(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4907927号
(P4907927)
(45)発行日　平成24年4月4日(2012.4.4)　(24)登録日　平成24年1月20日(2012.1.20)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 320C

請求項の数 3　(全 19 頁)

(21)出願番号　特願2005-266409(P2005-266409)
(22)出願日　平成17年9月14日(2005.9.14)
(65)公開番号　特開2007-79898(P2007-79898A)
(43)公開日　平成19年3月29日(2007.3.29)
審査請求日　平成20年8月1日(2008.8.1)

特許法第３０条第１項適用　平成１７年３月１５日　言語処理学会の中西印刷株式会社発行の「言語処理学会第１１回年次大会発表論文集」に発表

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100094662
弁理士　穂坂　和雄
(74)代理人　100096530
弁理士　今村　辰夫
(74)代理人　100119161
弁理士　重久　啓子
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　一井　康二
神奈川県横須賀市久里浜６－１０－２－４０５

最終頁に続く

(54)【発明の名称】データ表示装置、データ表示方法およびデータ表示プログラム

(57)【特許請求の範囲】
【請求項１】
キーワードに関するデータを表示するデータ表示装置であって、
複数のキーワードが入力キーワードとして入力されるキーワード入力手段と、
前記入力キーワードに基づいて、前記入力キーワードと同じ分野のキーワードを含む一定量のキーワード抽出用の文書データを格納したデータベースから抽出することで、前記入力キーワードの数より多いキーワードを抽出し、キーワードの総数を増加させるキーワード増加手段と、
前記出力された各キーワードに関するデータを表示データとして作成する表示データ作成手段と、
前記作成された表示データを画面表示するデータ表示手段とを備えると共に、
前記キーワード増加手段は、
前記入力キーワードを前記データベースで全文検索し、検索結果において前記入力キーワードの直前及び直後の文字列をパターンとして抽出するパターン抽出手段と、
前記パターン抽出手段で抽出したパターンを前記データベースで全文検索し、該パターンによって抽出される表現を抽出すると同時に、前記パターンで抽出される表現での前記入力キーワードの割合（p_i）によりスコアを算出し、前記抽出した表現を該スコアの大きい順にソートして、キーワードとして出力するキーワード抽出手段とを備える
ことを特徴とするデータ表示装置。
【請求項２】

キーワードに関するデータを表示するデータ表示方法であって、
複数のキーワードが入力キーワードとして入力するステップと、
前記入力キーワードに基づいて、前記入力キーワードと同じ分野のキーワードを含む一定量のキーワード抽出用の文書データを格納したデータベースから抽出することで、前記入力キーワードの数より多いキーワードを抽出し、キーワードの総数を増加させるステップと、
前記出力された各キーワードに関するデータを表示データとして作成するステップと、
前記作成された表示データを画面表示するステップとを有すると共に、
前記キーワードを増加させるステップは、
前記入力キーワードを前記データベースで全文検索し、検索結果において前記入力キーワードの直前及び直後の文字列をパターンとして抽出するステップと、
前記パターン抽出ステップで抽出したパターンを前記データベースで全文検索し、該パターンによって抽出される表現を抽出すると同時に、前記パターンで抽出される表現での前記入力キーワードの割合（p_i）によりスコアを算出し、前記抽出した表現を該スコアの大きい順にソートして、キーワードとして出力するステップとを有することを特徴とするデータ表示方法。

【請求項３】
キーワードに関するデータを表示するデータ表示装置が備えるコンピュータに実行させるためのプログラムであって、
前記コンピュータを、
複数のキーワードが入力キーワードとして入力されるキーワード入力手段と、
前記入力キーワードに基づいて、前記入力キーワードと同じ分野のキーワードを含む一定量のキーワード抽出用の文書データを格納したデータベースから抽出することで、前記入力キーワードの数より多いキーワードを抽出し、キーワードの総数を増加させるキーワード増加手段と、
前記出力された各キーワードに関するデータを表示データとして作成する表示データ作成手段と、
前記作成された表示データを画面表示するデータ表示手段と、
前記キーワード増加手段が備える、前記入力キーワードを前記データベースで全文検索し、検索結果において前記入力キーワードの直前及び直後の文字列をパターンとして抽出するパターン抽出手段と、
前記パターン抽出手段で抽出したパターンを前記データベースで全文検索し、該パターンによって抽出される表現を抽出すると同時に、前記パターンで抽出される表現での前記入力キーワードの割合（p_i）によりスコアを算出し、前記抽出した表現を該スコアの大きい順にソートして、キーワードとして出力するキーワード抽出手段として機能させるためのデータ表示プログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ表示技術に関し、特に、入力されたキーワードをキーワード抽出技術を用いて増加させた上で、増加したキーワードに関する数値データを表示するデータ表示装置、データ表示方法およびデータ表示プログラムに関する。より具体的には、本発明は、入力されたキーワードをキーワード抽出技術を用いて増加させた上で、増加後のキーワードを含む文書データの各年次の発表件数のデータ（年次発表データ）を画面表示する。
【背景技術】
【０００２】
大学、企業等の各研究機関は、有用な研究について、年次大会や論文誌において毎年文書の発表を行っている。
【０００３】
ここで、下記の非特許文献１に記載されている、入力されたデータを表形式で表示する

技術を用いれば、各キーワード（例えば、各研究機関や各研究分野）を含む文書の各年次の発表件数のデータ（年次発表データ）を表形式で表示することができる（非特許文献１参照）。

【０００４】

入力されたあるキーワードを含む文書の発表件数のデータを表形式で表示することは、従来から可能であった。

【非特許文献１】知りたい操作がすぐわかる　標準 Excel全機能Bible 2003，村田吉徳著，技術評論社，2004.2.1発行

【発明の開示】

【発明が解決しようとする課題】

【０００５】

しかし、従来技術では、入力されたキーワード以外のキーワードを含む文書についての年次発表データを表示することができないという問題があった。

【０００６】

例えば、従来技術では、キーワードを入力するユーザが思い付く数のキーワードについてしか、年次発表データを表示することができなかった。

【０００７】

本発明は、上記従来技術の問題点を解決し、入力されたキーワードに関するデータ（例えば数値データ）と、入力されたキーワード以外のキーワードに関するデータ（例えば、数値データ）とを表示するデータ表示装置、データ表示方法およびデータ表示プログラムの提供を目的とする。より具体的には、本発明は、例えば、入力されたキーワードを含む文書の年次発表データと入力されたキーワード以外のキーワードを含む文書の年次発表データとを表示することを目的とする。

