(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
2022年6月23日(23.06.2022)

WIPO | PCT

(10) 国際公開番号
WO 2022/130578 A1

(51) 国際特許分類:
G06F 16/35 (2019.01)

(21) 国際出願番号： PCT/JP2020/047218

(22) 国際出願日： 2020年12月17日(17.12.2020)

(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: 片江 伸之 (**KATAE, Nobuyuki**); 〒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: 真田 有，外 (**SANADA, Tamotsu** et al.); 〒1800004 東京都武蔵野市吉祥寺本町１丁目１０番３１号 ＮＭＦ吉祥寺本町ビル８階 Tokyo (JP).

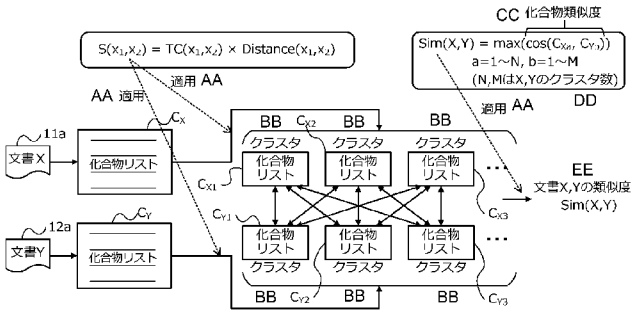
(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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：
— 国際調査報告（条約第21条(3)）

(54) **Title:** SIMILARITY DETERMINATION PROGRAM, SIMILARITY DETERMINATION DEVICE, AND SIMILARITY DETERMINATION METHOD

(54) 発明の名称： 類似度判定プログラム、類似度判定装置、及び、類似度判定方法



11a Document X
12a Document Y
C_X, C_{X1}, C_{X2}, C_{X3}, C_Y, C_{Y1}, C_{Y2}, C_{Y3} Compound list
AA Apply
BB Cluster
CC Compound similarity
DD N, M are respective numbers of clusters of X, Y
EE Similarity between documents X, Y

(57) **Abstract:** This similarity determination program causes a computer to execute processes for: acquiring a plurality of first groups (C_{X1}-C_{XN}) generated by classifying a plurality of first unique expressions on the basis of the positions of a plurality of respective first unique expressions included in a first document (X) and similarities of the plurality of respective unique expressions; acquiring a plurality of second groups (C_{Y1}-C_{YM}) generated by classifying a plurality of second unique expressions on the basis of the positions of a plurality of respective second unique expressions included in a second document (Y) and similarities of the plurality of respective second unique expressions; and determining a similarity (Sim) between the first document (X) and the second document (Y) on the basis of comparison between the plurality of first groups (C_{X1}-C_{XN}) and the plurality of second groups (C_{Y1}-C_{YM})

WO 2022/130578 A1

［続葉有］

(57) 要約：類似度判定プログラムは、第１の文書（Ｘ）に含まれる第１の複数の固有表現のそれぞれの位置と前記第１の複数の固有表現のそれぞれの類似度とに基づいて前記第１の複数の固有表現を分類することによって生成された第１の複数のグループ（C_{X1}～C_{XN}）を取得し、第２の文書（Ｙ）に含まれる第２の複数の固有表現のそれぞれの位置と前記第２の複数の固有表現のそれぞれの類似度とに基づいて前記第２の複数の固有表現を分類することによって生成された第２の複数のグループ（C_{Y1}～C_{YM}）を取得し、前記第１の複数のグループ（C_{X1}～C_{XN}）と前記第２の複数のグループ（C_{Y1}～C_{YM}）との比較に基づいて、前記第１の文書（Ｘ）と前記第２の文書（Ｙ）との類似度（Ｓｉｍ）を判定する、処理をコンピュータに実行させる。

明　細　書

発明の名称 ：

類似度判定プログラム、類似度判定装置、及び、類似度判定方法

技術分野

[0001]　本発明は、類似度判定プログラム、類似度判定装置、及び、類似度判定方法に関する。

背景技術

[0002]　複数文書を単語に分割し、単語の意味を表現するベクトル、及び、各単語の重みを算出して、ベクトル及び重みに基づき、各文書の文書ベクトルを算出することで、文書間の類似度を判定する手法が知られている。

先行技術文献

特許文献

[0003] 特許文献1：特開２００６－３３１２４５号公報

発明の概要

発明が解決しようとする課題

[0004]　文書内には様々な情報が記載されているため、文書全体を文書ベクトル化して文書間の類似度を判定すると、部分的に類似した文書間であっても類似度が低いと判定される可能性がある。

[0005]　１つの側面では、本発明は、部分的に類似した文書間の類似度の判定精度を向上させることを目的の１つとする。

課題を解決するための手段

[0006]　１つの側面では、類似度判定プログラムは、コンピュータに、以下の処理を実行させてよい。前記処理は、第１の文書に含まれる第１の複数の固有表現のそれぞれの位置と前記第１の複数の固有表現のそれぞれの類似度とに基づいて前記第１の複数の固有表現を分類することによって生成された第１の複数のグループを取得する処理を含んでよい。また、前記処理は、第２の文

書に含まれる第２の複数の固有表現のそれぞれの位置と前記第２の複数の固有表現のそれぞれの類似度とに基づいて前記第２の複数の固有表現を分類することによって生成された第２の複数のグループを取得する処理を含んでよい。さらに、前記処理は、前記第１の複数のグループと前記第２の複数のグループとの比較に基づいて、前記第１の文書と前記第２の文書との類似度を判定する処理を含んでよい。

発明の効果

[0007] １つの側面では、本発明は、部分的に類似した文書間の類似度の判定精度を向上させることができる。

図面の簡単な説明

[0008] [図1]比較例に係る類似度判定システムを説明するための図である。

[図2]図１に示す類似度判定システムによる化合物類似度の算出例を説明する図である。

[図3]第１実施形態に係る類似度判定システムを説明するための図である。

[図4]類似度判定システムの処理の一例を説明するための図である。

[図5]類似度判定システムの処理の一例を説明するための図である。

[図6]第１実施形態に係る類似度判定システムにおけるサーバの機能構成例を示すブロック図である。

[図7]サーバによる画面出力例を示す図である。

[図8]サーバの動作例を説明するフローチャートである。

[図9]サーバの機能を実現するコンピュータのハードウェア（ＨＷ）構成例を示すブロック図である。

[図10]第２実施形態に係る類似度判定システムを説明するための図である

[図11]第２実施形態に係る類似度判定システムにおけるサーバの機能構成例を示すブロック図である。

[図12]サーバによる画面出力例を示す図である。

[図13]サーバの文書ＤＢ（Database）構築処理の動作例を説明するフローチャートである。

[図14]サーバの文書検索処理の動作例を説明するフローチャートである。

[図15]第２実施形態の第１変形例及び第２変形例に係る類似度判定システムにおけるサーバの機能構成例を示すブロック図である。

[図16]サーバによる画面出力例を示す図である。

[図17]サーバによる画面出力例を示す図である。

[図18]第２実施例の第１変形例に係るサーバの動作例を説明するフローチャートである。

[図19]第２変形例に係るサーバの画面出力例を示す図である。

[図20]第３実施形態に係る類似度判定システムを説明するための図である。

[図21]第３実施形態に係る類似度判定システムの処理の一例を説明するための図である。

[図22]第３実施形態に係る類似度判定システムの処理の一例を説明するための図である。

[図23]第３実施形態に係る類似度判定システムにおけるサーバの機能構成例を示すブロック図である。

[図24]サーバの文書ＤＢ構築処理の動作例を説明するフローチャートである。

[図25]サーバの文書検索処理の動作例を説明するフローチャートである。

発明を実施するための形態

[0009] 以下、図面を参照して本発明の実施の形態を説明する。ただし、以下に説明する実施形態は、あくまでも例示であり、以下に明示しない種々の変形又は技術の適用を排除する意図はない。例えば、本実施形態を、その趣旨を逸脱しない範囲で種々変形して実施することができる。なお、以下の説明で用いる図面において、同一符号を付した部分は、特に断らない限り、同一若しくは同様の部分を表す。

[0010] 〔１〕第１実施形態

〔１－１〕比較例

上述したように、文書全体を文書ベクトル化して文書間の類似度を判定す

ると、部分的に類似した文書間であっても類似度が低いと判定される可能性がある。そこで、比較例として、文書中の固有表現に基づく類似度を判定することで文書間の類似度を判定する場合を説明する。比較例では、文書中の固有表現は、化合物名であるものとし、化合物名を含む化学分野の文書間の類似度を判定する場合を想定する。

[0011] 図１は、比較例に係る類似度判定システム１００を説明するための図である。図１に示すように、類似度判定システム１００は、クエリ文書（入力文書）の類似度の判定を要求するクエリ１０１と、１以上の比較対象文書を含む文書集合１０２とに基づき、固有表現に基づく類似度を算出する。

[0012] 例えば、類似度判定システム１００は、複数の文書、すなわち、クエリ１０１に含まれるクエリ文書及び文書集合１０２に含まれる比較対象文書のそれぞれから、固有表現として化合物名を抽出し（処理Ｐ１１０）、文書ごとに化合物リストを生成する。

[0013] 類似度判定システム１００は、文書ごとの化合物リストを比較することで、クエリ文書と比較対象文書の各々との間の化合物類似度を算出する（処理Ｐ１２０）。化合物類似度としては、例えば、コサイン類似度が挙げられる。

[0014] 類似度判定システム１００は、算出した化合物類似度に基づくランキング処理を行ない（処理Ｐ１３０）、クエリ文書との類似度が高い比較対象文書を類似度とともにランキング結果１０３として保存する。

[0015] 図２は、図１に示す類似度判定システム１００による化合物類似度の算出例を説明する図である。図２の例では、処理Ｐ１１０において、リチウムイオン電池に関するクエリ文書及び比較対象文書について、クエリ文書から化合物リストＣｘが生成され、比較対象文書から化合物リストＣｙが生成されたものとする。

[0016] 以下、化合物類似度の判定対象の文書の組について生成された化合物リストＣｘ及びＣｙを互いに区別しない場合には、単に「化合物リストＣ」と表記する。化合物リストＣは、化合物名と、当該化合物名の文書内での出現数

とを含んでよい。なお、図２の例において、化合物リストＣｘ及びＣｙ間で共通する共通化合物を下線太字で示す。

[0017] 類似度判定システム１００は、化合物リストＣに基づき、下記式（１）の演算により、化合物類似度としてのコサイン類似度を算出する。

[数1]

$$化合物類似度 = \cos(Cx, Cy) = \frac{\sum Cx_i Cy_i}{\sqrt{\sum Cx_i{}^2}+\sqrt{\sum Cy_i{}^2}} \quad （１）$$

[0018] 上記式（１）において、ｉは、化合物リストＣｘ及びＣｙに含まれる全ての化合物名を特定するためのインデックスであり、Cx_i及びCy_iは、化合物リストＣｘ及びＣｙ内のｉ番目の化合物名の出現数を示す。上記式（１）において、分母は、Ｃｘの化合物の出現数の２乗和の平方根と、Ｃｙの化合物の出現数の２乗和の平方根との和であり、分子は、ＣｘとＣｙとの間の共通化合物の出現数の積の総和である。

[0019] ところで、「文書」には、複数の要素についての記載を含む文書、一例として、複数の構成要素を備える装置、システム又は製造方法等について記載された特許文献又は論文等の文書がある。例えば、図２に示すリチウムイオン電池に関する文書には、「正極活物質」、「負極活物質」、「バインダー」、「電解質」、「電解液溶媒」等の、リチウムイオン電池の構成要素のそれぞれの区分（グループ）に関する化合物名が混在して記載されることがある。

[0020] このため、クエリ文書に記載された所定の要素に着目して比較対象文書との類似度を判定したい場合であっても、その他の要素、換言すれば調査対象ではない要素についての比較対象文書との差異が、文書間の類似度の判定結果に影響を与える場合がある。

[0021] 図２の例では、調査対象の要素が「正極活物質」である場合、「LiCoO2」等の「正極活物質」に関する化合物名が文書間で共通して出現する一方、他の要素に関する化合物名が文書間で相違するため、文書間の化合物類似度が低い値として算出される場合がある。

[0022] このように、調査対象の要素が文書間で類似する場合であっても、文書間の類似度が低いと判定される場合がある。

[0023] 〔１－２〕第１実施形態の説明

そこで、第１実施形態に係る類似度判定システム１は、文書に含まれる複数の固有表現のそれぞれを分類して複数のグループを生成し、文書間で、グループどうしの比較を行なうことにより、クエリ文書と比較対象文書との類似度を判定する。

[0024] 図３は、第１実施形態に係る類似度判定システム１を説明するための図であり、図４及び図５は、類似度判定システム１の処理の一例を説明するための図である。

[0025] 図３に示すように、第１実施形態に係る類似度判定システム１は、クエリ文書（入力文書）の類似度の判定を要求するクエリ１１と、判定対象となる１以上の比較対象文書を含む文書集合（文書群）１２とに基づき、固有表現に基づく類似度を算出する。

[0026] 例えば、類似度判定システム１は、比較例と同様に、複数の文書のそれぞれから、固有表現の一例としての化合物名を抽出し（処理Ｐ１）、文書ごとに、固有表現リスト、例えば化合物リストを生成する。

[0027] 図４及び図５の例では、類似度判定システム１は、クエリ１１に含まれるクエリ文書１１ａ（「文書Ｘ」と表記）から化合物名を抽出して化合物リストC_Xを生成する。また、類似度判定システム１は、文書集合１２に含まれる比較対象文書１２ａ（「文書Ｙ」と表記）から化合物名を抽出して化合物リストC_Yを生成する。クエリ文書１１ａは、第１の文書の一例であり、比較対象文書１２ａは、第２の文書の一例である。

[0028] 第１実施形態では、クエリ文書１１ａ及び比較対象文書１２ａは、リチウムイオン電池に関する文書であるものとする。以下、化合物類似度の判定対象の文書の組について生成された化合物リストC_X及びC_Yを互いに区別しない場合には、単に「化合物リストＣ」と表記する。

[0029] 第１実施形態に係る類似度判定システム１は、固有表現リストに基づき、

固有表現を分類及びグループ化するクラスタリングを実行する（図３の処理Ｐ２）。クラスタリングの手法としては、例えば、最短距離法等の既存の種々の手法が用いられてよい。

[0030] 次いで、類似度判定システム１は、クラスタごとの固有表現類似度、例えば化合物類似度を算出する（図３の処理Ｐ３）。固有表現類似度は、例えば、文書間のクラスタのペアごとに算出されてよい。

[0031] そして、類似度判定システム１は、固有表現類似度に基づき、クエリ文書１１ａとの類似度に応じて複数の比較対象文書１２ａの各々をランキング付けするランキング処理を行ない（図３の処理Ｐ４）、結果１３を出力する。結果１３は、ランキング結果を含んでもよい。

[0032] 以下、クラスタリング処理（処理Ｐ２）、固有表現類似度算出処理（処理Ｐ３）、及び、ランキング処理（処理Ｐ４）のそれぞれの一例を説明する。

[0033] （クラスタリング処理の一例）

図３の処理Ｐ２において、類似度判定システム１は、固有表現リストに基づき、固有表現リストに含まれる固有表現間の類似度スコアＳを、固有表現のペア（組）ごとに算出してよい。例えば、類似度判定システム１は、固有表現のペアについて、固有表現のそれぞれの位置と、固有表現間の類似度とに基づき類似度スコアＳを算出する。

[0034] 一例として、固有表現のペアを化合物x_1及びx_2と表記した場合、類似度判定システム１は、下記式（２）を用いて類似度スコアＳ（x_1，x_2）を算出してよい。

[数2]

$$\text{類似度スコア } S(x_1, x_2) = \mathrm{TC}(x_1, x_2) \times \mathrm{Distance}(x_1, x_2) \qquad (2)$$

[0035] 上記式（２）において、TC（x_1，x_2）は、MACCS KeyのTanimoto係数である。MACCS Keyは、化合物の特徴の表現手法（化合物記述子）の１つであり、Tanimoto係数は、MACCS Keyを用いて化合物間の構造類似度を示す指標の１つであり、固有表現が化合物名である場合の固有表現間の類似度の一例である。また、Distance（x_1，x_2）は、例えば、文書内での固有表現のそれぞれ

の出現位置の近さを数値化した値であり、一例として、以下の条件に応じた値である。

・化合物x_1及びx_2が同一文に出現し、且つ、並列関係にある場合：“1.0”

・化合物x_1及びx_2が同一文に出現する場合：“0.8”

・化合物x_1及びx_2が同一段落に出現する場合：“0.5”

・上記以外の場合：“0.1”

[0036] 類似度判定システム１は、化合物リストＣに含まれる複数の化合物名について、化合物名のペア（x_1，x_2）の組み合わせごとに上記式（２）を適用して、各ペア（x_1，x_2）の類似度スコアＳ（x_1，x_2）を算出してよい。

[0037] 類似度判定システム１は、算出した複数の類似度スコアＳ（x_1，x_2）に対して、例えば最短距離法等の手法を適用して、化合物リストＣに含まれる複数の化合物名を分類してグループ化することで、化合物名のクラスタリングを行なってよい。

[0038] 図４の例では、類似度判定システム１は、化合物リストC_Xに対するクラスタリングにより、化合物リストC_X内の化合物名をＮ（Ｎは２以上の整数）個のクラスタ（グループ）に分割し、クラスタごとの化合物リストC_{X1}～C_{XN}を生成する。また、類似度判定システム１は、化合物リストC_Yに対するクラスタリングにより、化合物リストC_Y内の化合物名をＭ（Ｍは２以上の整数）個のクラスタ（グループ）に分割し、クラスタごとの化合物リストC_{Y1}～C_{YM}を生成する。

[0039] 図５の例では、類似度判定システム１は、化合物リストC_X及びC_Yをそれぞれ４つのクラスタに分類し（Ｎ＝Ｍ＝４）、化合物リストC_{X1}～C_{X4}及びC_{Y1}～C_{Y4}を生成する。このようなクラスタリングにより、結果的に、化合物リストC_X及びC_Yを、以下のような４つの要素（特性）のクラスタに分類することができる。

・化合物リストC_{X1}及びC_{Y1}：

「正極活物質」の要素（特性）を有するクラスタ。

・化合物リストC_{X2}及びC_{Y2}：

「負極活物質」の要素（特性）を有するクラスタ。

・化合物リストC_{X3}及びC_{Y3}：

「バインダー」の要素（特性）を有するクラスタ。

・化合物リストC_{X4}及びC_{Y4}：

「電解液溶媒」の要素（特性）を有するクラスタ。

[0040] なお、ここまで、構造類似度としてMACCS KeyのTanimoto係数が用いられるものとして説明したが、これに限定されるものではない。例えば、化合物の特徴の表現手法としては、MACCS Key、換言すればMACCSフィンガープリントに限定されるものではなく、例えば、Morganフィンガープリント等の種々の化合物記述子が採用されてもよい。また、化合物間の構造類似度を示す指標としては、Tanimoto係数に限定されるものではなく、例えば、Dice係数等の種々の係数が用いられてもよい。

[0041] また、上記式（２）では、類似度判定システム１は、類似度スコアＳ（x_1，x_2）として、固有表現の文書内の出現位置の近さを数値化した値と、固有表現の類似度との積を算出するものとしたが、これに限定されるものではない。

[0042] 一例として、類似度判定システム１は、下記式（３）を用いて類似度スコアＳ（x_1，x_2）を算出してもよい。

[数3]

$$\text{類似度スコア } S(x_1, x_2) = W \times \mathrm{TC}(x_1, x_2) + (1 - W) \times \mathrm{Distance}(x_1, x_2) \qquad (3)$$

[0043] 上記式（３）において、Wは重みである。Wとしては、例えば、固有表現のそれぞれの位置と、固有表現間の類似度とが均等に考慮されるように“0.5”等の値がユーザ等により適宜定義及び設定されてもよい。或いは、Wは、検索クエリ及び正解例（正解データ）を含む訓練データに基づく機械学習により、正解例が上位に検索されるような値になるように訓練されたモデルに基づき設定されてもよい。

[0044] 例えば、化学構造は類似していないが１つの構成要素で同様に用いられる化合物（同一文で併記される可能性が高い）は、上記式（２）を用いると類

似度が過少評価される可能性がある。これに対し、上記式（３）のように、固有表現の文書内の出現位置の近さを数値化した値と、固有表現の類似度との重み付き和に基づき類似度スコアを算出することで、化合物の類似度を正当に評価することができる。

[0045] 以上のように、類似度判定システム１は、クエリ文書１１ａに含まれる第１の複数の化合物名のそれぞれの位置と第１の複数の化合物名のそれぞれの類似度とに基づいて第１の複数の化合物名を分類することで、第１クラスタ群を生成する。また、類似度判定システム１は、比較対象文書１２ａに含まれる第２の複数の化合物名のそれぞれの位置と第２の複数の化合物名のそれぞれの類似度とに基づいて第２の複数の化合物名を分類することで、第２クラスタ群を生成する。第１クラスタ群は、第１の複数のグループの一例であり、第２クラスタ群は、第２の複数のグループの一例である。

[0046] （固有表現類似度算出処理の一例）

図３の処理Ｐ３において、例えば、類似度判定システム１は、クエリ文書１１ａから生成した第１クラスタ群内の第１の複数のクラスタと、比較対象文書１２ａから生成した第２クラスタ群内の第２の複数のクラスタとをそれぞれ比較してよい。そして、類似度判定システム１は、第１の複数のクラスタと第２の複数のクラスタとの間の全てのクラスタのペアについて、下記式（４）の演算により、化合物類似度、一例としてコサイン類似度を算出してよい。

[数4]

$$\text{化合物類似度} = \cos(C_{Xa}, C_{Yb}) = \frac{\sum C_{Xai} C_{Ybi}}{\sqrt{\sum {C_{Xai}}^2} + \sqrt{\sum {C_{Ybi}}^2}} \qquad (4)$$

[0047] 上記式（４）において、ａは、１～Ｎの整数であり、ｂは、１～Ｍの整数である。ｉは、化合物リストC_{Xa}及びC_{Yb}に含まれる全ての化合物名を特定するためのインデックスであり、C_{Xai}及びC_{Ybi}は、化合物リストC_{Xa}及びC_{Yb}内のｉ番目の化合物名の出現数を示す。コサイン類似度の算出式は、上記式（１）と同様である。

[0048] 図４に示す化合物リストC_X及びC_Yの例では、類似度判定システム１は、化合物リストC_{X1}、C_{X2}、C_{X3}、・・・C_{XN}と、化合物リストC_{Y1}、C_{Y2}、C_{Y3}、・・・C_{YM}との全てのペア（組み合わせ）について、上記式（４）に従い化合物類似度を算出してよい。

[0049] （ランキング処理の一例）

例えば、類似度判定システム１は、ランキング処理において、クエリ文書１１ａと複数の比較対象文書１２ａの各々との間の文書類似度を取得し、文書類似度に基づき、クエリ文書１１ａとの類似度に応じた複数の比較対象文書１２ａのランキングを出力する。

[0050] 類似度判定システム１は、例えば、下記式（５）に従い、文書Ｘと１つの比較対象文書Ｙとの間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出してよい。

[数5]

$$\text{文書類似度 } \mathrm{Sim}(X,Y) = \max(\cos(C_{Xa}, C_{Yb})) \qquad (5)$$

[0051] 上記式（５）において、ｍａｘは、括弧内の全ての組み合わせのうちの最大の値を採用する関数である。

[0052] 類似度判定システム１は、上記式（５）に従い、上記式（４）で算出された固有表現類似度のうちの、値が最大となるクラスタのペア（ａ＝１～Ｎ、ｂ＝１～Ｍのいずれかの組み合わせ）を、文書Ｘ及びＹ間の文書類似度Ｓｉｍ（Ｘ，Ｙ）として採用してよい。

