(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2008-9671
(P2008-9671A)
(43)公開日　平成20年1月17日(2008.1.17)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	210D	5B075
		G06F	17/30	380E	
		G06F	17/30	170A	

審査請求 未請求　請求項の数 16　ＯＬ　（全 24 頁）

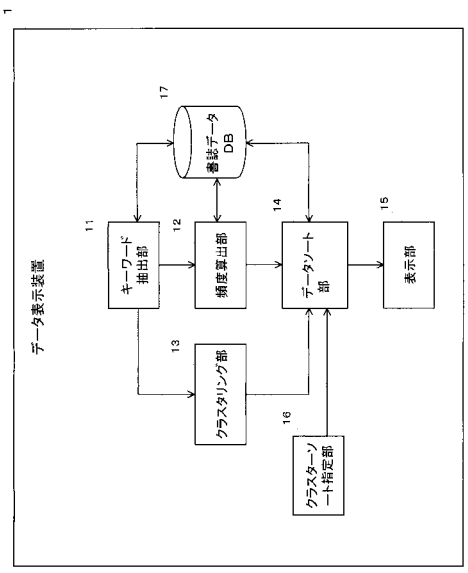
(21)出願番号　特願2006-178922 (P2006-178922)
(22)出願日　平成18年6月29日 (2006.6.29)

(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

Ｆターム(参考)　5B075 ND03　NK02　NK32　NR12　NR15
NS03　PQ02　PQ15　PQ46　PR04
QP01　UU06

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

(57)【要約】
【課題】文書の大凡の内容を一見して把握できるような表示を行うことを可能とする。
【解決手段】キーワード抽出部１１が書誌データＤＢ１７に蓄積されている文書群に含まれるキーワードを抽出し、頻度算出部１２がキーワード抽出部１１によって抽出された各キーワードの、書誌データＤＢ１７に蓄積されている文書群中に出現した頻度を算出し、データソート部１４が書誌データＤＢ１７に蓄積されている各文書の文書データを抽出し、各キーワードの頻度がより高いキーワードを多く含む文書の順に、各文書の文書データをソートし、クラスタリング部１３がキーワード抽出部１１によって抽出された各キーワードをクラスターにクラスタリングし、表示部１５がデータソート部１４によってソートされた文書データに対応付けて、上記クラスター中のキーワードのうち当該ソートされた文書データが持つ文書が含むキーワードを、クラスター毎に画面表示する。
【選択図】図１



【特許請求の範囲】
【請求項１】
データ表示装置であって、
複数の文書から構成される文書群中に含まれるキーワードを抽出するキーワード抽出手段と、
前記抽出されたキーワードを各キーワードが属するクラスターにクラスタリングするクラスタリング手段と、
前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を、前記各文書と前記クラスターとに対応付けて画面表示する表示手段とを備える
ことを特徴とするデータ表示装置。
【請求項２】
請求項１に記載のデータ表示装置において、
前記クラスタリング手段は、前記各キーワードのベクトル空間上の位置を示す位置ベクトルを生成し、生成された前記位置ベクトルが示す前記各キーワードの位置情報に基づいて、前記各キーワードが属するクラスターを決定する
ことを特徴とするデータ表示装置。
【請求項３】
請求項１又は請求項２に記載のデータ表示装置において、
前記表示手段は、予め定められた順序に基づいて、前記各クラスターを並び替えて画面表示する
ことを特徴とするデータ表示装置。
【請求項４】
請求項１又は請求項２に記載のデータ表示装置において、
前記表示手段は、各クラスターに属するキーワード又は前記各クラスターに属するキーワードが出現する文書の分布に基づいて、前記各クラスターを並び替えて画面表示する
ことを特徴とするデータ表示装置。
【請求項５】
請求項４に記載のデータ表示装置において、
前記表示手段は、各クラスターに属するキーワードが出現する文書数について昇順又は降順に前記各クラスターを並び替えて画面表示する
ことを特徴とするデータ表示装置。
【請求項６】
請求項２に記載のデータ表示装置において、
前記クラスタリング手段は、前記各キーワードの位置ベクトルの平均を前記各キーワードが属するクラスターの位置ベクトルとし、各クラスターの位置ベクトルに基づいて、クラスター同士の距離を求め、求めた距離が近いクラスター同士が近い位置に並び、求めた距離が遠いクラスター同士が離れた位置に並ぶように、前記各クラスターを並び替え、
前記表示手段は、前記並び替えられた各クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記並び替えられた各クラスターに属するキーワードの有無を示す情報を、前記各文書と前記並び替えられた各クラスターとに対応付けて画面表示する
ことを特徴とするデータ表示装置。
【請求項７】
請求項６に記載のデータ表示装置において、
前記クラスタリング手段は、ベクトル空間上に並んで配置された複数の基準の位置ベクトルを予め定義し、前記各クラスターの位置ベクトルを、前記各クラスターの位置ベクトルとの距離が最も近い基準の位置ベクトルの近くに配置することにより、前記各クラスターを並び替える
ことを特徴とするデータ表示装置。

【請求項８】
請求項６又は請求項７に記載のデータ表示装置において、
前記クラスタリング手段は、前記各文書がどのクラスターのどの単語を何個含んでいるかの情報を求め、求めた情報に基づいて、前記各文書のベクトル空間上の位置を示す位置ベクトルを求め、求めた前記各文書の位置ベクトルに基づいて、文書同士の距離を求め、求めた距離が近い文書同士が近い位置に並び、求めた距離が遠い文書同士が離れた位置に並ぶように、各文書を並び替える
ことを特徴とするデータ表示装置。
【請求項９】
請求項４乃至請求項８のいずれか１項に記載のデータ表示装置において、
前記表示手段は、並ぶ順序がより上位のクラスターに属するキーワードを含む文書の順に、前記文書に含まれるキーワード、又は、前記文書中における前記クラスターに属するキーワードの有無を示す情報を画面表示する
ことを特徴とするデータ表示装置。
【請求項１０】
請求項１乃至請求項９のいずれか１項に記載のデータ表示装置において、
前記表示手段は、前記クラスタリング手段によってクラスタリングされたクラスターを選択し、選択されたクラスターに対応付けて、前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を画面表示する
ことを特徴とするデータ表示装置。
【請求項１１】
請求項１に記載のデータ表示装置において、
前記クラスタリング手段は、前記各キーワードのベクトル空間上の位置を示す位置ベクトルを生成し、生成された前記位置ベクトルが示す前記各キーワードの位置情報に基づいて、キーワード同士の距離を求め、求めた距離が近いキーワード同士が近い位置に並び、求めた距離が遠いキーワード同士が離れた位置に並ぶように、各キーワードを並び替え、
前記表示手段は、並び替えられたキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記並び替えられたキーワードの有無を示す情報を、前記各文書に対応付けて画面表示する
ことを特徴とするデータ表示装置。
【請求項１２】
請求項１に記載のデータ表示装置において、
前記表示手段は、各クラスターを画面上に並ばせる順序を選択し、前記選択された順序がより上位のクラスターに属するキーワードを含む文書の順に、前記文書に含まれるキーワード、又は、前記文書中における前記クラスターに属するキーワードの有無を示す情報を画面表示する
ことを特徴とするデータ表示装置。
【請求項１３】
データ表示装置であって、
複数の文書から構成される文書群中に含まれるキーワードを抽出するキーワード抽出手段と、
前記抽出された各キーワードの、前記文書群中に出現する頻度を算出する頻度算出手段と、
前記各キーワードを前記各キーワードが属するクラスターにクラスタリングするクラスタリング手段と、
前記算出された頻度がより高いキーワードを含む文書の順に、前記文書群を構成する文書の文書データをソートするデータソート手段と、
前記データソート手段によってソートされた文書データに対応付けて、前記クラスターに属するキーワードのうち前記ソートされた文書データを持つ文書が含むキーワードを、

前記クラスター毎に画面表示する表示手段とを備える
ことを特徴とするデータ表示装置。
【請求項１４】
請求項１３に記載のデータ表示装置において、
前記データソート手段は、さらに、前記クラスターを前記クラスターに属するキーワードの数について降順又は昇順にソートする
ことを特徴とするデータ表示装置。
【請求項１５】
データ表示方法であって、
複数の文書から構成される文書群中に含まれるキーワードを抽出し、
前記各キーワードを前記各キーワードが属するクラスターにクラスタリングし、
前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を、前記クラスター毎に画面表示する
ことを特徴とするデータ表示方法。
【請求項１６】
データ表示プログラムであって、
コンピュータに、
複数の文書から構成される文書群中に含まれるキーワードを抽出する処理と、
前記各キーワードを前記各キーワードが属するクラスターにクラスタリングする処理と、
前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を、前記クラスター毎に画面表示する処理とを実行させる
ことを特徴とする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、文書の大凡の内容を一見して把握できるような表示を行うデータ表示装置、データ表示方法及びデータ表示プログラムに関する。
【背景技術】
【０００２】
従来から、文書中に含まれるキーワードの頻度を算出し、算出された頻度を表示する技術は存在した。
【０００３】
また、例えば、データを表形式で表示する技術について、下記の非特許文献１に記載されている。
【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】
【０００４】
しかし、上記従来技術は、文書群中に含まれるキーワードをキーワード群（クラスター）毎にクラスタリングし、当該クラスターに属するキーワードのうちの文書群を構成する文書に含まれるキーワード、又は、各文書中における当該クラスターに属するキーワードの有無を示す情報を、当該各文書と当該クラスターとに対応付けて画面表示することを行っていない。
【０００５】
従って、従来技術では、画面表示された各文書データが含むキーワードを参照して、当該文書データを持つ文書の大凡の内容を一見して把握することは困難である。

