(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4719921号
(P4719921)
(45)発行日　平成23年7月6日(2011.7.6)　(24)登録日　平成23年4月15日(2011.4.15)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 380E
G06F 17/30 170A
G06F 17/30 210A

請求項の数 10　(全 21 頁)

(21)出願番号　特願2005-330009(P2005-330009)
(22)出願日　平成17年11月15日(2005.11.15)
(65)公開番号　特開2007-140639(P2007-140639A)
(43)公開日　平成19年6月7日(2007.6.7)
審査請求日　平成20年10月15日(2008.10.15)

(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

審査官　波内　みさ

最終頁に続く

(54)【発明の名称】データ表示装置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】
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，
前記算出された頻度に基づいて，前記各キーワードのスコアを算出するスコア算出手段と，
前記算出された各キーワードのスコアの値の高いものから降順に前記各キーワードの優先度を設定し，各キーワードが文書に存在するとビット１，存在しないとビット０を割り当てて，各文書について前記設定された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順にソートするデータソート手段と，
前記ソートされた文書データを表示データとして画面表示するとともに，前記文書データが画面表示される画面と同一画面上において，前記各キーワードを前記優先度が高い順に表示データとして画面表示する表示手段とを備え，
前記表示手段は，さらに，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを含んでいるかを示す情報を表示データとして画面表示する

ことを特徴とするデータ表示装置。
【請求項２】
データ表示装置であって，
入力された，文書群中の文書と各文書に含まれるキーワードとに基づいて，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，
前記算出された頻度に基づいて，前記各キーワードのスコアを算出するスコア算出手段と，
前記算出された各キーワードのスコアの値の高いものから降順に前記各キーワードの優先度を設定し，各キーワードが文書に存在するとビット１，存在しないとビット０を割り当てて，各文書について前記設定された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順にソートするデータソート手段と，
前記ソートされた文書データを表示データとして画面表示するとともに，前記文書データが画面表示される画面と同一画面上において，前記各キーワードを前記優先度が高い順に表示データとして画面表示する表示手段とを備え，
前記表示手段は，さらに，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを含んでいるかを示す情報を表示データとして画面表示する
ことを特徴とするデータ表示装置。
【請求項３】
請求項１に記載のデータ表示装置において，
キーワードを選択するキーワード選択手段と，
前記キーワード抽出手段によって抽出された各キーワードの，前記選択されたキーワードを含む文書群中に出現する頻度である内部頻度を算出する内部頻度算出手段と，
前記内部頻度算出手段が算出した内部頻度に基づいて，前記各キーワードの内部スコアを算出する内部スコア算出手段とを備え，
前記データソート手段は，前記算出された内部スコアを，前記各キーワードの優先度として設定し，設定された優先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
ことを特徴とするデータ表示装置。
【請求項４】
請求項３に記載のデータ表示装置において，
前記内部頻度算出手段は，前記キーワード選択手段によって複数のキーワードが選択された場合に，前記キーワード抽出手段によって抽出された各キーワードの，前記選択された複数のキーワードの全てを含む文書群中に出現する頻度を，前記内部頻度として算出する
ことを特徴とするデータ表示装置。
【請求項５】
請求項３または請求項４に記載のデータ表示装置において，
前記スコア算出手段は，前記キーワード抽出手段によって抽出された各キーワードの文字数と，前記頻度算出手段によって算出された頻度とに基づいて，前記各キーワードのスコアを算出し，
前記データソート手段は，前記頻度算出手段によって算出された頻度と前記スコア算出手段によって算出されたスコアとに基づいて，前記各キーワードの優先度を設定し，前記内部頻度算出手段によって算出された内部頻度と前記内部スコア算出手段によって算出された内部スコアとに基づいて，前記各キーワードの優先度を更新し，各文書について前記更新された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを前記優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順に，前記文書群を構成する文書データをソートする