[0053] 図５の例では、類似度判定システム１は、化合物リストC_{X1}及びC_{Y1}のペア、換言すれば、「正極活物質」のクラスタどうしの化合物類似度が最大であると判定し、当該化合物類似度を文書Ｘ及びＹ間の文書類似度Ｓｉｍ（Ｘ，Ｙ）に決定する。

[0054] なお、図４及び図５並びに上記式（５）では、文書Ｘ（クエリ文書１１ａ）と、１つの文書Ｙ（比較対象文書１２ａ）との間の文書類似度を算出する例を示す。類似度判定システム１は、複数の比較対象文書１２ａ、例えば文書Y_1～Y_L（Ｌは２以上の整数であり、比較対象文書１２ａの文書数）それぞれについて上記処理を行ない、文書類似度Ｓｉｍ（Ｘ，Y_1）～Ｓｉｍ（Ｘ

，Y_L）を取得してよい。

[0055] そして、類似度判定システム１は、例えば、文書類似度Ｓｉｍ（Ｘ，Y_1）～Ｓｉｍ（Ｘ，Y_L）が高い文書Ｙから降順に検索対象の全文書Y_1～Y_Lをソートし、ソート結果を結果１３として出力してよい。結果１３には、ランク（順位）とともに文書Ｙの識別情報が含まれてよく、各文書Ｙの文書類似度Ｓｉｍ（Ｘ，Ｙ）が含まれてもよい。文書Ｙの識別情報には、文書番号又は文書コード等の識別子及び文書名等の書誌情報、並びに、要約及び所定の部分等の文書Ｙの少なくとも一部の内容、のうちの少なくとも１つを含んでもよい。

[0056] なお、類似度判定システム１は、結果１３として、特定の順位であると判定した文書Ｙ、例えば、クエリ文書１１ａとの間で最も文書類似度Ｓｉｍ（Ｘ，Ｙ）の高い文書Ｙの識別情報を出力してもよい。

[0057] 以上のように、第１実施形態に係る類似度判定システム１によれば、クラスタリング処理により分類したクラスタごとの固有表現類似度に基づき、文書間の類似度を判定することで、部分的に類似した文書間の類似度の判定精度を向上させることができる。

[0058] 〔１－３〕機能構成例

図６は、第１実施形態に係る類似度判定システム１におけるサーバ２の機能構成例を示すブロック図であり、図７は、サーバ２による画面出力例を示す図である。

[0059] サーバ２は、類似度判定装置、情報処理装置、又は、コンピュータの一例である。例えば、サーバ２は、類似度判定システム１において、図示しない端末装置、他のサーバ等との間で、クエリ文書１１ａ及び比較対象文書１２ａの受信、結果１３の送信等の種々の通信を行なってよい。

[0060] サーバ２は、例えば、端末装置に対して、アクセスを可能とするための機能を提供してよい。当該機能としては、例えば、端末装置によるアクセスに用いられる、ウェブページ等の画面の生成及び表示制御が挙げられる。例えば、端末装置は、ブラウザ等のアプリケーションを用いてサーバ２にアクセ

ス要求を送信し、サーバ２から受信する画面情報に基づきアプリケーションに表示されるウェブページを介して、サーバ２へのアクセスを行なってよい。例えば、サーバ２は、図７に示すように、クエリを指定するためのクエリ指定画面２１０、及び、判定結果を出力するための判定結果出力画面２４０の画面情報を出力してよい。

[0061] 類似度判定システム１による上述した類似度判定処理は、サーバ２により実現されてよい。図６に示すように、サーバ２は、例示的に、メモリ部２１、文書入力部２２、類似度算出部２３、及び、類似度出力部２４を備えてよい。メモリ部２１、文書入力部２２、類似度算出部２３、及び、類似度出力部２４は、制御部の一例である。

[0062] メモリ部２１は、類似度判定処理に係る種々のデータを記憶する記憶領域を有する。メモリ部２１は、例えば、図３に示すクエリ文書１１ａ、複数の比較対象文書１２ａ、結果１３等の情報を記憶してよい。また、メモリ部２１は、類似度判定処理における中間データとして、図４及び図５に示す文書ごとの化合物リストＣ、類似度スコアＳ、クラスタごとの化合物リストＣ、化合物類似度、文書類似度Ｓｉｍ等の情報を記憶してもよい。

[0063] 文書入力部２２は、図示しない端末装置又は他のサーバ等のコンピュータから、クエリ文書１１ａ及び比較対象文書１２ａの入力を受け付け、例えばメモリ部２１にＤＢ（Database）として蓄積してもよい。このように、文書入力部２２は、文書のＤＢを構築及び参照可能であってもよい。

[0064] また、文書入力部２２は、図示しない端末装置又は他のサーバ等のコンピュータから、類似判定要求に係るクエリ文書１１ａの入力を受け付け、メモリ部２１に格納してよい。クエリ文書１１ａは、例えばクエリ１１に含まれてもよい。

[0065] 文書入力部２２は、例えば、クエリ１１として、クエリ文書１１ａそのものではなく、クエリ文書１１ａの識別情報、例えば文書番号、文書コード等の情報を受け付けてもよい。この場合、文書入力部２２は、識別情報に基づき、例えばメモリ部２１のＤＢから、類似判定要求に係るクエリ文書１１ａ

を特定してよい。

[0066] 例えば、図７に示すように、文書入力部２２は、クエリ指定画面２１０の判定ボタン２１２が押下された際に入力欄２１１に設定されている文書番号を受け付けてよい。

[0067] 類似度算出部２３は、クエリ文書１１ａ及び比較対象文書１２ａの類似度を算出する。図６に例示するように、類似度算出部２３は、化合物名抽出部２３１、クラスタリング部２３２、及び、文書類似度算出部２３３を備えてよい。

[0068] 化合物名抽出部２３１は、メモリ部２１に格納されたクエリ文書１１ａ及び比較対象文書１２ａのそれぞれから化合物名を抽出し、化合物リストC_X及びC_Yを生成する。化合物名抽出部２３１の処理は、図３の処理Ｐ１の一例である。

[0069] クラスタリング部２３２は、化合物リストC_X及びC_Yのそれぞれに含まれる化合物名について類似度スコアＳを算出する。また、クラスタリング部２３２は、類似度スコアＳに基づき化合物名を複数のクラスタに分類して、化合物リストC_{X1}、C_{X2}、C_{X3}、・・・C_{XN}と、化合物リストC_{Y1}、C_{Y2}、C_{Y3}、・・・C_{YM}とを生成する。クラスタリング部２３２の処理は、図３の処理Ｐ２の一例である。

[0070] なお、クエリ文書１１ａ及び比較対象文書１２ａの一方又は双方が予めメモリ部２１に格納される場合、クラスタリング部２３２は、各文書について事前にクラスタリング処理を行ない、クラスタごとの化合物リストＣを生成しておいてもよい。一例として、クラスタリング部２３２は、メモリ部２１への文書の登録のタイミングで、登録される各文書についてクラスタリング処理を行なってもよい。

[0071] 文書類似度算出部２３３は、クラスタごとの化合物リストに基づき、クラスタごとの化合物類似度を算出し、文書内で最も化合物類似度の高いクラスタの化合物類似度を、当該文書の類似度Ｓｉｍ（Ｘ，Ｙ）として算出する。なお、文書類似度算出部２３３は、比較対象文書１２ａが複数（例えばＬ個

）存在する場合、比較対象文書１２ａごとの類似度Ｓｉｍ（Ｘ，$Ｙ_1$）～Ｓｉｍ（Ｘ，$Ｙ_L$）を算出してよい。文書類似度算出部２３３は、算出した類似度Ｓｉｍ（Ｘ，Ｙ）をメモリ部２１に格納してよい。

[0072] 類似度出力部２４は、類似度算出部２３が算出した類似度Ｓｉｍ（Ｘ，Ｙ）を出力する。なお、類似度出力部２４は、比較対象文書１２ａが複数（例えばＬ個）存在する場合、算出した類似度Ｓｉｍ（Ｘ，$Ｙ_1$）～Ｓｉｍ（Ｘ，$Ｙ_L$）が高い順に、比較対象文書１２ａ及び類似度Ｓｉｍ（Ｘ，Ｙ）の情報を出力してもよい。

[0073] 文書類似度算出部２３３及び類似度出力部２４の処理は、図３の処理Ｐ３及びＰ４の一例である。類似度出力部２４による出力には、例えば、図示しない端末装置等のコンピュータへの送信、メモリ部２１等のサーバ２の記憶領域への格納、等が含まれてよい。

[0074] 例えば、図７に示すように、類似度出力部２４は、判定結果出力画面２４０を出力してもよい。判定結果出力画面２４０には、クエリ文書１１ａの表示領域２４１と、比較対象文書１２ａの少なくとも１つ（図７では３つ）の表示領域２４５ａ～２４５ｃとが含まれてよい。表示領域２４１には、書誌情報及び要約等の表示領域２４２、及び、クエリ文書１１ａの全文を表示する画面に遷移するための全文参照ボタン２４３が含まれてよい。

[0075] 表示領域２４５ａ～２４５ｃには、書誌情報及び要約等の表示領域２４６ａ～２４６ｃ、全文参照ボタン２４７ａ～２４７ｃ、及び、類似すると判定されたクラスタの化合物リスト２４８ａ～２４８ｃが含まれてよい。なお、表示領域２４５ａ～２４５ｃには、類似度Ｓｉｍ（Ｘ，Ｙ）が表示されてよい。

[0076] このように、類似度出力部２４は、化合物リスト２４８ａ～２４８ｃの表示により、クエリ文書１１ａと比較対象文書１２ａとの間の類似度計算の結果、最も類似度が高いと判断されたクラスタに属する固有表現のリストをユーザに提示することができる。

[0077] 〔１－４〕動作例

図８は、サーバ２の動作例を説明するフローチャートである。図８に示すように、サーバ２は、クエリ文書１１ａに対する処理と、比較対象文書１２ａに対する処理とを互いに異なるタイミングで実施してもよい。

[0078] 図８に例示するように、文書入力部２２は、クエリ文書１１ａの入力を受け付ける（ステップＳ１）。化合物名抽出部２３１は、クエリ文書１１ａから化合物名を抽出する（ステップＳ２）。クラスタリング部２３２は、抽出した化合物名をクラスタリングする（ステップＳ３）。

[0079] また、文書入力部２２は、比較対象文書１２ａの入力を受け付ける（ステップＳ４）。化合物名抽出部２３１は、未選択の比較対象文書１２ａを選択し（ステップＳ５）、選択した比較対象文書１２ａから化合物名を抽出する（ステップＳ６）。クラスタリング部２３２は、抽出した化合物名をクラスタリングする（ステップＳ７）。

[0080] 文書類似度算出部２３３は、クエリ文書１１ａ及び比較対象文書１２ａの化合物クラスタを比較し、当該文書間の類似度Ｓｉｍを算出し（ステップＳ８）、メモリ部２１に格納する（ステップＳ９）。

[0081] 文書類似度算出部２３３は、未選択の比較対象文書１２ａがあるか否かを判定し（ステップＳ１０）、あると判定した場合（ステップＳ１０でＹＥＳ）、処理がステップＳ５に移行する。

[0082] 未選択の比較対象文書１２ａがないと文書類似度算出部２３３が判定した場合（ステップＳ１０でＮＯ）、類似度Ｓｉｍ（Ｘ，Ｙ）が高い順に、比較対象文書１２ａ及びその類似度Ｓｉｍ（Ｘ，Ｙ）を出力し（ステップＳ１１）、処理が終了する。

[0083] 〔１－５〕ハードウェア構成例

サーバ２は、仮想サーバ（ＶＭ；Virtual Machine）であってもよいし、物理サーバであってもよい。また、サーバ２の機能は、１台のコンピュータにより実現されてもよいし、２台以上のコンピュータにより実現されてもよい。さらに、サーバ２の機能のうちの少なくとも一部は、クラウド環境により提供されるＨＷ（Hardware）リソース及びＮＷ（Network）リソースを用いて

実現されてもよい。

[0084] 図９は、サーバ２の機能を実現するコンピュータ１０のハードウェア（ＨＷ）構成例を示すブロック図である。サーバ２の機能を実現するＨＷリソースとして、複数のコンピュータが用いられる場合は、各コンピュータが図９に例示するＨＷ構成を備えてよい。

[0085] 図９に示すように、コンピュータ１０は、ＨＷ構成として、例示的に、プロセッサ１０ａ、メモリ１０ｂ、記憶部１０ｃ、ＩＦ（Interface）部１０ｄ、Ｉ／Ｏ（Input / Output）部１０ｅ、及び読取部１０ｆを備えてよい。

[0086] プロセッサ１０ａは、種々の制御や演算を行なう演算処理装置の一例である。プロセッサ１０ａは、コンピュータ１０内の各ブロックとバス１０ｉで相互に通信可能に接続されてよい。なお、プロセッサ１０ａは、複数のプロセッサを含むマルチプロセッサであってもよいし、複数のプロセッサコアを有するマルチコアプロセッサであってもよく、或いは、マルチコアプロセッサを複数有する構成であってもよい。

[0087] プロセッサ１０ａとしては、例えば、ＣＰＵ、ＭＰＵ、ＧＰＵ、ＡＰＵ、ＤＳＰ、ＡＳＩＣ、ＦＰＧＡ等の集積回路（ＩＣ；Integrated Circuit）が挙げられる。なお、プロセッサ１０ａとして、これらの集積回路の２以上の組み合わせが用いられてもよい。ＣＰＵはCentral Processing Unitの略称であり、ＭＰＵはMicro Processing Unitの略称である。ＧＰＵはGraphics Processing Unitの略称であり、ＡＰＵはAccelerated Processing Unitの略称である。ＤＳＰはDigital Signal Processorの略称であり、ＡＳＩＣはApplication Specific ICの略称であり、ＦＰＧＡはField-Programmable Gate Arrayの略称である。

[0088] メモリ１０ｂは、種々のデータやプログラム等の情報を格納するＨＷの一例である。メモリ１０ｂとしては、例えばＤＲＡＭ（Dynamic Random Access Memory）等の揮発性メモリ、及び、ＰＭ（Persistent Memory）等の不揮発性メモリ、の一方又は双方が挙げられる。

[0089] 記憶部１０ｃは、種々のデータやプログラム等の情報を格納するＨＷの一

例である。記憶部１０ｃとしては、ＨＤＤ（Hard Disk Drive）等の磁気ディスク装置、ＳＳＤ（Solid State Drive）等の半導体ドライブ装置、不揮発性メモリ等の各種記憶装置が挙げられる。不揮発性メモリとしては、例えば、フラッシュメモリ、ＳＣＭ（Storage Class Memory）、ＲＯＭ（Read Only Memory）等が挙げられる。

[0090] また、記憶部１０ｃは、コンピュータ１０の各種機能の全部若しくは一部を実現するプログラム１０ｇ（類似度判定プログラム）を格納してよい。例えば、サーバ２のプロセッサ１０ａは、記憶部１０ｃに格納されたプログラム１０ｇをメモリ１０ｂに展開して実行することにより、図６に例示するサーバ２としての機能を実現できる。

[0091] 図６に示すメモリ部２１は、メモリ１０ｂ及び記憶部１０ｃの一方又は双方の記憶領域により実現されてよい。

[0092] ＩＦ部１０ｄは、ネットワークとの間の接続及び通信の制御等を行なう通信ＩＦの一例である。例えば、ＩＦ部１０ｄは、イーサネット（登録商標）等のＬＡＮ（Local Area Network）、或いは、ＦＣ（Fibre Channel）等の光通信等に準拠したアダプタを含んでよい。当該アダプタは、無線及び有線の一方又は双方の通信方式に対応してよい。例えば、サーバ２は、ＩＦ部１０ｄを介して、端末装置及び他のサーバのそれぞれと相互に通信可能に接続されてよい。また、例えば、プログラム１０ｇは、当該通信ＩＦを介して、ネットワークからコンピュータ１０にダウンロードされ、記憶部１０ｃに格納されてもよい。

[0093] Ｉ／Ｏ部１０ｅは、入力装置、及び、出力装置、の一方又は双方を含んでよい。入力装置としては、例えば、キーボード、マウス、タッチパネル等が挙げられる。出力装置としては、例えば、モニタ、プロジェクタ、プリンタ等が挙げられる。

[0094] 読取部１０ｆは、記録媒体１０ｈに記録されたデータやプログラムの情報を読み出すリーダの一例である。読取部１０ｆは、記録媒体１０ｈを接続可能又は挿入可能な接続端子又は装置を含んでよい。読取部１０ｆとしては、

例えば、ＵＳＢ（Universal Serial Bus）等に準拠したアダプタ、記録ディスクへのアクセスを行なうドライブ装置、ＳＤカード等のフラッシュメモリへのアクセスを行なうカードリーダ等が挙げられる。なお、記録媒体１０ｈにはプログラム１０ｇが格納されてもよく、読取部１０ｆが記録媒体１０ｈからプログラム１０ｇを読み出して記憶部１０ｃに格納してもよい。

[0095] 記録媒体１０ｈとしては、例示的に、磁気／光ディスクやフラッシュメモリ等の非一時的なコンピュータ読取可能な記録媒体が挙げられる。磁気／光ディスクとしては、例示的に、フレキシブルディスク、ＣＤ（Compact Disc）、ＤＶＤ（Digital Versatile Disc）、ブルーレイディスク、ＨＶＤ（Holographic Versatile Disc）等が挙げられる。フラッシュメモリとしては、例示的に、ＵＳＢメモリやＳＤカード等の半導体メモリが挙げられる。

[0096] 上述したコンピュータ１０のＨＷ構成は例示である。従って、コンピュータ１０内でのＨＷの増減（例えば任意のブロックの追加や削除）、分割、任意の組み合わせでの統合、又は、バスの追加若しくは削除等は適宜行なわれてもよい。例えば、サーバ２において、Ｉ／Ｏ部１０ｅ及び読取部１０ｆの少なくとも一方は、省略されてもよい。

[0097] 〔２〕第２実施形態

〔２－１〕第２実施形態の説明

次に、第２実施形態について説明する。第２実施形態では、第１実施形態に係る固有表現に基づく類似度に加えて、単語の意味ベクトルに基づく類似度を用いて、文書間の類似度を判定する手法を説明する。

[0098] なお、以下の第２実施形態の説明では、特に言及しない構成、処理又は機能は、既述の第１実施形態に係る構成、処理又は機能と同様であるものとする。

[0099] 図１０は、第２実施形態に係る類似度判定システム１Ａを説明するための図である。図１０に示すように、第２実施形態に係る類似度判定システム１Ａにおいて、クエリ１１及び文書集合１２に基づく処理Ｐ１～Ｐ３については、第１実施形態と同様である。

[0100] 図１０に例示するように、類似度判定システム１Ａにおいて、処理Ｐ１～Ｐ３と少なくとも一部の処理が並行又は前後して、処理Ｐ５～Ｐ８が実行されてよい。また、処理Ｐ３及びＰ８の結果に基づき、処理Ｐ９が実行され、結果１４が出力されてよい。以下、処理Ｐ５～Ｐ９について説明する。

[0101] 例えば、類似度判定システム１Ａは、複数の文書、例えばクエリ文書１１ａ及び複数の比較対象文書１２ａのそれぞれから、例えば形態素解析により単語を抽出する（処理Ｐ５）。

[0102] 類似度判定システム１Ａは、処理Ｐ５で得られた単語に基づき、複数の文書のそれぞれについて、統計的に単語の重みを算出する（処理Ｐ６）。例えば、類似度判定システム１Ａは、ｔｆ－ｉｄｆ（Term Frequency - Inverse Document Frequency）等の評価手法を用いて、文書内での単語の重要度を重みとして評価してよい。

[0103] また、類似度判定システム１Ａは、処理Ｐ６と少なくとも一部の処理が並行又は前後して、処理Ｐ７を実行する。例えば、類似度判定システム１Ａは、処理Ｐ５で得られた単語に基づき、複数の文書のそれぞれについて、単語ベクトルを算出する（処理Ｐ７）。単語ベクトルは、単語埋め込みベクトル又は意味ベクトルと称されてもよい。

[0104] 例えば、類似度判定システム１Ａは、単語の意味を表現するベクトルが格納されたベクトルデータベースを検索して単語ベクトルを取得してよい。一例として、類似度判定システム１Ａは、訓練済みのモデルに基づき、処理Ｐ５で得られた単語のそれぞれに対応する単語ベクトルを取得してよい。

[0105] 類似度判定システム１Ａは、文書ごとに、処理Ｐ７で取得した単語ベクトルと、処理Ｐ６で取得した単語の重みとを乗じた結果を文書内の全単語に亘って加算することで文書ベクトルを算出する。そして、類似度判定システム１Ａは、クエリ文書１１ａの文書ベクトル（第１のベクトル）と、比較対象文書１２ａの各々の文書ベクトル（第２のベクトル）との間の類似度を算出する。換言すれば、類似度判定システム１Ａは、単語の意味ベクトルに基づく、クエリ文書１１ａと比較対象文書１２ａとの間のテキスト類似度を算出

する（処理Ｐ８）。

[0106] 例えば、類似度判定システム１Ａは、クエリ文書１１ａと比較対象文書１２ａとについて、下記式（６）の演算により、テキスト類似度、一例としてコサイン類似度を算出してよい。

[数6]

$$\text{テキスト類似度} = \cos(\sum W_X, \sum W_Y) \qquad (6)$$

[0107] 上記式（６）において、W_Xは、文書Ｘに含まれる単語の分散ベクトルであり、W_Yは、文書Ｙに含まれる単語の分散ベクトルである。

[0108] 図４に示す化合物リストC_X及びC_Yの例では、類似度判定システム１Ａは、化合物リストC_{X1}、C_{X2}、C_{X3}、・・・C_{XN}と、化合物リストC_{Y1}、C_{Y2}、C_{Y3}、・・・C_{YM}との全てのペアについて、上記式（６）に従いテキスト類似度を算出してよい。

[0109] そして、類似度判定システム１Ａは、テキスト類似度及び固有表現類似度に基づき、クエリ文書１１ａとの類似度に応じて複数の比較対象文書１２ａの各々をランキング付けするランキング処理を行ない（処理Ｐ９）、結果１４を出力する。結果１４は、ランキング結果を含んでもよい。

[0110] 例えば、類似度判定システム１Ａは、ランキング処理において、テキスト類似度と固有表現類似度を統合した類似度を算出し、当該類似度に基づき、クエリ文書１１ａとの類似度に応じた複数の比較対象文書１２ａのランキングを出力する。

[0111] 類似度判定システム１Ａは、例えば、下記式（７）に従い、文書Ｘと１つの比較対象文書Ｙとの間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出してよい。

[数7]

$$\text{文書類似度 } \mathrm{Sim}(X,Y) = \max(\mathrm{fc}(C_{Xa}, C_{Yb}) \times \mathrm{ft}(X,Y)) \qquad (7)$$

[0112] 上記式（７）において、ｍａｘは、括弧内の全ての組み合わせのうちの最大の値を採用する関数である。ｆｃ及びｆｔは、それぞれ、下記式（８）及び（９）に示すように、固有表現類似度及びテキスト類似度である。

[数8]

$$\mathrm{fc}(C_{Xa}, C_{Yb}) = \cos(C_{Xa}, C_{Yb}) \quad (8)$$

$$\mathrm{ft}(X, Y) = \cos(\sum W_X, \sum W_Y) \quad (9)$$

[0113] 類似度判定システム１Ａは、上記式（７）～（９）に従い、上記式（４）で算出された固有表現類似度と、上記（６）で算出されたテキスト類似度に基づいて、文書Ｘ及びＹ間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を取得してよい。