【０００６】
本発明は、文書の大凡の内容を一見して把握できるような表示を行うデータ表示装置、データ表示方法及びデータ表示プログラムの提供を目的とする。
【課題を解決するための手段】
【０００７】
本発明のデータ表示装置は、複数の文書から構成される文書群中に含まれるキーワードを抽出するキーワード抽出手段と、前記抽出されたキーワードを各キーワードが属するクラスターにクラスタリングするクラスタリング手段と、前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を、前記各文書と前記クラスターとに対応付けて画面表示する表示手段とを備える。
【０００８】
好ましくは、本発明のデータ表示装置において、前記クラスタリング手段は、前記各キーワードのベクトル空間上の位置を示す位置ベクトルを生成し、生成された前記位置ベクトルが示す前記各キーワードの位置情報に基づいて、前記各キーワードが属するクラスターを決定する。
【０００９】
好ましくは、本発明のデータ表示装置において、前記表示手段は、予め定められた順序に基づいて、前記各クラスターを並び替えて画面表示する。
【００１０】
好ましくは、本発明のデータ表示装置において、前記表示手段は、各クラスターに属するキーワード又は前記各クラスターに属するキーワードが出現する文書の分布に基づいて、前記各クラスターを並び替えて画面表示する。
【００１１】
好ましくは、本発明のデータ表示装置において、表示手段は、各クラスターに属するキーワードが出現する文書数について昇順又は降順に前記各クラスターを並び替えて画面表示する。
【００１２】
好ましくは、本発明のデータ表示装置において、前記クラスタリング手段は、前記各キーワードの位置ベクトルの平均を前記各キーワードが属するクラスターの位置ベクトルとし、各クラスターの位置ベクトルに基づいて、クラスター同士の距離を求め、求めた距離が近いクラスター同士が近い位置に並び、求めた距離が遠いクラスター同士が離れた位置に並ぶように、前記各クラスターを並び替え、前記表示手段は、前記並び替えられた各クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記並び替えられた各クラスターに属するキーワードの有無を示す情報を、前記各文書と前記並び替えられた各クラスターとに対応付けて画面表示する。
【００１３】
好ましくは、本発明のデータ表示装置において、前記クラスタリング手段は、ベクトル空間上に並んで配置された複数の基準の位置ベクトルを予め定義し、前記各クラスターの位置ベクトルを、前記各クラスターの位置ベクトルとの距離が最も近い基準の位置ベクトルの近くに配置することにより、前記各クラスターを並び替える。
【００１４】
好ましくは、本発明のデータ表示装置において、前記クラスタリング手段は、前記各文書がどのクラスターのどの単語を何個含んでいるかの情報を求め、求めた情報に基づいて、前記各文書のベクトル空間上の位置を示す位置ベクトルを求め、求めた前記各文書の位置ベクトルに基づいて、文書同士の距離を求め、求めた距離が近い文書同士が近い位置に並び、求めた距離が遠い文書同士が離れた位置に並ぶように、各文書を並び替える。
【００１５】
好ましくは、本発明のデータ表示装置において、前記表示手段は、並ぶ順序がより上位のクラスターに属するキーワードを含む文書の順に、前記文書に含まれるキーワード、又

は、前記文書中における前記クラスターに属するキーワードの有無を示す情報を画面表示する。

【００１６】

好ましくは、本発明のデータ表示装置において、前記表示手段は、前記クラスタリング手段によってクラスタリングされたクラスターを選択し、選択されたクラスターに対応付けて、前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を画面表示する。

【００１７】

好ましくは、本発明のデータ表示装置において、前記クラスタリング手段は、前記各キーワードのベクトル空間上の位置を示す位置ベクトルを生成し、生成された前記位置ベクトルが示す前記各キーワードの位置情報に基づいて、キーワード同士の距離を求め、求めた距離が近いキーワード同士が近い位置に並び、求めた距離が遠いキーワード同士が離れた位置に並ぶように、各キーワードを並び替え、前記表示手段は、並び替えられたキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記並び替えられたキーワードの有無を示す情報を、前記各文書に対応付けて画面表示する。

【００１８】

好ましくは、本発明のデータ表示装置において、前記表示手段は、各クラスターを画面上に並ばせる順序を選択し、前記選択された順序がより上位のクラスターに属するキーワードを含む文書の順に、前記文書に含まれるキーワード、又は、前記文書中における前記クラスターに属するキーワードの有無を示す情報を画面表示する。

【００１９】

また、本発明のデータ表示装置は、複数の文書から構成される文書群中に含まれるキーワードを抽出するキーワード抽出手段と、前記抽出された各キーワードの、前記文書群中に出現する頻度を算出する頻度算出手段と、前記各キーワードを前記各キーワードが属するクラスターにクラスタリングするクラスタリング手段と、前記算出された頻度がより高いキーワードを含む文書の順に、前記文書群を構成する文書の文書データをソートするデータソート手段と、前記データソート手段によってソートされた文書データに対応付けて、前記クラスターに属するキーワードのうち前記ソートされた文書データを持つ文書が含むキーワードを、前記クラスター毎に画面表示する表示手段とを備える。

【００２０】

好ましくは、本発明のデータ表示装置において、前記データソート手段は、さらに、前記クラスターを前記クラスターに属するキーワードの数について降順又は昇順にソートする。

【００２１】

また、本発明のデータ表示方法は、複数の文書から構成される文書群中に含まれるキーワードを抽出し、前記各キーワードを前記各キーワードが属するクラスターにクラスタリングし、前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を、前記クラスター毎に画面表示する。

【００２２】

また、本発明のデータ表示プログラムは、コンピュータに、複数の文書から構成される文書群中に含まれるキーワードを抽出する処理と、前記各キーワードを前記各キーワードが属するクラスターにクラスタリングする処理と、前記クラスターに属するキーワードのうちの前記文書群を構成する文書に含まれるキーワード、又は、各文書中における前記クラスターに属するキーワードの有無を示す情報を、前記クラスター毎に画面表示する処理とを実行させる。

【発明の効果】

【００２３】

本発明のデータ表示装置、データ表示方法及びデータ表示プログラムは、複数の文書から構成される文書群中に含まれるキーワードを抽出し、上記各キーワードを上記各キーワードが属するクラスターにクラスタリングし、当該クラスターに属するキーワードのうちの上記文書群を構成する文書に含まれるキーワード、又は、各文書中における当該クラスターに属するキーワードの有無を示す情報を、当該各文書と当該クラスターとに対応付けて画面表示する。従って、本発明によれば、文書の大凡の内容を一見して把握できるような表示を行うことが可能となる。

【発明を実施するための最良の形態】

【００２４】

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、文書群中の文書の文書データをソートして画面表示する処理装置である。

【００２５】

データ表示装置１は、キーワード抽出部１１、頻度算出部１２、クラスタリング部１３、データソート部１４、表示部１５、クラスターソート指定部１６、書誌データデータベース（ＤＢ）１７を備える。

【００２６】

キーワード抽出部１１は、書誌データＤＢ１７に蓄積されている文書群に含まれるキーワードを抽出する。キーワード抽出部１１によるキーワードの抽出手法については、後述する。

【００２７】

頻度算出部１２は、キーワード抽出部１１によって抽出された各キーワードの、書誌データＤＢ１７に蓄積されている文書群中に出現した頻度を算出する。ここで、キーワードの頻度とは、例えば、キーワードが出現する文書の数を意味する。例えば、キーワード「本塁打記録」を含む文書数が２０である場合には、算出されるキーワード「本塁打記録」の頻度は２０である。また、本発明の実施の形態においては、文書群でのキーワードの出現回数をキーワードの頻度とする構成を採ることもできる。

【００２８】

クラスタリング部１３は、キーワード抽出部１１によって抽出されたキーワードを、１又は複数のクラスターにクラスタリング（分類）する。ここで、クラスターとは、１又は複数のキーワードで構成されるキーワード群である。クラスタリング部１３は、例えば、後述するように、各キーワードの位置（ベクトル空間における位置）を示すベクトル（位置ベクトル）を生成する。そして、クラスタリング部１３は、生成されたベクトルが示す各キーワードの位置情報に基づいて、各キーワードが属するクラスターを決定する。例えば、キーワード「リーグ」、「米大リーグ」、「大リーグ」をあるクラスターに属するキーワードとしてクラスタリングし、キーワード「試合」、「チーム」、「スタジアム」を、他のクラスターに属するキーワードとしてクラスタリングする。

【００２９】

データソート部１４は、書誌データＤＢ１７に蓄積されている文書から、文書データ（例えば、文書中のテキストデータ、文書のタイトル、著者名等）を抽出し、抽出した文書データをソートする。

【００３０】

すなわち、データソート部１４は、まず、抽出した文書データを図示しないバッファ中に格納する。そして、データソート部１４は、頻度算出部１２によって算出された各キーワードの頻度がより高いキーワードを多く含む文書の順に、上記バッファ中に格納された文書の文書データをソートする。データソート部１４は、後述するクラスターソート指定部１６によって指示された、各クラスター（キーワード群）のソート内容に従って、各クラスターをソートするようにしてもよい。