ことを特徴とするデータ表示装置。
【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，
前記表示手段は，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを何個含んでいるかを示す情報を表示データとして画面表示する
ことを特徴とするデータ表示装置。
【請求項７】
請求項１乃至請求項６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，前記頻度算出手段によって算出された頻度が予め定められた閾値未満であるキーワードを含む文書については，その文書が前記頻度が前記閾値未満であるキーワードを含むことを示す情報を，前記頻度が予め定められた閾値以上であるキーワードを画面表示する表示領域とは別の表示領域に表示データとして画面表示する
ことを特徴とするデータ表示装置。
【請求項８】
請求項１乃至請求項７のいずれか１項に記載のデータ表示装置において，
前記表示手段によって画面表示される表示データを指定する表示データ指定手段を備え，
前記表示手段は，前記表示データ指定手段によって指定された表示データのみを画面表示する
ことを特徴とするデータ表示装置。
【請求項９】
請求項１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，
前記データソート手段は，前記文書群を構成する文書の文書データを，各文書に関連する日付について降順また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１０】
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，
前記算出された頻度に基づいて，前記各キーワードのスコアを算出するスコア算出手段と，
前記算出された各キーワードのスコアの値の高いものから降順に前記各キーワードの優先度を設定し，各キーワードが文書に存在するとビット１，存在しないとビット０を割り当てて，各文書について前記設定された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順にソートするデータソート手段と，
前記ソートされた文書データを表示データとして画面表示するとともに，前記文書データが画面表示される画面と同一画面上において，前記各キーワードを前記優先度が高い順に表示データとして画面表示する表示手段として機能させるためのプログラムであって，
前記表示手段は，さらに，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを含んでいるかを示す情報を表示データとして画面表示する
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，あるキーワードを含む文書群において，出現

頻度の高い他のキーワード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およ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の出現頻度に基づいて、各キーワードに対して優先度を設定し、より優先度の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して表示することは行っていない。

【０００５】

従って、従来技術では、例えば、あるキーワードを含む文書群において、出現頻度の高い他のキーワードを一見して把握できるような表示を行うことはできない。従来技術では、あるキーワードを含む文書中にどのキーワードが共起して出現するかといった、キーワード間の共起関係を把握することができない。

【０００６】

本発明は，上記従来技術の問題点を解決し，あるキーワードを含む文書群において，出現頻度の高い他のキーワード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およ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】

上記課題を解決するため，本発明は，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，前記算出された頻度に基づいて，前記各キーワードのスコアを算出するスコア算出手段と，前記算出された各キーワードのスコアの値の高いものから降順に前記各キーワードの優先度を設定し，各キーワードが文書に存在するとビット１，存在しないとビット０を割り当てて，各文書について前記設定された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順にソートするデータソート手段と，前記ソートされた文書データを表示データとして画面表示するとともに，前記文書データが画面表示される画面と同一画面上において，前記各キーワードを前記優先度が高い順に表示データとして画面表示する表示手段とを備え，前記表示手段は，さらに，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを含んでいるかを示す情報を表示データとして画面表示することを特徴とする。

【０００８】

また，本発明は，データ表示装置であって，入力された，文書群中の文書と各文書に含まれるキーワードとに基づいて，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，前記算出された頻度に基づいて，前記各キーワードのスコアを算出するスコア算出手段と，前記算出された各キーワードのスコアの値の高いものから降順に前記各キーワードの優先度を設定し，各キーワードが文書に存在するとビット１，存在しないとビット０を割り当てて，各文書について前記設定された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データ

を優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順にソートするデータソート手段と，前記ソートされた文書データを表示データとして画面表示するとともに，前記文書データが画面表示される画面と同一画面上において，前記各キーワードを前記優先度が高い順に表示データとして画面表示する表示手段とを備え，前記表示手段は，さらに，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを含んでいるかを示す情報を表示データとして画面表示することを特徴とする。

【０００９】

また、本発明は、前記のデータ表示装置において、さらに、キーワードを選択するキーワード選択手段と、前記キーワード抽出手段によって抽出された各キーワードの、前記選択されたキーワードを含む文書群中に出現する頻度である内部頻度を算出する内部頻度算出手段と、前記内部頻度算出手段が算出した内部頻度に基づいて、前記各キーワードの内部スコアを算出する内部スコア算出手段とを備え、前記データソート手段は、前記算出された内部スコアを、前記各キーワードの優先度として設定し、設定された優先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、さらに、前記内部頻度算出手段は、前記キーワード選択手段によって複数のキーワードが選択された場合に、前記キーワード抽出手段によって抽出された各キーワードの、前記選択された複数のキーワードの全てを含む文書群中に出現する頻度を、前記内部頻度として算出することを特徴とする。

