(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2005-322121
(P2005-322121A)
(43)公開日　平成17年11月17日(2005.11.17)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/30	G06F 17/30	320D	5B075
	G06F 17/30	170A	
	G06F 17/30	360Z	

審査請求 未請求　請求項の数 7　OL　（全 20 頁）

(21)出願番号　特願2004-140841 (P2004-140841)
(22)出願日　平成16年5月11日 (2004.5.11)

特許法第３０条第１項適用申請有り　２００４年３月１９日　言語処理学会発行の「言語処理学会第１０回年次大会　併設ワークショップ「固有表現と専門用語」発表論文集」に発表

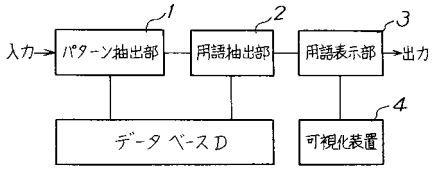
(71)出願人　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100103827
弁理士　平岡　憲一
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　馬　青
京都市伏見区深草塚本町６７　龍谷大学内
(72)発明者　白土　保
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】表示システム

(57)【要約】
【課題】ユーザが入力した少数の用語と同じ分野の用語を抽出し該抽出した用語を二次元の図に類似した用語が集まるように表示すること。
【解決手段】入力する複数の少数の用語である入力正例と、一定量の文書データを格納したデータベースDと、入力された前記入力正例を前記データベースDで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出部１と、前記パターン抽出部１で抽出したパターンを前記データベースD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用語抽出部２と、前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段４に入力して二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示部３とを備える。
【選択図】　図３

表示システムの説明図



【特許請求の範囲】
【請求項１】
入力する複数の少数の用語である入力正例と、
一定量の文書データを格納したデータベースと、
入力された前記入力正例を前記データベースで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出部と、
前記パターン抽出部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用語抽出部と、
前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段に入力して二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示部とを備えることを特徴とした表示システム。
【請求項２】
入力する複数の少数の用語である入力正例と、
一定量の文書データを格納したデータベースと、
入力された前記入力正例を前記データベースで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出部と、
前記パターン抽出部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入力正例の割合（p_i）によりスコアを算出し、該スコアの大きい順に抽出する用語抽出部と、
前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段に入力し二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示部とを備え、
前記用語表示部は、抽出の順序又は前記パターンで抽出される表現での入力正例の割合（p_i）も同時に表示して出力することを特徴とした表示システム。
【請求項３】
前記用語抽出部は、前記スコアの算出に前記パターンで抽出される表現での入力正例の割合（p_i）に前記パターンが出現した前記入力正例の個数（f_i）を前記入力正例の個数（n_i）で割った値を掛けた値（$p_i \times f_i / n_i$）を用いることを特徴とした請求項２記載の表示システム。
【請求項４】
前記用語抽出部は、前記入力正例になかった字種を含む表現を抽出しないようにすることを特徴とした請求項１～３のいずれかに記載の表示システム。
【請求項５】
前記入力正例として複数の少数の用語の対を入力し、前記用語表示部で用語の対の表現を表示することを特徴とした請求項１～４のいずれかに記載の表示システム。
【請求項６】
前記用語表示部は、前記入力正例がわかるような表示を行うことを特徴とした請求項１～５のいずれかに記載の表示システム。
【請求項７】
複数の少数の用語である入力正例を入力する手順と、
一定量の文書データをデータベースに格納する手順と、
入力された前記入力正例を前記データベースで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出手順と、
前記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用語抽出手順と、
前記入力正例になかった字種を含む表現を抽出しないようにする用語抽出手順と、
前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段に入力し二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示手順とを、

コンピュータに実行させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、ユーザが入力した少数の単語と同じ分野の用語を収拾して可視化して表示するシステムに関する。

【０００２】

近年、質問応答処理システム（下記文献(1) 参照）において、固有名詞を取り出すような固有表現抽出の技術は必然的に必要な技術となっており、生物情報処理分野でタンパク質表現の抽出が重要視されそれに関する研究が盛んになっている（下記文献(2) 参照）ように、固有表現抽出（固有名詞、数字等の抽出）、専門用語抽出などの研究は非常に重要なものとなってきている。また、固有表現、専門用語に関する技術・考え方は、新しい分野に適用されつつあり、また、これからも適用されるもので、用語の種類は極めて多様なもの、多彩なものとなりつつある。

【０００３】

文献(1):村田真樹“質問応答システムの現状と展望”電子情報通信学会学会誌, Vol.86, No.12, (2003),pp.959-963。

【０００４】

文献(2):Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi, “Boundary correction of protein names adapting heuristic rules ”Fifth International Conference on Intelligent Text Processing and Computational Linguistics(CICLing 2004),(2004)。

【０００５】

そこで、本発明では、多様な用語に関する、用語抽出の評価データを作成することができるものである。このデータでは、用語データも極力漏れの少ない形で作成されており、用語抽出の実験において、再現率・適合率を算出するなどの性能評価に用いることができるものである。本発明では、このデータの説明とこのデータを利用した簡単な用語抽出の評価実験について述べる。さらに、可視化機能を有する用語抽出の応用システムについても述べる。この応用システムは、ユーザが入力した数語の単語と同じ分野の用語を約２０秒で収集して可視化して提示するシステムで実用的でかつ有益なものである。