【００３１】

なお、本発明の一実施形態によれば、データソート部１４は、書誌データＤＢ１７から

抽出した各文書から各文書に関連する日付（例えば、発行日）のデータを抽出し、文書データ（例えば、文書のテキストデータ、文書のタイトル、著者名等）を日付について降順又は昇順にソートする構成を採ってもよい。

【００３２】

表示部１５は、データソート部１４によってソートされた文書データに対応付けて、上記クラスター中のキーワードのうち当該ソートされた文書データを持つ文書が含むキーワードを、クラスター毎に画面表示する。表示部１５は、データソート部１４によってソートされた文書データに対応付けて、データソート部１４によってソートされたクラスター中のキーワードのうち当該ソートされた文書データを持つ文書が含むキーワードを、データソート部１４によってソートされた各クラスター毎に画面表示するようにしてもよい。

【００３３】

クラスターソート指定部１６は、データソート部１４による各クラスターのソート内容を指定する。例えば、クラスターソート指定部１６は、各クラスターを、クラスター中のキーワードの数について降順にソートすることを指示する制御情報をデータソート部１４に対して送出する。クラスターソート指定部１６は、各クラスターを、クラスター中のキーワードの数について昇順にソートすることを指示する制御情報をデータソート部１４に対して送出するようにしてもよい。また、クラスターソート指定部１６は、例えば、後述する各クラスター間の距離に基づいて各クラスターをソートすることを指示する制御情報をデータソート部１４に対して送出するようにしてもよい。例えば、クラスターソート指定部１６は、後述する各クラスター間の距離が近いクラスター同士が直線上に結合するように各クラスターをソートすることを指示する制御情報をデータソート部１４に対して送出する。当該制御情報を受けたデータソート部１４は、例えば、距離が最も近いクラスター同士をまず結合し、結合したクラスターによって構成される、複数のクラスター群を作成する。そして、データソート部１４は、例えば、あるクラスター群におけるクラスターのいずれかと距離が最も近い他のクラスター群のクラスターを、当該距離が最も近いクラスターと結合する処理を繰り返して、クラスターを直線上に結合する。

【００３４】

書誌データＤＢ１７には、大量の文書（書誌データ）が蓄積されている。

【００３５】

本発明の一実施形態によれば、頻度算出部１２は、キーワード抽出部１１が抽出した各キーワードについて算出した頻度と、各キーワードの文字数とに基づいて、各キーワードのスコアを算出するようにしてもよい。各キーワードのスコアは、例えば、各キーワードの文字数に頻度を乗じた値として算出される。

【００３６】

本発明の実施の形態においては、頻度算出部１２が、キーワード抽出部１１によって抽出された各キーワードの文字数を用いずに、各キーワードについて算出された頻度に基づいて、所定の計算式を用いて、各キーワードのスコアを算出する構成を採ってもよい。

【００３７】

例えば、頻度算出部１２は、以下に示すような、ＴＦ／ＩＤＦ法を用いたスコアの算出方法又はＯｋａｐｉのウェイティング法を用いて、各キーワードのスコアを算出する。

【００３８】

（ＴＦ／ＩＤＦ法を用いたスコアの算出方法）

一般に、重要なキーワードを含む文書の検索には、主にＴＦ／ＩＤＦ法が用いられる。ここで、ＴＦとは、一般に、ある文書でのあるキーワードの出現回数を意味し、ＩＤＦとは、一般に、予め用意された多数の文書のうち、上記キーワードが出現する文書数の逆数を意味する。

【００３９】

一般に、ＴＦ／ＩＤＦ法では、以下の式で算出されるＳｃｏｒｅ（Ｄ）が高い文書を検索結果として出力する。

【００４０】

Ｓｃｏｒｅ（Ｄ）＝Σ（ｔｆ（ｗ，Ｄ）×ｌｏｇ（Ｎ／ｄｆ（ｗ））） 上記の式において、ｗは、ユーザが入力するキーワード、Σは、ｔｆ（ｗ，Ｄ）×ｌｏｇ（Ｎ／ｄｆ（ｗ））をｗ∈Ｗで加算することを意味する。また、ｔｆ（ｗ，Ｄ）は、文書Ｄでのｗの出現回数であり、ｄｆ（ｗ）は、全文書においてｗが出現した文書の数であり、Ｎは、文書の総数である。

【００４１】

ＴＦ／ＩＤＦ法の本発明への適用に当たっては、例えば、上記文書Ｄを、書誌データＤＢ１７に蓄積されている文書群として、ｔｆ（ｗ，Ｄ）を算出する。また、例えば、書誌データＤＢ１７とは別のデータベース（図示を省略）に蓄積されている大量の文書群を、上記ｄｆ（ｗ）の意味の説明において記述した「全文書」として、ｄｆ（ｗ）を算出する。

【００４２】

そして、算出されたｔｆ（ｗ，Ｄ）とｌｏｇ（Ｎ／ｄｆ（ｗ））との積を、各キーワードｗのスコアとして算出する。

【００４３】

（Ｏｋａｐｉのウェイティング法を用いたスコアの算出方法）

一般に、Ｏｋａｐｉのウェイティング法（下記の文献（１）参照）では、以下の式で算出されるＳｃｏｒｅ（Ｄ）が高い文書を検索結果として出力する。

【００４４】

文献（１）：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，自然言語処理（言語処理学会誌），２０００年４月，７巻，２号，ｐ．１４１～ｐ．１６０

【００４５】

【数１】

$$\mathrm{Score}(D)=\sum_{w\in W}\left(\frac{\mathrm{tf}(w,D)}{\frac{\mathrm{length}(D)}{\Delta}+\mathrm{tf}(w,D)}\times\log\frac{N}{\mathrm{df}(w)}\right)$$

【００４６】

ここで、ｗは、ユーザが入力するキーワード、Ｗは、ユーザが入力するキーワードの集合を意味する。また、ｔｆ（ｗ，Ｄ）は、文書Ｄでのｗの出現回数であり、ｄｆ（ｗ）は、全文書においてｗが出現した文書の数であり、Ｎは、文書の総数である。また、ｌｅｎｇｔｈ（Ｄ）は、文書Ｄの長さ（文字列単位）である。Δは、全文書における文書の長さの平均である。

【００４７】

Ｏｋａｐｉのウェイティング法の本発明への適用に当たっては、例えば、上記文書Ｄを、書誌データＤＢ１７に蓄積されている文書群として、

【００４８】

【数２】

$$\frac{\mathrm{tf}(w,D)}{\frac{\mathrm{length}(D)}{\Delta}+\mathrm{tf}(w,D)}$$

【００４９】

を算出する。算出された値をｔｆ項とする。

【００５０】
また、例えば、書誌データＤＢ１７とは別のデータベース（図示を省略）に蓄積されている大量の文書群を、上記ｄｆ（ｗ）の意味の説明において記述した「全文書」として、ｌｏｇ（Ｎ／ｄｆ（ｗ））を算出する。算出されたｌｏｇ（Ｎ／ｄｆ（ｗ））をｉｄｆ項とする。そして、算出されたｔｆ項とｉｄｆ項との積を、各キーワードｗのスコアとして算出する。
【００５１】
また、本発明の一実施形態によれば、データソート部１４は、頻度算出部１２によって算出された各キーワードのスコアがより高いキーワードを多く含む文書の順に、各文書の文書データをソートするようにしてもよい。
【００５２】
次に、クラスタリング部１３による、キーワードのクラスタリング処理について具体的に説明する。クラスタリングには、以下に示すような様々な方法がある。
（階層クラスタリング（ボトムアップクラスタリング）による方法）
距離が最も近い成員同士を結合していき、クラスターを作る。そして、距離が最も近いクラスター同士を結合する。成員とは、クラスタリングの対象となるキーワード（単語）であって、あるクラスターに属するキーワード（単語）である。クラスター間の距離の定義は様々ある。例えば、クラスターＡとクラスターＢとの距離を、クラスターＡの成員（すなわち、クラスターＡに属するキーワード）とクラスターＢの成員（すなわち、クラスターＢに属するキーワード）との距離の中で最も小さいものとしてもよい。ここで、成員と成員との距離とは、ベクトルで表現される後述する成員の位置間の距離である。また、例えば、クラスターＡとクラスターＢとの距離を、クラスターＡの成員とクラスターＢの成員との距離の中で最も大きいものとしてもよい。また、例えば、クラスターＡとクラスターＢとの距離を、全てのクラスターＡの成員とクラスターＢの成員との距離の平均としてもよい。また、全てのクラスターＡの成員の位置の平均をクラスターＡの位置とし、全てのクラスターＢの成員の位置の平均をクラスターＢの位置とし、当該クラスターＡの位置とクラスターＢの位置との距離をクラスターＡとクラスターＢとの距離としてもよい。
【００５３】
（ウォード法による方法）
以下に示すＷを定義する。
Ｗ　＝　ΣΣ（ｘ（ｉ，ｊ）－ａｖｅ＿ｘ（ｉ））＾２
＾は指数を意味する。例えば、上記の式における１つ目のΣは、ｉ＝１からｉ＝ｇまでの加算、２つ目のΣは、ｊ＝１からｊ＝ｎｉまでの加算を意味する。また、ｘ（ｉ，ｊ）は、ｉ番目のクラスターのｊ番目の成員の位置、ａｖｅ＿ｘ（ｉ）は、ｉ番目のクラスターの全ての成員の位置の平均を意味する。クラスター同士を結合していくと、Ｗの値が増加するが、ウォード法では、Ｗの値がなるべく大きくならないようにクラスター同士を結合していく。
【００５４】
（クラスタリングの終了条件）
予めクラスターの個数を決めておいて、クラスターの個数が当該予め決められた数になったときに、クラスター同士を結合するのをやめるようにしてもよい。また、予め距離の閾値を決めておいて、その閾値数以上離れているクラスター同士を結合するのをやめるようにしてもよい。
【００５５】
（各成員の位置）
各成員（単語）の位置は、後述するように、各成員に関する種々の情報（例えば、各成員の属性情報）を用いて求める。上記各成員に関する種々の情報に基づいて、ベクトルの次元を決定する。そして、各成員に関する種々の情報に基づいて、上記決定された次元を持つベクトルの要素の値を求めてベクトル（位置ベクトル）を生成する。生成したベクトルは、各成員の位置を示している。各成員（単語）に関する種々の情報としては、例えば