【課題を解決するための手段】

【０００８】

前記課題を解決するため、本発明は、次のように構成した。

(1)：キーワードに関するデータを表示するデータ表示装置であって、複数のキーワードが入力キーワードとして入力されるキーワード入力手段と、前記入力キーワードに基づいて、前記入力キーワードと同じ分野のキーワードを含む一定量のキーワード抽出用の文書データを格納したデータベースから抽出することで、前記入力キーワードの数より多いキーワードを抽出し、キーワードの総数を増加させるキーワード増加手段と、前記出力された各キーワードに関するデータを表示データとして作成する表示データ作成手段と、前記作成された表示データを画面表示するデータ表示手段とを備えると共に、前記キーワード増加手段は、前記入力キーワードを前記データベースで全文検索し、検索結果において前記入力キーワードの直前及び直後の文字列をパターンとして抽出するパターン抽出手段と、前記パターン抽出手段で抽出したパターンを前記データベースで全文検索し、該パターンによって抽出される表現を抽出すると同時に、前記パターンで抽出される表現での前記入力キーワードの割合（p_i）によりスコアを算出し、前記抽出した表現を該スコアの大きい順にソートして、キーワードとして出力するキーワード抽出手段とを備えることを特徴とする。

【００１２】

(2)：キーワードに関するデータを表示するデータ表示方法であって、複数のキーワードが入力キーワードとして入力するステップと、前記入力キーワードに基づいて、前記入力キーワードと同じ分野のキーワードを含む一定量のキーワード抽出用の文書データを格納したデータベースから抽出することで、前記入力キーワードの数より多いキーワードを抽出し、キーワードの総数を増加させるステップと、前記出力された各キーワードに関するデータを表示データとして作成するステップと、前記作成された表示データを画面表示するステップとを有すると共に、前記キーワードを増加させるステップは、前記入力キーワードを前記データベースで全文検索し、検索結果において前記入力キーワードの直前及び直後の文字列をパターンとして抽出するステップと、前記パターン抽出ステップで抽出

したパターンを前記データベースで全文検索し、該パターンによって抽出される表現を抽出すると同時に、前記パターンで抽出される表現での前記入力キーワードの割合（p_i）によりスコアを算出し、前記抽出した表現を該スコアの大きい順にソートして、キーワードとして出力するステップとを有することを特徴とする。

【００１３】

(3)：キーワードに関するデータを表示するデータ表示装置が備えるコンピュータに実行させるためのプログラムであって、前記コンピュータを、複数のキーワードが入力キーワードとして入力されるキーワード入力手段と、前記入力キーワードに基づいて、前記入力キーワードと同じ分野のキーワードを含む一定量のキーワード抽出用の文書データを格納したデータベースから抽出することで、前記入力キーワードの数より多いキーワードを抽出し、キーワードの総数を増加させるキーワード増加手段と、前記出力された各キーワードに関するデータを表示データとして作成する表示データ作成手段と、前記作成された表示データを画面表示するデータ表示手段と、前記キーワード増加手段が備える、前記入力キーワードを前記データベースで全文検索し、検索結果において前記入力キーワードの直前及び直後の文字列をパターンとして抽出するパターン抽出手段と、前記パターン抽出手段で抽出したパターンを前記データベースで全文検索し、該パターンによって抽出される表現を抽出すると同時に、前記パターンで抽出される表現での前記入力キーワードの割合（p_i）によりスコアを算出し、前記抽出した表現を該スコアの大きい順にソートして、キーワードとして出力するキーワード抽出手段として機能させるためのデータ表示プログラムであることを特徴とする。

【発明の効果】

【００１８】

本発明のデータ表示装置は、入力されたキーワードに基づいて、キーワードの総数を増加させた上で、増加後のキーワードに関するデータを画面表示する。より具体的には、本発明のデータ表示装置は、増加後の各キーワードを含む文書についての年次発表データを画面表示する。

【００１９】

従って、本発明によれば、例えば、ユーザは、思い付く少数のキーワードを入力するだけで、自分が入力したキーワード以外のキーワードを含む文書の発表件数の推移を知ることができる。

【発明を実施するための最良の形態】

【００２０】

以下に、図を用いて、本発明の実施の形態について説明する。図１は、本発明の実施の形態におけるシステム構成の一例を示す図である。データ表示装置１は、キーワードに関するデータを表示する処理装置である。データ表示装置１は、キーワード入力部１１、キーワード増加部１２、表示データ作成部１３、データ表示部１４、キーワード抽出用データベース（ＤＢ）１５を備える。また、図中、１６は大量の文書データ（書誌データ）が蓄積されている書誌データＤＢである。書誌データＤＢ１６に格納されている書誌データとしては、例えば、図２に示すような、文書のタイトル、文書のテキスト内容、発表年次について記述されたデータが挙げられる。

【００２１】

キーワード入力部１１には、複数の少数のキーワードが入力される。キーワードとしては、例えば、研究機関名や研究分野等、文書中に一般に含まれる任意の用語が挙げられる。キーワード増加部１２は、後述するキーワード抽出技術を用いて、入力されたキーワードと同じ分野のキーワードをキーワード抽出用ＤＢ１５から抽出する。キーワードの抽出の結果、キーワードの総数が増加する。

【００２２】

表示データ作成部１３は、増加した各キーワードに関するデータを表示データとして作成する。例えば、増加した各キーワードに関する数値データを表示データとして作成する。より具体的には、表示データ作成部１３は、増加した各キーワードと、書誌データＤＢ

１６中の書誌データとに基づいて、各キーワードをタイトルに含む文書の、各年次の発表件数をカウントして、年次発表データを作成し、作成した年次発表データを表示対象のデータ（表示データ）とする。

【００２３】

なお、表示データ作成部１３は、例えば、上記年次発表データを処理して等高線データに変換し、変換後の等高線データを表示データとする構成をとることもできる。また、例えば、表示データ作成部１３は、例えば、上記年次発表データに基づいて、後述するバブルチャート上に画面表示されるデータを表示データとして作成する構成を採ることもできる。

【００２４】

また、本発明においては、表示データ作成部１３が作成する表示データは、数値データに限られない。例えば、表示データ作成部１３は、書誌データＤＢ１６中の書誌データ中において、増加した各キーワードと共起して出現する回数が高い言語表現を表示データとして作成する構成を採ることもできる。また、例えば、増加した各キーワードによって構成される質問に対する解答を表示データとして作成する構成を採ることもできる。

