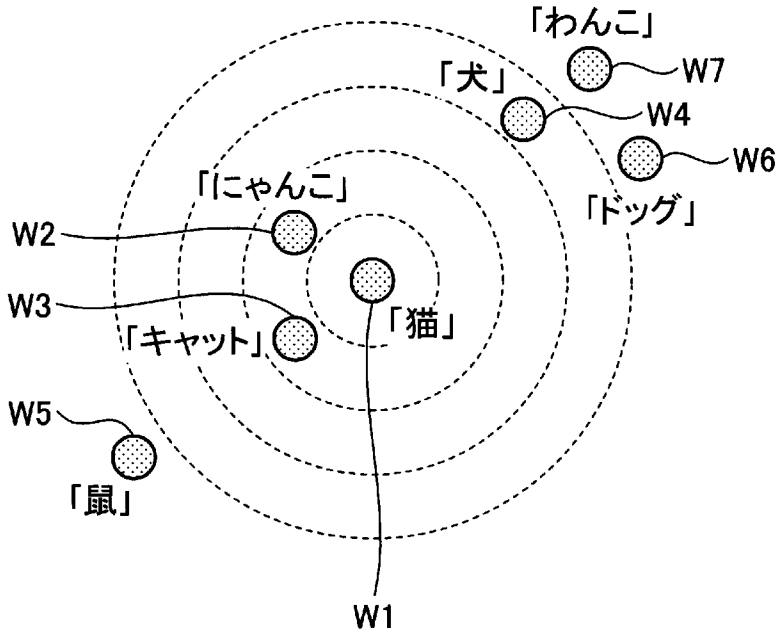
123a

	h_1	h_2	h_3	h_4	h_5	h_6	…
h_i(猫)	0	3	1	4	3	4	…
h_i(ごはん)	3	0	3	1	4	3	…
h_i(にゃんこ)	1	4	0	3	2	5	…
h_i(えさ)	2	1	2	0	5	2	…
h_i(花)	5	5	5	5	0	1	…
h_i(水)	4	2	4	2	1	0	…
：	：	：	：	：	：	：	

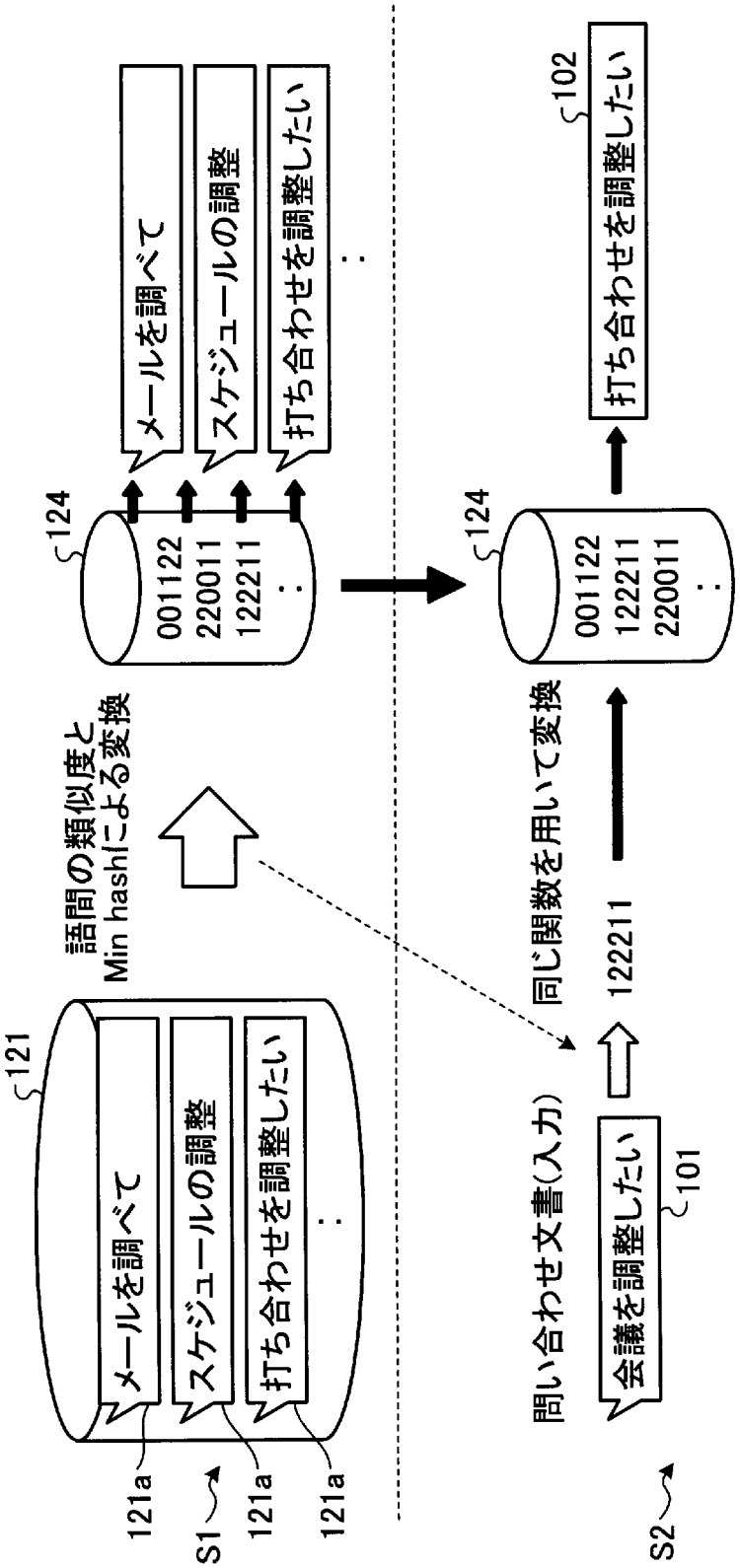
[図5]