、以下に示すものがある。
・単語に含まれる文字の種類（例えば、ひらがな、カタカナ、漢字、それ以外が、それぞれあるかないか）
・単語の長さ
・単語の語義
・単語の共起語
・単語の共起データ
（１）単語に含まれる文字の種類（例えば、ひらがな、カタカナ、漢字、それ以外が、それぞれあるかないか）
例えば、ある成員の文字の種類を表すために、ひらがなのみからなる単語か否か、カタカナのみからなる単語か否か、漢字のみからなる単語か否か、ひらがなのみ、又は、カタカナのみ、又は、漢字のみからなる単語以外の単語か否かという、４次元のベクトルの要素を用意（例えば、設定）し、当該ベクトルの要素に設定される値（例えば、１又は０）によって決まるベクトルを作成する。成員がひらがなのみからなる単語に該当すれば１の値を、該当しなければ０の値を対応するベクトルの要素に設定する。また、成員がカタカナのみからなる単語に該当すれば１の値を、該当しなければ０の値を対応するベクトルの要素に設定する。また、成員が漢字のみからなる単語に該当すれば１の値を、該当しなければ０の値を対応するベクトルの要素に設定する。また、成員がひらがなのみ、又は、カタカナのみ、又は、漢字のみからなる単語以外の単語に該当すれば１の値を、該当しなければ０の値を対応するベクトルの要素に設定する。
（２）単語の長さ
例えば、１次元のベクトルの要素を用意し、成員（単語）の文字の個数をカウントし、カウントされた成員（単語）の文字の個数を当該ベクトルの要素に設定して、単語の長さの情報を表現してもよい。
（３）単語の語義
例えば、予め記憶手段内に記憶された、以下のような分類語彙表を用意する。

あ，あ,4.310,1,10,*,
あ，亜,1.104,2,40,,
あ，亜,3.100,10,40,,
ああ，ああ,3.100,3,40,*,
ああ，ああ,4.310,1,20,*,
ああくとう，アーク燈,1.460,2,70,,
ああす，アース,1.462,6,10,,
ああち，アーチ,1.442,2,20,,
ああむほおる，アームホール,1.184,5,30,,
あある，アール,1.1961,4,10,,
あい，愛,1.3020,9,10,*,
あい，相,3.112,1,10,*,
あい，藍,1.502,6,40,,
あいいく，愛育,1.3642,1,40,,
あいいん，愛飲,1.3332,3,60,,
あいいん，合印,1.3114,1,30,Y,
あいうち，あい打ち,1.357,4,30,,
あいかぎ，合鍵,1.454,8,50,,
あいかわらず，相変らず,3.165,2,10,*,
あいかん，哀歓,1.3011,4,60,,
あいがん，哀願,1.366,1,100,,
あいがん，愛翫,1.3852,2,10,,
あいぎ，合着,1.421,4,40,,
あいきょう，愛郷,1.3020,11,170,,

あいきょう，愛嬌,1.3030,4,40,,

上記の"," で区切ってある情報は、それぞれ、単語の読み、単語の見出し語、単語の分類番号、単語の分類番号の下位番号１、単語の分類番号の下位番号２、標本使用頻度が７以上の単語かどうかを示す情報である。

【００５６】

上記の分類語彙表中の"," で区切ってある情報を各桁とし、例えば、上位３桁を意味分類と仮定して、その上位３桁の種類の数だけベクトルの次元を用意する。そして、各成員について、当該意味分類と合致したベクトルの要素の値を１に、それ以外を０に設定することによって、各成員の位置を示すベクトルを作成する。上述した分類語彙表中の情報を利用して各成員のベクトルを作成する方法では、上位３桁を意味分類としたが、他の桁を意味分類にしてもよいし、当該上位３桁と上記他の桁とをあわせた複数の桁の種類の数だけベクトルの次元を用意し、各成員について、当該意味分類と合致したベクトルの要素の値を１に、それ以外を０に設定することによって、各成員の位置を示すベクトルを作成するようにしてもよい。

【００５７】

各成員（単語）に関する種々の情報として、単語に含まれる文字の種類、単語の長さ、単語の語義を説明したが、本発明の一実施例によれば、各成員に関する他の情報も追加で利用して、各成員の位置を示すベクトルを生成してもよい。また、逆に、上述した情報（単語に含まれる文字の種類、単語の長さ、単語の語彙）の全てを用いるのではなく、それらの情報の一部を用いてベクトルを作成するようにしてもよい。例えば、単語の長さの情報を用いずに、ベクトルを作成するようにしてもよい。また、例えば、単語に含まれる文字の種類と単語の語義のみを用いて、ベクトルを作成するようにしてもよい。

（４）単語の共起語

成員（単語）の共起語を求めて、単語の種類の情報に基づいて決まるベクトルの次元を用意する。そして、当該単語の共起語に合致したベクトルの要素に１を、当該単語の共起語に合致しないベクトルの要素に０を設定する。単語の共起語としては、例えば、図１に示す書誌データＤＢ１７中に格納されている書誌データのうち、同じ書誌データ中に成員と共起して出現する単語を当該成員の共起語とする。本発明の一実施形態によれば、例えば、書誌データＤＢ１７中に格納されている書誌データとは異なるデータを用意し、当該データにおいて、成員が出現する領域と同一の領域（例えば、同一文、又は，同一段落、又は、同一データレコード等）に出現する単語を、当該成員の共起語とするようにしてもよい。

【００５８】

また、上述した単語の共起語を用いてベクトルを作成する方法では、共起語として出現していれば、ベクトルの要素に１を、共起語として出現していない場合に、ベクトルの要素に０を設定するが、ベクトルの要素には、共起語として出現した回数の値を設定するようにしてもよい。

（５）単語の共起データ

例えば、書誌データ１７ＤＢ中に格納された書誌データの行数だけのベクトルの次元を用意し、成員（単語）がその書誌データのある行に出現した場合、当該行に対応するベクトルの要素に１を、出現しない場合に、当該行に対応するベクトルの要素に０を入れる。

【００５９】

単語の語義、単語の共起語、単語の共起データについては，意味分類、共起語、共起データの数が多い場合もあるので、本発明の一実施形態によれば、既存のＬＳＩ（Latent semantic index ）などの次元圧縮の技術を使って、ベクトルの次元を減らすようにしてもよい。

【００６０】

次にトップダウンのクラスタリング（非階層クラスタリング）の方法を説明する。

（最大距離アルゴリズムによるクラスタリング）

ある成員と、当該成員と距離が最も離れた成員を求め、これらの成員をそれぞれのクラ

スターの中心とする。次に、それぞれのクラスターの中心と各成員との距離の最小値を各成員の距離とし、その距離が最も大きい成員を新たなクラスターの中心とする。当該クラスターの中心を求める処理を繰り返す。例えば、予め定めた数のクラスターになったときに、当該クラスターの中心を求める処理の繰り返しをやめる。また、例えば、クラスター間の距離が予め定めた数以下になったときに、当該クラスターの中心を求める処理の繰り返しをやめる。

【００６１】

また、クラスターの良さを例えばＡＩＣ情報量基準などで評価して、評価によって求まった値と予め定めた閾値との比較結果に基づいて、当該クラスターの中心を求める処理の繰り返しをやめるようにしてもよい。上記の最大距離アルゴリズムによるクラスタリングによれば、各成員は、各成員と最も近いクラスター中心を持つクラスターの成員となる。

（ｋ平均法）

例えば、以下に示すｋ平均法によって、予め定めた個数（ｋ個）にクラスタリングする。まず、ｋ個の成員をランダムに選択し、選択されたｋ個の成員をクラスターの中心とする。そして、各成員を、当該各成員に最も近いクラスター中心を持つクラスターの成員とする。