【００２５】

データ表示部１４は、表示データ作成部１３によって作成された表示データを画面表示する。キーワード抽出用ＤＢ１５は、一定量の文書データを格納したデータベースである。キーワード抽出用ＤＢ１５は、例えば、新聞、雑誌、Ｗｅｂデータ（ネットワーク上のデータ）等から抽出したデータ（一定量の文書データ）を格納している。

【００２６】

キーワード増加部１２は、パターン抽出部１２１とキーワード抽出部１２２とを備える。パターン抽出部１２１は、キーワード入力部１１に入力されたキーワードをキーワード抽出用ＤＢ１５で全文検索し、複数の入力キーワードの周辺に出現したパターンを抽出する。

【００２７】

キーワード抽出部１２２は、パターン抽出部１２１で抽出したパターンをキーワード抽出用ＤＢ１５で全文検索し、該パターンによって抽出される表現をキーワードとして出力する。

【００２８】

以下に、キーワード増加部１２によるキーワード抽出処理を説明する。パターン抽出部１２１は、入力された少数のキーワードをキーワード抽出用ＤＢ１５で全文検索し、該少数のキーワードの周辺に出現したパターンc_iを抽出する。キーワード抽出部１２２は、抽出したパターンc_iをキーワード抽出用ＤＢ１５で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅ（スコア；評価値）の値の大きい順にソートしてキーワードとして出力する。

【００２９】

（パターンの例の説明）

以下に、パターン抽出部１２１が抽出するパターンについて、該パターンが国名Ａである場合を例にとって説明する。

【００３０】

・入力キーワード：

日本
中国
朝鮮
タイ
韓国

・抽出パターンの例(1)：（両端とも利用、スピードは遅いが性能は良い）

日、Ａ軍
人のＡ人女性

日本はＡと
〔Ａ通信・
省。駐Ａ大使な
・抽出パターンの例(2) ：（片方のみ利用、片方は平仮名文字、スピードは早い）
［..Ａ国］。

【００３１】

語。Ａ
［..Ａ国］側
［..Ａ国］伝来
Ａ語入力

ただし、［..Ａ..］は、それ自体が国名Ａにマッチすることを意味する。例えば［Ａ国］だとそのマッチした用語の最後が国であることを意味する。

【００３２】

（キーワード抽出の具体的な説明）

入力する少数のキーワードとして、例えば、評価データの代表形で毎日新聞での頻度の多い方から有名そうな用語を五つ選択するものとする。また、例えば、ＣＤ毎日新聞（コンパクトディスクに記録された毎日新聞）１９９１－２０００年度版をキーワード抽出用ＤＢ１５とする。抽出の手順は以下のとおりである。

【００３３】

(1) 少数の複数のキーワードをキーワード抽出用ＤＢ１５で全文検索し、複数のキーワードの周辺に出現したパターンをc_iとして抽出する（キーワードの周辺に出現するパターンがそのキーワードだけ（一個）の場合は抽出しない）。（周辺に出現するパターンの定義は適宜行なう）。周辺に出現するパターンとして例えば、キーワードの前後（左右）３文字列を用いる場合は、前後それぞれ文字が１個、２個、３個の場合があるので、１個のキーワードで９通りのパターンができることになる。また、キーワード（自分自身）を含めたパターンとすることもできる。

【００３４】

(2) 次に抽出したパターンc_iをキーワード抽出用ＤＢ１５で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出する。

【００３５】

(3) 抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートして、キーワードとして出力する。

【００３６】

Ｓｃｏｒｅとして、以下のものがある。

【００３７】

・手法１（決定リスト法）

手法１は、抽出した表現ｅｘｐのＳｃｏｒｅとして、パターンc_iの中でp_iが最も大きかったパターンのp_iを使用する手法である。ここで、p_iはパターンc_iで抽出される表現ｅｘｐでの入力キーワードの割合（確からしさ、すなわち確信度となる）である。

【００３８】

例えば、パターンc_1についてキーワード抽出用ＤＢ１５で全文検索した結果、ｅｘｐ１、ｅｘｐ２、ｅｘｐ３、ｅｘｐ４、ｅｘｐ５までの５個のｅｘｐが抽出され、この５個のｅｘｐのうち、ｅｘｐ１～ｅｘｐ３までの３個が入力キーワードであった場合、p_1は３／５である。

【００３９】

【数１】

$$\mathrm{Score} = max_i p_i \qquad \text{式(1)}$$

・手法２（ベイズ法）

手法２は、抽出した表現ｅｘｐのＳｃｏｒｅとして、全てのパターンc_iのp_iを掛け合わせたものを使用する。

【００４０】

【数２】

$$\mathrm{Score} = \prod_i p_i \qquad \text{式(2)}$$

なお、実際にはp_i＝０の可能性が大きいため、本発明の実施の形態では、上記式（２）に代えて、以下の式（３）

Π（（１－Δ）／Δ＊p_i＋１）　　式（３）

を利用する構成をとることもできる。ここで、Δは微小値の定数であり、例えば、０．０００１を用いる。

【００４１】

例えば、Ｓｃｏｒｅを計算しているｅｘｐがパターンc_iから取れなかった場合は、p_i＝０として、上記の式（３）を用いて計算する。

【００４２】

・手法３（類似度に基づく方法）

手法３は、抽出した表現ｅｘｐのＳｃｏｒｅとして、抽出されたパターンの個数（総数）を用いる。つまり、多くのパターンで抽出されたものほどＳｃｏｒｅを大きくする。

【００４３】

【数３】

$$\mathrm{Score} = \sum_i 1 \qquad \text{式(4)}$$

・手法４（下記研究(3) 参照）

手法４は、抽出した表現ｅｘｐのＳｃｏｒｅとして、p_iの重みを加えた抽出されたパターンの個数を用いるものである。

【００４４】

【数４】

$$\mathrm{Score} = \sum_i (1 + 0.01 p_i log(f_i)) \qquad \text{式(5)}$$

ただし、f_iはパターンc_iが出現した入力キーワードの個数である。

【００４５】

研究(3):Ellen Riloff and Rosie Jones "Learning dictionaries for information extraction by multi-level bootstrapping" Proceedings of AAAI-99,(1999)。