【背景技術】

【０００６】

近年、質問応答の研究が重要視されつつあるが、質問応答システムでは、例えば、国名と首都名の対のデータのような二項データ（表２参照）をあらかじめ具備していれば、そういう関係の二項データでの質問応答を高性能に処理することができるものであった（例えば、非特許文献１参照）。二項データは、そういうシステムにも利用できるし、また、そういうシステムのために作成する二項データ抽出システムがあった（例えば、非特許文献２、３参照）。

【０００７】

しかし、従来のシステムは、入力した複数の用語と同じ分野の用語を高速にしかも正確に収集し、該収集した用語を二次元の図に、類似した用語が集まるように表示することは行われていなかった。

【非特許文献１】Michael Fleischman,Eduard Hovy,and Abdessamad Echihabi "Offline strategies for online question answering:Answering questions before they are asked" Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (eds. Erhard Hinrichs and Dan Roth),(2003),pp.1-7 。

【非特許文献２】Sergey Brin "Extracting patterns and relations from the world wide web" WebDB Workshop at 6th International Conference on Extending Database Technology,EDBT'98,(1998)。

【非特許文献３】安藤まや, 関根聡, 石崎俊 "定型表現を利用した新聞記事からの下位概

念単語の自動抽出" 情報処理学会自然言語処理研究会 2003-NL-157,(2003) 。

【発明の開示】

【発明が解決しようとする課題】

【０００８】

上記従来のものは、入力した複数の用語と同じ分野の用語を高速にしかも正確に収集し、該収集した用語を二次元の図に、類似した用語が集まるように表示するものではなかった。

【０００９】

本発明は上記問題点の解決を図り、ユーザが入力した少数の用語と同じ分野の用語を抽出し、該抽出した用語を二次元の図に類似した用語が集まるように表示し、不要な用語を容易に見つけることができるようにすることを目的とする。

【課題を解決するための手段】

【００１０】

図１は本発明の抽出システムである。図１中、１はパターン抽出部、２は用語抽出部、３は用語表示部、４は可視化装置（可視化手段）、Ｄはデータベースである。

【００１１】

本発明は、前記従来の課題を解決するため次のような手段を有する。

【００１２】

（１）：入力する複数の少数の用語である入力正例と、一定量の文書データを格納したデータベースＤと、入力された前記入力正例を前記データベースＤで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出部１と、前記パターン抽出部１で抽出したパターンを前記データベースＤ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用語抽出部２と、前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段４に入力して二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示部３とを備える。このため、ユーザが入力した少数の用語と同じ分野の用語を、二次元の図に類似した用語が集まるように表示することができる。

【００１３】

（２）：入力する複数の少数の用語である入力正例と、一定量の文書データを格納したデータベースＤと、入力された前記入力正例を前記データベースＤで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出部１と、前記パターン抽出部１で抽出したパターンを前記データベースＤ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入力正例の割合（p_i）によりスコアを算出し、該スコアの大きい順に抽出する用語抽出部２と、前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段４に入力し二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示部３とを備え、前記用語表示部３は、抽出の順序又は前記パターンで抽出される表現での入力正例の割合（p_i）も同時に表示して出力する。このため、スコアの高い用語を容易に見つけることができる。

【００１４】

（３）：前記（２）の表示システムにおいて、前記用語抽出部２は、前記スコアの算出に前記パターンで抽出される表現での入力正例の割合（p_i）に前記パターンが出現した前記入力正例の個数（f_i）を前記入力正例の個数（n_i）で割った値を掛けた値（$p_i \times f_i / n_i$）を用いる。このため、ユーザが入力した少数の用語と同じ分野の用語を確実に抽出して、二次元の図に類似した用語が集まるように表示することができる。

【００１５】

（４）：前記（１）～（３）の表示システムにおいて、前記用語抽出部２は、前記入力正例になかった字種を含む表現を抽出しないようにする。このため、ユーザが入力した少数の用語と同じ分野の用語を高速にしかも確実に抽出して、二次元の図に類似した用語が集まるように表示することができる。