【００６２】

次に、クラスター内の各成員の平均の位置に最も近い成員を、それぞれのクラスターの中心とする。そして、各成員を、当該各成員に最も近いクラスター中心を持つクラスターの成員とする。また、クラスター内の各成員の平均の位置に最も近い成員をそれぞれのクラスターの中心とする。上記のクラスターの中心を求める処理を繰り返し、クラスターの中心が移動しなくなったときに、クラスターの中心を求める処理の繰り返しをやめる。本発明の一実施形態によれば、予め定めた回数だけクラスターの中心を求める処理を繰り返してやめるようにしてもよい。そして、最終的なクラスター中心を持つクラスターを決定する。そして、各成員を、当該各成員が最も近いクラスター中心を持つクラスターの成員とする。上記の手法によって、成員のクラスタリングをする。本発明において用いるクラスタリングの方法は、上述した方法に限定されるものではない。本発明に係るデータ表示装置１は、上述したクラスタリングの方法以外の様々な方法を用いて、キーワードのクラスタリングをするようにしてもよい。例えば、本発明の一実施形態によれば、例えば、距離の近い成員を直線上に順に結合する。すなわち、距離が最も近い成員同士をまず結合し、結合した成員によって構成される、複数の成員群を作成する。ある成員群における成員のいずれかと距離が最も近い他の成員群の成員を、当該距離が最も近い成員と結合する処理を繰り返して、成員を直線上に結合する。そして、直線上に結合された成員からなるリストを、予め定めた数の成員毎に区切って、各成員のクラスタリングをするようにしてもよい。

【００６３】

また、本発明の一実施形態によれば、予めデータ表示装置１内の記憶手段（図示を省略）内に、単語と単語が属する分類（クラスター）との対応情報を予め記憶させておき、クラスタリング部１３が、当該記憶手段内の、単語と単語が属する分類（クラスター）との対応情報に基づいて、各単語をクラスタリングするようにしてもよい。

【００６４】

次に、データソート部１４による、抽出した文書の文書データのソート処理について、具体的に説明する。上述したように、データソート部１４は、例えば、頻度算出部１２によって算出された頻度がより高いキーワードを含む文書の順に、各文書の文書データをソートする。

【００６５】

例えば、文書Ａが、頻度が最も高いキーワード「本塁打記録」と、頻度が２番目に高いキーワード「本塁打」と、頻度が３番目に高いキーワード「リーグ」とを含んでいるものとし、また、例えば、文書Ｂが、頻度が最も高いキーワード「本塁打記録」と頻度が３番目に高いキーワード「リーグ」と、頻度が４番目に高いキーワード「マグワイア」とを含

んでいるものとする。文書Ａは、文書Ｂに含まれない、頻度が２番目に高いキーワードを含んでいる。この場合、文書Ａは、文書Ｂに比べて、頻度がより高いキーワードを含んでいる。

【００６６】

頻度がより高いキーワードを含んでいるということを、さらに具体的に説明する。例えば、各キーワードを頻度について降順に並べ、文書があるキーワードを含む場合に、そのキーワードにビット論理「１」を割り当て、文書があるキーワードを含まない場合に、そのキーワードにビット論理「０」を割り当てる。そして、各キーワードに割り当てられたビット論理によって構成される２進数を求める。

【００６７】

例えば、「本塁打記録」、「本塁打」、「リーグ」、「マグワイア」、・・・の順にキーワードが並ぶとすると、上記の文書Ａについて求められる２進数は、「１１１０・・・」であり、文書Ｂについて求められる２進数「１０１１・・・」より大きな数となる。

【００６８】

ある文書が、頻度がより高いキーワードを含んでいるということは、上記のように、例えば、頻度について降順に並んだ各キーワードを２進数の各桁とし、文書に含まれるキーワードにビット論理「１」を、文書に含まれないキーワードにビット論理「０」を割り当てた場合に構成される２進数が、より大きい数であることを意味している。

【００６９】

以下に、キーワード抽出部１１によるキーワードの抽出方法について説明する。

（１）形態素解析を用いた単語の認識による手法

まず、キーワード抽出部１１は、書誌データＤＢ１７に蓄積されている文書について、形態素解析を行い、単語の認識を行う。そして、特定の名詞の単語をキーワードとして取り出す。例えば、名詞だけをキーワードとして取り出す。但し、「こと」、「もの」などの一般的な名詞は、予め収集しておき、それらの名詞がキーワードとしては取り出されないようにしておく。キーワードとしては、名詞だけでなく、動詞などの他の品詞も取り出すこととしてもよい。