[0114] なお、上記式（７）では、文書Ｘ（クエリ文書１１ａ）と、１つの文書Ｙ（比較対象文書１２ａ）との間の文書類似度を算出する例を示す。類似度判定システム１Ａは、第１実施形態と同様に、文書Ｙの数に応じた文書類似度Ｓｉｍ（Ｘ，$Ｙ_1$）～Ｓｉｍ（Ｘ，$Ｙ_L$）を取得してよい。

[0115] そして、類似度判定システム１Ａは、例えば、第１実施形態と同様に、文書類似度Ｓｉｍ（Ｘ，$Ｙ_1$）～Ｓｉｍ（Ｘ，$Ｙ_L$）が高い文書Ｙから降順に検索対象の全文書$Ｙ_1$～$Ｙ_L$をソートすることで、ランキング処理を行なう（処理Ｐ９）。また、類似度判定システム１Ａは、ソート結果を結果１４として出力してよい。結果１４の内容及び出力手法は、第１実施形態に係る結果１３と同様である。

[0116] なお、文書類似度Ｓｉｍ（Ｘ，Ｙ）は、上記式（７）に基づき算出されるものとしたが、これに限定されるものではない。例えば、類似度判定システム１Ａは、文書類似度Ｓｉｍ（Ｘ，Ｙ）は、下記式（１０）に従い、文書Ｘと１つの比較対象文書Ｙとの間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を、固有表現似度とテキスト類似度との重み付き和として算出してもよい。

[数9]

$$\text{文書類似度 } \mathrm{Sim}(X, Y) = \max(w \times \mathrm{fc}(C_{Xa}, C_{Yb}) + (1 - w) \times \mathrm{ft}(X, Y)) \quad (10)$$

[0117] 上記式（１０）において、ｗは重みである。ｗとしては、例えば、固有表現類似度とテキスト類似度とが均等に考慮されるように“0.5”等の値がユーザ等により適宜定義及び設定されてもよい。或いは、ｗは、検索クエリ及び正解例（正解データ）を含む訓練データに基づく機械学習により、正解例が上位に検索されるような値になるように訓練されたモデルに基づき設定され

てもよい。

[0118] 以上のように、第２実施形態に係る類似度判定システム１Ａによれば、第１実施形態と同様の効果を奏することができる。また、類似度判定システム１Ａによれば、固有表現類似度に加えて、意味ベクトルに基づく類似度に基づき、文書間の類似度を判定することで、文書間の類似度の判定精度をより向上させることができる。

[0119] 〔２－２〕機能構成例

図１１は、第２実施形態に係る類似度判定システム１Ａにおけるサーバ３の機能構成例を示すブロック図であり、図１２は、サーバ３による画面出力例を示す図である。

[0120] サーバ３は、類似度判定装置、情報処理装置、又は、コンピュータの一例である。例えば、サーバ３は、類似度判定システム１Ａにおいて、図示しない端末装置、他のサーバ等との間で、クエリ文書１１ａ及び比較対象文書１２ａの受信、結果１４の送信等の種々の通信を行なってよい。

[0121] サーバ３は、サーバ２と同様に、例えば、端末装置に対して、アクセスを可能とするための機能を提供してよい。例えば、サーバ３は、図１２に示すように、検索クエリを指定するための検索クエリ指定画面３３０、及び、検索結果を出力するための検索結果出力画面３４０の画面情報を出力してよい。

[0122] 類似度判定システム１Ａによる上述した類似度判定処理は、サーバ３により実現されてよい。図１１に示すように、サーバ３は、例示的に、文書ＤＢ部３１、及び、文書検索部３２を備えてよい。文書ＤＢ部３１、及び、文書検索部３２は、制御部の一例である。なお、サーバ３は、図６に示す文書入力部２２を備えてもよい。

[0123] 文書ＤＢ部３１は、クエリ文書１１ａ及び比較対象文書１２ａを蓄積し、文書ＤＢを構築する文書ＤＢ構築処理を行なう。

[0124] 文書検索部３２は、クエリ１１の受け付けに応じて、文書ＤＢ部３１が記憶する情報に基づき、クエリ１１で指定されたクエリ文書１１ａと類似する

比較対象文書１２ａを検索する文書検索処理を行なう。文書検索処理は、類似度判定処理を含む処理であり、類似判定処理の利用例（応用例）である。

[0125] （文書ＤＢ部３１の説明）

図１１に示すように、文書ＤＢ部３１は、例示的に、文書蓄積部３１１、化合物名抽出部３１２、クラスタリング部３１３、化合物クラスタ蓄積部３１４、文書ベクトル算出部３１５、及び、文書ベクトル蓄積部３１６を備えてよい。

[0126] 文書蓄積部３１１は、第１実施形態に係るメモリ部２１（図６参照）の一例であり、複数の文書を蓄積する。文書は、クエリ文書１１ａ及び比較対象文書１２ａのいずれとしても用いられ得る文書である。従って、文書蓄積部３１１は、クエリ文書１１ａと、クエリ１１の対象となる複数の比較対象文書１２ａを含む文書集合（文書群）１２とを蓄積するといえる。文書蓄積部３１１は、クエリ１１の受付前に、予め複数の文書を蓄積してよい。なお、文書蓄積部３１１には、第１実施形態に係る文書入力部２２が受け付けた複数の文書が格納されてもよい。

[0127] 化合物名抽出部３１２は、図６に示す化合物名抽出部２３１の一例であり、文書蓄積部３１１が蓄積する複数の文書の各々から、固有表現の一例としての化合物名を抽出し、文書ごとの化合物リストC_X及びC_Yを生成する。化合物名抽出部３１２の処理は、図１０の処理Ｐ１の一例である。

[0128] クラスタリング部３１３は、図６に示すクラスタリング部２３２の一例である。クラスタリング部３１３は、文書ごとの化合物リストC_X及びC_Yに基づき、文書ごとに、化合物名の複数の化合物クラスタ、換言すれば、複数の化合物リストC_{X1}、C_{X2}、C_{X3}、・・・C_{XN}、C_{Y1}、C_{Y2}、C_{Y3}、・・・C_{YM}を生成する。クラスタリング部３１３の処理は、図１０の処理Ｐ２の一例である。

[0129] 化合物クラスタ蓄積部３１４は、図６に示すメモリ部２１の一例であり、クラスタリング部３１３が生成した複数の化合物クラスタ、換言すれば、複数の化合物リストC_{X1}、C_{X2}、C_{X3}、・・・C_{XN}、C_{Y1}、C_{Y2}、C_{Y3}、・

・・C_{YM}を蓄積する。

[0130] 文書ベクトル算出部３１５は、文書蓄積部３１１が蓄積する複数の文書の各々から単語を抽出し、単語に基づき単語の重み及び単語ベクトルを算出して、重み及び単語ベクトルに基づき、複数の文書の各々の文書ベクトルを算出する。文書ベクトル算出部３１５の処理は、図１０の処理Ｐ５～Ｐ７及び処理Ｐ８の少なくとも一部の一例である。

[0131] 文書ベクトル蓄積部３１６は、図６に示すメモリ部２１の一例であり、文書ベクトル算出部３１５が算出した文書ベクトルを蓄積する。

[0132] （文書検索部３２の説明）

図１１に示すように、文書検索部３２は、例示的に、検索クエリ指定部３２１、文書類似度算出部３２２、検索結果生成部３２３、及び、検索結果出力部３２４を備えてよい。

[0133] 検索クエリ指定部３２１は、図６に示す文書入力部２２の一例であり、図示しない端末装置又は他のサーバ等のコンピュータから、文書検索を要求するクエリ１１（以下、「検索クエリ１１」と表記する場合がある）の入力を受け付ける。

[0134] 例えば、図１２に示すように、検索クエリ指定部３２１は、検索クエリ指定画面３３０の検索ボタン３３２が押下された際に入力欄３３１に設定されている、クエリ文書１１ａの文書番号を受け付けてよい。

[0135] 文書類似度算出部３２２は、図６に示す文書類似度算出部２３３の一例である。文書類似度算出部３２２は、化合物クラスタ蓄積部３１４に蓄積された化合物クラスタ、及び、文書ベクトル蓄積部３１６に蓄積された文書ベクトルに基づき、検索クエリ１１で指定されたクエリ文書１１ａと他の文書との間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出する。

[0136] 例えば、文書類似度算出部３２２は、化合物クラスタ蓄積部３１４に蓄積された化合物クラスタのうちの、クエリ文書１１ａ及び比較対象文書１２ａにそれぞれ対応する化合物クラスタどうしを比較し、化合物類似度を算出してよい。

[0137] また、文書類似度算出部３２２は、文書ベクトル蓄積部３１６に蓄積された文書ベクトルのうちの、クエリ文書１１ａ及び比較対象文書１２ａにそれぞれ対応する文書ベクトルどうしを比較し、テキスト類似度を算出してよい。

[0138] そして、文書類似度算出部３２２は、化合物類似度及びテキスト類似度に基づき、文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出し、文書類似度Ｓｉｍ（Ｘ，Ｙ）の大きい順に比較対象文書１２ａをソートすることで、ランキングの結果１４を生成してよい。

[0139] 文書類似度算出部３２２の処理は、図１０の処理Ｐ３、処理Ｐ８の少なくとも一部、及び、処理Ｐ９の一例である。

[0140] 検索結果生成部３２３は、結果１４に基づき、出力するための検索結果を生成する。例えば、検索結果生成部３２３は、図１２に示す検索結果出力画面３４０を生成してよい。検索結果出力画面３４０は、図７に示す判定結果出力画面２４０における判定結果２４４を検索結果３４４に置き換えたものであってもよい。

[0141] 図１２に示すように、検索結果出力画面３４０には、クエリ文書１１ａの表示領域３４１と、比較対象文書１２ａの少なくとも１つ（図１２では３つ）の表示領域３４５ａ～３４５ｃとが含まれてよい。表示領域３４１には、クエリ文書１１ａの書誌情報及び要約等の表示領域３４２、及び、クエリ文書１１ａの全文参照ボタン３４３が含まれてよい。

[0142] 表示領域３４５ａ～３４５ｃには、比較対象文書１２ａの書誌情報及び要約等の表示領域３４６ａ～３４６ｃ、全文参照ボタン３４７ａ～３４７ｃ、及び、類似すると判定されたクラスタの化合物リスト３４８ａ～３４８ｃが含まれてよい。なお、表示領域３４５ａ～３４５ｃには、類似度Ｓｉｍ（Ｘ，Ｙ）が表示されてもよい。

[0143] 検索結果出力部３２４は、図示しない端末装置又は他のサーバ等のコンピュータに対して、検索結果出力画面３４０を出力する。

[0144] 〔２－３〕動作例

図１３は、サーバ３の文書ＤＢ構築処理の動作例を説明するフローチャートであり、図１４は、サーバ３の文書検索処理の動作例を説明するフローチャートである。

[0145] （文書ＤＢ構築処理の動作例）

図１３に例示するように、文書蓄積部３１１は、未選択の文書を選択し（ステップＳ２１）、文書ＤＢに文書を登録する（ステップＳ２２）。

[0146] 文書ベクトル算出部３１５は、文書のテキストの文書ベクトルを算出する（ステップＳ２３）。文書ベクトル蓄積部３１６は、算出した文書ベクトルを文書と対応付けて、例えば文書ＤＢ又は文書ベクトルＤＢ等に登録（蓄積）する（ステップＳ２４）。

[0147] 化合物名抽出部３１２は、文書のテキストから化合物名を抽出する（ステップＳ２５）。クラスタリング部３１３は、抽出した化合物名をクラスタリングする（ステップＳ２６）。化合物クラスタ蓄積部３１４は、化合物クラスタの情報を文書と対応付けて、例えば文書ＤＢ又は化合物クラスタＤＢ等に登録（蓄積）する（ステップＳ２７）。

[0148] 文書蓄積部３１１は、未選択の文書があるか否かを判定し（ステップＳ２８）、あると判定した場合（ステップＳ２８でＹＥＳ）、処理がステップＳ２１に移行する。文書蓄積部３１１が未選択の文書がないと判定した場合（ステップＳ２８でＮＯ）、処理が終了する。

[0149] なお、ステップＳ２３及びＳ２４の処理と、ステップＳ２５～Ｓ２７の処理とを入れ替えてもよいし、これらの処理の少なくとも一部が前後又は並行して実行されてもよい。

[0150] （文書検索処理の動作例）

図１４に例示するように、検索クエリ指定部３２１は、検索クエリ指定画面３３０からのクエリ文書１１ａの指定を受け付ける（ステップＳ３１）。

[0151] 文書類似度算出部３２２は、文書ベクトル蓄積部３１６からクエリ文書１１ａの文書ベクトルを取得し（ステップＳ３２）、化合物クラスタ蓄積部３１４からクエリ文書１１ａの化合物クラスタを取得する（ステップＳ３３）

。

[0152] 文書類似度算出部３２２は、未選択の文書を選択し（ステップＳ３４）、文書ベクトル蓄積部３１６から選択した文書の文書ベクトルを取得し（ステップＳ３５）、化合物クラスタ蓄積部３１４から選択した文書の化合物クラスタを取得する（ステップＳ３６）。

[0153] 文書類似度算出部３２２は、クエリ文書１１ａ及び選択した文書の文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出する（ステップＳ３７）。

[0154] 文書類似度算出部３２２は、未選択の文書があるか否かを判定し（ステップＳ３８）、ある場合（ステップＳ３８でＹＥＳ）、処理がステップＳ３４に移行する。文書類似度算出部３２２が未選択の文書がないと判定した場合（ステップＳ３８でＮＯ）、文書類似度算出部３２２は、文書類似度が高い順に所定の個数の文書と、各文書でクエリ文書１１ａとの文書類似度が最も高いクラスタとを抽出する（ステップＳ３９）。

[0155] 検索結果生成部３２３は、抽出したデータに基づき検索結果を生成し、検索結果出力部３２４が検索結果、例えば検索結果出力画面３４０を出力し（ステップＳ４０）、処理が終了する。

[0156] 〔２－４〕第１変形例

次に、第２実施形態の第１変形例について説明する。

[0157] （機能構成例）

図１５は、第２実施形態の第１変形例及び後述する第２変形例に係る類似度判定システム１Ｂにおけるサーバ４の機能構成例を示すブロック図であり、図１６及び図１７は、サーバ４による画面出力例を示す図である。

[0158] 第１変形例に係る類似度判定システム１Ｂは、クエリ文書１１ａのクラスタリングの結果であるクラスタを固有表現のリストとしてユーザに提示し、類似度計算に用いるクラスタをユーザに選択させた後に、選択されたクラスタを用いて類似度計算を行なう。これにより、クエリ文書１１ａ内の複数の構成要素のうち、ユーザの意図した構成要素による比較対象文書１２ａの検索を行なうことができ、文書間の類似度の判定精度をより向上させることが

できる。

[0159] 図１５に示すように、サーバ４は、例示的に、文書ＤＢ部３１及び文書検索部４２を備えてよい。文書ＤＢ部３１及び文書検索部４２は、制御部の一例である。文書ＤＢ部３１は、図１１に示す文書ＤＢ部３１と同様である。

[0160] 文書検索部４２は、例示的に、検索クエリ指定部４２１、文書類似度算出部４２２、検索結果生成部４２３、検索結果出力部４２４、化合物クラスタ取得部４２５、クラスタ提示部４２６、及び、クラスタ指定部４２７を備えてよい。

[0161] 検索クエリ指定部４２１、文書類似度算出部４２２、検索結果生成部４２３及び検索結果出力部４２４は、特に言及しない場合、図１１に示す検索クエリ指定部３２１、文書類似度算出部３２２、検索結果生成部３２３及び検索結果出力部３２４と同様である。

[0162] 化合物クラスタ取得部４２５は、検索クエリ指定部４２１が受け付けたクエリ文書１１ａの化合物クラスタを化合物クラスタ蓄積部３１４から取得し、クエリ文書１１ａとともにクラスタ提示部４２６に通知する。

[0163] クラスタ提示部４２６は、化合物クラスタ取得部４２５から取得したクエリ文書１１ａの化合物クラスタをユーザに提示する。例えば、クラスタ提示部４２６は、図１６に示すクラスタ指定画面４４０を生成し、端末装置又は他のサーバ等のコンピュータに出力する。

[0164] 図１６に示すように、クラスタ指定画面４４０は、クエリ文書１１ａの表示領域４４１と、当該クエリ文書１１ａに含まれる複数の化合物クラスタを提示する表示領域４４４とを含んでよい。表示領域４４１には、クエリ文書１１ａの書誌情報及び要約等の表示領域４４２、及び、クエリ文書１１ａの全文参照ボタン４４３が含まれてよい。

[0165] 表示領域４４４には、クエリ文書１１ａの複数のクラスタに対応する複数の化合物リスト４４５と、複数の化合物リスト４４５から類似度計算に用いる化合物クラスタを指定するためのチェックボックス４４６と、検索を実行するための検索ボタン４４７とを含んでよい。

[0166] クラスタ指定部４２７は、クラスタ指定画面４４０の検索ボタン４４７が押下された際にチェックボックス４４６が選択されている化合物リスト４４５の情報を、文書類似度算出部４２２に通知する。

[0167] 文書類似度算出部４２２は、クエリ文書１１ａと、選択した文書との間の文書類似度Ｓｉｍ（Ｘ，Ｙ）の算出の際に用いる化合物クラスタを、クラスタ指定部４２７から指定された化合物リストに制限（限定）する。例えば、文書類似度算出部４２２は、クエリ文書１１ａの複数の化合物クラスタのうちの指定された化合物クラスタに限定し、当該化合物クラスタの化合物リストと、選択した文書の複数の化合物リストとを比較してよい。

[0168] 検索結果生成部４２３及び検索結果出力部４２４は、文書類似度算出部４２２による結果１４に基づき、図１７に示すような検索結果出力画面４５０を生成及び出力してよい。ここで、クラスタ指定画面４４０において、類似度計算に用いる化合物クラスタが指定されている。このため、検索結果出力画面４５０では、化合物リスト（図１２の化合物リスト３４８ａ～３４８ｃ参照）の表示が省略されてよい。なお、図１２の例と同様に、検索結果出力画面４５０に化合物リストが表示されてもよい。

[0169] このように、第１変形例に係るサーバ４によれば、第２実施形態と同様の効果を奏することができるほか、類似度の判定に用いる化合物クラスタを適切な化合物クラスタに限定することができ、文書間の類似度の判定精度をより向上させることができる。また、類似度の判定に用いる化合物クラスタの数を制限できるため、文書検索処理の処理時間を短縮することができる。

[0170] （文書検索処理の動作例）

図１８は、サーバ４の文書検索処理の動作例を説明するフローチャートである。以下、サーバ４の文書検索処理のうちの図１４に示す動作例とは異なる処理を説明する。

[0171] 図１８に例示するように、化合物クラスタ取得部４２５は、化合物クラスタ蓄積部３１４から、クエリ文書１１ａの化合物クラスタ、換言すれば化合物リストを取得する（ステップＳ４１）。

[0172] クラスタ提示部４２６は、化合物クラスタ取得部４２５が取得した化合物リストを含むクラスタ指定画面４４０を生成し、ユーザに提示する（ステップＳ４２）。クラスタ指定部４２７は、クラスタ指定画面４４０における化合物クラスタの指定を受け付け（ステップＳ４３）、処理がステップＳ３４に移行する。なお、ステップＳ４１～Ｓ４３の処理は、ステップＳ３２の前、後、又は、並行して実行されてもよい。

[0173] 文書類似度算出部４２２は、選択した文書の文書ベクトル及び化合物クラスタの取得後、選択した文書の化合物クラスタを、クラスタ指定部４２７が受け付けた指定クラスタに限定する。

[0174] そして、文書類似度算出部４２２は、クエリ文書１１ａの化合物リストを、指定クラスタ、換言すれば、指定された化合物リストに限定して、クエリ文書１１ａと選択した文書との間の文書間類似度を算出し（ステップＳ４４）、処理がステップＳ３８に移行する。

[0175] 文書類似度算出部４２２は、全ての文書について文書類似度の算出が完了すると（ステップＳ３８でＮＯ）、類似度が高い順に、所定の個数の文書を抽出し（ステップＳ４５）、処理がステップＳ４０に移行する。

[0176] 〔２－５〕第２変形例

次に、第２実施形態の第２変形例について説明する。

[0177] 第２変形例に係る類似度判定システム１Ｂにおけるサーバ４の機能構成例は、図１５に示す第１変形例と同様である。図１９は、第２変形例に係るサーバ４の画面出力例を示す図である。図１９に示すように、第２変形例では、検索クエリ指定部４２１が、検索クエリ指定画面４６０に、クエリ文書１１ａの文書番号の入力欄４６１及び１以上のキーワード（ここでは化合物名）の入力領域４６２を含める。クラスタ指定部４２７は、検索ボタン４６３が押下された際に入力欄４６１に入力されているクエリ文書１１ａの文書番号及び入力領域４６２に入力されている１以上のキーワードの情報を、文書類似度算出部４２２に通知する。

[0178] 文書類似度算出部４２２は、指定されたクエリ文書１１ａと比較する文書

の化合物クラスタを、クラスタ指定部４２７が受け付けたキーワードを含む（例えば所定回数以上含む）クラスタに限定する。そして、文書類似度算出部４２２は、指定キーワードを含むクラスタ、換言すれば化合物リストに着目して、クエリ文書１１ａと当該文書との間の文書間類似度を算出する。

[0179] このように、第２変形例に係るサーバ４によれば、第１変形例と同様の効果を奏することができるほか、ユーザは、特定のクラスタに限定せず、所定のキーワードを含むクラスタを柔軟に指定することができ、利便性が高い。

[0180] 〔３〕第３実施形態

〔３－１〕第３実施形態の説明

次に、第３実施形態について説明する。第３実施形態では、第２実施形態に係るテキスト類似度の算出処理において、固有表現類似度の算出の過程で得られるクラスタリング結果を利用する手法を説明する。

[0181] なお、以下の第３実施形態の説明では、特に言及しない構成、処理又は機能は、既述の第１実施形態及び第２実施形態に係る構成、処理又は機能と同様であるものとする。

[0182] 図２０は、第３実施形態に係る類似度判定システム１Ｃを説明するための図であり、図２１及び図２２は、類似度判定システム１Ｃの処理の一例を説明するための図である。

[0183] 図２０に示すように、第３実施形態に係る類似度判定システム１Ｃは、図１０に示す類似度判定システム１Ｂの処理Ｐ８を処理Ｐ１０及びＰ１１に置き換え、処理Ｐ９を処理Ｐ１２に置き換え、処理Ｐ２の処理結果を処理Ｐ１０に渡すものである。以下、処理Ｐ１０～Ｐ１２について説明する。

[0184] （部分文書クラスタリング処理；処理Ｐ１０）

類似度判定システム１Ｃは、処理Ｐ１０において、文書ごとに、文書を分割することによって複数の部分文書（部分テキスト）を取得する。部分文書、換言すれば、文書の分割単位としては、例えば、文、段落、章又は節等が挙げられる。以下、部分文書が段落であるものとする。

[0185] 図２１及び図２２の例では、類似度判定システム１Ｃは、クエリ１１に含

まれる文書Ｘを分割して複数の段落$Ｐ_{X}$を取得し、文書集合１２に含まれる文書Ｙを分割して複数の段落$Ｐ_{Y}$を取得する。以下、段落$Ｐ_{X}$及び$Ｐ_{Y}$を互いに区別しない場合には、単に「段落Ｐ」と表記する。

[0186] 類似度判定システム１Ｃは、段落Ｐを、処理Ｐ２で得られる固有表現（例えば化合物）クラスタに基づきクラスタリングすることによって、部分文書クラスタを取得する。例えば、類似度判定システム１Ｃは、固有表現クラスタに含まれる固有表現と、複数の段落Ｐに含まれる固有表現との間の一致度に基づいて、段落Ｐをクラスタリングしてよい。