【００１６】
（５）：前記（１）～（４）の表示システムにおいて、前記入力正例として複数の少数の用語の対を入力し、前記用語表示部で用語の対の表現を表示する。このため、ユーザが入力した少数の用語の対と同じ分野の用語の対を、二次元の図に類似した用語の対が集まるように表示することができる。
【００１７】
（６）：前記（１）～（５）の表示システムにおいて、前記用語表示部は、前記入力正例がわかるような表示を行う。このため、ユーザが入力した少数の用語と用語抽出部が抽出した同じ分野の用語を容易に区別することができる。
【００１８】
（７）：複数の少数の用語である入力正例を入力する手順と、一定量の文書データをデータベースに格納する手順と、入力された前記入力正例を前記データベースで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出するパターン抽出手順と、前記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用語抽出手順と、前記入力正例になかった字種を含む表現を抽出しないようにする用語抽出手順と、前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段に入力し二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力する用語表示手順とをコンピュータに実行させるためのプログラムとする。このため、このプログラムをコンピュータにインストールすることで、入力した少数の用語と同じ分野の用語を二次元の図に類似した用語が集まるように表示する表示システムを容易に提供することができる。
【発明の効果】
【００１９】
本発明によれば次のような効果がある。
【００２０】
（１）：パターン抽出部で入力正例をデータベースで全文検索し、複数の前記入力正例の周辺に出現したパターンを抽出し、用語抽出部で前記パターン抽出部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し、用語表示部で前記抽出した表現である各用語に対して、その用語の抽出に使われたパターンを文脈とし、該用語と文脈の対の集合を可視化手段に入力して二次元マップでの各用語の座標を定め、該求まった座標に用語を表示した図を出力するため、ユーザが入力した少数の用語と同じ分野の用語を、二次元の図に類似した用語が集まるように表示することができ、不要な用語を容易に見つけることができる。
【００２１】
（２）：用語表示部で抽出の順序又はパターンで抽出される表現での入力正例の割合（p_i）も同時に表示して出力するため、スコアの高い用語を容易に見つけることができる。
【００２２】
（３）：用語抽出部でスコアの算出にパターンで抽出される表現での入力正例の割合（p_i）に前記パターンが出現した前記入力正例の個数（f_i）を前記入力正例の個数（n_i）で割った値を掛けた値（$p_i \times f_i / n_i$）を用いるため、ユーザが入力した少数の用語と同じ分野の用語を確実に抽出して、二次元の図に類似した用語が集まるように表示することができる。
【００２３】
（４）：用語抽出部で入力正例になかった字種を含む表現を抽出しないようにするため、ユーザが入力した少数の用語と同じ分野の用語を高速にしかも確実に抽出して、二次元の図に類似した用語が集まるように表示することができる。
【００２４】
（５）：入力正例として複数の少数の用語の対を入力し、用語表示部で用語の対の表現を表示するため、ユーザが入力した少数の用語の対と同じ分野の用語の対を、二次元の図

に類似した用語の対が集まるように表示することができる。
【００２５】
（６）：用語表示部で入力正例がわかるような表示を行うため、ユーザが入力した少数の用語と用語抽出部が抽出した同じ分野の用語を容易に区別することができる。
【発明を実施するための最良の形態】
【００２６】
§１：用語抽出の説明
（１）：抽出システムの説明
図１は本発明の抽出システムの説明図である。図１において、用語を抽出する抽出システムには、パターン抽出部１、用語抽出部２、データベースＤが設けてある。
【００２７】
パターン抽出部１は、（入力手段等から）入力された少数の正例（用語）をデータベースＤで全文検索し、該少数の正例の周辺に出現したパターンc_iを抽出するものである。用語抽出部２は、抽出したパターンc_iをデータベースＤ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（出力手段等に）出力するものである。データベースＤ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）を格納するものである。
【００２８】
（フローチャートによる説明）
図２は用語抽出処理フローチャートである。以下図２の処理Ｓ１～Ｓ３に従って説明する。
【００２９】
Ｓ１：パターン抽出部１に、入力手段から少数の入力正例が入力される。
【００３０】
Ｓ２：パターン抽出部１で、入力正例をデータベースＤで全文検索し、複数の入力正例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：用語抽出部２で、パターン抽出部１で抽出したパターンc_iをデータベースＤ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出力手段に出力する。
【００３１】
（パターンの例の説明）
1) 国名Ａと首都名Ｂの対（二項データ）の場合の例
・パターン抽出部１に入力される入力正例の例：
日本＿東京
中国＿北京
韓国＿ソウル
シンガポール＿シンガポール
ロシア＿モスクワ
・パターン抽出部１が抽出する抽出パターンの例：
、Ａの首都Ｂ
Ａの首都Ｂ
Ａ・Ｂ市に
る。〔Ｂ支局〕Ａ
Ｂ支局〕Ａ外務省
ただし、左右にＡ、Ｂがある場合は、さらにその左右は平仮名文字であることが条件となる。
【００３２】
2) 国名Ａの場合の例

・入力正例：
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］。

【００３３】

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。

【００３４】

（２）：用語抽出用評価データの作成の説明

本発明では、用語抽出用評価データ（正解データ）を作成した。作成したデータの例を表１と表２に示している。表１は、国名に関するデータで国名を国ごとに行に分けて格納しており、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。表２は、国名と首都名の対のデータで表１と同じく国ごとにデータを行に分けて先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。

【００３５】

本発明では、表１と表２をそれぞれ一項データ、二項データと呼ぶ。それぞれここで示したようなものを５８種類作った。全データの規模を表３に示している。代表形数と代表形＋異表記数はそれらの延べ数である。

【００３６】

作った５８種類のデータは、「太陽系惑星」「衛星」「十二支」「祝日」「スペースシャトル」「大河ドラマ」「相撲関連」「花の名称」「サッカー守備位置」「プロ野球選手名」「世界遺産」「村名と県名の対」「祝日と日付の対」「太陽系惑星と衛星の対」「作曲家と音楽作品の対」などと多様なものである。

表１：一項データの例(国名データ)

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
...