【００７０】

形態素解析には、例えば、奈良先端大で開発されている形態素解析システムである ChaSen （下記の文献（２）参照）を用いる。

【００７１】

文献（２）：形態素解析システム茶筌（http://chasen.aist-nara.ac.jp/index.html.ja）

ChaSen 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。

【００７２】

例えば、「学校へ行く」を入力すると、以下の結果を得る。

【００７３】

```
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。

【００７４】

また、英語の品詞タグつけシステムとしては、 Brill（下記の文献（３）参照）のものが有名である。このシステムを用いれば、英語文の各単語の品詞を推定することができる。

【００７５】

文献（３）： Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging,Computational Ling

uistics, Vol. 21, No. 4, p.543-565, 1995.
（２）ＴＦ／ＩＤＦ法などを利用した方法
書誌データＤＢ１７に蓄積されている文書について、形態素解析を行い、例えば、名詞だけを取り出す。そして、取り出された各名詞について、前述したＴＦ／ＩＤＦ法に基づいて算出される所定のスコアを求め、求まったスコアが所定の値よりも大きいものか、スコアが所定の値よりも大きいものから順に所定の値の個数だけ取り出したものをキーワードとする。なお、上記のスコアは、前述したＯｋａｐｉのウェイティング法を用いて算出されるスコアを用いてもよい。
（３）高精度な既存のキーワード抽出のツールを利用する方法
一般に文書中では複数の単語の組み合わせで複雑な概念を表す場合が多く、文書の内容が専門的な事項に特化すれば、その傾向はさらに顕著なものとなる。そこで、例えば、（ａ）形態素解析プログラムによる単語分割、（ｂ）複合語の作成、（ｃ）文書中における重要度の計算、という３つのステップを踏むことで、複合語により複雑な概念を表すことが多い専門用語をキーワードとして文書中から抽出することができる。
【００７６】
例えば、下記の文献（４）に記載されている手法は、文書から取り出した単名詞について、単名詞の左右に連接する単語の種類数あるいは頻度を用いたスコアを算出し、これら左右のスコアを組み合わせて、単名詞のスコアを算出する。単名詞のスコアに基づいて、単名詞から生成される複合名詞のスコアを算出する。そして、算出された複合名詞のスコアが所定の値より大きいものを、キーワードとして取り出す。本発明においても、文献（４）に記載された手法を用いて、キーワードを抽出する構成を採ることができる。
【００７７】
文献（４）：中川裕志、森辰則、湯本紘彰：" 出現頻度と連接頻度に基づく専門用語抽出",自然言語処理、Vol.10 No.1, pp. 27 - 45, ２００３年１月
なお、本発明の実施の形態において、キーワード抽出部１１によるキーワードの抽出方法は、上述した３つの方法に限定されるものではない。キーワード抽出部１１は、他の任意のキーワードの抽出方法を用いてキーワードを抽出することができる。
【００７８】
本発明の一実施形態によれば、キーワード抽出部１１が、以下に述べる固有表現抽出技術を用いてキーワードを抽出してもよい。固有表現とは、人名、地名、組織名などの固有名詞、金額などの数値表現といった、特定の事物・数量を意味する言語表現のことで、固有表現抽出とは、そういった固有表現を文章中から計算機で自動で抽出する技術である。例えば、「日本の首相は小泉純一郎である」という文に対して固有表現抽出を行なうと、固有表現の「日本」と「小泉純一郎」が地名、人名として、抽出される。キーワード抽出部１１は、抽出された固有表現をキーワードとして出力する。
【００７９】
そして、本発明の一実施形態によれば、例えば、予めデータ表示装置１内の記憶手段（図示を省略）内に、キーワード（地名、人名等の固有表現）と当該キーワードが属する分類（クラスター）との対応情報を予め記憶させておき、クラスタリング部１３が、キーワード抽出部１１によって抽出されたキーワード（固有表現）と、当該記憶手段内の、キーワードと当該キーワードが属する分類（クラスター）との対応情報に基づいて、各キーワードをクラスタリングするようにしてもよい。
【００８０】
以下に、固有表現抽出の一般的な手法の例について説明する。
（１）機械学習を用いる手法
機械学習を用いて固有表現を抽出する手法がある（例えば、以下の参考文献（１）参照）。
【００８１】
参考文献（１）：浅原正幸, 松本裕治, 日本語固有表現抽出における冗長的な形態素解析の利用情報処理学会自然言語処理研究会 NL153-7 2002

まず、例えば、「日本の首相は小泉さんです。」という文を、各文字に分割し、分割した文字について、以下のように、 B－LOCATION、 I－LOCATION等の正解タグを付与することによって、正解を設定する。以下の一列目は、分割された各文字であり、各文字の正解タグは二列目である。

日 B－LOCATION
本 I－LOCATION
の 0
首 0
相 0
は 0
小 B－PERSON
泉 I－PERSON
さ 0
ん 0
で 0
す 0
。 0

上記において、B －？？？は、ハイフン以下の固有表現の種類の始まりを意味するタグである。例えば、 B－LOCATIONは、地名という固有表現の始まりを意味しており、 B－PERSONは、人名という固有表現の始まりを意味している。また、I －？？？は、ハイフン以下の固有表現の種類の始まり以外を意味するタグであり、0 はこれら以外である。従って、例えば、文字「日」は、地名という固有表現の始まりに該当する文字であり、文字「本」までが地名という固有表現である。

【００８２】

このように、各文字の正解を設定しておき、このようなデータから学習し、新しいデータでこの正解を推定し、この正解のタグから、各固有表現の始まりと、どこまでがその固有表現かを認識して、固有表現を推定する。

【００８３】

この各文字に設定された正解のデータから学習するときには、システムによってさまざまな情報を素性という形で利用する。例えば、

日 B－LOCATION

の部分は、

日本－B 名詞－B

などの情報を用いる。日本－B は、日本という単語の先頭を意味し、名詞－B は、名詞の先頭を意味する。単語や品詞の認定には、例えば前述したChaSenによる形態素解析を用いる。ChaSenを用いれば、入力された日本語を単語に分割することができる。例えば、ChaSenは、前述したように、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。

【００８４】

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。

【００８５】

なお、例えば、上記の参考文献（１）では、素性として、入力文を構成する文字の、文字自体（例えば、「小」という文字）、字種（例えば、ひらがなやカタカナ等）、品詞情報、タグ情報（例えば、「 B－PERSON」等）を利用している。

【００８６】

これら素性を利用して学習する。タグを推定する文字やその周辺の文字にどういう素性が出現するかを調べ、どういう素性が出現しているときにどういうタグになりやすいかを学習し、その学習結果を利用して新しいデータでのタグの推定を行なう。機械学習には、例えばサポートベクトルマシンを用いる。

【００８７】

固有表現抽出には、上記の手法の他にも種々の手法がある。例えば、最大エントロピーモデルと書き換え規則を用いて固有表現を抽出する手法がある（参考文献（２）参照）。

【００８８】

参考文献（２）：内元清貴，馬青，村田真樹，小作浩美，内山将夫，井佐原均，最大エントロピーモデルと書き換え規則に基づく固有表現抽出，言語処理学会誌，Vol.7，No.2，2000

また、例えば、以下の参考文献（３）に、サポートベクトルマシンを用いて日本語固有表現抽出を行う手法について記載されている。

【００８９】

参考文献（３）：山田寛康，工藤拓，松本裕治，Support Vector Machineを用いた日本語固有表現抽出，情報処理学会論文誌，Vol.43，No.1"，2002

（２）作成したルールを用いる手法

人手でルールを作って固有表現を取り出すという方法もある。

【００９０】

例えば、

名詞＋「さん」だと人名とする
名詞＋「首相」だと人名とする
名詞＋「町」だと地名とする
名詞＋「市」だと地名とする

などである。

【００９１】

本発明の一実施形態によれば、データ表示装置１は、キーワード抽出部１１を用いない構成を採ることもできる。例えば、文書と文書に含まれるキーワードとが対応付けられたデータを所定のデータベース（図１では図示を省略）内に蓄積しておき、上記データベース内に蓄積されているデータから、頻度算出部１２が、各キーワードの、上記データベース中の文書群中に出現した頻度を算出する構成を採ることもできる。

【００９２】

本発明の一実施形態によれば、キーワード抽出部１１が、書誌データＤＢ１７に蓄積されている文書群から、単位表現を抽出し、抽出された単位表現が上記文書群において出現する箇所を特定し、特定された箇所において、当該単位表現と隣接して出現している数値を抽出する。そして、キーワード抽出部１１は、抽出された数値と当該単位表現とから構成される数値情報をキーワードとして抽出する。例えば、４０号等といった数値情報がキーワードとして抽出される。そして、クラスタリング部１３が、抽出された数値情報をクラスタリングするようにしてもよい。

【００９３】

本発明の一実施形態によれば、データ表示装置１は、頻度算出部１２、データソート部１４、クラスターソート指定部１６が省略された構成を採ってもよい。例えば、キーワード抽出部１１が、書誌データＤＢ１７中の文書群中に含まれるキーワードを抽出し、クラスタリング部１３が、抽出されたキーワードを各キーワードが属するクラスターにクラスタリングし、表示部１５が、当該クラスターに属するキーワードのうちの上記文書群を構成する文書に含まれるキーワード、又は、各文書中における当該クラスターに属するキーワードの有無を示す情報（例えば、符号）を、当該各文書と当該クラスターとに対応付けて画面表示するようにしてもよい。また、データ表示装置１が上記の構成を採る場合において、クラスタリング部１３が、各キーワードのベクトル空間上の位置を示す位置ベクトルを生成し、生成された当該位置ベクトルが示す各キーワードの位置情報に基づいて、各

キーワードが属するクラスターを決定するようにしてもよい。

【００９４】

また、データ表示装置１が上記の構成を採る場合において、表示部１５は、予め定められた順序に基づいて、各クラスターを並び替えて画面表示するようにしてもよい。また、表示部１５が、各クラスターに属するキーワード又は各クラスターに属するキーワードが出現する文書の分布に基づいて、各クラスターを画面表示するようにしてもよい。ここで、各クラスターに属するキーワードの分布とは、例えば、各クラスターに属するキーワードが出現する文書数の分布、各クラスターに属するキーワード数の分布、又は、各クラスターに属するキーワードが出現する文書数を各クラスターに属する全てのキーワードについて合計した数の分布である。また、表示部１５は、各クラスターに属するキーワードが出現する文書数について昇順又は降順に各クラスターを並び替えて画面表示するようにしてもよい。