【００４６】

・手法５（下記文献(4) 参照）

手法５は、抽出した表現ｅｘｐのＳｃｏｒｅとして、少なくとも一つは確からしくなる値を用いるものである。

【００４７】

【数５】

$$\mathrm{Score} = 1 - \prod_i (1 - p_i) \qquad \text{式(6)}$$

上記式（６）は、確からしくない（１－p_i）を掛け合わせることで一つも確からしく

ないことになり、そして、これを１から引くと少なくとも一つは確からしくなる。
【００４８】
文献(4):村田真樹，井佐原均 "同義テキストの照合に基づくパラフレーズに関する知識の自動獲得" 情報処理学会自然言語処理研究会 2001-NL-142,(2001)。
【００４９】
上記手法１、２、４、５では、Ｓｃｏｒｅが同じときは、手法３のＳｃｏｒｅでソートし、手法３では手法５のＳｃｏｒｅでソートする。
【００５０】
図３は、パターンとしてキーワードの左と先頭のいずれかを含む１～３文字と右側のそれの組み合わせを用いて行ったキーワードの抽出結果に対して、予め用意した所定の種類数の正解データを使って、適合率・再現率を求めた結果の一例を示す図である。ここで、正解データとしては、例えば、図４に示すようなデータ例を用意する（図４は、国名データの例を示しており、国名を国ごとに行に分けて格納し、行頭を代表形としてそれ以外は代表形の異表記として同じ行に格納している）。図４に示すデータ形式と同様のデータ形式を持つ正解データを、例えば、国名データの他に、衛星、祝日、太陽系惑星、世界遺産等に関するデータのように、多種類用意する。
【００５１】
図３において、ＡＰは、情報検索（下記文献(5) 参照）で用いるaverage precision の平均であり、正解記事を上位から取ったたびに求めた適合率の平均である。本願の内容の場合は、正解キーワード分を上位から取ったたびに求めた適合率の平均（ただし、入力キーワードは正解キーワードから除く）である。
【００５２】
文献(5):村田真樹，馬青，内元清貴，小作浩美，内山将夫，井佐原均 "位置情報と分野情報を用いた情報検索" 言語処理学会誌，Vol.7,No.2,(2000)。
【００５３】
ＲＰは、r-precision の平均であり、正解記事数分だけを検索した時に正解の記事が含まれている割合である。本願の内容の場合は、正解キーワード分だけを抽出した時に正解キーワードが含まれている割合である。なお、適合率は正解率と同じであり、正解キーワードが含まれる割合のことである。ＴＰは、上位５個での精度の平均である。
【００５４】
（制約に基づく抽出方法の説明）
（ａ）字種とＫＲを利用する方法
図３に示す例で、抽出方法には、さらに字種とＫＲを利用する方法を用いた。ここで、字種とは、漢字、カタカナ、ひらがな、記号、数字などであり、例えば英語だと、アルファベット、数字、記号、単語の先頭が大文字かどうかなどである。
【００５５】
字種を利用する方法では、入力した少数（例えば、５個）のキーワードになかった字種を含む表現を抽出しない方法である。例えば、入力した５個のキーワードにひらがなが無かった場合は、ひらがなを含む表現を抽出しないようにするものである。
【００５６】
ＫＲを利用する方法では、p_i を $p_i * f_i / n_i$ に置き換えた方法である。この方法の利点は、p_i が同じでも f_i / n_i の値により確信度を変えることができるものである。ただし、n_i は入力キーワードの個数で、手法３のときはＫＲの場合は１を f_i に置き換えた。なお、評価では抽出した結果でキーワードの異表記は除いた。また、字種による方法以外にも次のような方法もある。
【００５７】
（ｂ）品詞に基づく方法
品詞に基づく方法では、例えば、入力表現に名詞しかない場合は出力時に名詞以外の表現を省く、また、入力表現に形容詞しかない場合は出力時に形容詞以外の表現を省くというものである。さらに、表現が複数の単語で構成されている場合は、末尾の単語（形態素

）の品詞の情報を使うようにすることができる。
【００５８】
（例による説明１）
入力キーワードとして次のものであった場合、
「楽しい」「哀しい」「嬉しい」「とても嬉しい」「とても哀しい」
抽出物として次のものが得られる場合、
「とても」「新しい」「美しい」「とても美しい」「とても難しい」
上記抽出物の表現中の末尾の単語の品詞を推定し、上記入力キーワードでは、末尾の単語の品詞は「形容詞」しかないので、抽出物の中で、末尾の単語の品詞が「形容詞」でない、副詞（「とても」）を除いて出力するようにする。
【００５９】
（例による説明２）
入力キーワードとして次のものであった場合、
「楽しい」「歓喜」「悲痛」「悲しい」
上記入力キーワードでは、「形容詞」と「名詞」のように複数種類があった場合は、それらの品詞は出力し、それらの品詞以外の表現は出力しないようにする。
【００６０】
なお、前述のような末尾の単語（形態素）の品詞の推定等の品詞情報を得るためには、次のような形態素解析システム（形態素解析手段）が必要になる。
【００６１】
・形態素解析システムの説明
日本語を単語に分割するために、キーワード抽出部１２２で形態素解析システムを利用することが必要になる。ここではChaSenについて説明する（奈良先端大で開発されている形態素解析システム茶筌。http://chasen.aist-nara.ac.jp/index.html.jp で公開されている）。
【００６２】
これは、日本語文を分割し、さらに、各単語の品詞も推定してくれる。例えば、「学校へ行く」を入力すると以下の結果を得ることができる。
【００６３】
学校　ガッコウ　学校　名詞－一般
へ　へ　へ　助詞－格助詞－一般
行く　イク　行く　動詞－自立　五段・カ行促音便　基本形
ＥＯＳ
このように各行に一個の単語が入るように分割され、各単語に読みや品詞の情報が付与される。
【００６４】
（ｃ）共通部分文字列に基づく方法
例えば、入力表現がすべて同じ「しい」という共通末尾表現を持っている場合、出力時に「しい」を持たない表現を省くものである。なお、これは末尾だけでなく、先頭の文字列でも同様にできる。
【００６５】
（例による説明）
入力キーワードとして次のものであった場合、
「悲しい」「楽しい」「嬉しい」
抽出されるものが次の場合、
「歓喜」「悲痛」「美しい」「新しい」
上記入力キーワードの共通部分文字列が「しい」なので、「しい」を持たない「歓喜」と「悲痛」を削除して出力するものである。
【００６６】
（ｄ）ユーザによる制約の指定