【００１１】

また，本発明は，前記のデータ表示装置において，前記スコア算出手段は，前記キーワード抽出手段によって抽出された各キーワードの文字数と，前記頻度算出手段によって算出された頻度とに基づいて，前記各キーワードのスコアを算出し，前記データソート手段は，前記頻度算出手段によって算出された頻度と前記スコア算出手段によって算出されたスコアとに基づいて，前記各キーワードの優先度を設定し，前記内部頻度算出手段によって算出された内部頻度と前記内部スコア算出手段によって算出された内部スコアとに基づいて，前記各キーワードの優先度を更新し，各文書について前記更新された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを前記優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順に，前記文書群を構成する文書データをソート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、前記表示手段は、前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを何個含んでいるかを示す情報を表示データとして画面表示することを特徴とする。

【００１３】

また、本発明は、前記のデータ表示装置において、前記表示手段は、前記頻度算出手段によって算出された頻度が予め定められた閾値未満であるキーワードを含む文書については、その文書が前記頻度が前記閾値未満であるキーワードを含むことを示す情報を、前記頻度が予め定められた閾値以上であるキーワードを画面表示する表示領域とは別の表示領域に表示データとして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、さらに、前記表示手段によって画面表示される表示データを指定する表示データ指定手段を備え、前記表示手段は、前記表示データ指定手段によって指定された表示データのみを画面表示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、前記データソート手段は、前記文書群を構成する文書の文書データを、各文書に関連する日付について降順または昇順にソー

トすることを特徴とする。

【００１７】

また，本発明は，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，前記算出された頻度に基づいて，前記各キーワードのスコアを算出するスコア算出手段と，前記算出された各キーワードのスコアの値の高いものから降順に前記各キーワードの優先度を設定し，各キーワードが文書に存在するとビット１，存在しないとビット０を割り当てて，各文書について前記設定された優先度の降順にキーワードの存否を表す複数ビットから成る２進数で表し，前記文書群を構成する各文書の文書データを優先度が高いキーワードを含む文書の順にするため，各文書を表す前記２進数の大きい数値の順にソートするデータソート手段と，前記ソートされた文書データを表示データとして画面表示するとともに，前記文書データが画面表示される画面と同一画面上において，前記各キーワードを前記優先度が高い順に表示データとして画面表示する表示手段として機能させるためのプログラムであって，前記表示手段は，さらに，前記画面表示された各文書データに対応する文書が前記画面表示された各キーワードを含んでいるかを示す情報を表示データとして画面表示す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、文書群中に含まれるキーワードの出現頻度に基づいて、各キーワードに対して優先度を設定し、より優先度の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して画面表示する。

【００１９】

また、本発明は、文書データが画面表示される画面と同一画面上において、各キーワードを、優先度の高い順に画面表示するとともに、画面表示された各文書データに対応する文書が、画面表示された各キーワードを含んでいるかを示す情報を画面表示する。

【００２０】

従って、本発明によれば、あるキーワードを含む文書群において、出現頻度の高い他のキーワードを一見して把握できるような表示を行うことができる。また、本発明によれば、あるキーワードを含む文書中にどのキーワードが共起して出現するかといった、キーワード間の共起関係を容易に把握することが可能となる。

【００２１】

また、本発明によれば、表示された各文書データに含まれるキーワードを見ることによって、各文書の概略の内容を推測することが可能となる。

【００２２】

また、本発明は、例えば、ユーザが、優先度の高い順に画面表示されたキーワードを選択すれば、選択されたキーワードを含む文書群中に出現する各キーワードの頻度に基づいて算出される内部スコアに基づいて、キーワードの優先度を設定し、設定された優先度がより高いキーワードを含む文書の順に、文書群を構成する文書の文書データをソートし直す。従って、本発明によれば、ユーザは、ユーザが思い付いたキーワードを自ら入力する必要がなく、画面表示されたキーワードを選択するだけで、文書データをソートし直すことが可能となる。特に、画面表示されたキーワードは、優先度の高い順に並んでいるため、ユーザは、画面上において、優先度の高いキーワードから順に各キーワードを見ていくことで、ユーザにとって有用なキーワードを容易に見つけて、選択することが可能となる。

【発明を実施するための最良の形態】

【００２３】

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。

【００２４】
データ表示装置１は、キーワード抽出部１１、頻度算出部１２、スコア算出部１３、データソート部１４、表示部１５、キーワード選択部１６、内部頻度算出部１７、内部スコア算出部１８、書誌データデータベース（ＤＢ）１９、表示データ指定部２０を備える。
【００２５】
キーワード抽出部１１は、書誌データＤＢ１９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。
【００２６】
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ＤＢ１９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「日本語」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「日本語」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。
【００２７】
スコア算出部１３は、キーワード抽出部１１が抽出した各キーワードの文字数と頻度算出部１２が算出した頻度とに基づいて、各キーワードのスコアを算出する。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。
【００２８】
本発明の実施の形態においては、キーワード抽出部１１が抽出した各キーワードの文字数を用いずに、頻度算出部１２によっ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。
【００２９】
例えば、スコア算出部１３は、以下に示すような、ＴＦ／ＩＤＦ法を用いたスコアの算出方法または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。
【００３０】
（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。
【００３１】
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。
【００３２】
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。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。
【００３３】
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ＤＢ１９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ＤＢ１９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。
【００３４】
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。