【００３７】

【数１】

表 2: 二項データの例 (国名 - 首都名対データ)

アイスランド アイスランド共和国 ＩＳＬ	レイキャビク レイキャヴィーク レイキャヴィク
アイルランド アイルランド共和国 ＩＲＬ	ダブリン
アゼルバイジャン アゼルバイジャン共和国 ＡＺＥ	バクー
アゾレス諸島	アングラ
アドゥイゲ アドゥイゲ共和国	マイコープ マイコプ
アフガニスタン アフガニスタン共和国	カブール ガブール
アメリカ アメリカ合衆国 米国 米 ＵＳＡ	ワシントン ワシントンＤ．Ｃ． ワシントンＤＣ
...	...

表３：データの規模

	一項データ	二項データ
データの種類の数	58	58
代表形数	17696	19387
代表形＋異表記数	26728	106850

（データの作成方法は以下の方法をとった）

ａ）単一の辞書・参考書などから手入力する。（例：太陽系惑星、衛星、十二支、祝日、スペースシャトル）

ｂ）Ｗｅｂのサイトから入手する。（例：大河ドラマの名称）

ｃ）複数の辞書・参考書またＷｅｂのサイトから得た情報を組み合わせる。（例：世界の山、花の名称、商品名）

ｄ）その分野の知識が豊富な人間が知識と記憶によって作成する。（例：サッカー守備位置）

（データの補充、異表記の作成には以下の方法をとった）

ａ）規則性を持った異表記を自動で生成する。（例：人名から姓を取り出す）

ｂ）規則性がない異表記をＷｅｂなどから取得する。（例：世界の山，花の名称，商品名）

ｃ）規則性がない異表記を思いつく範囲で入力する。（例：相撲決り手、星座）

ｄ）その分野の知識が豊富な人間が知識と記憶によって作成する。（例：サッカー守備位置）

データの作成の際には、それぞれのデータごとにその収集方法、異表記作成方法、代表形の基準の定義、異表記の基準の定義、その他のコメント、代表形の網羅度の情報を作成している。例えば、「国名データ」の代表形の基準の定義は「正式名称ではなく最も一般的に使用されるもの（例：フランス共和国→代表形「フランス」）」と記載されている。

【００３８】

代表形の網羅度としては、ほぼ１００％網羅している、網羅していないなどの情報を与えている。本発明で扱った用語の分類は国名や首都名や太陽系惑星など、用語の個数に限りがあることがわかっているものが多く、それらの場合はいつの時点での用語であるのかをはっきりさせておけば、ほぼ１００％網羅しているであろうデータを作ることができる。

【００３９】

このため、本発明のデータは、用語抽出の適合率・再現率の精度の計算に用いることができる。データの種類の方は５８種類と少なく、世の中の用語の種類すべてから比べるとかなり小さいが、そのデータの種類を犠牲にして、容易に収集できかつある特定の分野内

の用語に特定することでその分野内での網羅性をあげて適合率・再現率の精度計算可能なデータを作成しているのである。

【００４０】

また、本発明の用語データは普遍的なものであり、用語の自動抽出対象のデータがいかなるものであっても利用できる評価用データである。例えば、毎日新聞から抽出した結果をこのデータで評価をしてもよいし、またＷｅｂから抽出した結果をこのデータで評価をしてもよい。

【００４１】

一項データは普通の用語リストであるが、二項データは用語の対のデータであり、用語リストというよりは知識のようなものに近い。近年、質問応答の研究が重要視されつつあるが、質問応答システムでもこういう二項データをあらかじめ具備していればそういう関係の二項データでの質問応答を高性能に処理することができる。二項データはそういうシステムにも利用できるし、またそういうシステムのために作成する二項データ抽出システムの評価データにも利用できるのである。

【００４２】

（３）：用語抽出の具体的な説明

前記（２）により評価データ（正解データ）ができたので、これを使った簡単な用語抽出実験を行なってみた。この実験では網羅性が「ほぼ１００％網羅している」となっているデータのうち実験できるように代表形が１０個以上あったデータ（一項データで４０種類、二項データで４４種類）を用いた。ここで行なう実験では、少数の正例を使って学習し多くの正例を取ってくる正例のみによる学習を利用した（例えば、アメリカ、日本等の代表的な少数の国を入力し、他の国を抽出できるかどうかの実験）。

【００４３】

ここでの実験では異表記は正例とせず代表形のみを正例とした。入力の少数の正例としては、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人手で五つ選んだ。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年度版を正例の取得対象のデータ（データベース）Ｄとした。抽出の手順は以下のとおりである。

【００４４】

(1) 少数の正例をデータベースＤで全文検索し、複数の正例の周辺に出現したパターンをc_iとして抽出する（正例の周辺に出現するパターンがその正例だけ（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、正例の前後（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個の正例で９通りのパターンができることになる。また、正例を含めたパターンとすることもできる。