上記では、入力表現から自動で制約を得る方法を説明したが、この制約はユーザにさせることもできる。例えば、ユーザが「漢字のみ」というオプションを選択すると出力では漢字以外の字種を用いた表現を出力しないことができる。また、ユーザが末尾は「しい」というオプションを選択すると出力では「しい」を末尾に持たない表現を出力しないようにすることができる。さらに、ユーザが品詞は名詞というオプションを選択すると出力では名詞以外の表現を出力しないようにする。

【００６７】

（フローチャートによる説明）

図５は、本発明の実施の形態におけるデータ表示処理フローの一例を示す図である。以下図５の処理Ｓ１～Ｓ５に従って説明する。図５に示すデータ表示処理フローは、表示データ作成部１３が、キーワード抽出部１２２によって出力されたキーワードに関する数値データを表示データとして作成する場合の例である。

【００６８】

Ｓ１：キーワード入力部１１に、少数のキーワードを入力する。例えば、キーワードとして、京都大、東工大、ＮＥＣ、通信総研、ニューヨーク大という５つのキーワードを入力する。

【００６９】

Ｓ２：キーワード増加部１２のパターン抽出部１２１で、入力キーワードをキーワード抽出用ＤＢ１５で全文検索し、複数の入力キーワードの周辺に出現したパターンをc_iとして抽出する。（周辺に出現するパターンの定義は適宜行なう。）

Ｓ３：キーワード増加部１２のキーワード抽出部１２２で、パターン抽出部１２１で抽出したパターンc_iをキーワード抽出用ＤＢ１５で全文検索し、パターンc_iによって抽出される表現ｅｘｐを抽出すると同時に、抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートし、キーワードとして出力する。

【００７０】

キーワード抽出部１２２は、例えば、京都大、東工大、ＮＥＣ、通信総研、ニューヨーク大という入力キーワードの他、横浜国大、ＮＴＴ、徳島大、日立、奈良先端大、電通大、鳥取大学、東京大学・・・といった多くの研究機関名をキーワードとして出力する。

【００７１】

Ｓ４：表示データ作成部１３で、キーワード抽出部１２２によって出力されたキーワードに関する数値データを表示データとして作成する。表示データ作成部１３は、例えば、キーワード抽出部１２２によって出力されたキーワードと書誌データＤＢ１６中の書誌データとに基づいて、各キーワードをタイトルに含む文書の年次発表データを表示データとして作成する。すなわち、表示データ作成部１３は、例えば、各キーワードをタイトルに含む文書の、各年次の発表件数をカウントして、年次発表データを作成する。例えば、図６（Ａ）に示すような年次発表データが作成される。

【００７２】

図６（Ａ）に示す年次発表データは、例えば、キーワードの一つであるＡ大学については、第３年次に１件、第４年次に５件、第６年次に１０件、第７年次に１件の文書発表があり、Ｂ大学については、第１年次に５件、第２年次に３件、第３年次に１０件、第８年次に１件の文書発表があり、Ｃシステムズについては、第４年次に２件、第７年次に４件、第８年次に１２件、第９年次に５件、第１０年次に１３件の文書発表があることを示している。

【００７３】

表示データ作成部１３は、上記定期発表データを等高線データに変換し、変換後の等高線データを表示データとする構成をとることもできる。

【００７４】

Ｓ５：データ表示部１４で、表示データ作成部１３によって作成された表示データを画面表示する。データ表示部１４は、例えば図７に示すように、各研究機関の各年次における文書の発表件数のデータが等高線表示される画面を表示する。発表件数の度合いによっ

て等高線の表示色が異なっている。例えば、８～１０件の発表件数に対応する等高線の表示色は一番濃い色で表示される。

【００７５】

なお、データ表示部１４は、例えば、図８に示すように、各研究機関の各年次における文書の発表件数のデータをバブルチャートとして画面表示する構成を採ることもできる。なお、バブルチャートとは、一般に、ある事象を示す（円）を２つの軸を持つ図上に配置した図のことを言う。図８に示すバブルチャートでは、円の大きさが発表件数の度合いを示している。

【００７６】

本発明の実施の形態においては、表示データ作成部１３は、キーワード増加部１２による処理によって数が増加したキーワードの第１の組と前記数が増加したキーワードの第２の組の双方に関する数値データを表示データとして作成し、データ表示部１４が、作成された表示データを２次元画面上に画面表示する構成を採ることもできる。

【００７７】

例えば、キーワード入力部１１に入力された、京都大、東工大という２つのキーワード（研究機関名）からなるキーワードの組（第１のキーワード群）と、意味、知識という２つのキーワード（研究分野）からなるキーワードの組（第２のキーワード群）のそれぞれを入力キーワードとして、上記ステップＳ１～ステップＳ３の処理を行う。

【００７８】

そして、表示データ作成部１３が、例えば、図６（Ｂ）に示すような表示データを作成する。図６（Ｂ）に示す表示データでは、第１のキーワード群のキーワード入力部１１への入力に基づいてキーワード増加部１２から出力された、京都大、東工大、ＮＥＣ、通信総研、ニューヨーク大という５つの第１のキーワード（研究機関名）が縦軸に、第２のキーワード群のキーワード入力部１１への入力に基づいてキーワード増加部１２から出力された、意味、知識、辞書、支援、用例という５つの第２のキーワード（研究分野）が横軸に並べられている。

【００７９】

そして、図６（Ｂ）に示す表示データにおいて、第１のキーワード群中のあるキーワード（例えば、「ＮＥＣ」）に対応する行と、第２のキーワード群中のあるキーワード（例えば、「意味」）に対応する列とが交差する枡目には、例えば、表示データ作成部１３によって書誌データＤＢ１６中の書誌データから抽出された、双方のキーワード（例えば、「ＮＥＣ」と「意味」）を含む文書の発表件数のデータ（例えば、「７」件）が格納される。

【００８０】

図９は、本発明の別の実施の形態におけるシステム構成の一例を示す図である。データ表示装置２は、キーワードに関するデータを表示する処理装置である。図９中に示すデータ表示装置２が備える構成要素のうち、図１に示すデータ表示装置１が備える構成要素と同一の符号が付けられたものは、当該データ表示装置１が備える構成要素と同様の機能を有する。

【００８１】

データ表示装置２のキーワード増加部２１は、キーワード入力部１１に入力されたキーワードを増加させる。単語データデータベース（ＤＢ）２２には、単語と単語の分野との対応情報が格納されている。例えば、図１０に示すような、単語と単語の分野との対応情報が格納されている。例えば、「研究分野」という分野に対応する単語として、意味、知識、辞書、支援、用例といった単語が格納されている。