【００３５】

（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。

【００３６】

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

【００３７】

【数１】

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

【００３８】

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。

【００３９】

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ＤＢ１９に蓄積されている文書群として、

【００４０】

【数２】

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

【００４１】

を算出する。算出された値をｔｆ項とする。

【００４２】

また、例えば、書誌データＤＢ１９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。

【００４３】

データソート部１４は、書誌データＤＢ１９に蓄積されている文書から、文書データ（例えば、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。

【００４４】

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と、スコア算出部１３によって算出された各キーワードのスコアとに基づいて、各キーワードの優先度を設定する。

【００４５】
データソート部１４は、頻度算出部１２によって算出された頻度が高いキーワードほど高い優先度を設定する。また、データソート部１４は、頻度が同じであるキーワードについては、スコア算出部１３によって算出されたスコアが高いキーワードほど高い優先度を設定する。
【００４６】
本発明の実施の形態においては、データソート部１４は、スコア算出部１３によって算出された各キーワードのスコアを、各キーワードの優先度として設定する構成を採ってもよい。
【００４７】
各キーワードの優先度は、後述する表示部１５によって文書データとともに表示される各キーワードの表示の順序を規定する。
【００４８】
そして、データソート部１４は、設定した優先度がより高いキーワードを含む文書の順に、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。
【００４９】
また、データソート部１４は、後述する内部頻度算出部１７によって算出された内部頻度と、後述する内部スコア算出部１８によって算出された内部スコアとに基づいて、各キーワードの優先度を更新し、更新された優先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。
【００５０】
各キーワードの優先度を更新する場合、データソート部１４は、後述する内部頻度算出部１７によって算出される内部頻度が高いキーワードほど高い優先度を設定する。内部頻度が同じであるキーワードについては、後述する内部スコア算出部１８によって算出される内部スコアが高いキーワードほど高い優先度を設定する。
【００５１】
本発明の実施の形態においては、データソート部１４は、後述する内部スコア算出部１８によって算出される内部スコアを各キーワードの優先度として設定する構成を採ってもよい。
【００５２】
なお、本発明の実施の形態においては、データソート部１４は、書誌データＤＢ１９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、文書のタイトル、著者名等）を日付について降順または昇順にソートする構成を採ってもよい。
【００５３】
表示部１５は、データソート部１４によってソートされた各文書データを画面表示する。また、表示部１５は、各文書データが画面表示される画面と同一画面上において、優先度が高い順に各キーワードを画面表示する。また、表示部１５は、画面表示された各文書データに対応する文書が、画面表示された各キーワードを含んでいるかを示す情報を画面表示する。なお、表示部１５は、画面表示された各文書データに対応する文書が、画面表示された各キーワードを何個含んでいるかを示す情報を画面表示する構成を採ってもよい。
【００５４】
また、表示部１５は、書誌データＤＢ１９に蓄積されている文書群中に出現する頻度が予め定められた閾値未満であるキーワードを含む文書については、その文書が、上記頻度が閾値未満であるキーワードを含むことを示す情報を、頻度が予め定められた閾値以上であるキーワードを画面表示する表示領域とは別の表示領域に画面表示する構成を採ることもできる。
【００５５】
また、表示部１５は、文書データの画面表示後に、後述する表示データ指定部２０によ

って指定されたデータ以外のデータを画面から消去する構成を採ることができる。また、表示部１５は、文書データの画面表示後に、後述する表示データ指定部２０によって指定されたデータを画面から消去する構成を採ることができる。

【００５６】

キーワード選択部１６は、キーワードを選択する。内部頻度算出部１７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、上記選択されたキーワードを含む文書群中に出現する頻度である内部頻度を算出する。ここで、各キーワードの内部頻度とは、例えば、選択されたキーワードを含む文書群に含まれる文書のうち、各キーワードが出現する文書の数を意味する。また、本発明の実施の形態においては、選択されたキーワードを含む文書群での各キーワードの出現回数を内部頻度とする構成を採ることもできる。