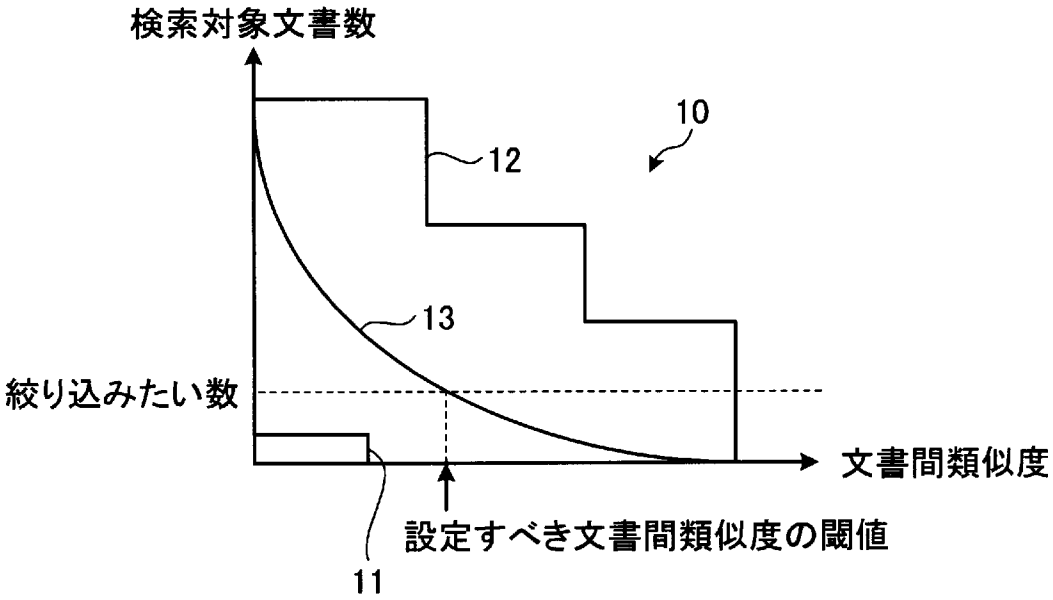
[図6]