【００４５】

(2) 次に抽出したパターンc_iをデータベースＤ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。

【００４６】

(3) 抽出した表現ｅｘｐをＳｃｏｒｅの値の大きい順にソートして出力する。

【００４７】

Ｓｃｏｒｅとして、以下のものがある。

【００４８】

・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を使用するもの。

【００４９】

【数２】

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

・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ｃｉのｐｉを掛け合わせたものを使用する。

【００５０】

【数３】

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

・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。

【００５１】

【数４】

$$\text{Score} = \sum_i 1 \qquad (3)$$

・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、ｐｉ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。

【００５２】

【数５】

$$\text{Score} = \sum_i (1 + 0.01 p_i log(f_i)) \qquad (4)$$

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

【００５３】

・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。

【００５４】

【数６】

$$\text{Score} = 1 - \prod_i (1 - p_i) \qquad (5)$$

上記式(5) は、確からしくない（１－ｐｉ）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。

【００５５】

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) 。

【００５６】

ただし、f_i はパターンc_i が出現した入力正例の個数で、p_i はパターンc_i で抽出される表現での入力正例の割合（確からしさ、すなわち確信度となる）である。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。

【００５７】

一項データを使って、パターンとしては、正例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使って実験を行なった。その結果の抽出精度の比較を表４に示す。

【００５８】

【数７】

表 4: 一項データの比較実験

	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

表４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、正解正例分を上位から取ったたびに求めた適合率の平均（ただし、入力正例は正解正例から除く）である。

【００５９】

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、正解正例分だけを抽出した時に正解正例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、正解正例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。

【００６０】

（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

表４の例で抽出方法には、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。

【００６１】

字種を利用する方法では、入力した少数（この例では５個）の用語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個の用語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。

【００６２】

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 は入力正例の個数で、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正例の異表記は除いた。また、字種による方法以外にも次のような方法もある。

【００６３】

（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。

【００６４】

（例による説明１）

入力正例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、上記入力正例では、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。

【００６５】

（例による説明２）

入力正例として次のものであった場合、

「楽しい」「歓喜」「悲痛」「悲しい」

上記入力正例では、「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。

【００６６】

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。

【００６７】

・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、用語抽出部２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）。

【００６８】

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。

【００６９】

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。

【００７０】

（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。

【００７１】

（例による説明）

入力正例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」

上記入力正例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。

【００７２】
（ｄ）ユーザによる制約の指定
今までは、入力表現から自動で制約を得る方法でしたが、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。
【００７３】
なお、少数の正例を入力する研究には既に前記研究(3)があり、この研究(3)では抽出方法にブートストラップのアルゴリズムをくっつけて、入力正例を増やすことで精度高くより多くの用語を抽出する方法を使っている。本発明の方法の精度もブートストラップを使うことで改善されるだろう。研究(3)として手法４があるが、その式を少し変えた式やいろいろな工夫をした手法の結果も、評価用データがあるので比較できるのである。本発明の比較では、手法５で字種・ＫＲを利用する方法が最も良いことがわかる。
【００７４】
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)。
【００７５】
次に二項データを使った実験をした。パターンとしては、二項データのどちらが先に出現したかを示す正例の順序と一つ目に出現した正例の左と先頭のいずれかを含む１～３文字と二つの正例の間の表現の組み合わせと正例の順序と二つ目に出現した正例の右と後部のいずれかを含む１～３文字と二つの正例の間の表現の組み合わせを用いた。
【００７６】
正例の単語の境界が定まるように正例の左側か右側でパターンの文字列を取り出さない場合は、それらは平仮名文字であることをパターンの条件とした。こちらの実験では手法３で字種・ＫＲを利用する方法の精度が最も良くＡＰは0.026，ＲＰは0.040，ＴＰは0.141であった。
【００７７】
なお、抽出システムを高速化するため、パターンc_iを「入力正例の左と先頭のいずれかを含む１～３文字でかつ入力正例の右が平仮名文字であるもの、または、入力正例の右と後部のいずれかを含む１～３文字でかつ入力正例の左が平仮名文字であること」とすることができる。
【００７８】
§２：二次元表示の説明
（１）：表示システムの説明
図３は本発明の表示システムの説明図である。図３において、表示システムには、パターン抽出部１、用語抽出部２、用語表示部３、可視化装置４、データベースＤが設けてある。
【００７９】
パターン抽出部１は、（入力手段等から）入力された少数の正例（用語）をデータベースＤで全文検索し、該少数の正例の周辺に出現したパターンc_iを抽出するものである。用語抽出部２は、抽出したパターンc_iをデータベースＤ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）の値の大きい順にソートして（出力手段等に）出力するものである。用語表示部３は、各用語に対して、その用語の抽出に使われたパターンを文脈とし、それらの文脈の情報から各用語間の類似度を求めて、用語間の意味的距離を意味する行列を作るものである。可視化装置４は、自己組織化マップ等（二次元表示手段）を用いて二次元の図に表示（可視化）するものである。データベースＤは、例えば、新聞、雑誌、Ｗｅｂデータ等から抽出したデータ（一定量の文書データ）を格納するものである。

