

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2007-79898

(P2007-79898A)

(43) 公開日 平成19年3月29日(2007.3.29)

(51) Int. Cl.

G06F 17/30 (2006.01)

F I

G06F 17/30 320C

テーマコード(参考)

5B075

審査請求 未請求 請求項の数 10 O L (全 20 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)

特許法第30条第1項適用申請有り 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

(71) 出願人 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 100119161
 弁理士 重久 啓子
 (74) 代理人 100121511
 弁理士 小田 直
 (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

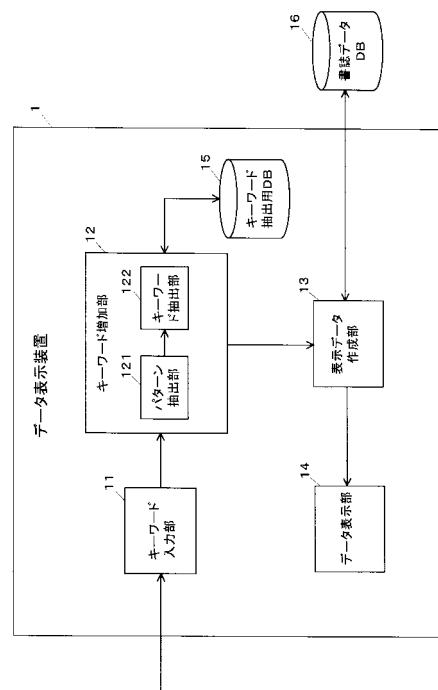
(54) 【発明の名称】 データ表示装置、データ表示方法およびデータ表示プログラム

(57) 【要約】

【課題】 入力されたキーワードに関するデータと、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とを表示することを可能とする。

【解決手段】 パターン抽出部121が、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 c_i として抽出し、キーワード抽出部122が、抽出されたパターン 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て、キーワードとして出力する。表示データ作成部13が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し、データ表示部14が作成された表示データを画面表示する。

【選択図】 図1



【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
ことを特徴とするデータ表示装置。

10

【請求項 2】

請求項 1 に記載のデータ表示装置において、
前記キーワード増加手段は、前記入力キーワードと同じ分野のキーワードを、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
ことを特徴とするデータ表示装置。

【請求項 3】

請求項 1 または請求項 2 に記載のデータ表示装置において、
前記キーワード増加手段は、
前記入力キーワードを前記キーワード抽出用データベースで全文検索し、前記入力キー
ワードの周辺に出現したパターンを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キーワード抽出用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。

20

【請求項 4】

請求項 1 に記載のデータ表示装置において、
前記キーワード増加手段は、前記入力キーワードと同じ分野の単語を、単語と単語の分
野との対応情報が格納されたデータベースから抽出し、キーワードとして出力する
ことを特徴とするデータ表示装置。

30

【請求項 5】

請求項 1 に記載のデータ表示装置において、
前記キーワード増加手段は、
予めデータベース中に記憶された、意味的類似による単語の分類情報であるシソーラス
データに基づいて、前記入力されたキーワードと、前記シソーラスデータ中の単語との類
似度を算出する類似度算出手段と、
前記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。

40

【請求項 6】

請求項 1 乃至請求項 5 のいずれか 1 項に記載のデータ表示装置において、
前記表示データ作成手段は、前記増加したキーワードの第 1 の組と前記増加したキー
ワードの第 2 の組の双方に関するデータを表示データとして作成し、
前記データ表示手段は、前記表示データを 2 次元画面上に画面表示する
ことを特徴とするデータ表示装置。

【請求項 7】

請求項 1 乃至請求項 5 のいずれか 1 項に記載のデータ表示装置において、
前記表示データ作成手段は、前記増加した各キーワードに関する数値データをカウント
して取得し、取得した数値データに基づいて表示データを作成する

50

ことを特徴とするデータ表示装置。

【請求項 8】

請求項 7 に記載のデータ表示装置において、

前記表示データ作成手段は、前記増加したキーワードの第 1 の組と前記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カウントして取得し、取得した数値データに基づいて表示データを作成し、

前記データ表示手段は、前記表示データを 2 次元画面上に画面表示する

ことを特徴とするデータ表示装置。

【請求項 9】

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、

10

複数のキーワードを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

ことを特徴とするデータ表示方法。

【請求項 10】

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、

20

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】

30

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】

40

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】

50

【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、10

【課題を解決するための手段】

【0008】

上記課題を解決するため、本発明は、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20

【0009】

また、本発明は、前記のデータ表示装置において、前記キーワード増加手段は、前記入力キーワードと同じ分野のキーワードを、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を特徴とする。

【0010】

また、本発明は、前記のデータ表示装置において、前記キーワード増加手段は、前記入力キーワードを前記キーワード抽出用データベースで全文検索し、前記入力キーワードの周辺に出現したパターンを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キーワード抽出用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。30