[0187] 図２１の例では、類似度判定システム１Ｃは、文書Ｘについて、下記式（１１）に従い、クラスタごとの化合物リスト$Ｃ_{X1}$～$Ｃ_{XN}$のそれぞれと、複数の段落$Ｐ_{X}$のそれぞれとの一致度に基づき、部分文書クラスタ$Ｐ_{X1}$～$Ｐ_{XN}$を生成する。また、類似度判定システム１Ｃは、文書Ｙについて、下記式（１２）に従い、クラスタごとの化合物リスト$Ｃ_{Y1}$～$Ｃ_{YN}$のそれぞれと、複数の段落$Ｐ_{Y}$のそれぞれとの一致度ｃｏｓ（$Ｃ_{PX,}$ $Ｃ_{Xa}$）に基づき、部分文書クラスタ$Ｐ_{Y1}$～$Ｐ_{YN}$を生成する。

[数10]

$$\text{段落}P_X\text{が属するクラスタ}\ \mathrm{Cluster}(P_X) = argmax(\cos(C_{PX}, C_{Xa})) \qquad (11)$$

$$\text{段落}P_Y\text{が属するクラスタ}\ \mathrm{Cluster}(P_Y) = argmax(\cos(C_{PY}, C_{Yb})) \qquad (12)$$

[0188] 上記式（１１）及び式（１２）において、$Ｃ_{PX}$は、段落$Ｐ_{X}$に含まれる化合物リストであり、$Ｃ_{PY}$は、段落$Ｐ_{Y}$に含まれる化合物リストである。$Ｃ_{Xa}$及び$Ｃ_{Yb}$は、処理Ｐ２で得られる文書Ｘ及びＹの化合物リストである。ａｒｇｍａｘは、括弧内の要素が最大となるときの条件（ここではクラスタ）を抽出する関数である。上記式（１１）及び式（１２）によれば、段落Ｐに含まれる化合物名の各々と、化合物クラスタ内の化合物名の各々との間のコサイン類似度が最大となる、例えば出現数が最多となる要素（化合物クラスタ）に、段落Ｐを振り分けることができる。

[0189] 図２２の例では、類似度判定システム１Ｃは、段落$Ｐ_{X}$及び$Ｐ_{Y}$をそれぞれ

４つのクラスタに分類し（Ｎ＝Ｍ＝４）、部分文書クラスタ$Ｐ_{Ｘ１}$～$Ｐ_{Ｘ４}$及び$Ｐ_{Ｙ１}$～$Ｐ_{Ｙ４}$を生成する。このようなクラスタリングにより、結果的に、段落$Ｐ_{Ｘ}$及び$Ｐ_{Ｙ}$を、以下のような４つの要素（特性）の部分文書クラスタに分類することができる。

・部分文書クラスタ$Ｐ_{Ｘ１}$及び$Ｐ_{Ｙ１}$：

「負極活物質」について記載された段落。

・部分文書クラスタ$Ｐ_{Ｘ２}$及び$Ｐ_{Ｙ２}$：

「正極活物質」について記載された段落。

・部分文書クラスタ$Ｐ_{Ｘ３}$及び$Ｐ_{Ｙ３}$：

「バインダー」について記載された段落。

・部分文書クラスタ$Ｐ_{Ｘ４}$及び$Ｐ_{Ｙ４}$：

「電解液溶媒」について記載された段落。

[0190] そして、類似度判定システム１Ｃは、部分文書クラスタのそれぞれに含まれる単語に基づいて、複数の部分文書クラスタのそれぞれに対応する複数の部分文書ベクトルを算出する。例えば、類似度判定システム１Ｃは、部分文書クラスタごとに、処理Ｐ７で取得した単語ベクトルと、処理Ｐ６で取得した単語の重みとを乗じた結果を部分文書クラスタ内の全単語に亘って加算することで、部分文書ベクトルを算出してよい。

[0191] （テキスト類似度算出処理；処理Ｐ１１）

類似度判定システム１Ｃは、処理Ｐ１１において、クエリ文書１１ａの部分文書ベクトルと、比較対象文書１２ａの各々の部分文書ベクトルとの間の類似度、換言すれば、単語の意味ベクトルに基づく、部分文書クラスタ間のテキスト類似度を算出する。クエリ文書１１ａの部分文書ベクトルは、第１の複数のベクトルの一例であり、比較対象文書１２ａの部分文書ベクトルは、第２の複数のベクトルの一例である。

[0192] 例えば、類似度判定システム１Ｃは、クエリ文書１１ａの部分文書クラスタと比較対象文書１２ａの部分文書クラスタとについて、下記式（１３）の演算により、テキスト類似度、一例としてコサイン類似度を算出してよい。

[数11]

$$\text{テキスト類似度 ft}(P_{Xa}, P_{Yb}) = \cos(\sum WP_{Xa}, \sum WP_{Yb}) \quad (13)$$

[0193] 上記式（１３）において、ＷＰ$_{Xa}$は、段落Ｐ$_{Xa}$に含まれる単語の分散ベクトルであり、ＷＰ$_{Yb}$は、段落Ｐ$_{Yb}$に含まれる単語の分散ベクトルである。

[0194] 図２１に示す例では、類似度判定システム１Ｃは、部分文書クラスタＰ$_{X1}$、Ｐ$_{X2}$、Ｐ$_{X3}$、・・・Ｐ$_{XN}$と、部分文書クラスタＰ$_{Y1}$、Ｐ$_{Y2}$、Ｐ$_{Y3}$、・・・Ｐ$_{YM}$との全てのペアについて、上記式（１３）に従いテキスト類似度を算出してよい。

[0195] （ランキング処理；処理Ｐ１２）

そして、類似度判定システム１Ｃは、テキスト類似度及び固有表現類似度に基づき、クエリ文書１１ａとの類似度に応じて複数の比較対象文書１２ａの各々をランキング付けするランキング処理を行ない（処理Ｐ１２）、結果１４を出力する。

[0196] 例えば、類似度判定システム１Ｃは、ランキング処理において、テキスト類似度と固有表現類似度を統合した類似度を算出し、当該類似度に基づき、クエリ文書１１ａとの類似度に応じた複数の比較対象文書１２ａのランキングを出力する。

[0197] 類似度判定システム１Ｃは、例えば、下記式（１４）に従い、文書Ｘと１つの比較対象文書Ｙとの間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出してよい。

[数12]

$$\text{文書類似度 Sim}(X, Y) = max(\text{fc}(C_{Xa}, C_{Yb}) \times \text{ft}(P_{Xa}, P_{Yb})) \quad (14)$$

[0198] 上記式（１４）において、ｆｃは前述の（８）式に従った固有表現類似度、ｆｔは上記式（１３）に従ったテキスト類似度である。

[0199] なお、上記式（１４）では、文書Ｘ（クエリ文書１１ａ）と、１つの文書Ｙ（比較対象文書１２ａ）との間の文書類似度を算出する例を示す。類似度判定システム１Ｃは、第２実施形態と同様に、文書Ｙの数に応じた文書類似度Ｓｉｍ（Ｘ，Ｙ$_1$）～Ｓｉｍ（Ｘ，Ｙ$_L$）を取得してよい。

[0200] そして、類似度判定システム１Ｃは、例えば、第２実施形態と同様に、文書類似度Ｓｉｍ（Ｘ，$Ｙ_1$）～Ｓｉｍ（Ｘ，$Ｙ_L$）が高い文書Ｙから降順に検索対象の全文書$Ｙ_1$～$Ｙ_L$をソートすることで、ランキング処理を行なう。また、類似度判定システム１Ｃは、ソート結果を結果１４として出力してよい。

[0201] なお、類似度判定システム１Ｃは、第２実施形態と同様に、下記式（１５）に従い、文書Ｘと１つの比較対象文書Ｙとの間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を、固有表現似度とテキスト類似度との重み付き和として算出してもよい。

[数13]

$$\text{文書類似度 } \mathrm{Sim}(X,Y) = \max(w \times \mathrm{fc}(C_{Xa}, C_{Yb}) + (1-w) \times \mathrm{ft}(P_{Xa}, P_{Yb})) \qquad (15)$$

[0202] 以上のように、第３実施形態に係る類似度判定システム１Ｃによれば、第１及び第２実施形態と同様の効果を奏することができる。

[0203] また、図２２に示すように、部分文書クラスタどうしの比較を行なうことで、例えば、「正極活物質」についての意味ベクトルが類似しているから文書Ｘ及びＹの類似度が高い、と判断することができる。図２２では、便宜上、意味ベクトル空間を２次元で示すが、実際には数百次元のベクトルとなり得る。第３実施形態によれば、部分文書クラスタどうしの比較により、部分的に類似する文書間の類似度の判定精度を向上させることができる。

[0204] 〔３－２〕機能構成例

図２３は、第３実施形態に係る類似度判定システム１Ｃにおけるサーバ５の機能構成例を示すブロック図である。サーバ５は、特に言及しない場合、図１１に示すサーバ３、又は、図１５に示すサーバ４と同様であってよい。

[0205] 類似度判定システム１Ｃによる上述した類似度判定処理は、サーバ５により実現されてよい。図２３に示すように、サーバ５は、例示的に、文書ＤＢ部５１及び文書検索部５２を備えてよい。文書ＤＢ部５１及び文書検索部５２は、制御部の一例である。

[0206] 文書ＤＢ部５１は、図１１に示す文書ベクトル算出部３１５及び文書ベクトル蓄積部３１６に代えて、文書クラスタベクトル算出部５１５及び文書クラスタベクトル蓄積部５１６を備えてよい。また、文書ＤＢ部５１では、クラスタリング部３１３から文書クラスタベクトル算出部５１５に、クラスタリング結果である化合物クラスタが出力されてよい。

[0207] 文書検索部５２は、図１１に示す文書類似度算出部３２２に代えて、文書類似度算出部５２２を備えてよい。

[0208] （文書ＤＢ部５１の説明）

例えば、文書クラスタベクトル算出部５１５は、クラスタリング部３１３からの化合物クラスタの情報に基づき、部分文書クラスタごとの文書ベクトルを算出してよい。文書クラスタベクトル算出部５１５の処理は、図２０に示す処理Ｐ１０の処理の一例である。

[0209] 文書クラスタベクトル蓄積部５１６は、文書クラスタベクトル算出部５１５が算出した部分文書クラスタごとの文書ベクトルを蓄積する。

[0210] （文書検索部５２の説明）

文書類似度算出部５２２は、クエリ文書１１ａの部分文書ベクトルと、比較対象文書１２ａの各々の部分文書ベクトルとの間の文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出し、文書類似度Ｓｉｍ（Ｘ，Ｙ）から結果１４を生成してよい。文書類似度算出部５２２の処理は、図２０の処理Ｐ３、処理Ｐ１１、及び、処理Ｐ１２の一例である。

[0211] なお、第３実施形態に係る文書検索部５２は、図１２、図１６、図１７、図１９等に示す上述した種々の画面を出力してよい。

[0212] 〔３－３〕動作例

図２４は、サーバ５の文書ＤＢ構築処理の動作例を説明するフローチャートであり、図２５は、サーバ５の文書検索処理の動作例を説明するフローチャートである。

[0213] （文書ＤＢ構築処理の動作例）

図２４は、図１３に示すステップＳ２３及びＳ２４を削除し、ステップＳ

２７とＳ２８との間にステップＳ５１～Ｓ５４を追加したものである。

[0214] 図２４に例示するように、文書クラスタベクトル算出部５１５は、文書のテキストを所定単位に分割し（ステップＳ５１）、化合物クラスタ蓄積部３１４が蓄積する化合物クラスタに基づき、各分割単位（段落Ｐ）をクラスタリングする（ステップＳ５２）。

[0215] また、文書クラスタベクトル算出部５１５は、各部分文書クラスタの文書ベクトルを算出する（ステップＳ５３）。文書クラスタベクトル蓄積部５１６は、各部分文書クラスタの文書ベクトルを蓄積し（ステップＳ５４）、処理がステップＳ２８に移行する。

[0216] （文書検索処理の動作例）

図２５は、図１４に示すステップＳ３２、Ｓ３５、Ｓ３７を、それぞれステップＳ６１、Ｓ６２、ステップＳ６３に置き換えたものである。

[0217] 文書類似度算出部５２２は、ステップＳ６１において、文書クラスタベクトル蓄積部５１６からクエリ文書１１ａの部分文書クラスタの文書ベクトルを取得する。

[0218] 文書類似度算出部５２２は、ステップＳ６２において、文書クラスタベクトル蓄積部５１６から選択した文書の部分文書クラスタの文書ベクトルを取得する。

[0219] 文書類似度算出部５２２は、ステップＳ６３において、ステップＳ６１及びＳ６２でそれぞれ取得した文書ベクトルと、化合物クラスタとに基づき、文書類似度Ｓｉｍ（Ｘ，Ｙ）を算出する。

[0220] 〔４〕その他

上述した第１～第３実施形態、並びに、第２実施形態の第１及び第２変形例に係る技術は、以下のように変形、変更して実施することができる。

[0221] 例えば、上述した第１～第３実施形態、並びに、第２実施形態の第１及び第２変形例では、固有表現として、化合物名が用いられる場合を例に挙げて説明したが、これに限定されるものではない。固有表現としては、例えば遺伝子配列（ゲノム）等、自然言語処理において固有表現抽出処理の対象とな

り得る種々の用語が用いられてもよい。

[0222] また、例えば、図６、図１１、図１５、図２０に示すサーバ２～サーバ５のそれぞれが備える機能構成は、任意の組み合わせで併合してもよく、それぞれ分割してもよい。また、第１～第３実施形態、並びに、第２実施形態の第１及び第２変形例を適宜組み合わせて実施してもよい。さらに、サーバ２～サーバ５のそれぞれは、図７、図１２、図１６、図１７、図１９のいずれの画面の画面情報を生成してもよく、画面に応じた機能構成を備えてよい。

[0223] さらに、図６、図１１、図１５、図２０に示すサーバ２～サーバ５のそれぞれは、複数の装置がネットワークを介して互いに連携することにより、各処理機能を実現する構成であってもよい。一例として、メモリ部２１はＤＢサーバ、文書ＤＢ部３１及び５１はアプリケーションサーバ及びＤＢサーバの組み合わせ、文書入力部２２、類似度算出部２３、類似度出力部２４、文書検索部３２、４２及び５２はアプリケーションサーバ及びＷｅｂサーバの組み合わせ、等であってもよい。これらの場合、コンピュータ、アプリケーションサーバ及びＤＢサーバが、ネットワークを介して互いに連携することにより、サーバ２～５としての各処理機能を実現してもよい。

[0224] また、サーバ３～５のそれぞれは、図９に例示するコンピュータ１０のＨＷ構成を備えてよい。

符号の説明

[0225] １、１Ａ～１Ｃ　類似度判定システム

１０　コンピュータ

１１　クエリ

１１ａ　クエリ文書

１２　文書集合

１２ａ　比較対象文書

１３、１４　結果

２～５　サーバ

２１　メモリ部

２２　　文書入力部

２３　　類似度算出部

２４　　類似度出力部

２３１、３１２　　化合物名抽出部

２３２、３１３　　クラスタリング部

２３３　　文書類似度算出部

３１、５１　　文書ＤＢ部

３１１　　文書蓄積部

３１４　　化合物クラスタ蓄積部

３１５　　文書ベクトル算出部

３１６　　文書ベクトル蓄積部

３２、４２、５２　　文書検索部

３２１、４２１　　検索クエリ指定部

３２２、４２２、５２２　　文書類似度算出部

３２３、４２３　　検索結果生成部

３２４、４２４　　検索結果出力部

４２５　　化合物クラスタ取得部

４２６　　クラスタ提示部

４２７　　クラスタ指定部

５１５　　文書クラスタベクトル算出部

５１６　　文書クラスタベクトル蓄積部

請求の範囲

[請求項1] 第１の文書に含まれる第１の複数の固有表現のそれぞれの位置と前記第１の複数の固有表現のそれぞれの類似度とに基づいて前記第１の複数の固有表現を分類することによって生成された第１の複数のグループを取得し、

第２の文書に含まれる第２の複数の固有表現のそれぞれの位置と前記第２の複数の固有表現のそれぞれの類似度とに基づいて前記第２の複数の固有表現を分類することによって生成された第２の複数のグループを取得し、

前記第１の複数のグループと前記第２の複数のグループとの比較に基づいて、前記第１の文書と前記第２の文書との類似度を判定する、処理をコンピュータに実行させる、類似度判定プログラム。

[請求項2] 前記第１の複数のグループを取得する処理は、前記第１の複数の固有表現の各々の前記第１の文書内の出現位置の近さを数値化した値と、前記第１の複数の固有表現の各々の類似度とを用いたクラスタリング処理を含み、

前記第２の複数のグループを取得する処理は、前記第２の複数の固有表現の各々の前記第２の文書内の出現位置の近さを数値化した値と、前記第２の複数の固有表現の各々の類似度とを用いたクラスタリング処理を含む、

請求項１に記載の類似度判定プログラム。

[請求項3] 前記類似度を判定する処理は、前記第１の複数のグループの各々と前記第２の複数のグループの各々との組み合わせの中で、グループの類似度が最大となる組み合わせの前記グループの類似度を、前記第１の文書と前記第２の文書との類似度と判定する処理を含む、

請求項１又は請求項２に記載の類似度判定プログラム。

[請求項4] 前記第２の複数のグループのうちの前記グループの類似度が最大となるグループに属する固有表現のリストを含む画面情報を出力する、

処理を前記コンピュータに実行させる、

請求項３に記載の類似度判定プログラム。

[請求項5] 前記第１の文書に含まれる単語に基づいて、前記第１の文書に対応する第１のベクトルを算出し、

前記第２の文書に含まれる単語に基づいて、前記第２の文書に対応する第２のベクトルを算出する、

処理を前記コンピュータに実行させ、

前記類似度を判定する処理は、前記第１の複数のグループと前記第２の複数のグループとの比較と、前記第１のベクトルと前記第２のベクトルとの比較とに基づいて、前記第１の文書と前記第２の文書との前記類似度を判定する処理を含む、

請求項１～請求項４のいずれか１項に記載の類似度判定プログラム。

[請求項6] 前記第１のベクトルを算出する処理は、前記第１の文書を分割することによって得られた第１の複数の部分文書を前記第１の複数のグループに基づいて分類することによって得られた第１の複数の部分文書グループについて、前記第１の複数の部分文書グループのそれぞれに含まれる単語に基づいて、前記第１の複数の部分文書グループのそれぞれに対応する第１の複数のベクトルを算出する処理を含み、

前記第２のベクトルを算出する処理は、前記第２の文書を分割することによって得られた第２の複数の部分文書を前記第２の複数のグループに基づいて分類することによって得られた第２の複数の部分文書グループについて、前記第２の複数の部分文書グループのそれぞれに含まれる単語に基づいて、前記第２の複数の部分文書グループのそれぞれに対応する第２の複数のベクトルを算出する処理を含む、

請求項５に記載の類似度判定プログラム。

[請求項7] 前記第１の複数のグループの各々の情報を含む画面情報を出力する、

処理を前記コンピュータに実行させ、

前記類似度を判定する処理は、前記画面情報に応じて選択されたグループと、前記第２の複数のグループとの比較に基づいて、前記第１の文書と前記第２の文書との前記類似度を判定する処理を含む、

請求項１～請求項６のいずれか１項に記載の類似度判定プログラム。

[請求項8] 前記類似度を判定する処理は、前記第１の複数のグループのうちの指定されたキーワードを含むグループと、前記第２の複数のグループとの比較に基づいて、前記第１の文書と前記第２の文書との前記類似度を判定する処理を含む、

請求項１～請求項７のいずれか１項に記載の類似度判定プログラム。

[請求項9] 前記第１の文書が検索クエリで指定された文書であり、

前記第２の文書が前記検索クエリの検索対象となる文書群に含まれる複数の前記第２の文書のうちの１つであり、

前記第１の文書と前記複数の第２の文書の各々との複数の前記類似度に応じて判定した第２の文書の情報を、前記検索クエリの検索結果として出力する、

処理を前記コンピュータに実行させる、

請求項１～請求項８のいずれか１項に記載の類似度判定プログラム。

[請求項10] 前記第１の複数の固有表現、及び、前記第２の複数の固有表現の各々は化合物名であり、

前記第１の複数の固有表現のそれぞれの類似度、及び、前記第２の複数の固有表現のそれぞれの類似度の各々は、化合物の構造類似度である、

請求項１～請求項９のいずれか１項に記載の類似度判定プログラム。

[請求項11] 第１の文書に含まれる第１の複数の固有表現のそれぞれの位置と前記第１の複数の固有表現のそれぞれの類似度とに基づいて前記第１の複数の固有表現を分類することによって生成された第１の複数のグループを取得し、

第２の文書に含まれる第２の複数の固有表現のそれぞれの位置と前

記第２の複数の固有表現のそれぞれの類似度とに基づいて前記第２の複数の固有表現を分類することによって生成された第２の複数のグループを取得し、

前記第１の複数のグループと前記第２の複数のグループとの比較に基づいて、前記第１の文書と前記第２の文書との類似度を判定する、制御部を備える、

類似度判定装置。

[請求項12] 前記制御部は、

前記第１の複数のグループの取得において、前記第１の複数の固有表現の各々の前記第１の文書内の出現位置の近さを数値化した値と、前記第１の複数の固有表現の各々の類似度とを用いたクラスタリング処理を行ない、

前記第２の複数のグループの取得において、前記第２の複数の固有表現の各々の前記第２の文書内の出現位置の近さを数値化した値と、前記第２の複数の固有表現の各々の類似度とを用いたクラスタリング処理を行なう、

請求項１１に記載の類似度判定装置。

[請求項13] 前記制御部は、前記類似度の判定において、前記第１の複数のグループの各々と前記第２の複数のグループの各々との組み合わせの中で、グループの類似度が最大となる組み合わせの前記グループの類似度を、前記第１の文書と前記第２の文書との類似度と判定する

請求項１１又は請求項１２に記載の類似度判定装置。

[請求項14] 前記制御部は、前記第２の複数のグループのうちの前記グループの類似度が最大となるグループに属する固有表現のリストを含む画面情報を出力する、

請求項１３に記載の類似度判定装置。

[請求項15] 前記制御部は、

前記第１の文書に含まれる単語に基づいて、前記第１の文書に対応

する第１のベクトルを算出し、

前記第２の文書に含まれる単語に基づいて、前記第２の文書に対応する第２のベクトルを算出し、

前記類似度の判定において、前記第１の複数のグループと前記第２の複数のグループとの比較と、前記第１のベクトルと前記第２のベクトルとの比較とに基づいて、前記第１の文書と前記第２の文書との前記類似度を判定する、

請求項１１～請求項１４のいずれか１項に記載の類似度判定装置。

[請求項16] 前記制御部は、

前記第１のベクトルを算出する処理において、前記第１の文書を分割することによって得られた第１の複数の部分文書を前記第１の複数のグループに基づいて分類することによって得られた第１の複数の部分文書グループについて、前記第１の複数の部分文書グループのそれぞれに含まれる単語に基づいて、前記第１の複数の部分文書グループのそれぞれに対応する第１の複数のベクトルを算出し、

前記第２のベクトルを算出する処理において、前記第２の文書を分割することによって得られた第２の複数の部分文書を前記第２の複数のグループに基づいて分類することによって得られた第２の複数の部分文書グループについて、前記第２の複数の部分文書グループのそれぞれに含まれる単語に基づいて、前記第２の複数の部分文書グループのそれぞれに対応する第２の複数のベクトルを算出する、

請求項１５に記載の類似度判定装置。

[請求項17] 第１の文書に含まれる第１の複数の固有表現のそれぞれの位置と前記第１の複数の固有表現のそれぞれの類似度とに基づいて前記第１の複数の固有表現を分類することによって生成された第１の複数のグループを取得し、