【００８０】
（フローチャートによる説明）
図４は表示処理フローチャートである。以下図４の処理Ｓ１１～Ｓ１４に従って説明する。
【００８１】
Ｓ１１：パターン抽出部１に、入力手段から少数の入力正例が入力される。
【００８２】
Ｓ１２：パターン抽出部１で、入力正例をデータベースＤで全文検索し、複数の入力正例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。）
Ｓ１３：用語抽出部２で、パターン抽出部１で抽出したパターンc_iをデータベースＤで全文検索し、パターンc_iによって抽出される表現（用語）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用語表示部３に渡す。
【００８３】
このとき、抽出された各用語にはどのパターンが使われたかのデータもくっつけておく。
【００８４】
Ｓ１４：用語表示部３で、各用語に対して、その用語の抽出に使われたパターンを文脈とし、それらの文脈の情報から各用語間の類似度を求め（同じパターンを持つ用語は類似度を高くする）て、用語間の意味的距離を意味する行列を作りこの行列を可視化装置４（自己組織化マップＳＯＭ＿ＰＡＫツール）に入力し、二次元マップでの各用語の座標を定める。そして、求まった座標に用語を表示した図を出力する。
【００８５】
（２）：可視化装置（自己組織化マップＳＯＭ＿ＰＡＫツール）の説明
意味マップの自動構築マシンとしてはKohonen の自己組織化型神経回路網モデルである自己組織化マップ（Self-Organizing Map,略してＳＯＭ）（Kohonen, T.: Self-organizing maps, 2nd edition,Springer,1997.）を用いる。ＳＯＭは高次元入力を持つ２次元配列のノードで構成され、以下に述べる自己組織化によって、高次元データをその特徴を反映するように２次元空間にマッピングすることができる。
【００８６】
【数８】

$$\text{入力 } x=[\xi_1,\xi_2,\cdots,\xi_N]^T\in\Re^N \text{ ならば、個々のノード } i \text{ はそれぞれ}$$
$$\text{参照ベクトル } m_i=[\mu_{i1},\mu_{i2},\cdots,\mu_{iN}]^T\in\Re^N \text{ を持つものとする。}$$

但し、参照ベクトルの要素μ_{ij}はノードｉと入力要素ξ_jの間の重みであり、自己組織過程において少しずつ修正される。入力ベクトルｘが与えられたとき、まず、その入力をすべてのノードの参照ベクトルと比較し、ユークリッド距離の一番短いノードを活性化する。マッピング処理段階ではこのノードのみ活性化される。このノードを勝者ノードと呼ぶ。即ち、勝者ノードｃは以下の式(6) のように選ばれる。
【００８７】
【数９】

$$c=argmin_i\{||x-m_i||\} \qquad (6)$$

一方、自己組織化過程では、グローバルに自己組織化が行われるように、勝者ノードだけでなくその近傍のノードも活性化させ、リラックス処理を行う。即ち、活性化されたす

べてのノードに対し、それらの参照ベクトルを入力ベクトルに近づくように修正を行う。
【００８８】
【数１０】

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \qquad (7)$$

ここで、ｔは学習回数で、$h_{ci}(t)$は、例えば以下の式(8)のように定義された近傍関数である。
【００８９】
【数１１】

$$h_{ci}(t) = \alpha(t) \cdot exp(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}) \qquad (8)$$

但し、$r_c \in \Re^2$ と $r_i \in \Re^2$ はそれぞれ勝者ノード c と近傍ノード i の位置ベクトルである。

従って、項$\| r_c - r_i \|$は近傍ノードｉが勝者ノードｃから離れて行くにつれ、h_{ci}が小さくなり$m_i(t)$の修正量が小さくなることを意味する。また、α(t)は学習率で、σ(t)は近傍の大きさ（半径）である。これらは時間と共に単調に減少していく関数であればよい。
【００９０】
通常、学習過程は「整列」フェーズと「微調整」フェーズからなる。「整列」フェーズにおいてはα(t)とσ(t)の初期値を共に大きく取り、時間と共に減少して行く。ノードの配置の基本形はこのフェーズで形成される。一方、残りのフェーズでは、α(t)とσ(t)は小さい値のまま長時間をかけて、初期フェーズで形成された基本形を微調整する。
【００９１】
本発明者らがこれまでに提案してきた可視化装置の例として、用語（ノード）間の意味的に近い単語どうしは近いところに、意味的に遠い単語どうしは離れたところに配置されるような、単言語の意味マップの自動構築手法がある（例えば、以降に記載の文献(6)．及び Ma, Q., Zhang, M., Murata, M., Zhou, M., Isahara, H.: Self-Organizing Chinese and Japanese Semantic Maps, The 19th International Conference on Computational Linguistics (COLING'2002), Taiwan, pp. 605-611, August,2002. 参照）。
【００９２】
（３）：具体例による説明
前記の用語抽出手法は、文字列の全文検索を使うので高速に計算できる特徴を持つ。この文字列の全文検索には、suffix arrayという高速検索アルゴリズムがある。そこで前記の手法を用いた簡易な応用システムを開発した。まず、ユーザは、好きな数語を入力する。この数語を少数の正例として前記の字種・ＫＲを利用する手法５を利用して入力の数語と同じ分野の語を収集する。
【００９３】
この時高速化のため前記§１の（３）（用語抽出の具体的な説明）のアルゴリズムでパターンc_iは、より簡易なものにし、また途中の計算過程で何回かデータを閾値で足切りして減らしている。なお、正確には前記の実験のプログラムも少々は足切りなどをしている。それでも前記のものは一つ１０時間ぐらいかかる。
【００９４】
得られた同じ分野の語を見やすいように可視化して出力する。可視化には、下記文献(6)の意味マップを用いた。意味マップの利用では高速化のため学習回数を１／１０にして