【００８２】

また、シソーラスデータベース（ＤＢ）２３には、意味的類似による単語の分類情報であるシソーラスデータが格納されている。例えば、シソーラスＤＢ２３には、図１１に示すような、単語と単語に振られた１０桁の数字（分類番号）との対応情報がシソーラスデータとして格納されている。図１１に示す例では、シソーラスデータが分類語彙表の形式

で示されている。

【００８３】

なお、分類語彙表とは、一般に、単語を意味に基づいて整理した表であり、各単語に対して分類番号という数字が付与されている。この１０桁の分類番号は、７レベルの階層構造を示しており、上位５レベルは分類番号の最初の５桁で表現され、６レベル目は次の２桁、最下層のレベルは最後の３桁で表現されている。

【００８４】

類似度算出部２１１は、シソーラスＤＢ２３中のシソーラスデータに基づいて、キーワード入力部１１に入力されたキーワードとシソーラスデータ中の単語との類似度を算出する。キーワード抽出部２１２は、算出された類似度が予め定めた閾値以上の単語をキーワードとして抽出し、出力する。

【００８５】

本発明の実施の形態においては、キーワード抽出部２１２は、単語データＤＢ２２中に格納された、単語と単語の分野との対応情報に基づいて、キーワード入力部１１に入力されたキーワードと同じ分野の単語をキーワードとして抽出し、出力する構成を採ることもできる。

【００８６】

図１２は、本発明の別の実施の形態におけるデータ表示処理フローの一例を示す図である。図１２に示すデータ表示処理フローは、表示データ作成部１３が、キーワード抽出部２１２によって出力されたキーワードに関する数値データを表示データとして作成する場合の例である。

【００８７】

Ｓ１１：キーワード入力部１１に、少数のキーワードを入力する。

【００８８】

Ｓ１２：キーワード増加部２１のキーワード抽出部２１２で、キーワード入力部１１に入力されたキーワードと同じ分野の単語を単語データＤＢ２２中から抽出し、キーワードとして出力する。例えば、キーワード入力部１１にキーワード「知識」が入力されたとすると、図１０に示す単語データＤＢ２２から、単語「知識」が対応する「研究分野」という分野に属する（対応する）単語である「意味」、「知識」、「辞書」、「支援」、「用例」を抽出し、キーワードとして出力する。

【００８９】

Ｓ１３：表示データ作成部１３で、キーワード抽出部２１２によって出力されたキーワードに関する数値データを表示データとして作成する。表示データ作成部１３は、例えば、キーワード抽出部２１２によって出力されたキーワードと書誌データＤＢ１６中の書誌データとに基づいて、各キーワードをタイトルに含む文書の年次発表データを表示データとして作成する。すなわち、表示データ作成部１３は、例えば、各キーワードをタイトルに含む文書の、各年次の発表件数をカウントして、上述した図６（Ａ）に示すような年次発表データを作成する。表示データ作成部１３は、上述したように、上記定期発表データを等高線データに変換し、変換後の等高線データを表示データとする構成をとることもできる。

【００９０】

Ｓ１４：データ表示部１４で、表示データ作成部１３によって作成された表示データを画面表示する。データ表示部１４は、例えば上述した図７に示すように、各研究機関の各年次における文書の発表件数のデータが等高線表示される画面を表示する。

【００９１】

なお、データ表示部１４は、例えば、上述した図８に示すように、各研究機関の各年次における文書の発表件数のデータをバブルチャートとして画面表示する構成を採ることもできる。

【００９２】

また、上記Ｓ１３、Ｓ１４において、表示データ作成部１３が、キーワード増加部２１

による処理によって数が増加したキーワードの第１の組と前記数が増加したキーワードの第２の組の双方に関する数値データを表示データとして作成し、データ表示部１４が、作成された表示データを２次元画面上に画面表示する構成を採ることもできる。

【００９３】

図１３は、本発明の更に別の実施の形態におけるデータ表示処理フローの一例を示す図である。

【００９４】

Ｓ２１：キーワード入力部１１に、少数のキーワードを入力する。

【００９５】

Ｓ２２：キーワード増加部２１の類似度算出部２１１が、キーワード入力部１１に入力されたキーワードとシソーラスＤＢ２３中の単語との類似度を算出する。類似度算出部２１１は、例えば、類似度を以下のようにして算出する。

【００９６】

図１１に示すシソーラスＤＢ２３内に格納されたシソーラスデータ（分類語彙表）中の各単語に振られた、１０桁の分類番号における各桁の数字の一致の割合を用いて、類似度を求める。すなわち、例えば、分類語彙表中の各単語に振られた分類番号について、キーワード入力部１１に入力されたキーワードと同一の単語に振られた分類番号との間での、各桁の数字の一致の割合を算出し、算出された値を類似度とする。なお、例えば、分類番号の６桁目と７桁目、および、８桁目と９桁目と１０桁目は、それぞれ連続した１つの数字として考える。

【００９７】

例えば、キーワード入力部１１に入力されたキーワードが「日本」である場合、図１１に示す分類語彙表中の単語「日本」と「ソ連」には、それぞれ以下のような分類番号が振られている。以下では、分類番号の上位５レベルと、６レベル目と、最下層のレベルとの間を空白で区切って示す。

【００９８】

日本：１２５９０　０１　０１２

ソ連：１２５９０　０４　１９２

例えば、両単語の分類番号の上位５レベルにおいて、最初の５桁が一致するので、算出されるキーワード「日本」と分類語彙表中の単語「ソ連」との類似度は、類似度５である。

【００９９】

また、例えば、キーワード入力部１１に入力されたキーワードが「母校」である場合、分類語彙表中の単語「母校」と「学校」には、それぞれ以下のような分類番号が振られている。

【０１００】

母校：１２６３０　１３　０１５

学校：１２６３０　１０　０１２

例えば、両単語の分類番号の上位５レベルにおいて、最初の５桁が一致するので、算出されるキーワード「母校」と分類語彙表中の単語「学校」との類似度は、類似度５である。

【０１０１】

また、例えば、キーワード入力部１１に入力されたキーワードが「学校」である場合、分類語彙表中の単語「学校」と「学園」には、それぞれ以下のような分類番号が振られている。