【００５７】

また、本発明の実施の形態においては、キーワード選択部１６によって複数のキーワードが選択された場合には、内部頻度算出部１７は、選択された複数のキーワードを全て含む文書群中に各キーワードが出現する頻度を内部頻度として算出する構成を採ってもよい。

【００５８】

内部スコア算出部１８は、内部頻度算出部１７が算出した内部頻度と、キーワード選択部１６によって選択されたキーワードを含む文書に含まれる各キーワードの文字数とに基づいて、各キーワードの内部スコアを算出する。各キーワードの内部スコアは、例えば、各キーワードの文字数に内部頻度を乗じた値として算出される。

【００５９】

本発明の実施の形態においては、各キーワードの文字数を用いずに、内部頻度算出部１７によって算出された内部頻度に基づいて各キーワードの内部スコアを算出する構成を採ってもよい。例えば、内部スコア算出部１８は、上述したＴＦ／ＩＤＦ法や、Ｏｋａｐｉのウェイティング法を用いて内部スコアを算出する構成を採ってもよい。

【００６０】

書誌データＤＢ１９には、大量の文書（書誌データ）が蓄積されている。表示データ指定部２０は、表示部１５によって画面表示されるデータを指定する。

【００６１】

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、設定された優先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。

【００６２】

例えば、文書Ａが、優先度が最も高いキーワード「日本語」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「情報」とを含んでいるものとし、また、例えば、文書Ｂが、優先度が最も高いキーワード「日本語」と３番目の優先度であるキーワード「情報」と、４番目の優先度であるキーワード「自動」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、２番目の優先度である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、優先度がより高いキーワードを含んでいる。

【００６３】

優先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、各キーワードを優先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。

【００６４】

例えば、「日本語」、「解析」、「情報」、「自動」、・・・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。

【００６５】
ある文書が、優先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、優先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。
【００６６】
なお、本発明のデータ表示装置１の構成は、図１に示すものに限定されない。本発明の実施の形態においては、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。
【００６７】
また、本発明の実施の形態においては、例えば、文書と文書に含まれるキーワードの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、その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。
【００６８】
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ＤＢ１９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。
【００６９】
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。
【００７０】
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。
【００７１】
例えば、「学校へ行く」を入力すると、以下の結果を得る。
【００７２】
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。
【００７３】
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。
【００７４】
文献（３）： Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging,Computational Linguistics, 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ＤＢ１９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。
【００７５】
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。
【００７６】
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。
【００７７】
図２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。まず、キーワード抽出部１１が、書誌データＤＢ１９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「日本語」、「解析」、「情報」、「自動」、「翻訳」、「表現」、「モデル」、「抽出」、「手法」、「名詞」、「要約」、「検索」、・・・といったキーワードを抽出する。
【００７８】
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ＤＢ１９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）。
【００７９】
例えば、図３の表に示すように、算出されるキーワード「日本語」の頻度は２０、キーワード「解析」の頻度は１５、キーワード「情報」の頻度は１２、キーワード「自動」の頻度は１０、キーワード「翻訳」の頻度は９、キーワード「表現」の頻度は８、キーワード「モデル」の頻度は７、キーワード「抽出」の頻度は７、キーワード「手法」の頻度は６、キーワード「名詞」の頻度は５、キーワード「要約」の頻度は４、キーワード「検索」の頻度は３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３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２や頻度１についても算出され得る。
【００８０】
次に、スコア算出部１３が、キーワード抽出部１１が抽出した各キーワードの文字数と頻度算出部１２が算出した頻度とに基づいて、各キーワードのスコアを算出する（ステップＳ３）。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

して算出する。なお、スコアの算出に用いる文字数は、例えば、半角１文字を単位とする。従って、例えば、全角の文字については、１文字の文字数は２である。

【００８１】

例えば、図３の表に示すように、算出されるキーワード「日本語」のスコアは、頻度２０に文字数６を乗じた値である１２０となる。同様にして、キーワード「解析」のスコアは６０、キーワード「情報」のスコアは４８、キーワード「自動」のスコアは４０、キーワード「翻訳」のスコアは３６、キーワード「表現」のスコアは３２、キーワード「モデル」のスコアは４２、キーワード「抽出」のスコアは２８、キーワード「手法」のスコアは２４、キーワード「名詞」のスコアは２０、キーワード「要約」のスコアは１６、キーワード「検索」のスコアは１２である。