第２の文書に含まれる第２の複数の固有表現のそれぞれの位置と前記第２の複数の固有表現のそれぞれの類似度とに基づいて前記第２の

複数の固有表現を分類することによって生成された第２の複数のグループを取得し、

前記第１の複数のグループと前記第２の複数のグループとの比較に基づいて、前記第１の文書と前記第２の文書との類似度を判定する、処理をコンピュータが実行する、類似度判定方法。

[請求項18] 前記第１の複数のグループを取得する処理は、前記第１の複数の固有表現の各々の前記第１の文書内の出現位置の近さを数値化した値と、前記第１の複数の固有表現の各々の類似度とを用いたクラスタリング処理を含み、

前記第２の複数のグループを取得する処理は、前記第２の複数の固有表現の各々の前記第２の文書内の出現位置の近さを数値化した値と、前記第２の複数の固有表現の各々の類似度とを用いたクラスタリング処理を含む、

請求項１７に記載の類似度判定方法。

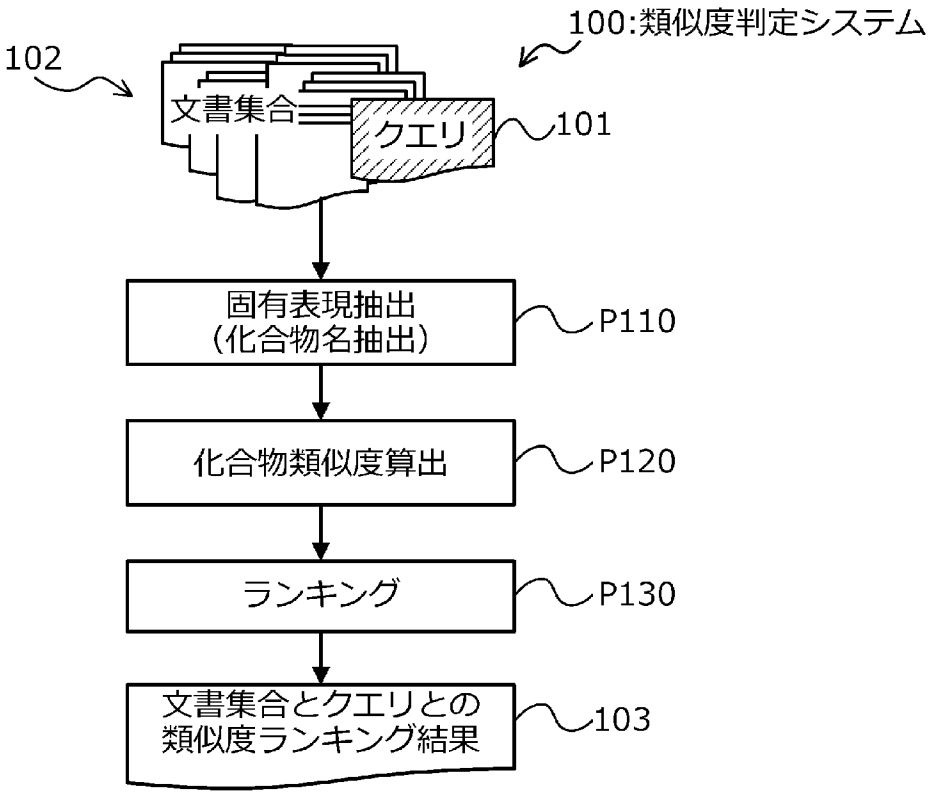
[請求項19] 前記類似度を判定する処理は、前記第１の複数のグループの各々と前記第２の複数のグループの各々との組み合わせの中で、グループの類似度が最大となる組み合わせの前記グループの類似度を、前記第１の文書と前記第２の文書との類似度と判定する処理を含む、

請求項１７又は請求項１８に記載の類似度判定方法。

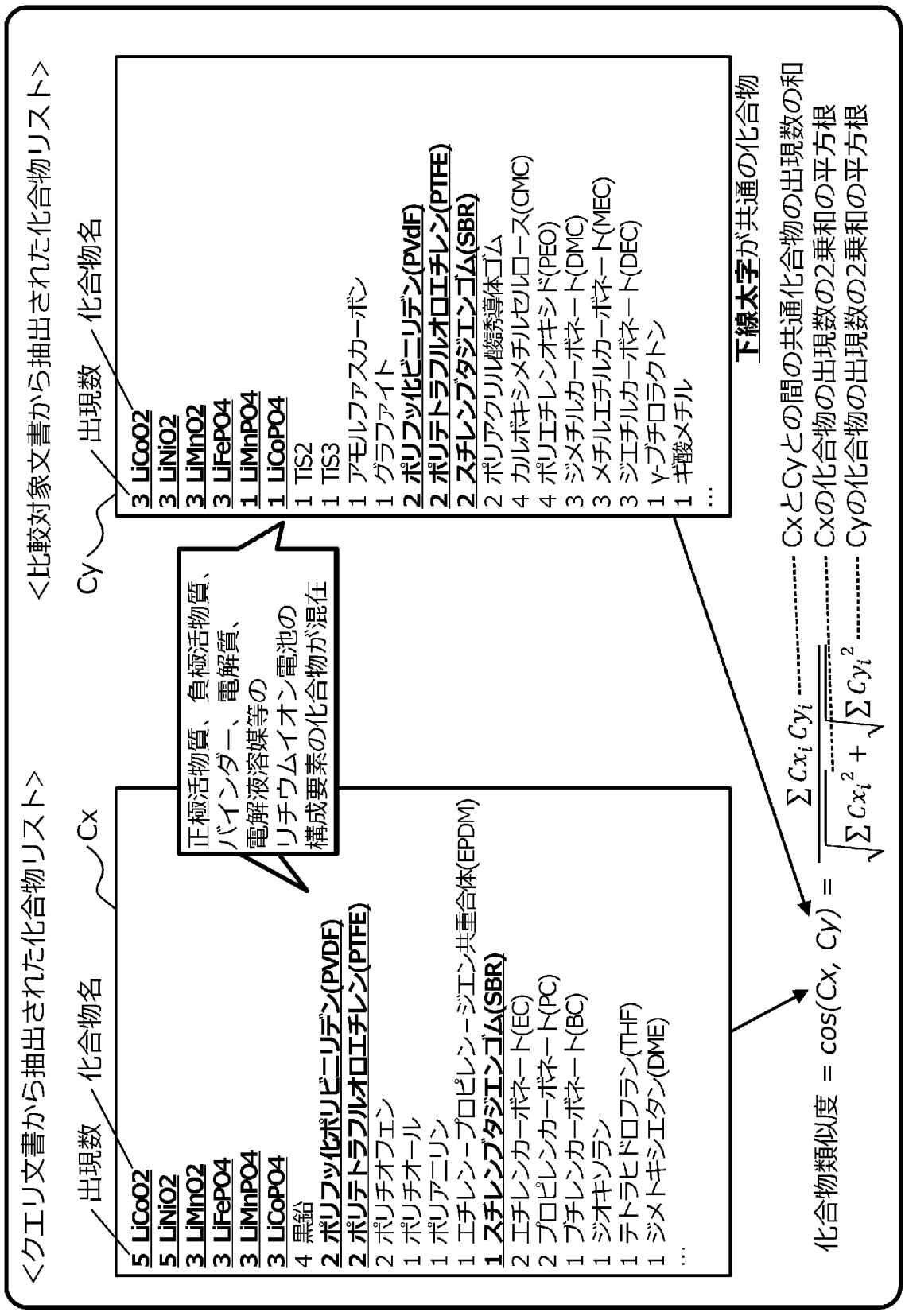
[請求項20] 前記第２の複数のグループのうちの前記グループの類似度が最大となるグループに属する固有表現のリストを含む画面情報を出力する、処理を前記コンピュータが実行する、

請求項１９に記載の類似度判定方法。

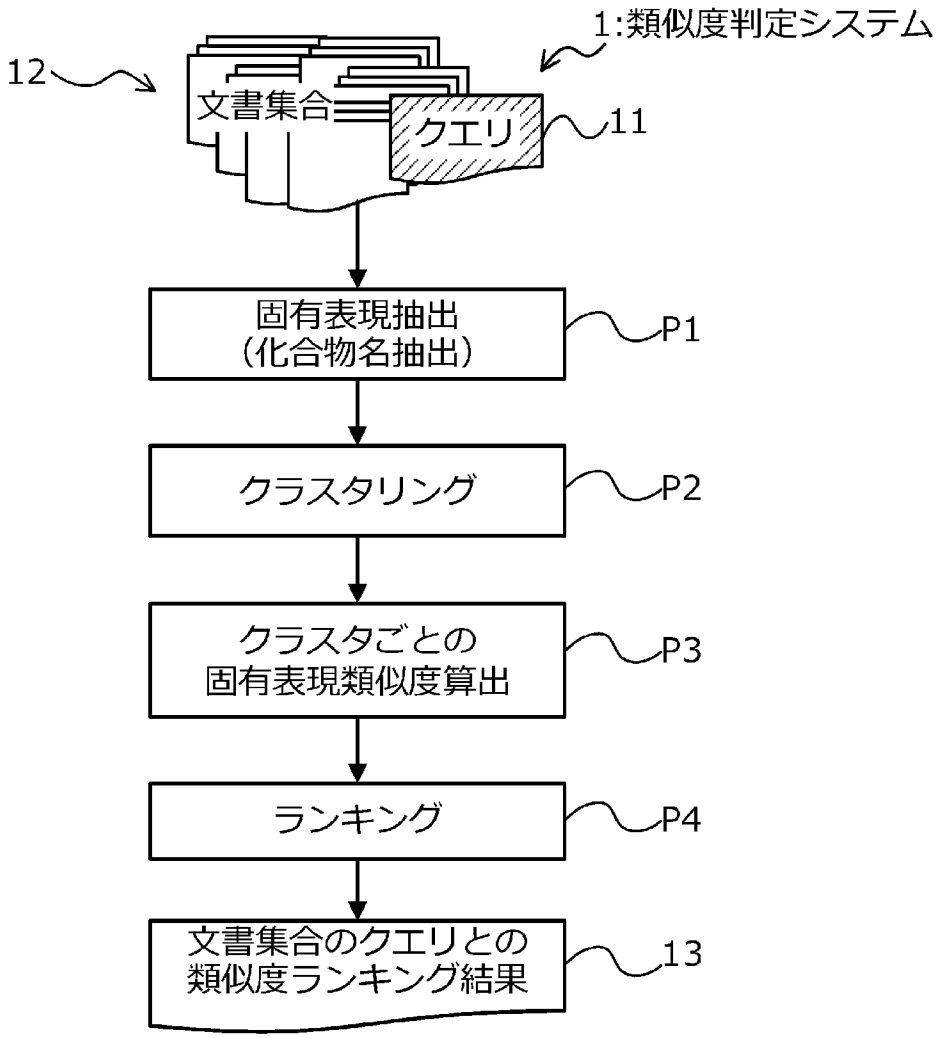
[図1]


100:類似度判定システム
102
文書集合
クエリ
101
固有表現抽出
(化合物名抽出)
P110
化合物類似度算出
P120
ランキング
P130
文書集合とクエリとの
類似度ランキング結果
103

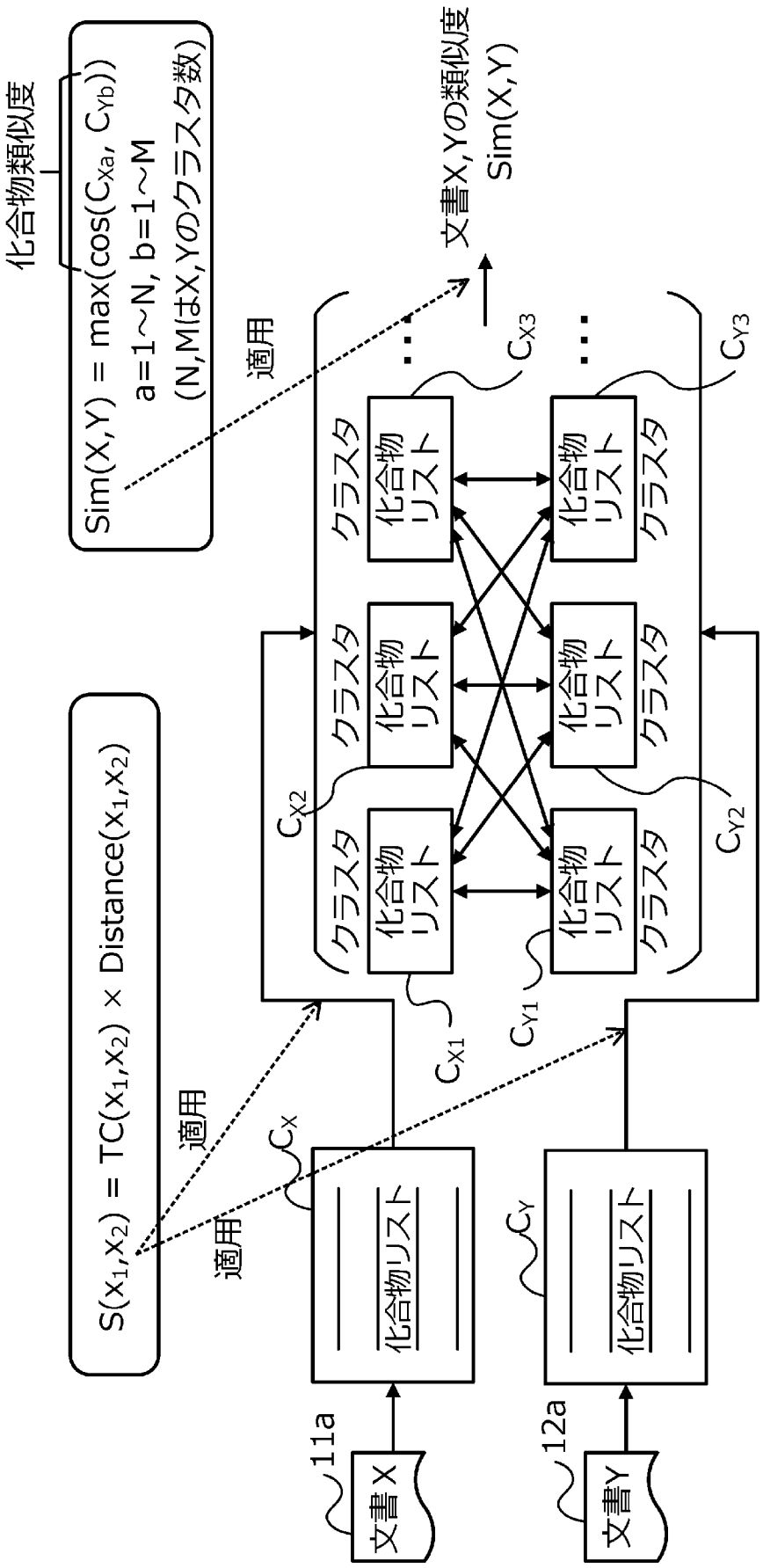
[図2]


<クエリ文書から抽出された化合物リスト>
出現数
化合物名
Cx
5 LiCoO2
5 LiNiO2
3 LiMnO2
3 LiFePO4
3 LiMnPO4
3 LiCoPO4
4 黒鉛
2 ポリフッ化ポリビニリデン(PVDF)
2 ポリテトラフルオロエチレン(PTFE)
2 ポリアクリル酸
1 ポリチオフェン
1 ポリチオール
1 ポリアニリン
1 エチレン－プロピレン－ジエン共重合体(EPDM)
1 スチレンブタジエンゴム(SBR)
2 エチレンカーボネート(EC)
2 プロピレンカーボネート(PC)
1 ブチレンカーボネート(BC)
1 ジオキソラン
1 テトラヒドロフラン(THF)
1 ジメトキシエタン(DME)
…
正極活物質、負極活物質、
バインダー、電解質、
電解液溶媒等の
リチウムイオン電池の
構成要素の化合物が混在
<比較対象文書から抽出された化合物リスト>
出現数
化合物名
Cy
3 LiCoO2
3 LiNiO2
3 LiMnO2
3 LiFePO4
1 LiMnPO4
1 LiCoPO4
1 TiS2
1 TiS3
1 アモルファスカーボン
1 グラファイト
2 ポリフッ化ビニリデン(PVdF)
2 ポリテトラフルオロエチレン(PTFE)
2 スチレンブタジエンゴム(SBR)
2 ポリアクリル酸誘導体ゴム
4 カルボキシメチルセルロース(CMC)
4 ポリエチレンオキシド(PEO)
3 ジメチルカーボネート(DMC)
3 メチルエチルカーボネート(MEC)
3 ジエチルカーボネート(DEC)
1 γ-ブチロラクトン
1 ギ酸メチル
…
下線太字が共通の化合物
化合物類似度 = cos(Cx, Cy) = $\frac{\sum Cx_i\, Cy_i}{\sqrt{\sum Cx_i^2} + \sqrt{\sum Cy_i^2}}$
CxとCyとの間の共通化合物の出現数の2乗和の平方根
Cxの化合物の出現数の2乗和の平方根
Cyの化合物の出現数の2乗和の平方根

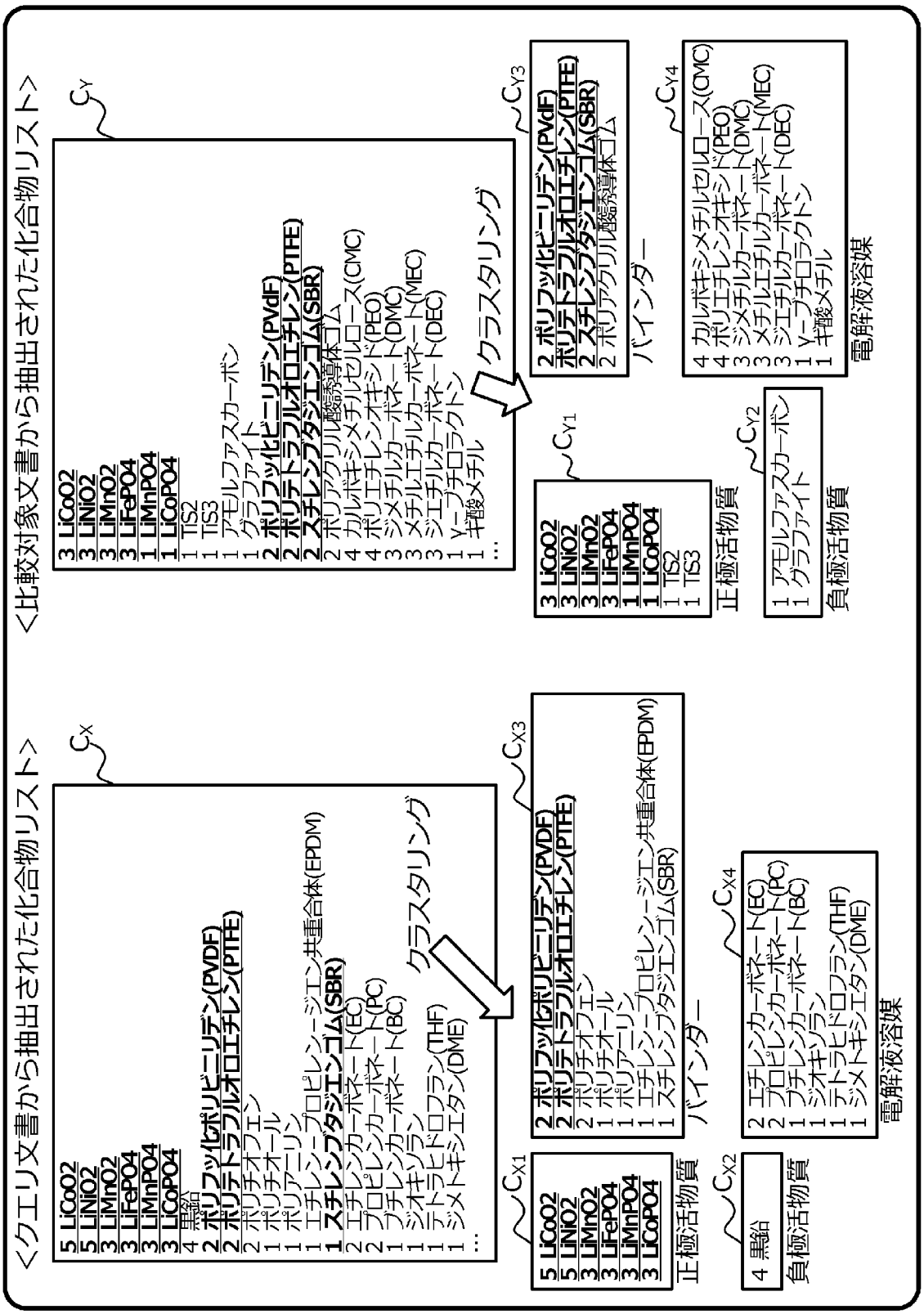
[図3]


1:類似度判定システム
12
文書集合
クエリ
11
固有表現抽出
（化合物名抽出）
P1
クラスタリング
P2
クラスタごとの
固有表現類似度算出
P3
ランキング
P4
文書集合のクエリとの
類似度ランキング結果
13

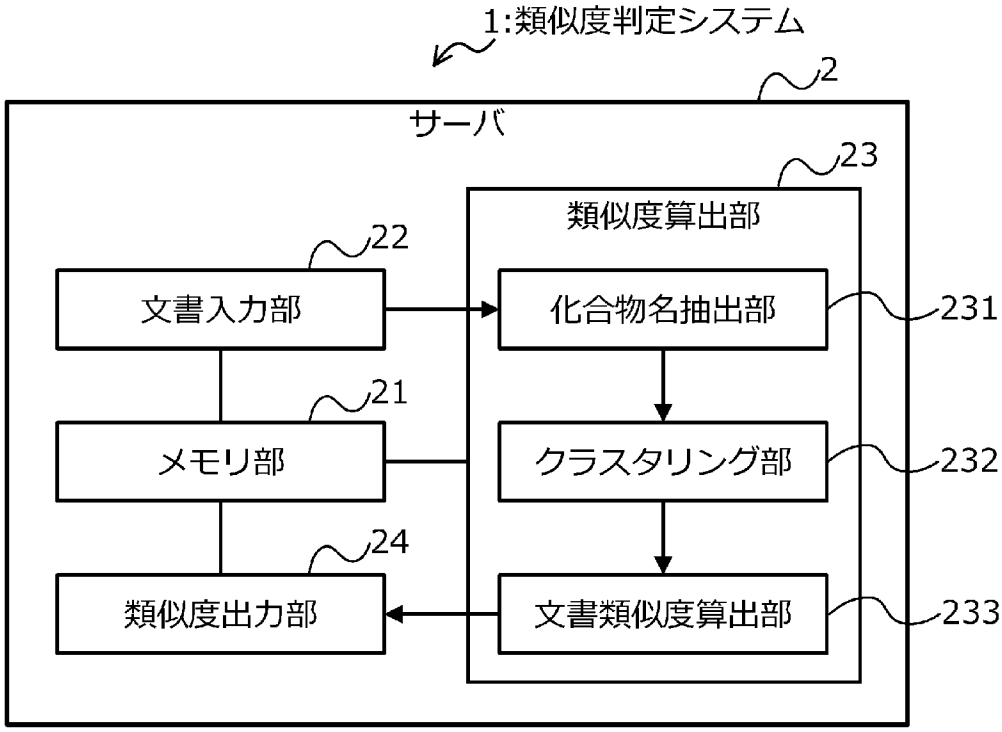
[図4]


化合物類似度
Sim(X,Y) = max(cos(CXa, CYb))
a=1～N, b=1～M
(N,MはX,Yのクラスタ数)
適用
文書X,Yの類似度
Sim(X,Y)
CX3
CY3
クラスタ
化合物リスト
CX2
CY2
CX1
CY1
S(x1,x2) = TC(x1,x2) × Distance(x1,x2)
適用
適用
CX
CY
化合物リスト
11a
文書X
12a
文書Y