【0011】

また、本発明は、前記のデータ表示装置において、前記キーワード増加手段は、前記入力キーワードと同じ分野の単語を、単語と単語の分野との対応情報が格納されたデータベースから抽出し、キーワードとして出力することを特徴とする。

【0012】

また、本発明は、前記のデータ表示装置において、前記キーワード増加手段は、予めデータベース中に記憶された、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に基づいて、前記入力されたキーワードと、前記シソーラスデータ中の単語との類似度を算出する類似度算出手段と、前記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、出力するキーワード抽出手段とを備えることを特徴とする。40

【0013】

また、本発明は、前記のデータ表示装置において、前記表示データ作成手段は、前記増加したキーワードの第1の組と前記増加したキーワードの第2の組の双方に関するデータを表示データとして作成し、前記データ表示手段は、前記表示データを2次元画面上に画面表示することを特徴とする。

【0014】

また、本発明は、前記のデータ表示装置において、前記表示データ作成手段は、前記増加した各キーワードに関する数値データをカウントして取得し、取得した数値データに基づいて表示データを作成することを特徴とする。

【0015】

また、本発明は、前記のデータ表示装置において、前記表示データ作成手段は、前記増加したキーワードの第1の組と前記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カウントして取得し、取得した数値データに基づいて表示データを作成し、前記データ表示手段は、前記表示データを2次元画面上に画面表示することを特徴とする。

【0016】

また、本発明は、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、複数のキーワードを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ことを特徴とする。

10

【0017】

また、本発明は、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して機能させるためのデータ表示プログラムである。

20

【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。

30

【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。

40

【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】

50

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。

10

【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 DB 16 中の書誌データ中において、増加した各キーワードと共に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 DB 15 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 DB 15 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。

20

【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 DB 15 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 DB 15 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。

30

【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 DB 15 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 DB 15 で全文検索し、パターン c_i によって抽出される表現 $e x p$ を抽出すると同時に、抽出した表現 $e x p$ を $S c o r e$ （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】

40

（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

50

- ・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

10

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。

20

【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。

30

【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。

40

【0038】

例えば、パターン c_i についてキーワード抽出用DB15で全文検索した結果、 exp_1 、 exp_2 、 exp_3 、 exp_4 、 exp_5 までの5個の exp が抽出され、この5個の exp のうち、 $exp_1 \sim exp_3$ までの3個が入力キーワードであった場合、 p_i は3/5である。

【0039】

【数 1】

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

・手法 2 (ベイズ法)

手法 2 は、抽出した表現 $e \times p$ の $S \text{ c o r e}$ 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】

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

10

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

$$\left(\frac{1 - \epsilon}{p_i + 1} \right) \quad \text{式(3)}$$

を利用する構成をとることもできる。ここで、 ϵ は微小値の定数であり、例えば、0.0001 を用いる。

【0041】

例えば、 $S \text{ c o r e}$ を計算している $e \times p$ 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。

20

【0042】

・手法 3 (類似度に基づく方法)

手法 3 は、抽出した表現 $e \times p$ の $S \text{ c o r e}$ 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 $S \text{ c o r e}$ を大きくする。

【0043】

【数 3】

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

30

・手法 4 (下記研究(3) 参照)

手法 4 は、抽出した表現 $e \times p$ の $S \text{ c o r e}$ 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】

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

40

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 $e \times p$ の $S \text{ c o r e}$ 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】

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

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)。

10

【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。

20

【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とき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とき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)。

30

【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。

40

【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

50

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」

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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桁目には、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

10

20

30

40

50

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ができる。

【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ができる。

【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。

10

20

30

40

50

【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

10

20

30

40

50

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。 10

【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。 20

【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。 30

【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 40
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】