【００８２】

次に、データソート部１４が、書誌データＤＢ１９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Ｓ４）。例えば、文書データとして、文書のタイトル、著者名等のデータがバッファ中に格納される。

【００８３】

また、データソート部１４が、頻度算出部１２が算出した頻度とスコア算出部１３が算出したスコアとに基づいて、各キーワードの優先度を設定する（ステップＳ５）。データソート部１４は、頻度が高いキーワードほど高い優先度を設定する。また、データソート部１４は、例えば、頻度が同じであるキーワードについては、算出されたスコアが高いキーワードほど高い優先度を設定する。

【００８４】

従って、例えば、図３の表中に示す各キーワードについては、「日本語」、「解析」、「情報」、「自動」、「翻訳」、「表現」、「モデル」、「抽出」、「手法」、「名詞」、「要約」、「検索」、・・・といったキーワードの順に、より高い優先度が設定される。

【００８５】

データソート部１４は、優先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（ステップＳ６）。そして、表示部１５が、ステップＳ４においてデータソート部１４によってソートされた各文書の文書データを画面表示するとともに、各キーワードを優先度が高い順に画面表示する（ステップＳ７）。上記ステップＳ７の処理においては、表示部１５は、さらに、各文書がどのキーワードを含んでいるかを示す情報を画面表示する。

【００８６】

ステップＳ７の処理の結果、例えば、図４に示すような画面が表示される。図４の画面表示例では、優先度がより高いキーワードを含む文書の順に、論文名、著者名という文書の文書データが表示されている。また、図４の画面表示例では、矩形の枠で囲ったキーワードが、優先度が高い順に左から表示されている。なお、図４の画面左端に示す番号「１」、「２」、・・・は、行番号を示しており、画面中の矩形で囲った各キーワードの上部に示す番号「１」、「２」、・・・は、列番号を示している。

【００８７】

例えば、論文名が「Ａ」で著者名が「ａ」である文書は、優先度が最も高いキーワード「日本語」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「情報」と、４番目の優先度であるキーワード「自動」と、５番目の優先度であるキーワード「翻訳」と、６番目の優先度であるキーワード「表現」とを含んでいるとする。

【００８８】

また、例えば、論文名が「Ｂ」で著者名が「ｂ」である文書は、優先度が最も高いキーワード「日本語」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「情報」と、４番目の優先度であるキーワード「自動」と、５番目の優先度であるキーワード「翻訳」とを含んでいるが、６番目の優先度であるキーワード「表現」

は含んでいないとする。

【００８９】

また、例えば、論文名が「Ｃ」で著者名が「ｃ」である文書は、優先度が最も高いキーワード「日本語」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「情報」と、４番目の優先度であるキーワード「自動」とを含んでいるが、５番目の優先度であるキーワード「表現」は含んでいないとする。

【００９０】

本発明の実施の形態においては、優先度がより高いキーワードを含む文書の順に、各文書のデータがソートされ、画面表示されることから、図４の画面表示例では、上の行から、優先度がより高いキーワードを含む、論文名が「Ａ」で著者名が「ａ」という文書データ、論文名が「Ｂ」で著者名が「ｂ」という文書データ、論文名が「Ｃ」で著者名が「ｃ」という文書データの順に表示されている。

【００９１】

また、上述したように、ステップＳ７の処理においては、各文書がどのキーワードを含んでいるかを示す情報が画面表示される。例えば、図４に示すように、各文書に係る文書データが画面表示されている行と同じ行において、各文書が含んでいる矩形の枠で囲ったキーワードと同一の単語が、当該キーワードが画面表示されている列と同じ列に画面表示される。

【００９２】

図４に示す画面が、例えばセルで構成されている場合を想定すると、文書データが配置された行と、当該文書データに係る文書が含んでいる矩形の枠で囲ったキーワードが配置された列とが交差するセルに、当該矩形の枠で囲ったキーワードと同一の単語が配置される。

【００９３】

図４に示す画面を見ると、論文名が「Ａ」で著者名が「ａ」という文書データが画面表示されている第１行目において、この文書データに係る文書が含んでいる、矩形の枠で囲った各キーワード「日本語」、「解析」、「情報」、「自動」、「翻訳」、「表現」と同一の各単語（「日本語」、「解析」、「情報」、「自動」、「翻訳」、「表現」）が、矩形の枠で囲った各キーワードが表示されている列と同じ列に画面表示されている。