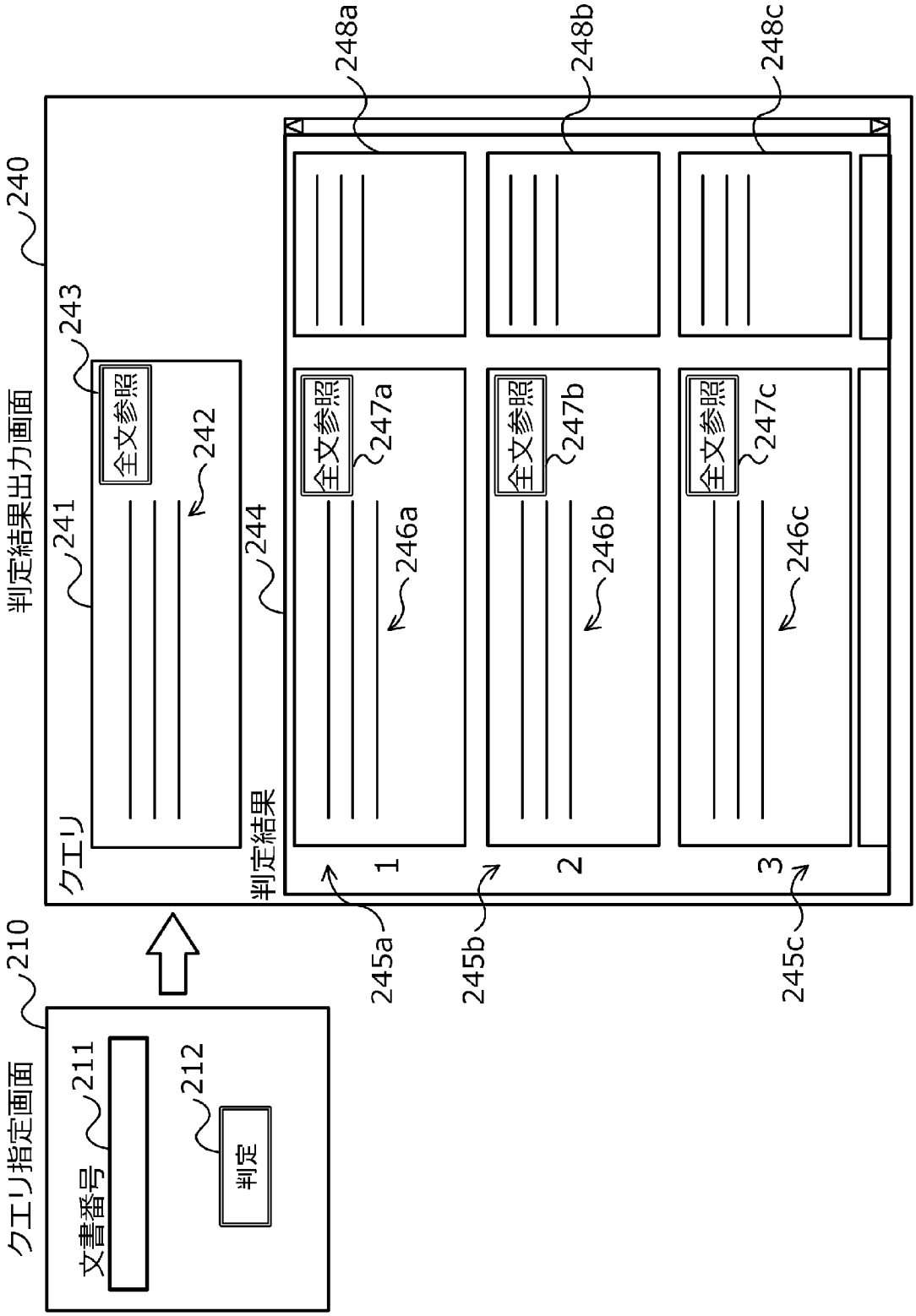
[図5]


<クエリ文書から抽出された化合物リスト>
5 LiCoO2
5 LiNiO2
3 LiMnO2
3 LiFePO4
3 LiMnPO4
3 LiCoPO4
4 黒鉛
2 ポリフッ化ビニリデン(PVdF)
2 ポリテトラフルオロエチレン(PTFE)
2 ポリチオフェン
1 ポリチオオール
1 ポリアニリン
1 エチレン－プロピレン－ジエン共重合体(EPDM)
1 スチレンブタジエンゴム(SBR)
2 エチレンカーボネート(EC)
2 プロピレンカーボネート(PC)
1 ブチレンカーボネート(BC)
1 ジオキソラン
1 テトラヒドロフラン(THF)
1 ジメトキシエタン(DME)
…
クラスタリング
CX
CX1
5 LiCoO2
5 LiNiO2
3 LiMnO2
3 LiFePO4
3 LiMnPO4
3 LiCoPO4
正極活物質
CX2
4 黒鉛
負極活物質
CX3
2 ポリフッ化ビニリデン(PVdF)
2 ポリテトラフルオロエチレン(PTFE)
2 ポリチオフェン
1 ポリチオオール
1 ポリアニリン
1 エチレン－プロピレン－ジエン共重合体(EPDM)
1 スチレンブタジエンゴム(SBR)
バインダー
CX4
2 エチレンカーボネート(EC)
2 プロピレンカーボネート(PC)
1 ブチレンカーボネート(BC)
1 ジオキソラン
1 テトラヒドロフラン(THF)
1 ジメトキシエタン(DME)
電解液溶媒
<比較対象文書から抽出された化合物リスト>
3 LiCoO2
3 LiNiO2
3 LiMnO2
3 LiFePO4
1 LiMnPO4
1 LiCoPO4
1 TiS2
1 TiS3
1 アモルファスカーボン
1 グラファイト
2 ポリフッ化ビニリデン(PVdF)
2 ポリテトラフルオロエチレン(PTFE)
2 スチレンブタジエンゴム(SBR)
2 ポリアクリル酸誘導体ゴム
4 カルボキシメチルセルロース(CMC)
4 ポリエチレンオキシド(PEO)
3 ジメチルカーボネート(DMC)
3 メチルエチルカーボネート(MEC)
3 ジエチルカーボネート(DEC)
1 γ－ブチロラクトン
1 ギ酸メチル
…
クラスタリング
CY
CY1
3 LiCoO2
3 LiNiO2
3 LiMnO2
3 LiFePO4
1 LiMnPO4
1 LiCoPO4
1 TiS2
1 TiS3
正極活物質
CY2
1 アモルファスカーボン
1 グラファイト
負極活物質
CY3
2 ポリフッ化ビニリデン(PVdF)
2 ポリテトラフルオロエチレン(PTFE)
2 スチレンブタジエンゴム(SBR)
2 ポリアクリル酸誘導体ゴム
バインダー
CY4
4 カルボキシメチルセルロース(CMC)
4 ポリエチレンオキシド(PEO)
3 ジメチルカーボネート(DMC)
3 メチルエチルカーボネート(MEC)
3 ジエチルカーボネート(DEC)
1 γ－ブチロラクトン
1 ギ酸メチル
電解液溶媒

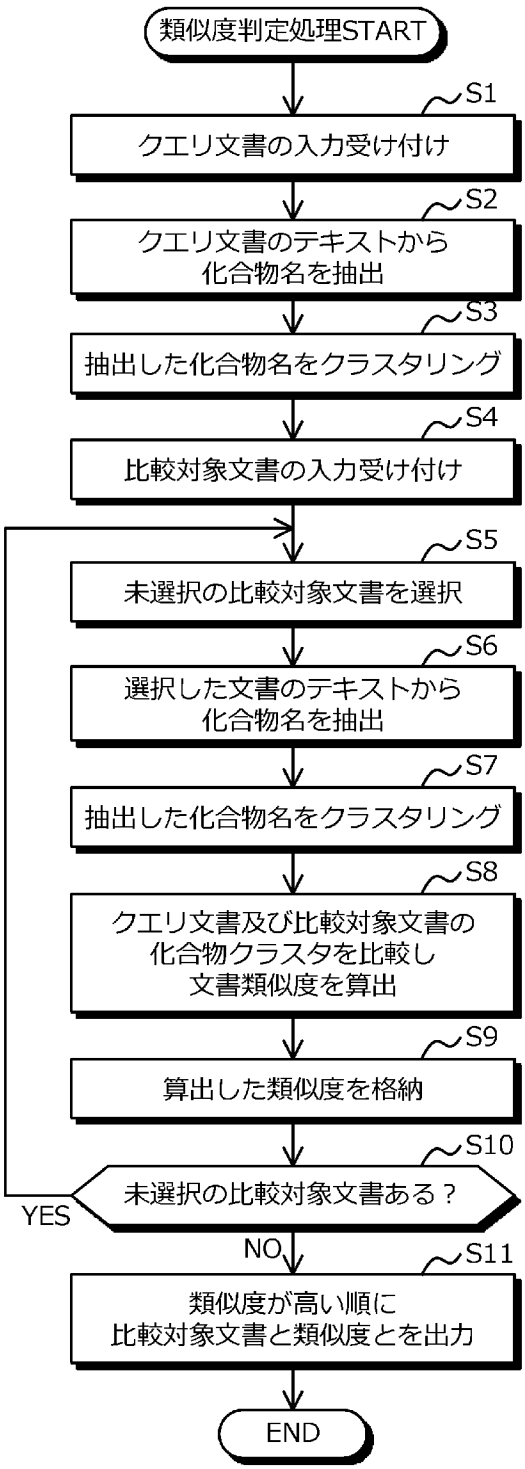
[図6]


1:類似度判定システム
2
サーバ
23
類似度算出部
22
文書入力部
化合物名抽出部
231
21
メモリ部
クラスタリング部
232
24
類似度出力部
文書類似度算出部
233

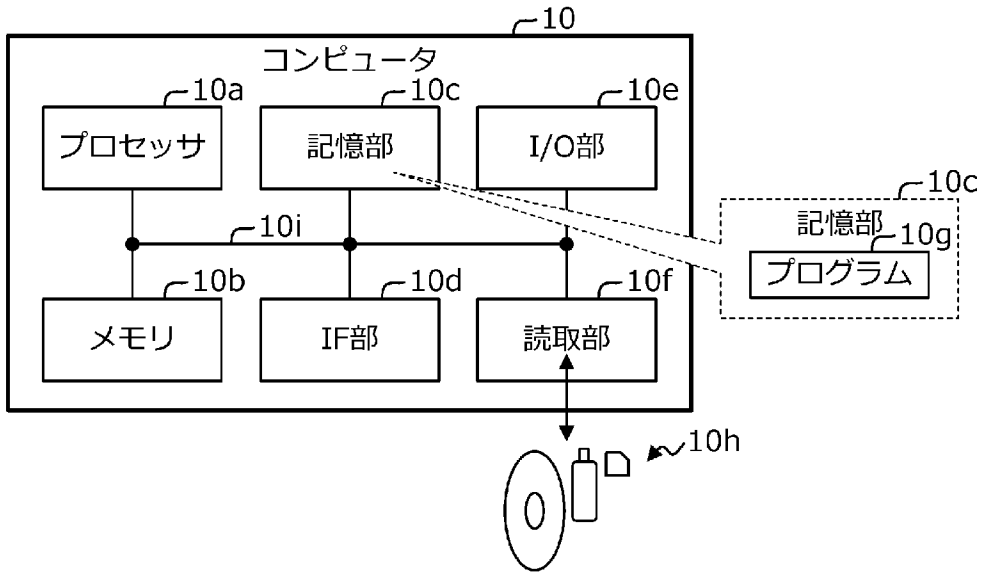
[図7]


クエリ指定画面
210
文書番号
211
判定
212
判定結果出力画面
240
クエリ
241
全文参照
242
243
判定結果
244
1
2
3
全文参照
247a
246a
全文参照
247b
246b
全文参照
247c
246c
248a
248b
248c
245a
245b
245c

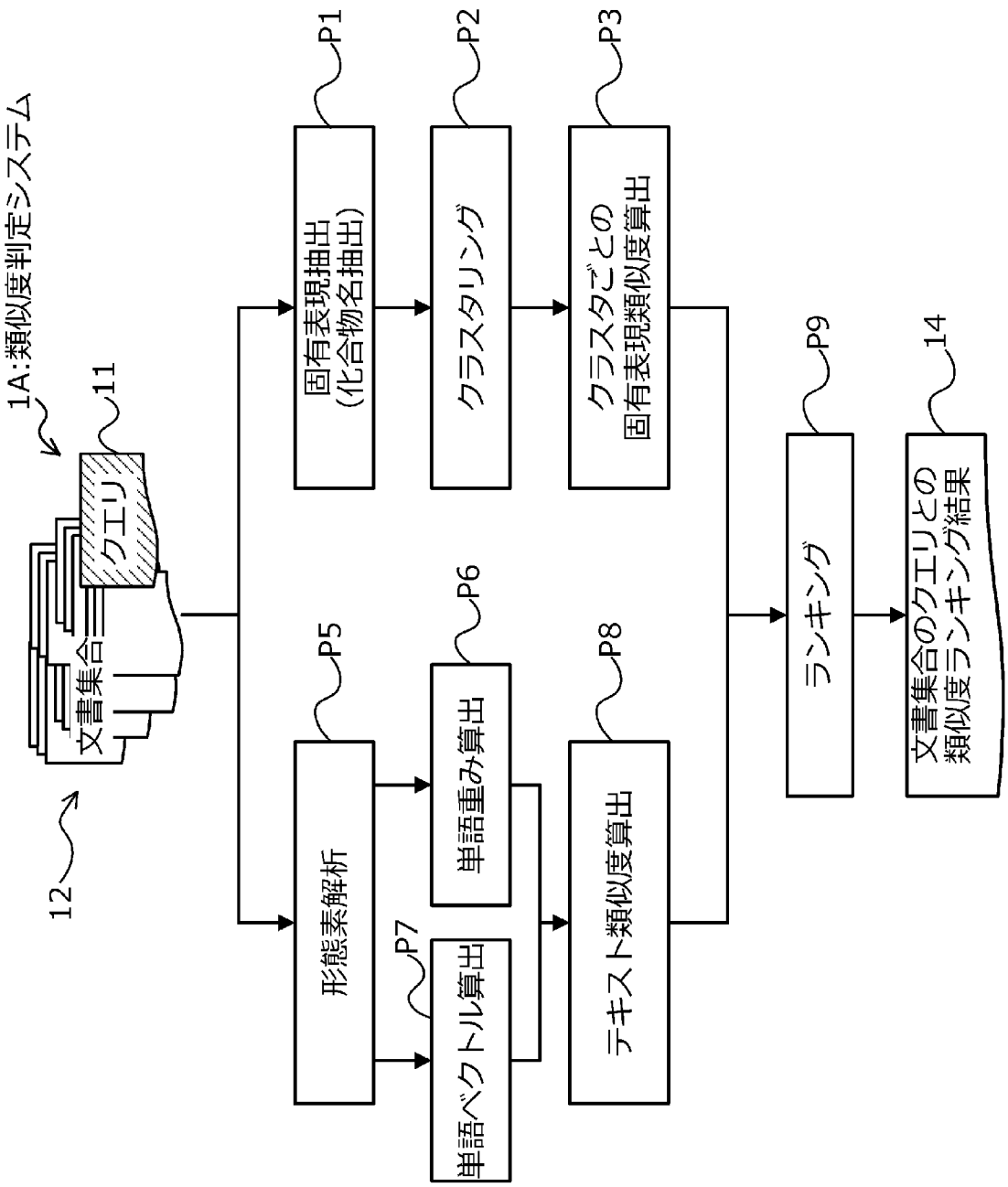
[図8]


類似度判定処理START
S1
クエリ文書の入力受け付け
S2
クエリ文書のテキストから
化合物名を抽出
S3
抽出した化合物名をクラスタリング
S4
比較対象文書の入力受け付け
S5
未選択の比較対象文書を選択
S6
選択した文書のテキストから
化合物名を抽出
S7
抽出した化合物名をクラスタリング
S8
クエリ文書及び比較対象文書の
化合物クラスタを比較し
文書類似度を算出
S9
算出した類似度を格納
S10
未選択の比較対象文書ある？
YES
NO
S11
類似度が高い順に
比較対象文書と類似度とを出力
END

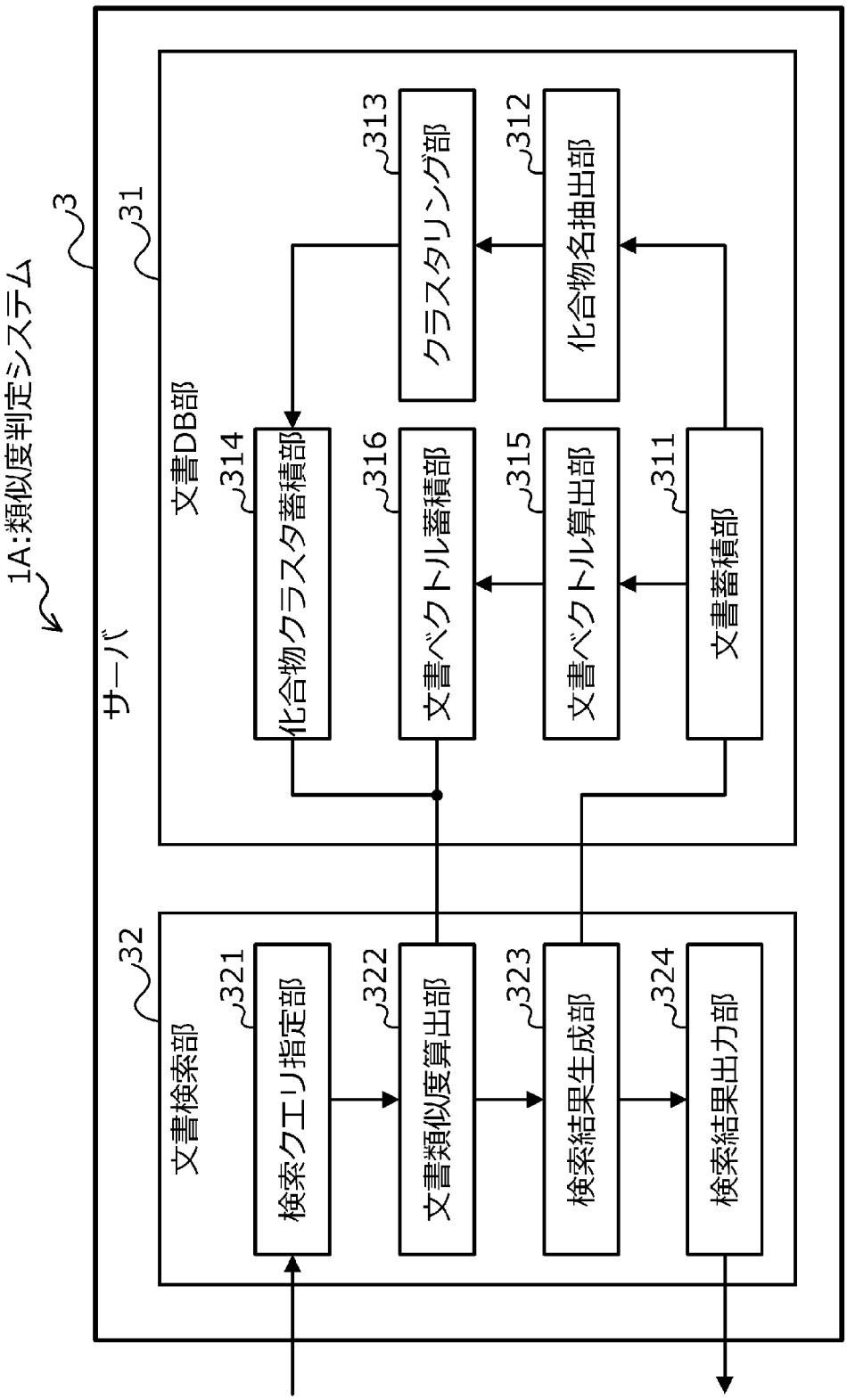
[図9]


10
コンピュータ
10a
プロセッサ
10c
記憶部
10e
I/O部
10i
10b
メモリ
10d
IF部
10f
読取部
10h
10c
記憶部
10g
プログラム

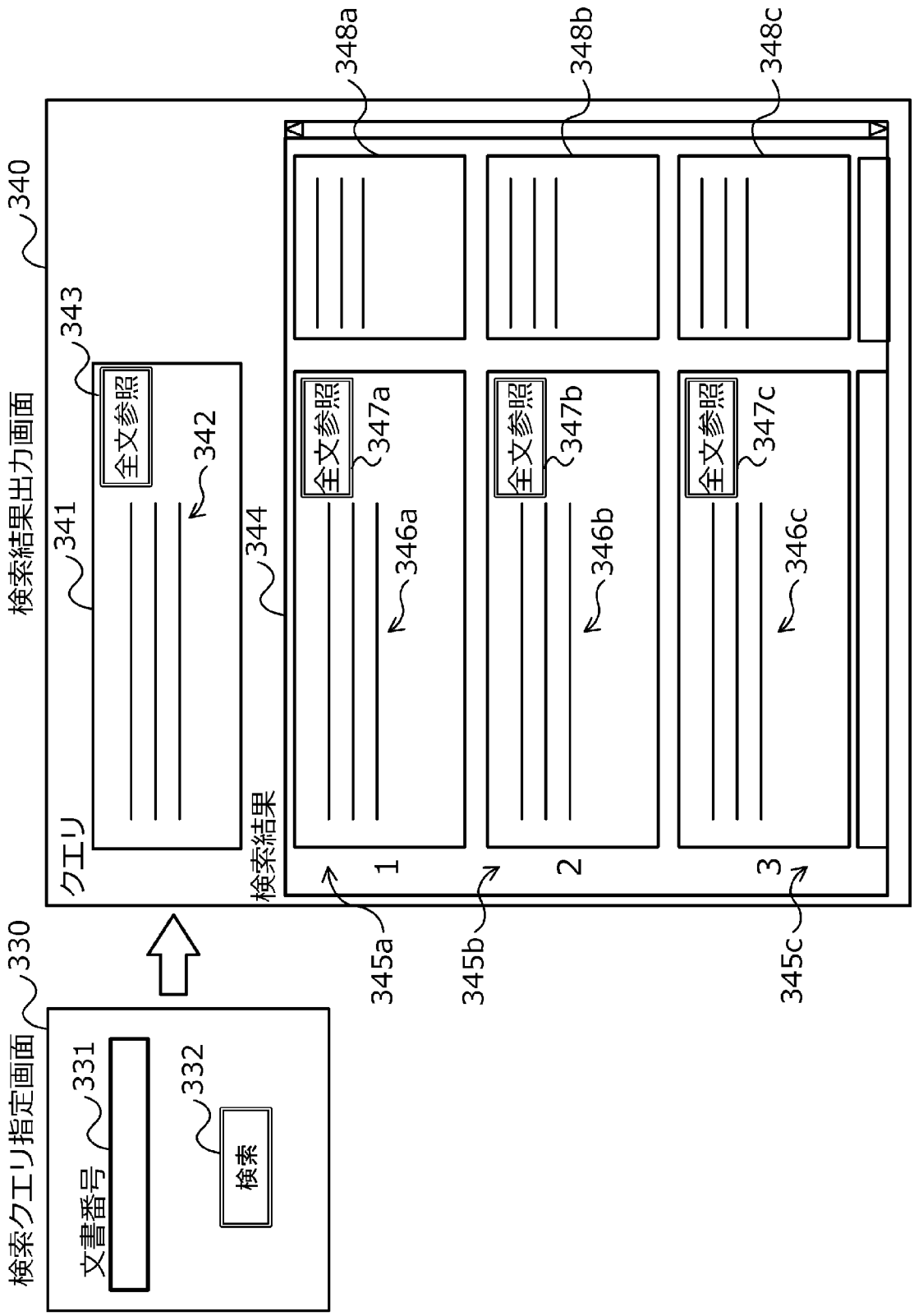
[図10]