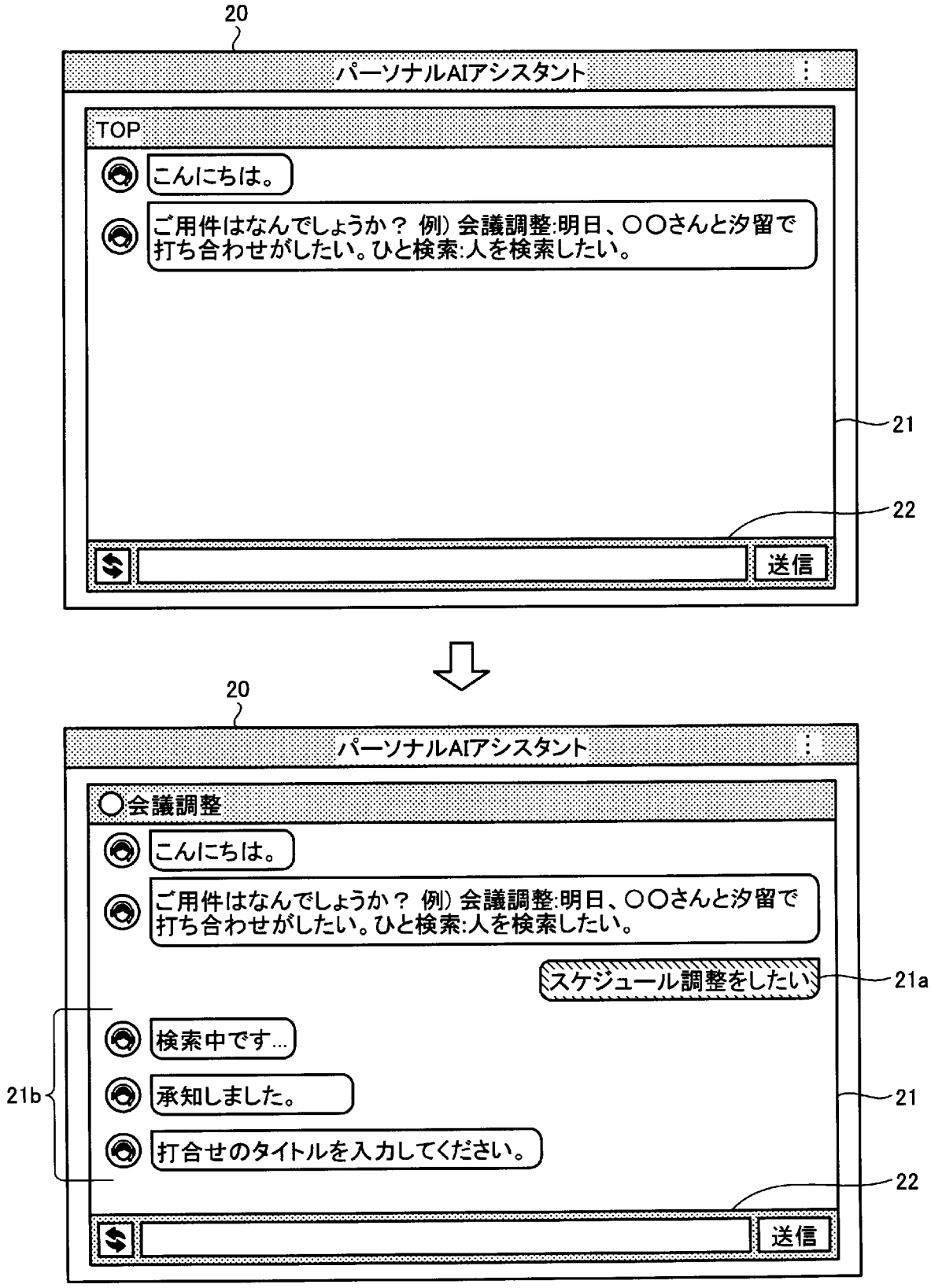
[図7]


121
メールを調べて
スケジュールの調整
打ち合わせを調整したい
121a
121a
121a
S1
話間の類似度と
Min hashによる変換
124
001122
220011
122211
124
001122
122211
220011
102
打ち合わせを調整したい
同じ関数を用いて変換
122211
問い合わせ文書(入力)
会議を調整したい
101
S2

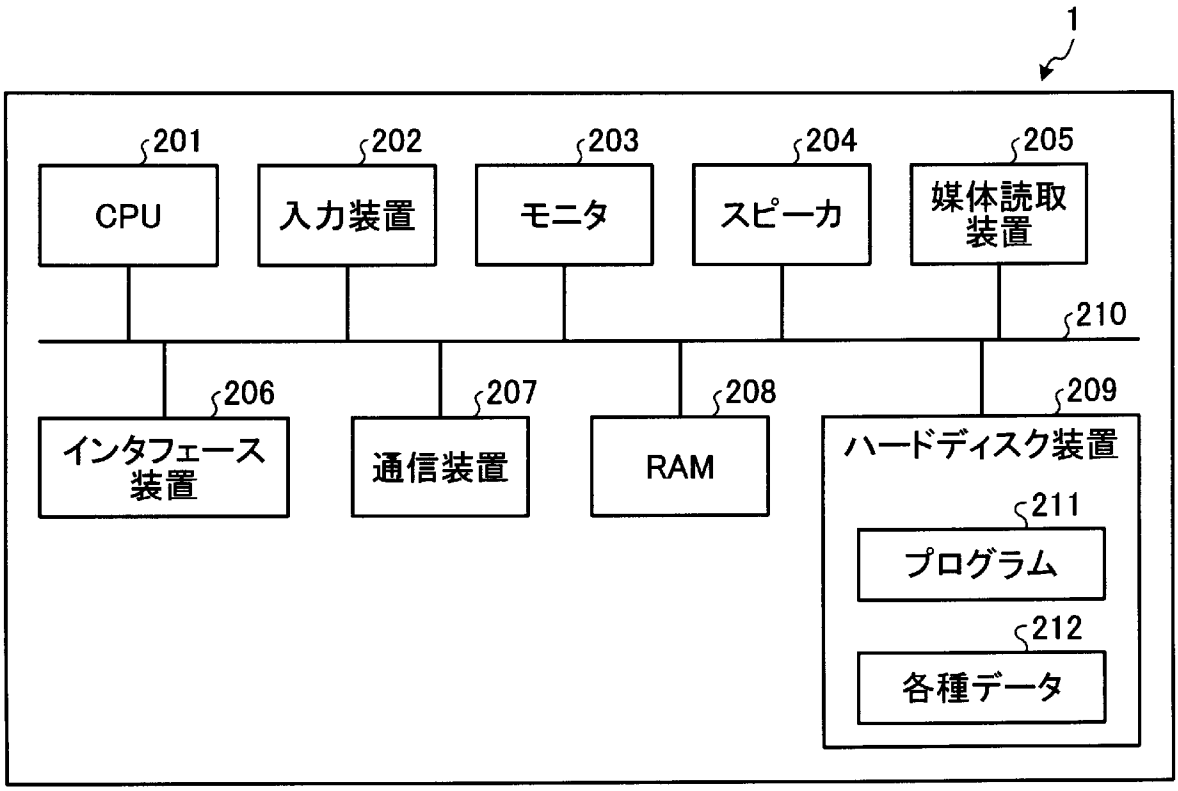
[図8]