【０１０２】

学校：１２６３０　１０　０１２

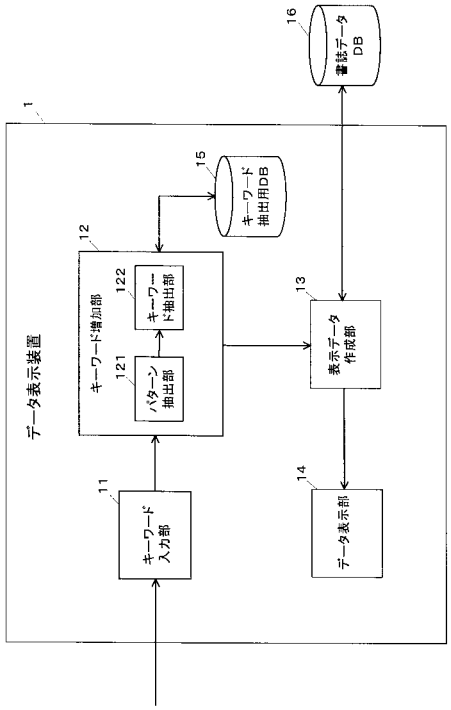
学園：１２６３０　１０　０１５

例えば、両単語の分類番号の上位５レベルにおいて、最初の５桁が一致し、また、６レベル目の２桁の数字「１０」が一致するので、算出されるキーワード「学校」と分類語彙

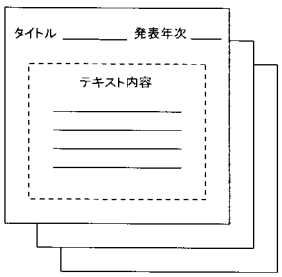
表中の単語「学園」との類似度は、類似度７である。
【０１０３】
また、例えば、キーワード入力部１１に入力されたキーワードが「学校」である場合、分類語彙表中の単語「学校」と「ソ連」には、それぞれ以下のような分類番号が振られている。
【０１０４】
学校：１２６３０　１０　０１２
ソ連：１２５９０　０４　１９２
例えば、両単語の分類番号の上位５レベルにおいて、最初の２桁が一致するため、算出されるキーワード「学校」と分類語彙表中の単語「ソ連」との類似度は、類似度２である。
【０１０５】
Ｓ２３：キーワード増加部２１のキーワード抽出部２１２が、算出された類似度が予め定めた閾値以上の単語をキーワードとして出力する。
【０１０６】
Ｓ２４：表示データ作成部１３で、キーワード抽出部２１２によって出力されたキーワードに関する数値データを表示データとして作成する。表示データ作成部１３は、例えば、キーワード抽出部２１２によって出力されたキーワードと書誌データＤＢ１６中の書誌データとに基づいて、各キーワードをタイトルに含む文書の年次発表データを表示データとして作成する。すなわち、表示データ作成部１３は、例えば、各キーワードをタイトルに含む文書の、各年次の発表件数をカウントして、前述した図６（Ａ）に示すような年次発表データを作成する。表示データ作成部１３は、前述したように、上記定期発表データを等高線データに変換し、変換後の等高線データを表示データとする構成をとることもできる。
【０１０７】
Ｓ２５：データ表示部１４で、表示データ作成部１３によって作成された表示データを画面表示する。データ表示部１４は、例えば前述した図７に示すように、各研究機関の各年次における文書の発表件数のデータが等高線表示される画面を表示する。
【０１０８】
なお、データ表示部１４は、例えば、前述した図８に示すように、各研究機関の各年次における文書の発表件数のデータをバブルチャートとして画面表示する構成を採ることもできる。
【０１０９】
また、上記Ｓ２４、Ｓ２５において、表示データ作成部１３が、キーワード増加部２１による処理によって数が増加したキーワードの第１の組と前記数が増加したキーワードの第２の組の双方に関する数値データを表示データとして作成し、データ表示部１４が、作成された表示データを２次元画面上に画面表示する構成を採ることもできる。
【０１１０】
なお、本発明は、コンピュータにより読み取られ実行されるプログラムとして実施することもできる。本発明を実現するプログラムは、コンピュータが読み取り可能な、可搬媒体メモリ、半導体メモリ、ハードディスクなどの適当な記録媒体に格納することができ、これらの記録媒体に記録して提供され、または、通信インタフェースを介してネットワークを利用した送受信により提供されるものである。
【図面の簡単な説明】
【０１１１】
【図１】システム構成の一例を示す図である。
【図２】書誌データの一例を示す図である。
【図３】キーワードの抽出結果に対する適合率・再現率の一例を示す図である。
【図４】正解データの一例を示す図である。
【図５】データ表示処理フローの一例を示す図である。

【図６】表示データの一例を示す図である。
【図７】表示データの画面表示例を示す図である。
【図８】表示データの画面表示例を示す図である。
【図９】システム構成の一例を示す図である。
【図１０】単語データＤＢの一例を示す図である。
【図１１】シソーラスＤＢの一例を示す図である。
【図１２】データ表示処理フローの一例を示す図である。
【図１３】データ表示処理フローの一例を示す図である。
【符号の説明】
【０１１２】
１、２　データ表示装置
１１　キーワード入力部
１２、２１　キーワード増加部
１３　表示データ作成部
１４　データ表示部
１５　キーワード抽出用ＤＢ
１６　書誌データＤＢ
２２　単語データＤＢ
２３　シソーラスＤＢ
１２１　パターン抽出部
１２２、２１２　キーワード抽出部
２１１　類似度算出部

【図１】



【図２】



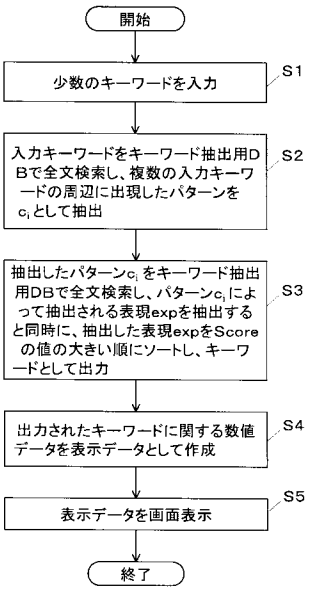
【図３】

	字種とKRを利用せず			字種の利用			KRの利用			字種とKRの利用		
	AP	RP	TP	AP	RP	TP	AP	RP	TP	AP	RP	TP
手法1	0.102	0.170	0.305	0.142	0.220	0.365	0.096	0.161	0.255	0.154	0.231	0.360
手法2	0.174	0.235	0.445	0.182	0.244	0.475	0.177	0.235	0.465	0.185	0.247	0.490
手法3	0.171	0.234	0.435	0.178	0.242	0.460	0.182	0.239	0.475	0.189	0.251	0.490
手法4	0.172	0.234	0.435	0.179	0.244	0.465	0.172	0.235	0.435	0.179	0.245	0.465
手法5	0.174	0.236	0.410	0.192	0.264	0.450	0.190	0.246	0.460	0.206	0.272	0.490