1A:類似度判定システム
11
クエリ
文書集合
12
固有表現抽出
(化合物名抽出)
P1
クラスタリング
P2
クラスタごとの
固有表現類似度算出
P3
形態素解析
P5
単語重み算出
P6
単語ベクトル算出
P7
テキスト類似度算出
P8
ランキング
P9
文書集合のクエリとの
類似度ランキング結果
14

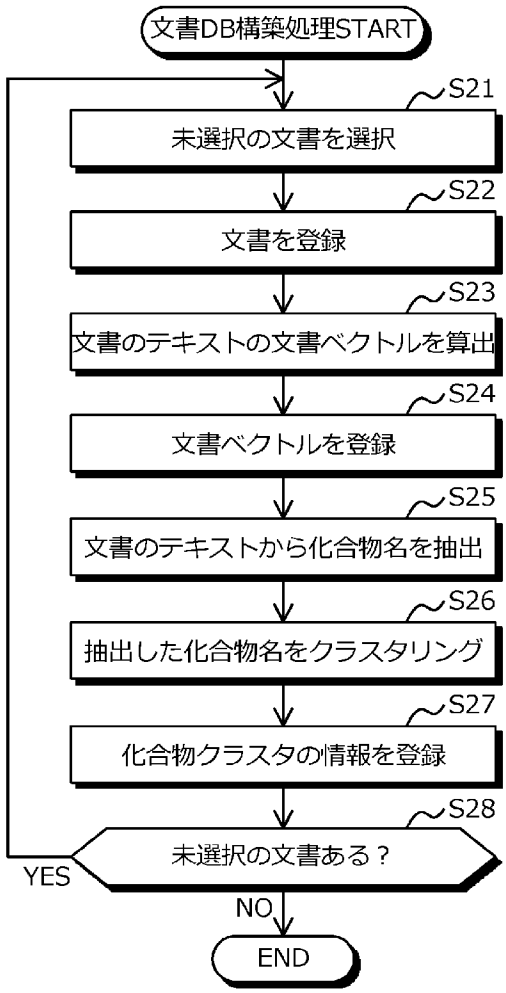
[図11]


1A:類似度判定システム
3
サーバ
31
文書DB部
313
クラスタリング部
312
化合物名抽出部
314
化合物クラスタ蓄積部
316
文書ベクトル蓄積部
315
文書ベクトル算出部
311
文書蓄積部
32
文書検索部
321
検索クエリ指定部
322
文書類似度算出部
323
検索結果生成部
324
検索結果出力部

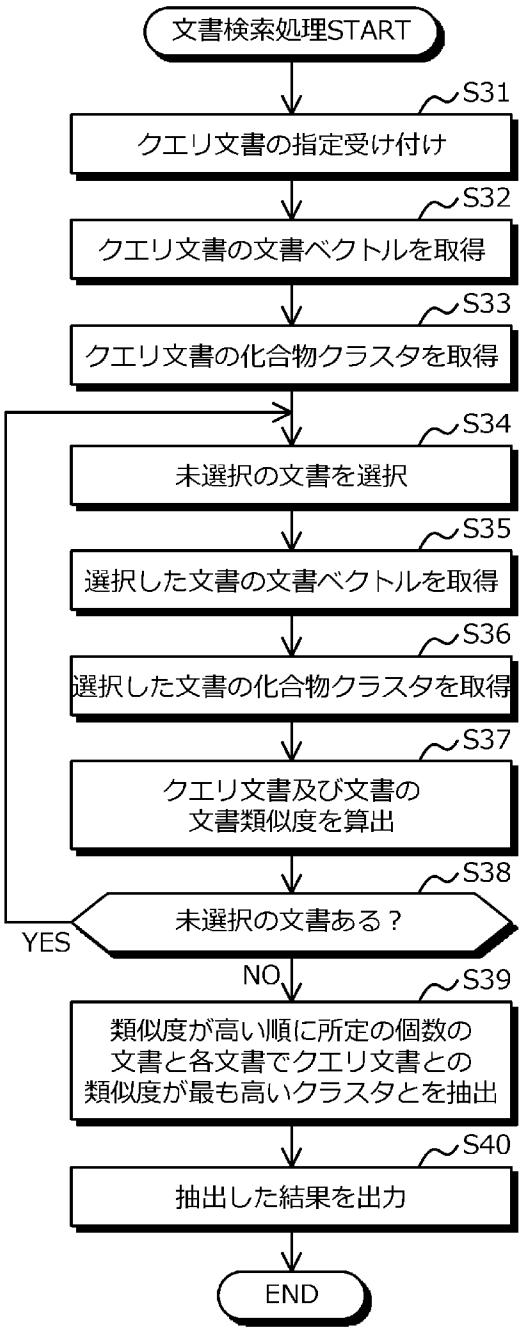
[図12]


検索クエリ指定画面
330
文書番号
331
検索
332
検索結果出力画面
340
クエリ
341
全文参照
342
343
検索結果
344
全文参照
347a
346a
1
345a
348a
全文参照
347b
346b
2
345b
348b
全文参照
347c
346c
3
345c
348c

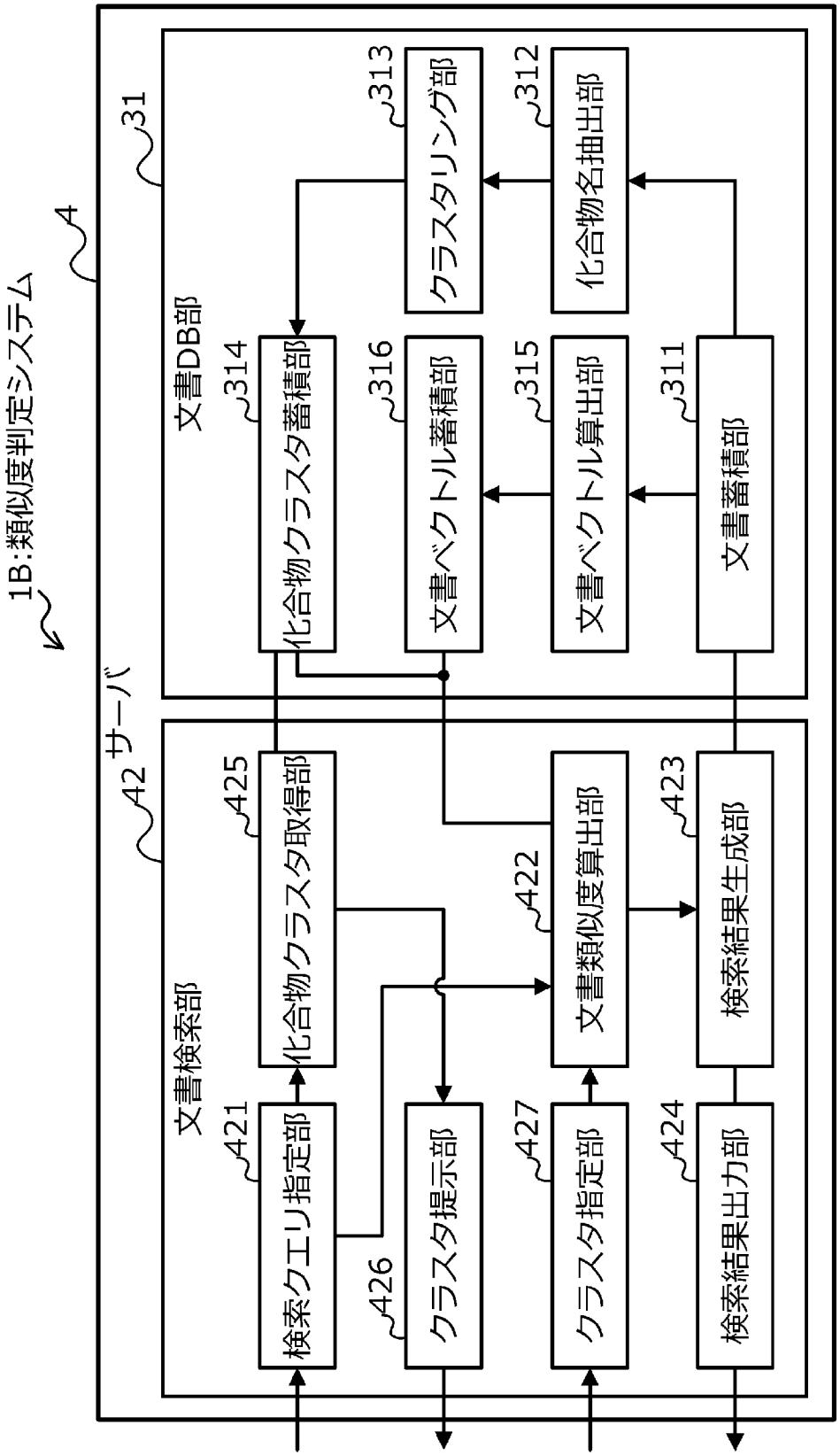
[図13]


文書DB構築処理START
S21
未選択の文書を選択
S22
文書を登録
S23
文書のテキストの文書ベクトルを算出
S24
文書ベクトルを登録
S25
文書のテキストから化合物名を抽出
S26
抽出した化合物名をクラスタリング
S27
化合物クラスタの情報を登録
S28
未選択の文書ある？
YES
NO
END

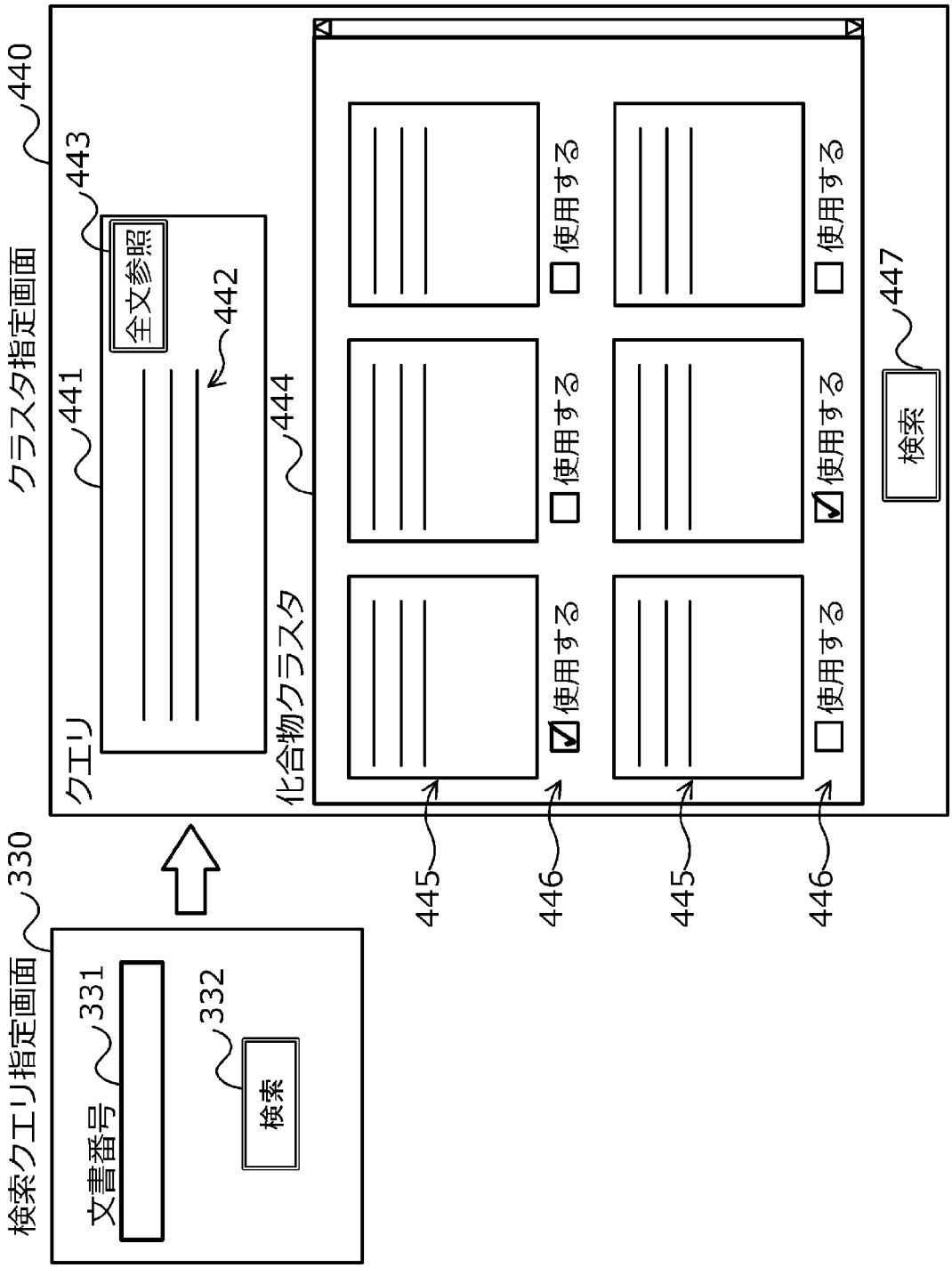
[図14]


文書検索処理START
S31
クエリ文書の指定受け付け
S32
クエリ文書の文書ベクトルを取得
S33
クエリ文書の化合物クラスタを取得
S34
未選択の文書を選択
S35
選択した文書の文書ベクトルを取得
S36
選択した文書の化合物クラスタを取得
S37
クエリ文書及び文書の
文書類似度を算出
S38
未選択の文書ある？
YES
NO
S39
類似度が高い順に所定の個数の
文書と各文書でクエリ文書との
類似度が最も高いクラスタとを抽出
S40
抽出した結果を出力
END

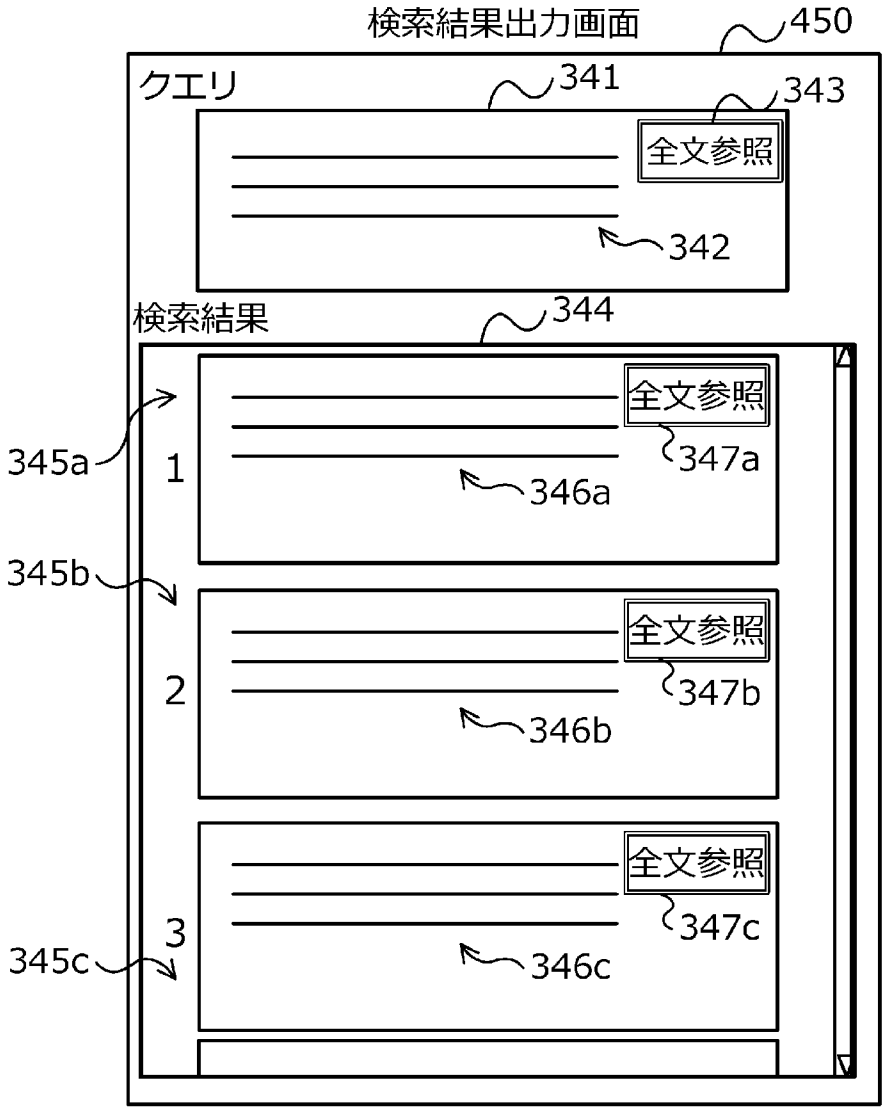
[図15]


1B:類似度判定システム
4
31
文書DB部
313
クラスタリング部
312
化合物名抽出部
314
化合物クラスタ蓄積部
316
文書ベクトル蓄積部
315
文書ベクトル算出部
311
文書蓄積部
42 サーバ
文書検索部
425
化合物クラスタ取得部
422
文書類似度算出部
423
検索結果生成部
421
検索クエリ指定部
426
クラスタ提示部
427
クラスタ指定部
424
検索結果出力部

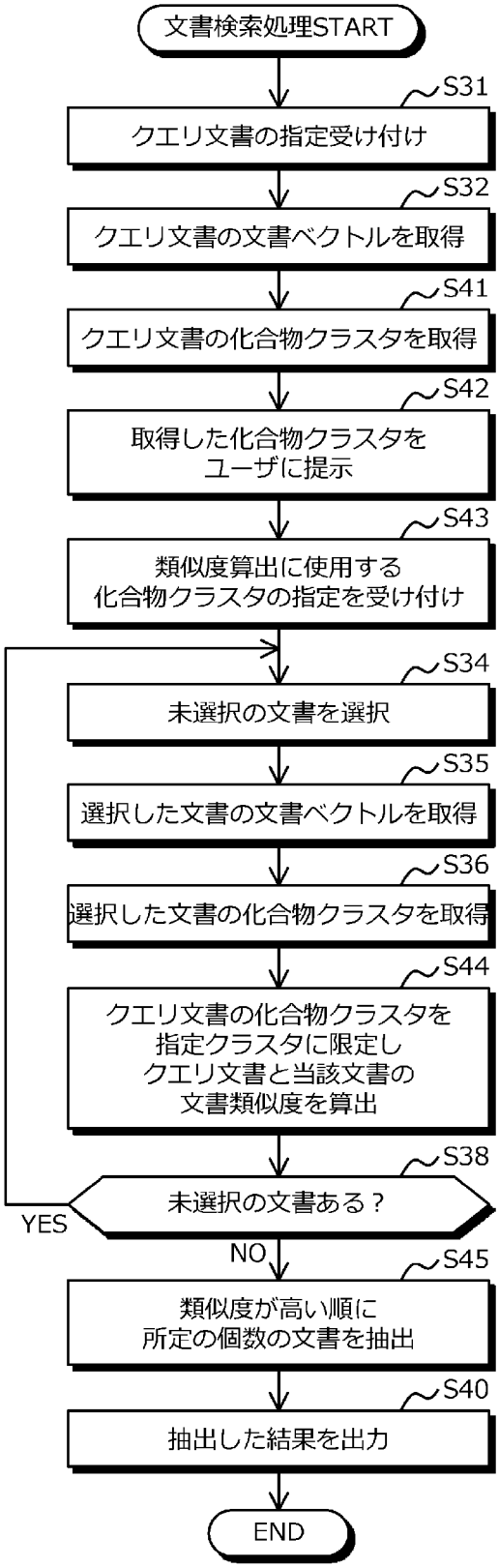
[図16]


検索クエリ指定画面
330
文書番号
331
検索
332
クラスタ指定画面
440
クエリ
441
全文参照
442
443
化合物クラスタ
444
使用する
445
446
447
検索

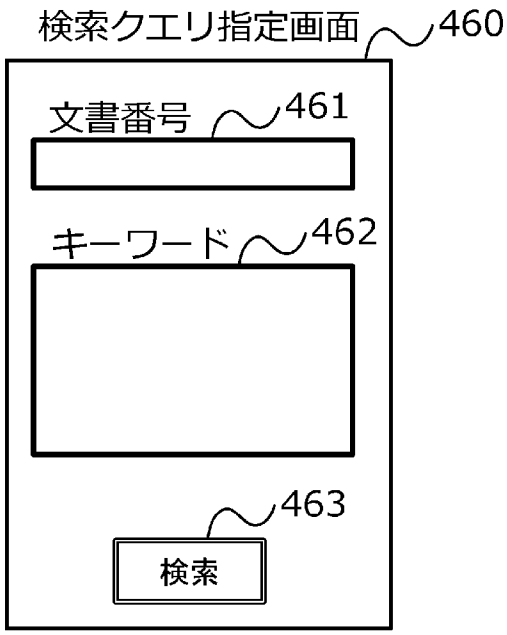
[図17]


検索結果出力画面
450
クエリ
341
343
全文参照
342
検索結果
344
345a
1
全文参照
347a
346a
345b
2
全文参照
347b
346b
3
全文参照
347c
345c
346c

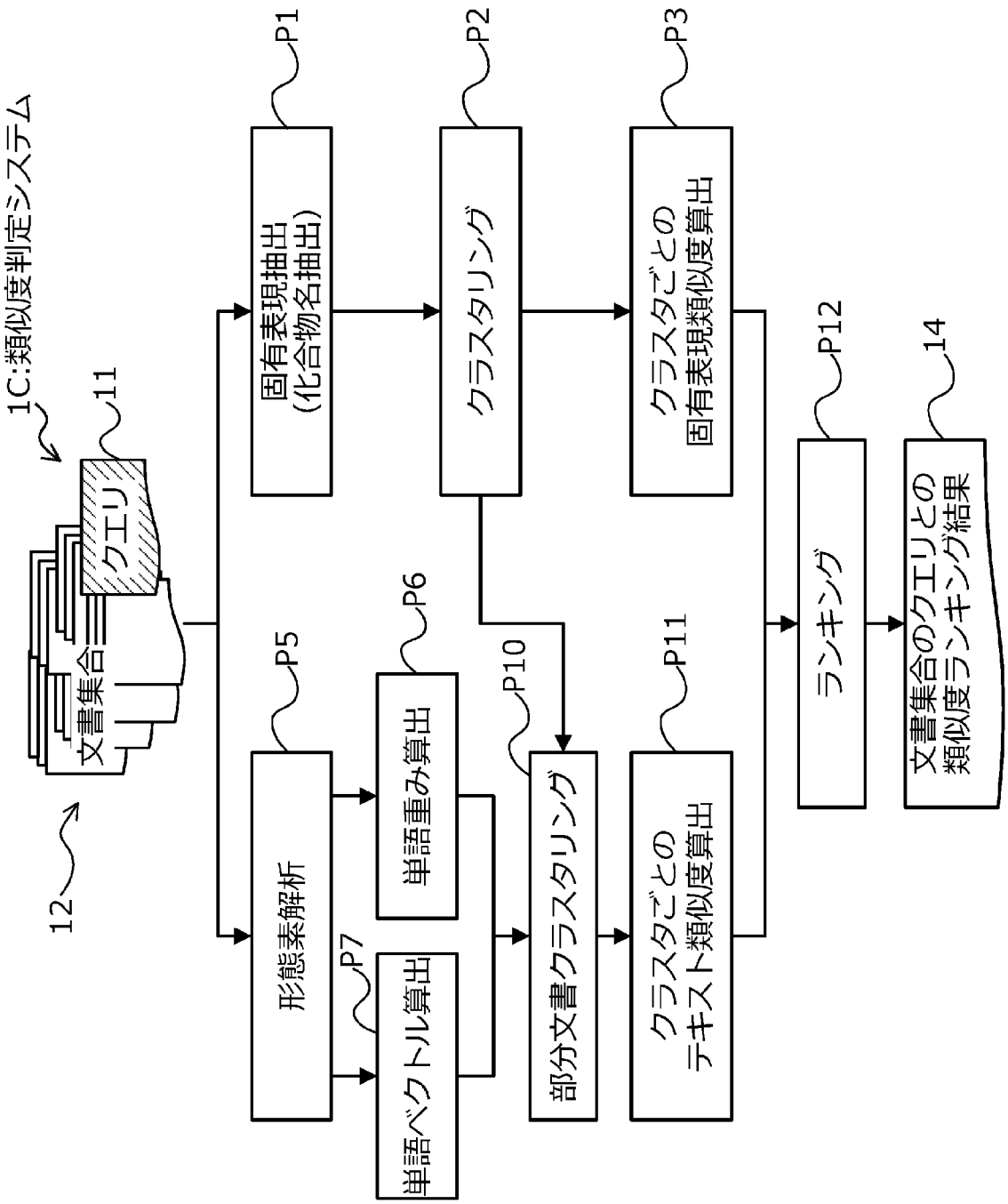
[図18]