【００９４】

なお、本発明の実施の形態においては、表示部１５が表示する、各文書がどのキーワードを含んでいるかを示す情報は、画面表示されている各キーワードと同一の単語に限られない。例えば、各文書に係る文書データが画面表示されている行と同じ行において、各文書が含んでいる矩形の枠で囲った各キーワードが画面表示されている列と同じ列に、※印等を画面表示することによって、各文書がどのキーワードを含んでいるかが分かるようにしてもよい。

【００９５】

また、本発明の実施の形態では、例えば、ステップＳ２において算出された、頻度２や頻度１に係るキーワードを含む文書については、その文書に係る文書データが表示される行と同じ行に、当該文書が頻度２や頻度１に係るキーワードを含むことを示す情報を表示する構成を採ることもできる。

【００９６】

例えば、図４に示す画面表示例では、論文名が「Ａ」で著者名が「ａ」という文書データが表示されている行と同じ行に、頻度２に係るキーワード「尺度」と頻度１に係るキーワード「揺れ」が表示されている。従って、論文名が「Ａ」で著者名が「ａ」という文書データに係る文書は、頻度２に係るキーワード「尺度」と頻度１に係るキーワード「揺れ」を含んでいることが分かる。

【００９７】

図４に示す画面表示を見れば、例えば、キーワード「日本語」を含む文書群中において、キーワード「日本語」の他に、「解析」や「情報」といったキーワードを含む文書が多

く見られることがわかる。言い換えると、図４に示す画面表示を見れば、例えば、キーワード「日本語」を含む文書中において、「解析」や「情報」といったキーワードがキーワード「日本語」と共起して出現する割合が高いことが一見してわかる。

【００９８】

また、図４に示す画面表示を見れば、例えば、論文名が「Ａ」で著者名が「ａ」という文書データに係る文書は、「日本語」、「解析」、「情報」、「自動」、「翻訳」、「表現」というキーワードに関連する内容の文書であることが一見してわかる。

【００９９】

次に、キーワード選択部１６が、キーワードを選択する（ステップＳ８）。例えば、図４に示す画面上において、矩形の枠で囲ったキーワード「情報」が、左クリック等されると、キーワード選択部１６によってキーワード「情報」が選択される。

【０１００】

内部頻度算出部１７が、内部頻度を算出する（ステップＳ９）。例えば、キーワード選択部１６によって選択されたキーワードを含む文書群に含まれる文書のうち、上記ステップＳ１においてキーワード抽出部１１によって抽出された各キーワードが出現する文書の数を、内部頻度として算出する。

【０１０１】

例えば、図４に示す画面を参照すると、選択されたキーワード「情報」を含む１２個の文書からなる文書群において、キーワード「情報」が出現する頻度は１２、キーワード「解析」が出現する頻度は１０である。従って、例えば、図５の表に示すように、キーワード「情報」の内部頻度は１２、キーワード「解析」の内部頻度は１０である。

【０１０２】

同様にして、図５の表に示すように、例えば、キーワード「自動」の内部頻度は８、キーワード「日本語」の内部頻度は７、キーワード「表現」の内部頻度は６、キーワード「翻訳」の内部頻度は５、キーワード「モデル」の内部頻度は４、キーワード「抽出」の内部頻度は４、キーワード「手法」の内部頻度は３、キーワード「名詞」の内部頻度は２、キーワード「要約」の内部頻度は２、キーワード「検索」の内部頻度は１、・・・である。

【０１０３】

内部スコア算出部１８が、内部頻度算出部１７が算出した内部頻度と各キーワードの文字数とに基づいて、各キーワードの内部スコアを算出する（ステップＳ１０）。内部スコア算出部１８は、例えば、各キーワードの文字数に内部頻度を乗じて、各キーワードの内部スコアを算出する。なお、内部スコアの算出に用いる文字数は、例えば、半角１文字を単位とする。従って、例えば、全角の文字については、１文字の文字数は２である。

【０１０４】

例えば、図５の表に示すように、算出されるキーワード「情報」の内部スコアは、内部頻度１２に文字数４を乗じた値である４８となる。同様にして、キーワード「解析」の内部スコアは４０、キーワード「自動」の内部スコアは３２、キーワード「日本語」の内部スコアは４２、キーワード「表現」の内部スコアは２４、キーワード「翻訳」の内部スコアは２０、キーワード「モデル」の内部スコアは２４、キーワード「抽出」の内部スコアは１６、キーワード「手法」の内部スコアは１２、キーワード「名詞」の内部スコアは８、キーワード「要約」の内部スコアは８、キーワード「検索」の内部スコアは４である。