[図9]


20
パーソナルAIアシスタント
TOP
こんにちは。
ご用件はなんでしょうか？ 例）会議調整:明日、○○さんと汐留で打ち合わせがしたい。ひと検索:人を検索したい。
21
22
送信
20
パーソナルAIアシスタント
○会議調整
こんにちは。
ご用件はなんでしょうか？ 例）会議調整:明日、○○さんと汐留で打ち合わせがしたい。ひと検索:人を検索したい。
スケジュール調整をしたい
21a
検索中です...
承知しました。
打合せのタイトルを入力してください。
21b
21
22
送信

[図10]


1
201
CPU
202
入力装置
203
モニタ
204
スピーカ
205
媒体読取装置
210
206
インタフェース装置
207
通信装置
208
RAM
209
ハードディスク装置
211
プログラム
212
各種データ

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/034306

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl. G06F16/383(2019.01)i, G06F16/31(2019.01)i, G06F16/90(2019.01)i, G06F16/93(2019.01)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. G06F16/383, G06F16/31, G06F16/90, G06F16/93

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	1971-2019
Registered utility model specifications of Japan	1996-2019
Published registered utility model applications of Japan	1994-2019

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.
A	JP 2010-267108 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 25 November 2010, entire text, all drawings (Family: none)	1-21
A	JP 2015-201042 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 12 November 2015, entire text, all drawings (Family: none)	1-21
A	US 2011/0087669 A1 (STRATIFY, INC.) 14 April 2011, entire text, all drawings (Family: none)	1-21

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

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"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)
"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
"T" 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
"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
"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
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 October 2019 (08.10.2019)	15 October 2019 (15.10.2019)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2019/034306

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106156154 A (ALIBABA GROUP HOLDING LTD.) 23 November 2016, entire text, all drawings (Family: none)	1-21
A	CN 107784110 A (BEIJING RUN TECHNOLOGY CO., LTD.) 09 March 2018, entire text, all drawings (Family: none)	1-21

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１９／０３４３０６

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F16/383(2019.01)i, G06F16/31(2019.01)i, G06F16/90(2019.01)i, G06F16/93(2019.01)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.　G06F16/383, G06F16/31, G06F16/90, G06F16/93

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	１９７１－２０１９年
日本国実用新案登録公報	１９９６－２０１９年
日本国登録実用新案公報	１９９４－２０１９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
A	JP 2010-267108 A（日本電信電話株式会社）2010.11.25, 全文、全図（ファミリーなし）	1-21
A	JP 2015-201042 A（日本電信電話株式会社）2015.11.12, 全文、全図（ファミリーなし）	1-21
A	US 2011/0087669 A1 (STRATIFY, INC) 2011.04.14, 全文、全図（ファミリーなし）	1-21

☑　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	国際調査報告の発送日
０８．１０．２０１９	１５．１０．２０１９

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
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）　５Ｎ　８３８４
松尾　真人
電話番号　０３－３５８１－１１０１　内線　３５８６

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	CN 106156154 A (ALIBABA GROUP HOLDING LTD) 2016.11.23, 全文、全図（ファミリーなし）	1-21
A	CN 107784110 A (BEIJING RUN TECHNOLOGY CO LTD) 2018.03.09, 全文、全図（ファミリーなし）	1-21