【００９５】

また、データ表示装置１において、クラスタリング部１３が、各キーワードの位置ベクトルの平均を各キーワードが属するクラスターの位置ベクトルとし、各クラスターの位置ベクトルに基づいて、クラスター同士の距離を求め、求めた距離が近いクラスター同士が近い位置に並び、求めた距離が遠いクラスター同士が離れた位置に並ぶように、各クラスターを並び替え、表示部１５が、並び替えられた各クラスターに属するキーワードのうちの書誌データＤＢ１７中の文書群を構成する文書に含まれるキーワード、又は、各文書中における上記並び替えられた各クラスターに属するキーワードの有無を示す情報を、各文書と当該並び替えられた各クラスターとに対応付けて画面表示するようにしてもよい。

【００９６】

本発明の一実施形態によれば、クラスタリング部１３は、例えば、以下に述べる手法によって、各クラスターを並び替えるようにしてもよい。すなわち、クラスタリング部１３は、距離が最も近いクラスター同士を結合し、結合したクラスターのリストの端のクラスターのいずれかと距離が最も近いクラスターを、当該リストの端のクラスターと結合する。そして、クラスタリング部１３は、さらに、当該クラスターの結合によって得られるリストの端のクラスターのいずれかと距離が最も近いクラスターを結合する。クラスタリング部１３は、上記の処理を繰り返すことによって得られるリストにおけるクラスターの並び順に、各クラスターを並び替える。

【００９７】

また、クラスタリング部１３が、ベクトル空間上に並んで配置された複数の基準の位置ベクトルを予め定義し、各クラスターの位置ベクトルを、各クラスターの位置ベクトルとの距離が最も近い基準の位置ベクトルの近くに配置することにより、各クラスターを並び替えるようにしてもよい。すなわち、クラスタリング部１３は、例えば基準ベクトルＡ、Ｂ、Ｃの順に並んで配置された３つの基準ベクトルを定義し、あるクラスターの位置ベクトルとの距離が最も近い基準ベクトルが基準ベクトルＡである場合、当該クラスターの位置ベクトルを基準ベクトルＡの近くに配置する。同様にして、他のクラスターの位置ベクトルを、距離が最も近い基準ベクトルの近くに配置する。そして、クラスタリング部１３は、各クラスターの位置ベクトルが上記配置後の位置ベクトルとなるように各クラスターを並び替える。

【００９８】

また、データ表示装置１において、クラスタリング部１３は、各文書がどのクラスターのどの単語を何個含んでいるかの情報を求め、求めた情報に基づいて、各文書のベクトル空間上の位置を示す位置ベクトルを求め、求めた各文書の位置ベクトルに基づいて、文書同士の距離を求め、求めた距離が近い文書同士が近い位置に並び、求めた距離が遠い文書同士が離れた位置に並ぶように、各文書を並び替えるようにしてもよい。例えば、クラスタリング部１３は、各文書がどのクラスターのどの単語を何個含んでいるかの情報に基づいて、当該各文書の位置ベクトルの次元数と、当該位置ベクトルのベクトル要素の値を求めることによって、各文書の位置ベクトルを求める。

【００９９】
本発明の一実施形態によれば、クラスタリング部１３は、例えば、以下に述べる手法によって、各文書を並び替えるようにしてもよい。すなわち、クラスタリング部１３は、距離が最も近い文書同士を結合し、結合した文書のリストの端の文書のいずれかと距離が最も近い文書を、当該リストの端の文書と結合する。そして、クラスタリング部１３は、さらに、当該文書の結合によって得られるリストの端の文書のいずれかと距離が最も近い文書を結合する。クラスタリング部１３は、上記の処理を繰り返すことによって得られるリストにおける文書の並び順に、各文書を並び替える。
【０１００】
また、表示部１５は、並ぶ順序がより上位のクラスターに属するキーワードを含む文書の順に、当該文書に含まれるキーワード、又は、当該文書中における上記クラスターに属するキーワードの有無を示す情報（例えば、符号）を画面表示するようにしてもよい。例えば、クラスターＡ、クラスターＢ、クラスターＣの順にクラスターが並ぶ場合、文書Ａがクラスターに属するキーワードとクラスターＢに属するキーワードとクラスターＣに属するキーワードを含み、文書Ｂがクラスターに属するキーワードとクラスターＢに属するキーワードとを含むがクラスターＣに属するキーワードを含まないとき、表示部１５は、文書Ａに含まれるキーワード、文書Ｂに含まれるキーワードの順に画面表示する。
【０１０１】
また、表示部１５は、クラスタリング部１３によってクラスタリングされたクラスターを、ユーザの指定入力に従って選択し、選択されたクラスターに対応付けて、当該クラスターに属するキーワードのうちの、文書群を構成する文書に含まれるキーワード、又は、各文書中における当該クラスターに属するキーワードの有無を示す情報を画面表示するようにしてもよい。すなわち、表示部１５は、例えば、選択されないクラスターについては、当該クラスターに属するキーワードのうちの、文書群を構成する文書に含まれるキーワード、又は、各文書中における当該クラスターに属するキーワードの有無を示す情報を画面表示しないようにしてもよい。
【０１０２】
また、データ表示装置１において、クラスタリング部１３は、各キーワードのベクトル空間上の位置を示す位置ベクトルを生成し、生成された当該位置ベクトルが示す各キーワードの位置情報に基づいて、キーワード同士の距離を求め、求めた距離が近いキーワード同士が近い位置に並び、求めた距離が遠いキーワード同士が離れた位置に並ぶように、各キーワードを並び替え、表示部１５は、並び替えられたキーワードのうちの当該文書群を構成する文書に含まれるキーワード、又は、各文書中における当該並び替えられたキーワードの有無を示す情報を、各文書に対応付けて画面表示するようにしてもよい。
【０１０３】
本発明の一実施形態によれば、クラスタリング部１３は、例えば、以下に述べる手法によって、各キーワードを並び替えるようにしてもよい。すなわち、クラスタリング部１３は、距離が最も近いキーワード同士を結合し、結合したキーワードのリストの端のキーワードのいずれかと距離が最も近いキーワードを、当該リストの端のキーワードと結合する。そして、クラスタリング部１３は、さらに、当該キーワードの結合によって得られるリストの端のキーワードのいずれかと距離が最も近いキーワードを結合する。クラスタリング部１３は、上記の処理を繰り返すことによって得られるリストにおけるキーワードの並び順に、各キーワードを並び替える。
【０１０４】
また、表示部１５は、各クラスターを画面上に並ばせる順序を、ユーザの指定入力に従って選択し、選択された順序がより上位のクラスターに属するキーワードを含む文書の順に、文書に含まれるキーワード、又は、文書中における当該クラスターに属するキーワードの有無を示す情報を画面表示するようにしてもよい。
【０１０５】
図２は、データ表示処理フローの一例を示す図である。まず、キーワード抽出部１１が

、書誌データＤＢ１７に蓄積されている文書群に含まれるキーワードを抽出する（ステップＳ１）。例えば、キーワード「本塁打記録」、「本塁打」、「リーグ」、「マグワイア」、「カージナルス」、「米大リーグ」、「大リーグ」、「試合」、「ロジャー」、「単独トップ」、・・・といったキーワードを抽出する。

【０１０６】

次に、頻度算出部１２が、キーワード抽出部１１によって抽出された各キーワードの、書誌データＤＢ１７に蓄積されている文書群中に出現した頻度を算出する（ステップＳ２）。

【０１０７】

例えば、図３の表に示すように、算出されるキーワード「本塁打記録」の頻度は２０、キーワード「本塁打」の頻度は１５、キーワード「リーグ」の頻度は１２、キーワード「マグワイア」の頻度は１０、キーワード「カージナルス」の頻度は９、キーワード「米大リーグ」の頻度は８、キーワード「大リーグ」の頻度は７、キーワード「試合」の頻度は６、キーワード「ロジャー」の頻度は５、キーワード「単独トップ」の頻度は４である。なお、図３中に示す頻度は、各キーワードが出現する文書の数である。また、図３中では、頻度が４であるキーワードまでしか示していないが、本発明の実施の形態では、ステップＳ２において、例えば、頻度３、頻度２、頻度１についても算出され得る。

【０１０８】

次に、データソート部１４が、書誌データＤＢ１７に蓄積されている各文書の文書データを抽出し、バッファ中に格納する（ステップＳ３）。例えば、文書データとして、文書のテキストデータが抽出され、バッファ中に格納される。

【０１０９】

また、データソート部１４が、頻度算出部１２によって算出された各キーワードの頻度がより高いキーワードを多く含む文書の順に、バッファ中に格納された文書の文書データをソートする（ステップＳ４）。ステップＳ４においては、例えば、データソート部１４は、各キーワードをステップＳ２において算出された頻度について降順又は昇順にソートし、ソートした各キーワードの情報をバッファ中に保持するようにしてもよい。ステップＳ４の処理によって、例えば、各文書と各文書に含まれるキーワードとの対応情報がバッファ中に保持される。

【０１１０】

図４は、データソート部１４によってバッファ中に保持される、各文書と各文書に含まれるキーワードとの対応情報の例を示す図である。図４中において、矩形の枠で囲ったキーワードは、データソート部１４によって、例えば上記頻度について降順にソートされたキーワードである。

【０１１１】

例えば、論文名が「Ａ」である文書は、図３中に示される頻度が最も高いキーワード「本塁打記録」と、頻度が２番目に高いキーワード「本塁打」と、頻度が３番目に高いキーワード「リーグ」と、頻度が４番目に高いキーワード「マグワイア」と、頻度が５番目に高いキーワード「カージナルス」と、頻度が６番目に高いキーワード「米大リーグ」とを含んでいるものとする。

【０１１２】

また、例えば、論文名が「Ｂ」である文書は、頻度が最も高いキーワード「本塁打記録」と、頻度が２番目に高いキーワード「本塁打」と、頻度が３番目に高いキーワード「リーグ」と、頻度が４番目に高いキーワード「マグワイア」と、頻度が５番目に高いキーワード「カージナルス」とを含んでいるが、頻度が６番目に高いキーワード「米大リーグ」を含んでいないものとする。

【０１１３】

また、例えば、論文名が「Ｃ」である文書は、頻度が最も高いキーワード「本塁打記録」と、頻度が２番目に高いキーワード「本塁打」と、頻度が３番目に高いキーワード「リーグ」と、頻度が４番目に高いキーワード「マグワイア」とを含んでいるが、頻度が５番

目に高いキーワード「カージナルス」を含んでいないものとする。

【０１１４】