文書検索処理START
S31
クエリ文書の指定受け付け
S32
クエリ文書の文書ベクトルを取得
S41
クエリ文書の化合物クラスタを取得
S42
取得した化合物クラスタを
ユーザに提示
S43
類似度算出に使用する
化合物クラスタの指定を受け付け
S34
未選択の文書を選択
S35
選択した文書の文書ベクトルを取得
S36
選択した文書の化合物クラスタを取得
S44
クエリ文書の化合物クラスタを
指定クラスタに限定し
クエリ文書と当該文書の
文書類似度を算出
S38
未選択の文書ある？
YES
NO
S45
類似度が高い順に
所定の個数の文書を抽出
S40
抽出した結果を出力
END

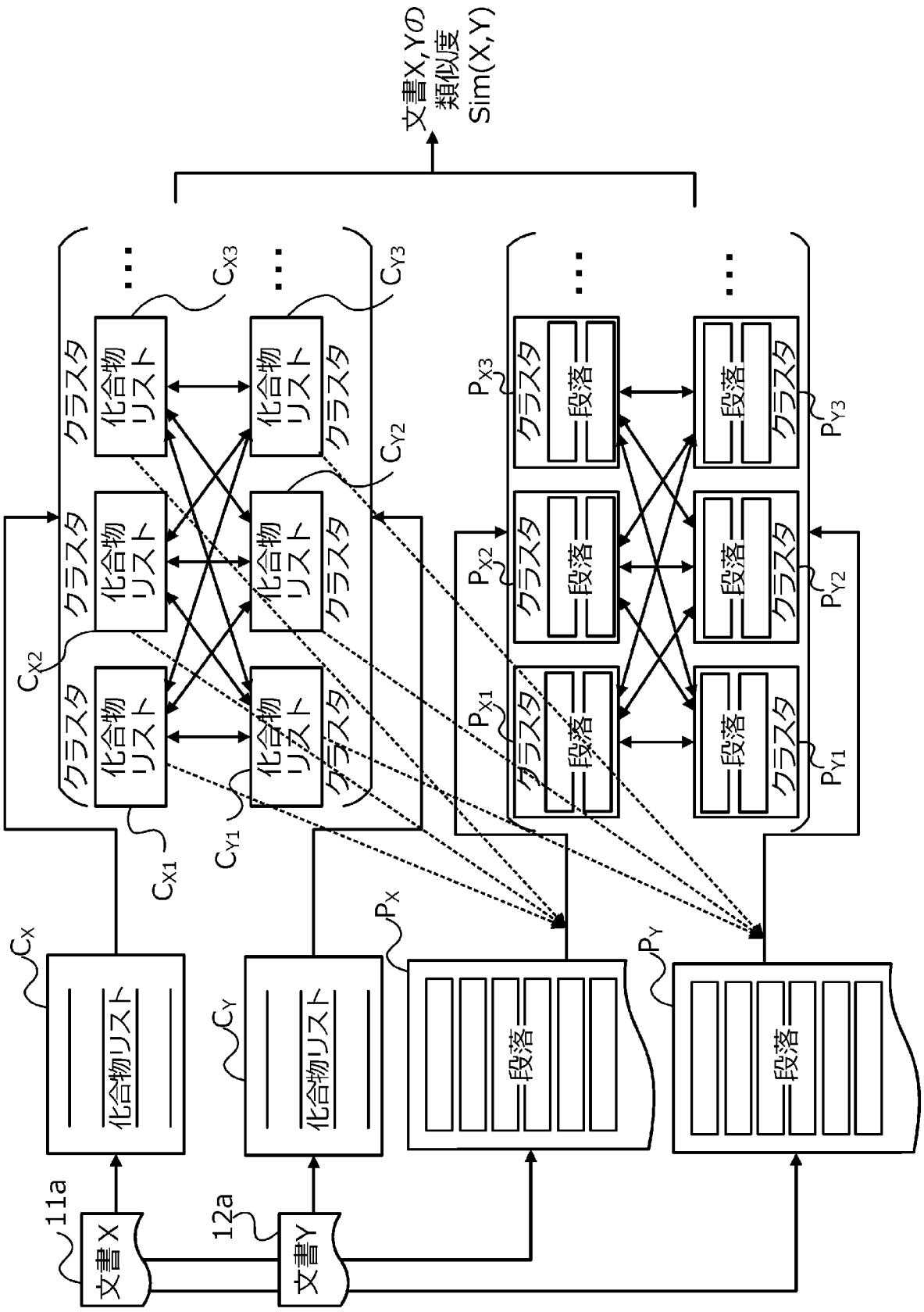
[図19]



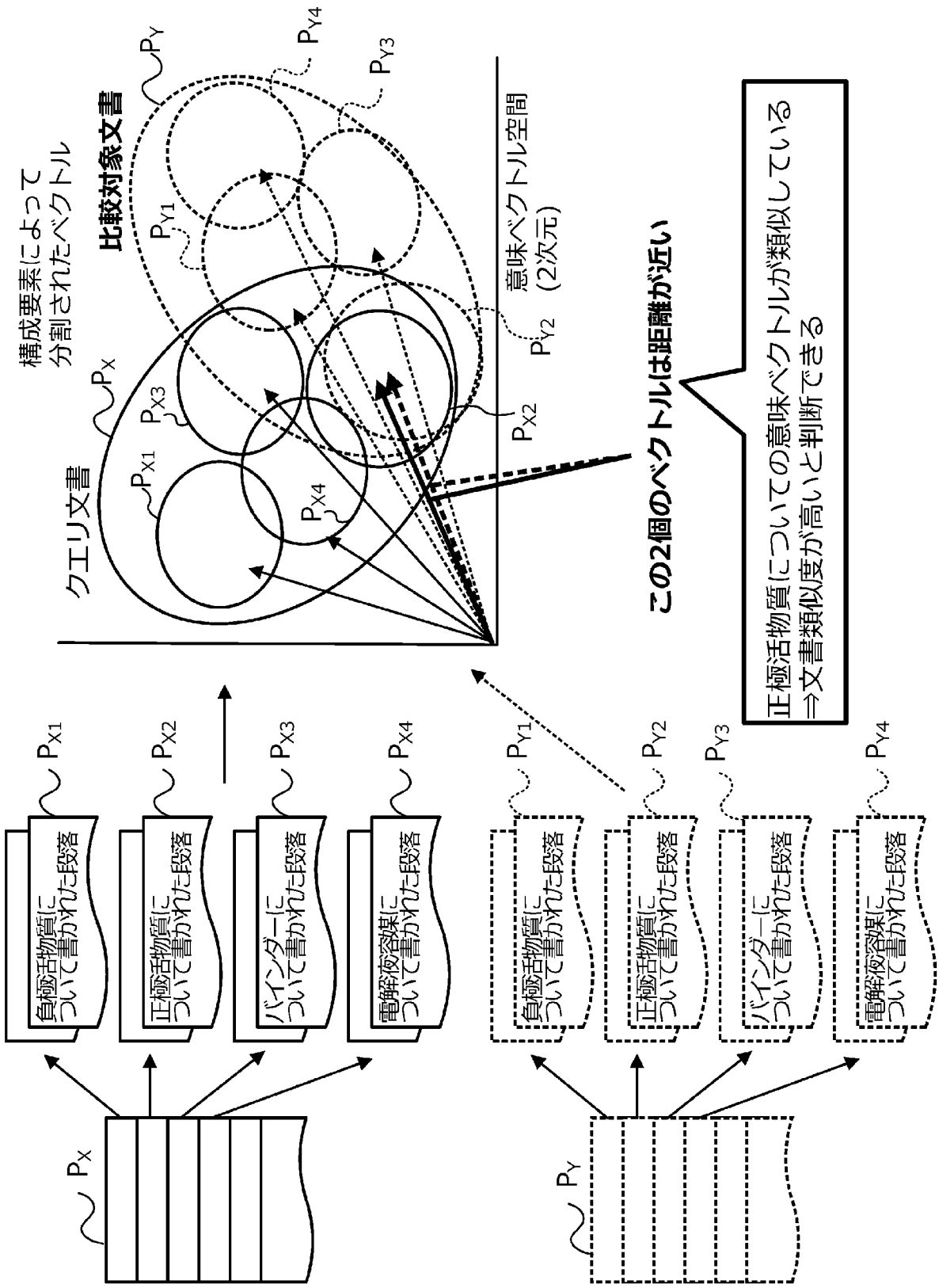
[図20]


1C:類似度判定システム
11
クエリ
文書集合
12
固有表現抽出
(化合物名抽出)
P1
クラスタリング
P2
クラスタごとの
固有表現類似度算出
P3
形態素解析
P5
単語重み算出
P6
単語ベクトル算出
P7
部分文書クラスタリング
P10
クラスタごとの
テキスト類似度算出
P11
ランキング
P12
文書集合のクエリとの
類似度ランキング結果
14

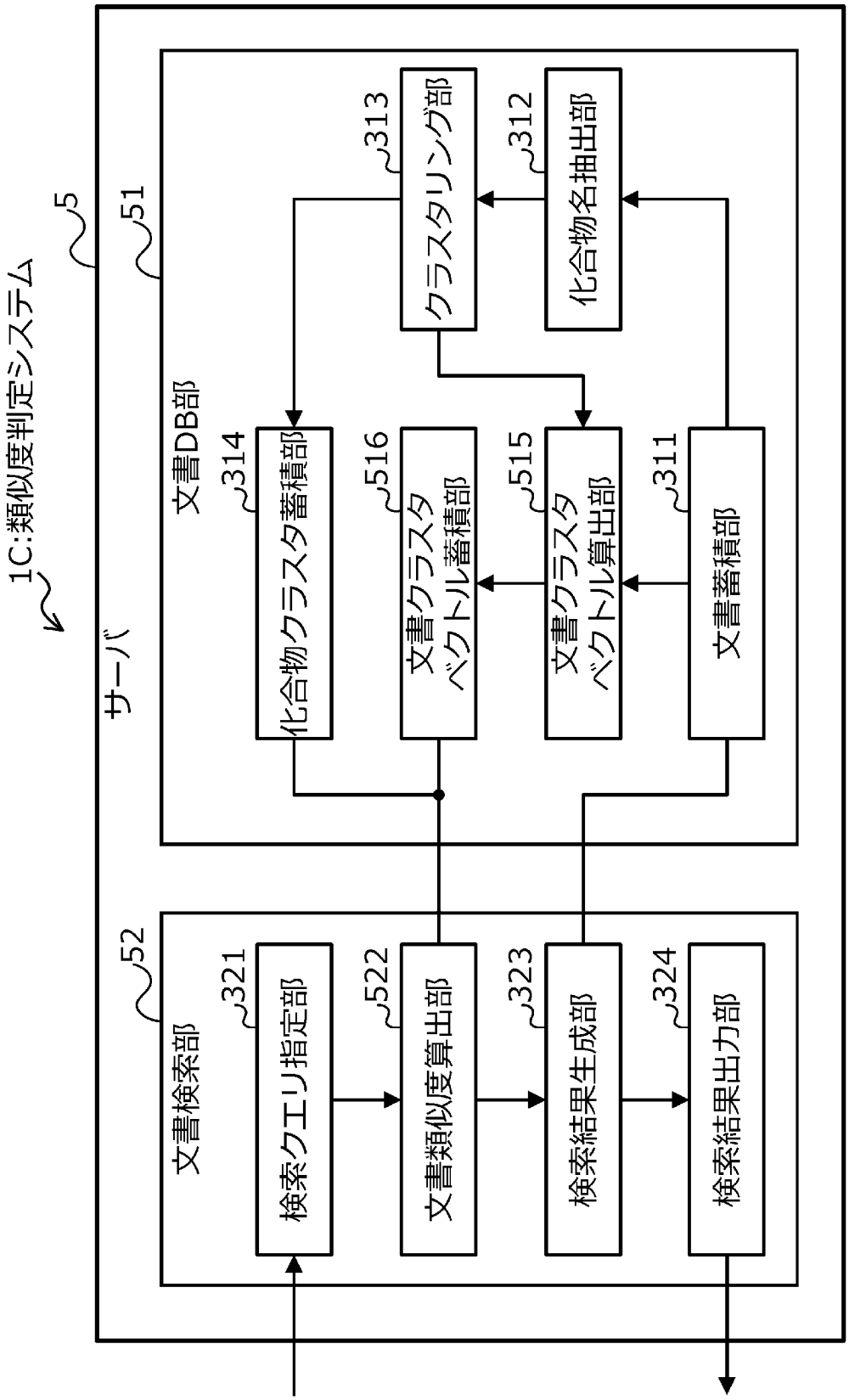
[図21]



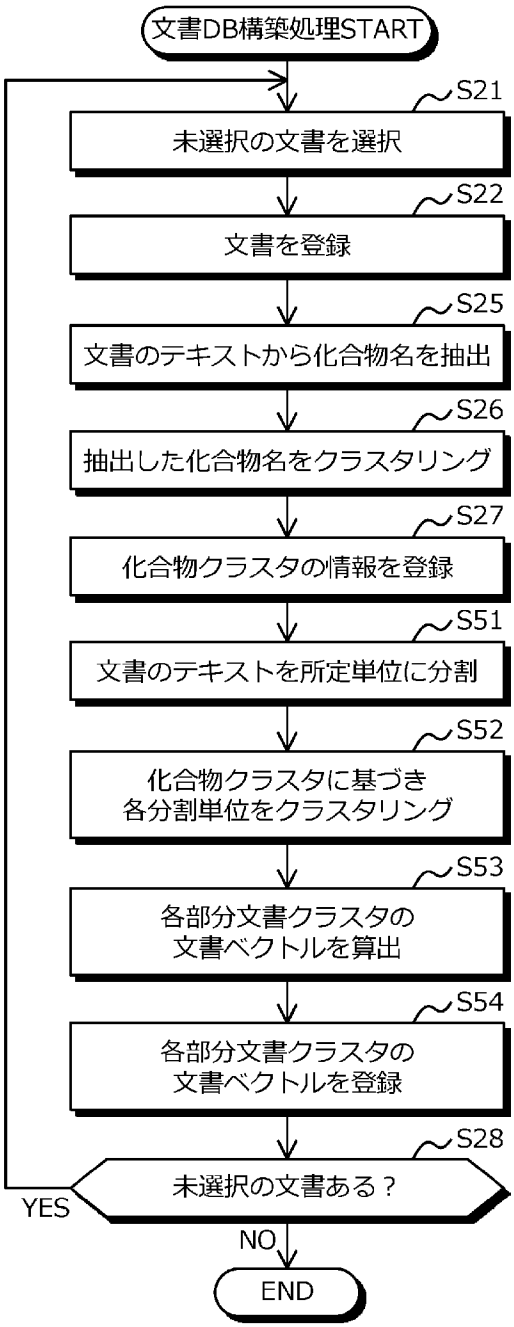
[図22]


構成要素によって
分割されたベクトル
比較対象文書
クエリ文書
意味ベクトル空間
(2次元)
この2個のベクトルは距離が近い
正極活物質についての意味ベクトルが類似している
⇒文書類似度が高いと判断できる
負極活物質について書かれた段落
正極活物質について書かれた段落
バインダーについて書かれた段落
電解液溶媒について書かれた段落

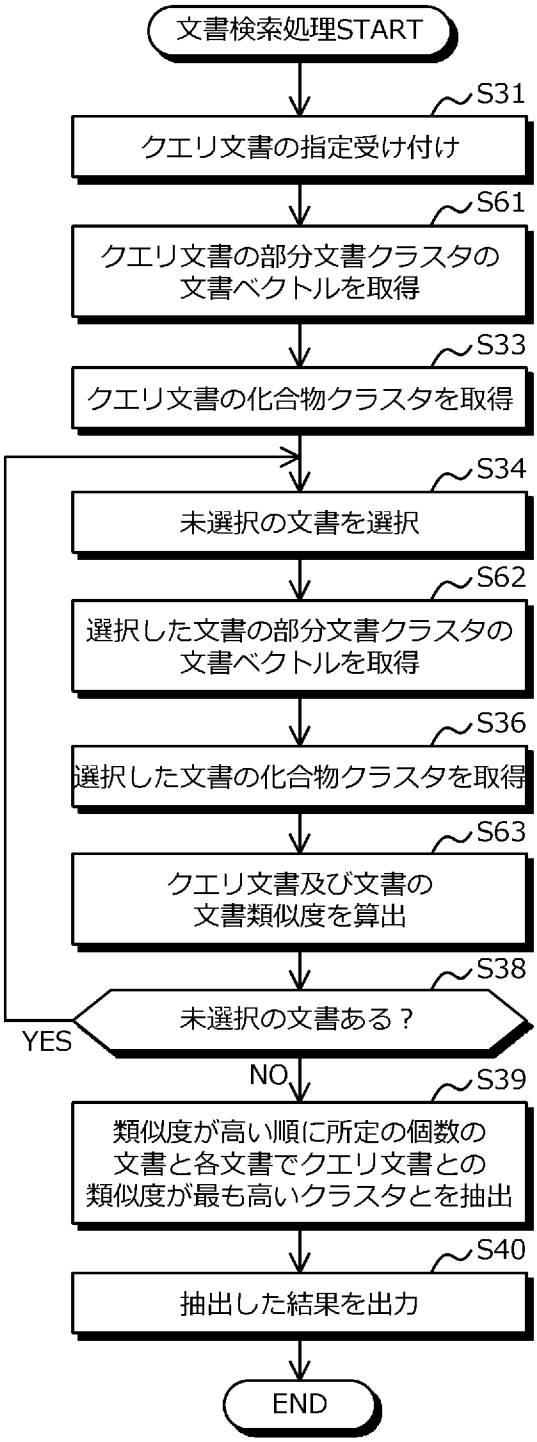
[図23]


1C:類似度判定システム
5
サーバ
51
文書DB部
313
クラスタリング部
312
化合物名抽出部
314
化合物クラスタ蓄積部
516
文書クラスタ
ベクトル蓄積部
515
文書クラスタ
ベクトル算出部
311
文書蓄積部
52
文書検索部
321
検索クエリ指定部
522
文書類似度算出部
323
検索結果生成部
324
検索結果出力部

[図24]


文書DB構築処理START
S21
未選択の文書を選択
S22
文書を登録
S25
文書のテキストから化合物名を抽出
S26
抽出した化合物名をクラスタリング
S27
化合物クラスタの情報を登録
S51
文書のテキストを所定単位に分割
S52
化合物クラスタに基づき
各分割単位をクラスタリング
S53
各部分文書クラスタの
文書ベクトルを算出
S54
各部分文書クラスタの
文書ベクトルを登録
S28
未選択の文書ある？
YES
NO
END

[図25]


文書検索処理START
S31
クエリ文書の指定受け付け
S61
クエリ文書の部分文書クラスタの
文書ベクトルを取得
S33
クエリ文書の化合物クラスタを取得
S34
未選択の文書を選択
S62
選択した文書の部分文書クラスタの
文書ベクトルを取得
S36
選択した文書の化合物クラスタを取得
S63
クエリ文書及び文書の
文書類似度を算出
S38
未選択の文書ある？
YES
NO
S39
類似度が高い順に所定の個数の
文書と各文書でクエリ文書との
類似度が最も高いクラスタとを抽出
S40
抽出した結果を出力
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2020/047218

A. CLASSIFICATION OF SUBJECT MATTER
G06F 16/35(2019.01)i
FI: G06F16/35
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)
G06F16/00-16/958

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	1971-2021
Registered utility model specifications of Japan	1996-2021
Published registered utility model applications of Japan	1994-2021

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 2008-9671 A (NATIONAL INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY) 17 January 2008 (2008-01-17) entire text, all drawings	1-20
A	JP 2002-259411 A (NEC CORP.) 13 September 2002 (2002-09-13) entire text, all drawings	1-20
A	JP 2013-20431 A (NEC CORP.) 31 January 2013 (2013-01-31) entire text, all drawings	1-20
A	JP 2000-112949 A (FUJI XEROX CO., LTD.) 21 April 2000 (2000-04-21) entire text, all drawings	1-20
A	JP 2016-45552 A (FUJITSU LTD.) 04 April 2016 (2016-04-04) paragraphs [0015]-[0016], [0057]	1-20
A	JP 11-272680 A (FUJITSU LTD.) 08 October 1999 (1999-10-08) paragraphs [0009]-[0010]	1-20

☐ 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
26 February 2021 (26.02.2021)	09 March 2021 (09.03.2021)

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
Information on patent family members

International application no.
PCT/JP2020/047218

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
JP 2008-9671 A	17 Jan. 2008	(Family: none)	
JP 2002-259411 A	13 Sep. 2002	(Family: none)	
JP 2013-20431 A	31 Jan. 2013	(Family: none)	
JP 2000-112949 A	21 Apr. 2000	(Family: none)	
JP 2016-45552 A	04 Apr. 2016	(Family: none)	
JP 11-272680 A	08 Oct. 1999	(Family: none)	

国際調査報告

国際出願番号
PCT/JP2020/047218

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））

G06F 16/35(2019.01)i
FI: G06F16/35

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））

G06F16/00-16/958

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	１９７１－２０２１年
日本国実用新案登録公報	１９９６－２０２１年
日本国登録実用新案公報	１９９４－２０２１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 2008-9671 A (独立行政法人情報通信研究機構) 17.01.2008 (2008-01-17) 全文、全図	1-20
A	JP 2002-259411 A (日本電気株式会社) 13.09.2002 (2002-09-13) 全文、全図	1-20
A	JP 2013-20431 A (日本電気株式会社) 31.01.2013 (2013-01-31) 全文、全図	1-20
A	JP 2000-112949 A (富士ゼロックス株式会社) 21.04.2000 (2000-04-21) 全文、全図	1-20
A	JP 2016-45552 A (富士通株式会社) 04.04.2016 (2016-04-04) 段落[0015]-[0016], [0057]	1-20
A	JP 11-272680 A (富士通株式会社) 08.10.1999 (1999-10-08) 段落[0009]-[0010]	1-20

☐ 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抵触するものではなく、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
26.02.2021	09.03.2021

名称及びあて先

日本国特許庁(ISA/JP)
〒100-8915
日本国
東京都千代田区霞が関三丁目４番３号

権限のある職員（特許庁審査官）

▲はま▼中　信行 5N 3240

電話番号 03-3581-1101 内線 3586

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2020/047218

引用文献	公表日	パテントファミリー文献	公表日
JP 2008-9671 A	17.01.2008	（ファミリーなし）	
JP 2002-259411 A	13.09.2002	（ファミリーなし）	
JP 2013-20431 A	31.01.2013	（ファミリーなし）	
JP 2000-112949 A	21.04.2000	（ファミリーなし）	
JP 2016-45552 A	04.04.2016	（ファミリーなし）	
JP 11-272680 A	08.10.1999	（ファミリーなし）	