【 0 1 1 1 】

【 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。

10

【 図 1 0 】 単語データ D B の一例を示す図である。

【 図 1 1 】 シソーラス D B の一例を示す図である。

【 図 1 2 】 データ表示処理フローの一例を示す図である。

【 図 1 3 】 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 】

【 0 1 1 2 】

1、 2 データ表示装置

1 1 キーワード入力部

1 2、 2 1 キーワード増加部

1 3 表示データ作成部

1 4 データ表示部

1 5 キーワード抽出用 D B

1 6 書誌データ D B

2 2 単語データ D B

2 3 シソーラス D B

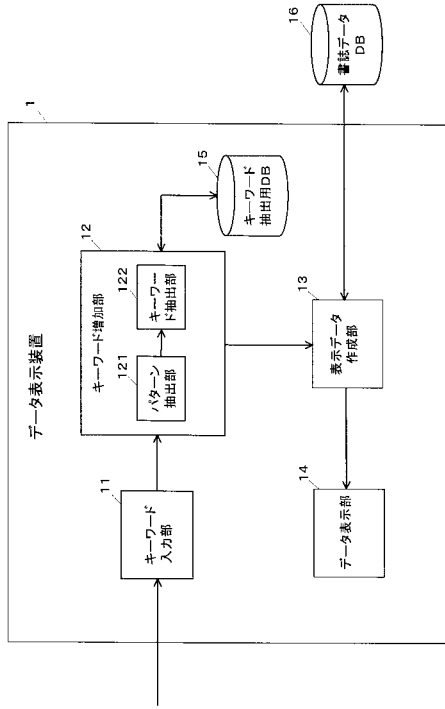
20

1 2 1 パターン抽出部

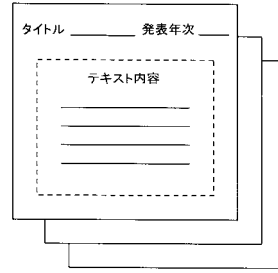
1 2 2、 2 1 2 キーワード抽出部

2 1 1 類似度算出部

【 図 1 】



【 図 2 】



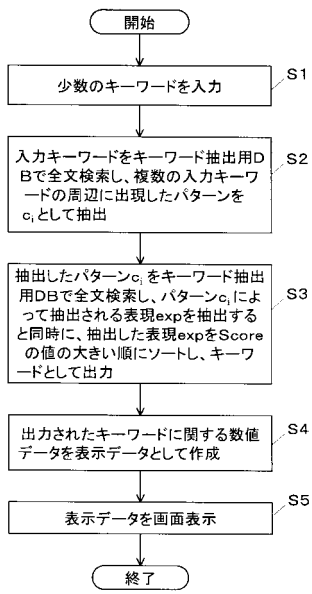
【 図 3 】

	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	0.465	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 】

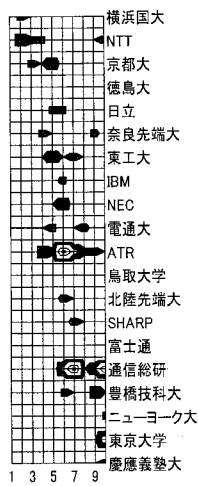
(A)

	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	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

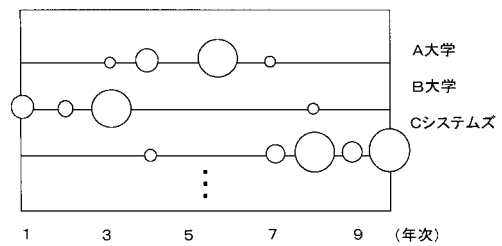
(B)

研究機関名	研究分野				
	意味	知識	特書	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

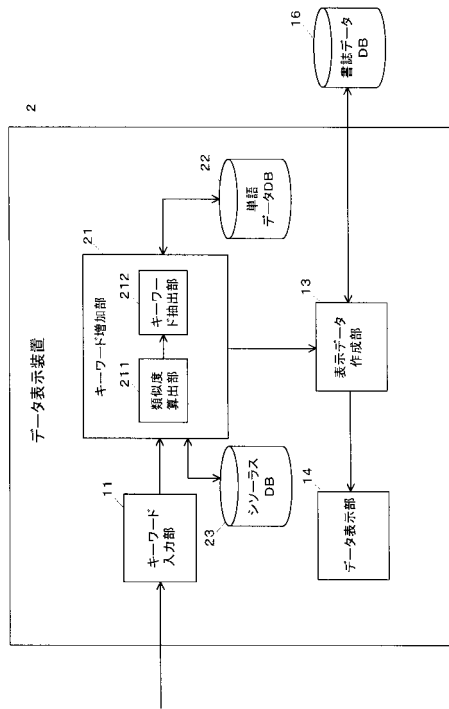
【 図 7 】



【 図 8 】



【図9】



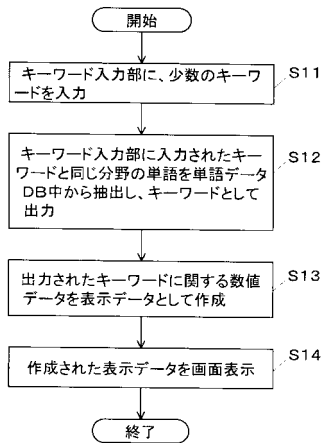
【図10】

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
⋮	⋮

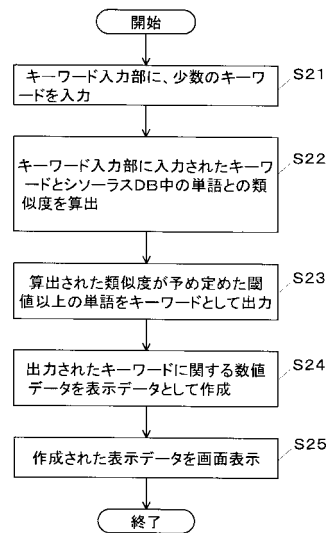
【図11】

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

【図12】



【図13】



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
- Fターム(参考) 5B075 NK37