いる。それ以外はすべて文献(6)と同じ方法を用いた。文脈にはパターンc_iから正例の内部表現を使う素性を除いたものを利用した。なお、本発明では用語抽出に用いたパターンc_iを意味マップの文脈に用いたが、パターンc_i以外に各用語の文脈を大規模データから抽出し直して意味マップの表示をするという方法もありえる。

【００９５】

この実装で、現在約２０秒での動作を実現している。（現在Perlでの実装部分もあるし今後のアルゴリズムの改良によっても速度をあげることは可能と思われる。）このシステムの性能は高速化で削った情報のため、一項データでは表４より精度が下がりＡＰは0.111、ＲＰは0.164、ＴＰは0.310であった。

【００９６】

文献(6):馬青, 神崎享子, 村田真樹, 内元清貴, 井佐原均“日本語名詞の意味マップの自己組織化”情報処理学会論文誌, Vol.42, No.10,(2001),pp.2379-2391。

【００９７】

（具体例(1)）

図５はユーザによる「色」に関する用語抽出の説明図である。図５において、実際にユーザがこのシステムに「赤色」「青色」「黄色」「紫色」「茶色」を入力した場合の出力（上位２０個の出力）を示してある。ここでは入力（入力正例）には“●”の記号を付けている。また各表現につけている数字は抽出手法で何番目に得られたかを示している。一般には意味マップにより類似したパターンを持つ表現が近くに配置され見やすくなる。この例だと左上の「赤色」の近くに「朱色」「紅色」が出現しており、類似した表現が近くに出現している場合があることがわかる。また、右の上から真ん中辺りに「音色」「異色」「同色」という色ではない表現が固まっているが、意味マップでの可視化ではこのように不適合な表現をどこかに集めてくれて（順位では集まらない）それを除けばよいということを見やすく示す効果もある。このように、不適合な表現は入力正例である“●”の用語と離れた位置に集まる傾向がある。

【００９８】

（具体例(2)）

前記の例等では、入力正例として名詞を用いる説明をしたが、形容詞などの評価表現も扱うことができる。例えば、パターン抽出部へ入力する入力正例として、「●悲しい」「●楽しい」「●哀しい」「●嬉しい」だと、用語抽出部の出力として次の用語を得ることができる。

【００９９】

「新しい」「美しい」「難しい」「厳しい」「激しい」「優しい」「寂しい」「珍しい」「苦しい」「正しい」「貧しい」「詳しい」「乏しい」「涼しい」「親しい」「欲しい」「悔しい」「真新しい」「忙しい」「著しい」「等しい」「重苦しい」「美味しい」「惜しい」「礼儀正しい」「心優しい」「生易しい」「堅苦しい」

このように、入力する用語は、名詞ばかりでなく、どのようなものも扱うことができる。

【０１００】

なお、可視化装置として自己組織化マップの方法について説明したが、主成分分析を用いる方法など他の方法を使用することもできる。また、入力正例に“●”の記号を付けているが、これ以外に、入力正例がわかるように何か別の印を付けるか色をかえるようにしてもよい。

【０１０１】

ここまでの例では、文脈の情報から、意味的距離を意味する行列を作り、これを可視化装置に入力する例を示したが、可視化装置自体は、他の行列を入力しても可視化することができる。例えば、ある語に対する文脈の情報を右に並べたような次の行列を入力して用語を可視化することも可能である。

文脈１　文脈２　文脈３　・・・・　文脈Ｎ

用語１	2	0	1	1
用語２	2	1	1	2
・・・・				
用語Ｍ	0	0	1	0

上記行列の各要素の数字は、その文脈でのその用語の出現回数を意味する。また、主成分分析による手法でも、このような形式のものを入力として可視化することが可能である。

【０１０２】

（４）：まとめ

以上のように、本発明では、用語抽出の際の評価に用いることができるデータを作成した。この評価データとして「国名」や「国名と首都名の組」など一項データと二項データを作成した。本発明のデータは用語の網羅性が高く、どのようなデータからの用語抽出においても再現率・適合率を算出するなどの性能評価ができるものである。また、本願で作成した二項データは、用語リストというよりは知識に近いもので、質問応答などの知識処理の研究にも用いることができるものである。