上記ステップＳ４において、データソート部１４は、例えば、論文名が「Ａ」という文書のテキストデータ、論文名が「Ｂ」という文書のテキストデータ、論文名が「Ｃ」という文書のテキストデータの順に並ぶようにソートする。ステップＳ４における処理の結果、例えば図４中に示すような順番で、各文書のテキストデータがソートされる。なお、図４中の、各論文名が記述された行と同一の行に記述されたキーワードは、各論文名を持つ文書が含むキーワードを示している。また、図４中に示す日付は、各文書の発行日を示す情報である。

【０１１５】

次に、クラスタリング部１３が、キーワード抽出部１１によって抽出された各キーワードをクラスター毎にクラスタリングする（ステップＳ５）。例えば、クラスタリング部１３は、図４中の矩形の枠で囲った各キーワードに含まれるキーワードのうち、キーワード「本塁打記録」、「本塁打」、「年間最多本塁打記録」、「年間ホームラン数」、「年間ホームラン」、「号本塁打」、「ホームラン」、「４６本」、「５０本」、「５４本」、「４７号」、「５０号」、「５１号」、「５２号」、「５４号」（本塁打記録、本塁打以外のキーワードは、図示を省略）を、クラスターＡとしてクラスタリングする。また、クラスタリング部１３は、例えば、キーワード「リーグ」、「米大リーグ」、「大リーグ」を、クラスターＢとしてクラスタリングする。また、クラスタリング部１３は、例えば、キーワード「マグワイア」、「マグワイア一塁手」、「マーク」、「ロジャー」、「ソーサ」、「ソーサ外野手」、「サミー」（マグワイア一塁手、マーク、ソーサ、ソーサ外野手、サミーについては、図示を省略）を、クラスターＣとしてクラスタリングする。また、クラスタリング部１３は、例えば、キーワード「カージナルス」、「ヤンキース」（ヤンキースについては、図示を省略）を、クラスターＤとしてクラスタリングする。また、クラスタリング部１３は、例えば、キーワード「試合」、「チーム」、「スタジアム」（チーム、スタジアムについては、図示を省略）を、クラスターＥとしてクラスタリングする。また、クラスタリング部１３は、例えば、キーワード「単独トップ」、「トップ」、「９位タイ」（トップ、９位タイについては、図示を省略）を、クラスターＦとしてクラスタリングする。

【０１１６】

そして、表示部１５が、データソート部１４によってソートされた文書データに対応付けて、上記クラスター中のキーワードのうち当該ソートされた文書データを持つ文書が含むキーワードを、クラスター毎に画面表示する（ステップＳ６）。

【０１１７】

例えば、図５に示す表示画面に示すように、表示部１５は、各文書（図５では論文名が「Ａ」である文書～論文名が「Ｅ」である文書）のテキストデータと、太線の矩形の枠で囲ったクラスター（図５では、クラスターＡ～クラスターＦ）とを画面表示する。また、図５に示す表示画面が、例えばセルで構成されている場合を想定すると、表示部１５は、例えば、論文名が「Ａ」である文書のテキストデータ（「米大リーグ、カージナルスのマーク・マグワイア一塁手が・・・」）が表示されるセルに対応する行とクラスターＡが表示されるセルに対応する列とが交差するセルの位置に、当該クラスターＡ中のキーワードのうち、当該文書が含むキーワード（「本塁打記録、本塁打、年間最多本塁打記録、号本塁打、４６本、４７号）を画面表示する。また、表示部１５は、例えば、論文名が「Ａ」である文書のテキストデータが表示されるセルに対応する行とクラスターＢが表示されるセルに対応する列とが交差するセルの位置に、当該クラスターＢ中のキーワードのうち、当該文書が含むキーワード（「リーグ」、「米大リーグ」、「大リーグ」）を画面表示する。また、表示部１５は、例えば、論文名が「Ａ」である文書のテキストデータが表示されるセルに対応する行とクラスターＣが表示されるセルに対応する列とが交差するセルの位置に、当該クラスターＣ中のキーワードのうち、当該文書が含むキーワード（「マグワイア」、「マグワイア一塁手」、「マーク」、「ロジャー」、「ソーサ」、「ソーサ外野

手」、「サミー」）を画面表示する。また、表示部１５は、例えば、論文名が「Ａ」である文書のテキストデータが表示されるセルに対応する行とクラスターＤが表示されるセルに対応する列とが交差するセルの位置に、当該クラスターＤ中のキーワードのうち、当該文書が含むキーワード（「カージナルス」、「ヤンキース」）を画面表示する。また、表示部１５は、例えば、論文名が「Ａ」である文書のテキストデータが表示されるセルに対応する行とクラスターＥが表示されるセルに対応する列とが交差するセルの位置に、当該クラスターＥ中のキーワードのうち、当該文書が含むキーワード（「試合」、「チーム」、「スタジアム」）を画面表示する。また、表示部１５は、例えば、論文名が「Ａ」である文書のテキストデータが表示されるセルに対応する行とクラスターＦが表示されるセルに対応する列とが交差するセルの位置に、当該クラスターＦ中のキーワードのうち、当該文書が含むキーワード（「単独トップ」）を画面表示する。

【０１１８】

同様にして、表示部１５は、論文名が「Ｂ」という文書、論文名が「Ｃ」という文書についてのテキストデータに対応付けて、クラスター中のキーワードのうち各文書が含むキーワードを、クラスター毎に画面表示する。

【０１１９】

本発明の一実施形態によれば、例えば上記ステップＳ６の処理の後に、クラスターソート指定部１６が、各クラスターのソート内容をデータソート部１４に対して指示し、データソート部１４が、指示されたソート内容に従って各クラスターをソートし、表示部１５が、データソート部１４によってソートされた文書データに対応付けて、データソート部１４によってソートされたクラスター中のキーワードのうち当該ソートされた文書データを持つ文書が含むキーワードを、データソート部１４によってソートされた各クラスター毎に画面表示するようにしてもよい。

【０１２０】

図５に示す表示画面を参照すれば、例えば、論文名が「Ａ」という文書は、米大リーグの試合における本塁打記録に関する文書であることを把握することができる。従って、本発明によれば、文書の大凡の内容を一見して把握できるような表示を行うことが可能となる。

【０１２１】

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、又は、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】

【０１２２】

【図１】本発明のシステム構成の一例を示す図である。

【図２】データ表示処理フローの一例を示す図である。

【図３】各キーワードの頻度を示す図である。

【図４】各文書と各文書に含まれるキーワードとの対応情報を示す図である。

【図５】画面表示例を示す図である。

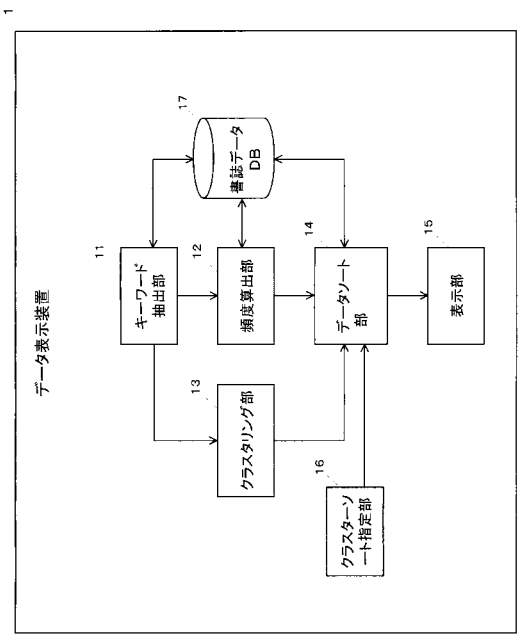
【符号の説明】

【０１２３】

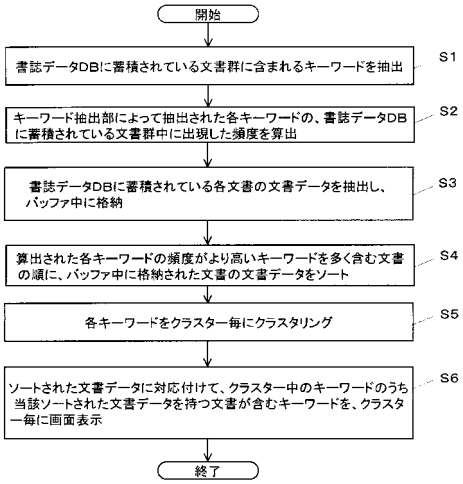
１　データ表示装置
１１　キーワード抽出部
１２　頻度算出部
１３　クラスタリング部
１４　データソート部
１５　表示部
１６　クラスターソート指定部

１７　書誌データＤＢ

【図１】



【図２】



【図３】

	本塁打記録	本塁打	リーグ	マグワイア	カージナルス	米大リーグ	大リーグ	試合	ロジャー	単独トップ	…
頻度	20	15	12	10	9	8	7	6	5	4	…

【図４】

	(論文名)	(テキストデータ)	(日付)	1 本塁打記録	2 本塁打	3 リーグ	4 マグワイア	5 カージナルス	6 米大リーグ	7 大リーグ	8 試合	9 ロジャー	10 単独トップ	…
1	A	……	980813	本塁打記録	本塁打	リーグ	マグワイア	カージナルス	米大リーグ					
2	B	……	980818	本塁打記録	本塁打	リーグ	マグワイア	カージナルス						
3	C	……	⋮	本塁打記録	本塁打	リーグ	マグワイア							
4	D	……		本塁打記録	本塁打	リーグ					試合	ロジャー	単独トップ	
5	E	……		本塁打記録	本塁打	リーグ					試合			
6	F	……		本塁打記録	本塁打			カージナルス						
7	G	……		本塁打記録	本塁打									
8	H			本塁打記録	本塁打									
9	I	⋮		本塁打記録		リーグ	マグワイア	カージナルス	米大リーグ	大リーグ				
10	J			本塁打記録		リーグ			米大リーグ	大リーグ				
11	K			本塁打記録			マグワイア							
12	L			本塁打記録					米大リーグ		試合	ロジャー	単独トップ	
13	M			本塁打記録							試合			
14	N			本塁打記録								ロジャー		
15	O			本塁打記録									単独トップ	
16	P			本塁打記録										
17	Q			本塁打記録										
18	R			本塁打記録										
19	S			本塁打記録										
20	T			本塁打記録										
21	U				本塁打	リーグ	マグワイア	カージナルス	米大リーグ	大リーグ	試合		単独トップ	
22	V				本塁打	リーグ	マグワイア		米大リーグ		試合	ロジャー		
23	W				本塁打	リーグ	マグワイア							
24	X				本塁打			カージナルス	米大リーグ	大リーグ		ロジャー		
25	Y				本塁打	リーグ			米大リーグ					
26	Z				本塁打	リーグ	マグワイア	カージナルス		大リーグ				
27	A−1				本塁打									
28	B−1						マグワイア	カージナルス		大リーグ				
29	C−1							カージナルス		大リーグ				
	⋮													

【図５】

(論文名)	(テキストデータ)	(日付)	クラスターA	クラスターB	クラスターC	クラスターD	クラスターE	クラスターF
			本塁打記録, 本塁打, 年間最多本塁打記録, 年間ホームラン数, 年間ホームラン, 号本塁打, ホームラン, 46本, 50本, 54本, 47号, 50号, 51号, 52号, 54号	リーグ, 米大リーグ, 大リーグ	マグワイア, マグワイアー塁手, マーク, ロジャー, ソーサ, ソーサ外野手, サミー	カージナルス, ヤンキース	試合, チーム, スタジアム	単独トップ, トップ, 9位タイ
A	米大リーグ、カージナルスのマーク・マグワイアー塁手が11日…	980813	本塁打記録, 本塁打, 年間最多本塁打記録, 号本塁打, 46本, 47号	リーグ, 米大リーグ, 大リーグ	マグワイア, マグワイアー塁手, マーク, ロジャー, ソーサ, ソーサ外野手, サミー	カージナルス, ヤンキース	試合, チーム, スタジアム	単独トップ
B	米大リーグ、カブスのサミー・ソーサ外野手が16日…	980818	本塁打記録, 本塁打, 年間最多本塁打記録, 号本塁打, 47号	リーグ, 米大リーグ, 大リーグ	マグワイア, マグワイアー塁手, マーク, ロジャー, ソーサ, ソーサ外野手, サミー	カージナルス, ヤンキース	試合, チーム, スタジアム	トップ
C	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	⋮		⋮	⋮	⋮	⋮	⋮	⋮