【図４】

アイスランド アイスランド共和国 ＩＳＬ
アイルランド アイルランド共和国 ＩＲＬ
アゼルバイジャン アゼルバイジャン共和国 ＡＺＥ
アゾレス諸島
アドゥイゲ アドゥイゲ共和国
アフガニスタン アフガニスタン共和国
アメリカ アメリカ合衆国 米国 米 ＵＳＡ
…

【図５】



【図６】

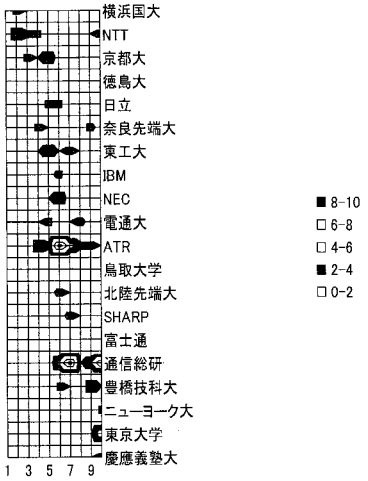
(A)

研究機関名	1	2	3	4	5	6	7	8	9	10	年次
A大学	0	0	1	5	0	10	1	0	0	0	
B大学	5	3	10	0	0	0	0	1	0	0	
Cシステムズ	0	0	0	2	0	0	4	12	5	13	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

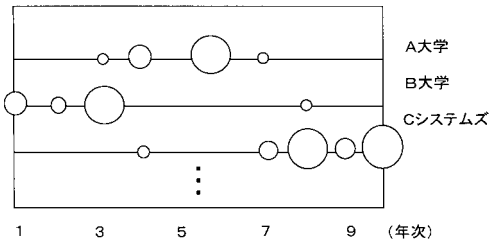
(B)

研究機関名	研究分野				
	意味	知識	辞書	支援	用例
京都大	0	0	1	0	2
東工大	2	2	0	0	0
NEC	7	0	1	0	1
通信総研	0	4	2	0	0
ニューヨーク大	3	1	2	0	0

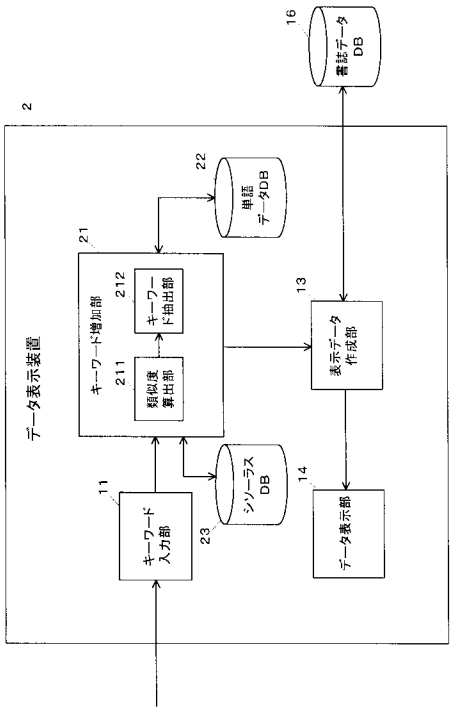
【図７】



【図８】



【図９】



【図１０】

分野	単語
研究機関名	京都大、東工大、NEC、通信総研、ニューヨーク大
研究分野	意味、知識、辞書、支援、用例
⋮	⋮

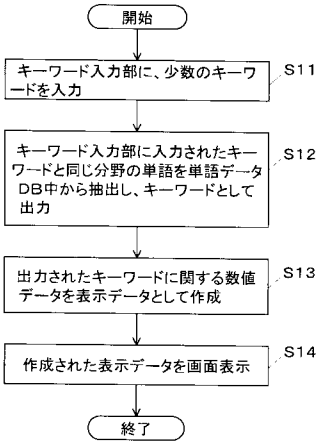
22

【図１１】

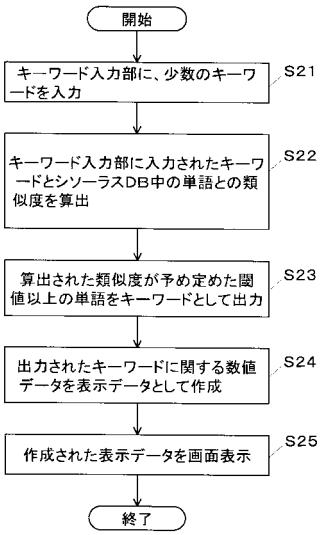
単語	分類番号
日本	1259001012
ソ連	1259004192
母校	1263013015
学校	1263010012
学園	1263010015
⋮	⋮

23

【図１２】


開始
キーワード入力部に、少数のキーワードを入力
S11
キーワード入力部に入力されたキーワードと同じ分野の単語を単語データDB中から抽出し、キーワードとして出力
S12
出力されたキーワードに関する数値データを表示データとして作成
S13
作成された表示データを画面表示
S14
終了

【図１３】


開始
キーワード入力部に、少数のキーワードを入力
S21
キーワード入力部に入力されたキーワードとシソーラスDB中の単語との類似度を算出
S22
算出された類似度が予め定めた閾値以上の単語をキーワードとして出力
S23
出力されたキーワードに関する数値データを表示データとして作成
S24
作成された表示データを画面表示
S25
終了

フロントページの続き

(72)発明者　馬　青
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　白土　保
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

審査官　長　由紀子

(56)参考文献　特開２００６－１１３７３３（ＪＰ，Ａ）
特開２０００－３３１０１２（ＪＰ，Ａ）
特開２０００－３１５２０６（ＪＰ，Ａ）
特開２００２－１３２８０８（ＪＰ，Ａ）
中渡瀬秀一，複合語からの類義語抽出法，情報処理学会研究報告，日本，社団法人情報処理学会，２００２年　3月１５日，Vol.2002 No.28，p.p.39-46

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０