【０１０３】

また、本願では簡単な用語抽出実験を行ない、種々の用語抽出方法の比較を行なった。一項データでは字種・ＫＲを利用する手法５（文献(4)）が最も良く、二項データでは字種・ＫＲを利用する手法３（類似度に基づく方法）が最も良かった。さらに、ユーザは好きな数語を入力すると約２０秒でその数語と同じ分野の語を収集して可視化して表示する応用システムを示した。このシステムでは類似した用語が集まり、また不要な用語も集まる傾向があり、抽出結果の用語の表示としては便利な特徴を持っている。

【０１０４】

なお、上記の例では、入力正例として２つの用語の対（二項データ）を用いた説明をしたが、３つ以上（三項以上）の用語の対を用いても同様に実施することができる。

【０１０５】

（５）：プログラムインストールの説明

パターン抽出部１、用語抽出部２、用語表示部３、可視化装置（可視化手段）４等は、プログラムで構成でき、主制御部（ＣＰＵ）が実行するものであり、主記憶に格納されているものである。このプログラムは、一般的な、コンピュータで処理されるものである。このコンピュータは、主制御部、主記憶、ファイル装置、表示装置、キーボード等の入力手段である入力装置などのハードウェアで構成されている。

【０１０６】

このコンピュータに、本発明のプログラムをインストールする。このインストールは、フロッピィ、光磁気ディスク等の可搬型の記録（記憶）媒体に、これらのプログラムを記憶させておき、コンピュータが備えている記録媒体に対して、アクセスするためのドライブ装置を介して、或いは、ＬＡＮ等のネットワークを介して、コンピュータに設けられたファイル装置にインストールされる。そして、このファイル装置から処理に必要なプログラムステップを主記憶に読み出し、主制御部が実行するものである。

【図面の簡単な説明】

【０１０７】

【図１】本発明の抽出システムの説明図である。

【図２】本発明の用語抽出処理フローチャートである。

【図３】本発明の表示システムの説明図である。

【図４】本発明の表示処理フローチャートである。

【図５】本発明のユーザによる「色」に関する用語抽出の説明図である。

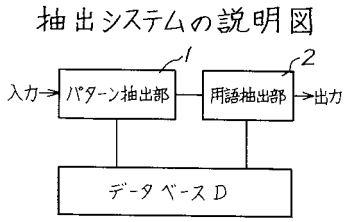
【符号の説明】

【０１０８】

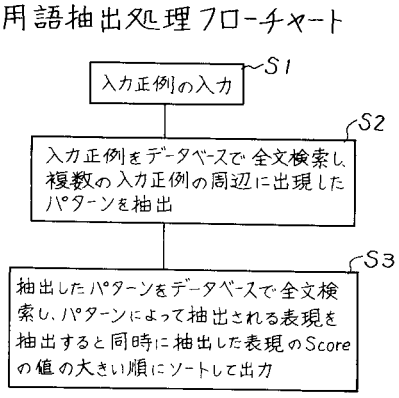
１　パターン抽出部

2　用語抽出部
3　用語表示部
4　可視化装置（可視化手段）
D　データベース

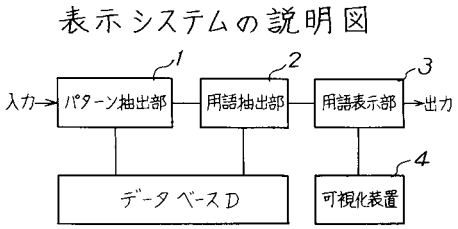
【図１】



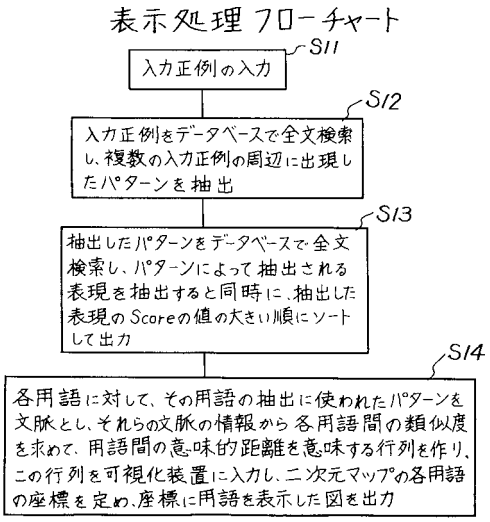
【図２】



【図３】

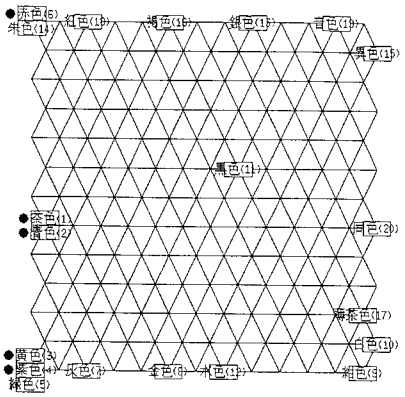


【図４】



【図５】

ユーザによる「色」に関する用語抽出の説明図



フロントページの続き

(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 ND03 NK32 NK35 NR06 NS10 PP02 PP22 PQ02 PQ12 PQ13
PQ74 PR06 QM05 QM08 QP01 QP03 UU06