【０１０５】

データソート部１４が、内部頻度算出部１７が算出した内部頻度と内部スコア算出部１８が算出した内部スコアとに基づいて、各キーワードの優先度を設定する（ステップＳ１１）。ステップＳ１１の処理によって、上記ステップＳ５において設定された優先度が更新される。データソート部１４は、内部頻度が高いキーワードほど高い優先度を設定する。また、データソート部１４は、内部頻度が同じであるキーワードについては、算出された内部スコアが高いキーワードほど高い優先度を設定する。

【０１０６】

従って、例えば、図５の表中に示す各キーワードについては、「情報」、「解析」、「自動」、「日本語」、「表現」、「翻訳」、「モデル」、「抽出」、「手法」、「名詞」、「要約」、「検索」、・・・といったキーワードの順に、高い優先度が設定される。

【０１０７】

そして、データソート部１４は、ステップＳ１１において設定された優先度がより高いキーワードを含む文書の順に、ステップＳ４においてバッファ中に格納された文書のデータをソートする（ステップＳ１２）。

【０１０８】

ステップＳ７に戻って、表示部１５が、データソート部１４によってソートされた各文書の文書データを画面表示するとともに、各キーワードを優先度が高い順に画面表示する（ステップＳ７）。

【０１０９】

例えば、図６に示すような画面が表示される。ここで、例えば、論文名が「Ａ」で著者名が「ａ」である文書は、優先度が最も高いキーワード「情報」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「自動」と、４番目の優先度であるキーワード「日本語」と、５番目の優先度であるキーワード「表現」と、６番目の優先度であるキーワード「翻訳」とを含んでおり、従って、優先度がより高いキーワードを最も多く含んでいるとする。

【０１１０】

また、例えば、論文名が「Ｂ」で著者名が「ｂ」である文書は、優先度が最も高いキーワード「情報」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「自動」と、４番目の優先度であるキーワード「日本語」と、６番目の優先度であるキーワード「翻訳」と１１番目の優先度であるキーワード「要約」とを含んでいるが、５番目の優先度であるキーワード「表現」は含んでいないとする。

【０１１１】

また、例えば、論文名が「Ｃ」で著者名が「ｃ」である文書は、優先度が最も高いキーワード「情報」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「自動」と、４番目の優先度であるキーワード「日本語」とを含んでいるが、５番目の優先度であるキーワード「表現」や、６番目の優先度であるキーワード「翻訳」は含んでいないとする。

【０１１２】

また、例えば、論文名が「Ｕ」で著者名が「ｕ」である文書は、優先度が最も高いキーワード「情報」と、２番目の優先度であるキーワード「解析」と、３番目の優先度であるキーワード「自動」と、５番目の優先度であるキーワード「表現」と、６番目の優先度であるキーワード「翻訳」と、７番目の優先度であるキーワード「モデル」と、８番目の優先度であるキーワード「抽出」と、１０番目の優先度であるキーワード「名詞」と、１１番目の優先度であるキーワード「要約」とを含んでいるが、４番目の優先度であるキーワード「日本語」は含んでいないとする。

【０１１３】

本発明の実施の形態においては、優先度がより高いキーワードを含む文書の順に、各文書のデータがソートされ、画面表示されることから、図６の画面表示例では、上の行から、論文名が「Ａ」で著者名が「ａ」という文書データ、論文名が「Ｂ」で著者名が「ｂ」という文書データ、論文名が「Ｃ」で著者名が「ｃ」という文書データ、論文名が「Ｕ」で著者名が「ｕ」という文書データの順に表示されている。

【０１１４】

また、例えば、図６の画面表示例に示すように、各文書に係る文書データが表示されている行と同じ行において、各文書が含んでいる矩形の枠で囲った各キーワードと同一の単語が、各キーワードが画面表示されている列と同じ列に表示される。

【０１１５】

図６に示す画面表示を見れば、例えば、選択されたキーワード「情報」を含む文書群中

において、キーワード「情報」の他に、「解析」や「自動」といったキーワードを含む文書が多く見られることが一見してわかる。言い換えると、図６に示す画面表示を見れば、例えば、キーワード「情報」を含む文書中において、「解析」や「自動」といったキーワードがキーワード「情報」と共起して出現する割合が高いことが一見してわかる。

【０１１６】

本発明におけるデータ表示処理フローは、図２に示す処理フローに限られるものではない。例えば、図２のステップＳ７の直後に、表示データ指定部２０が、画面表示されるデータを表示データとして指定する処理を行い、表示部１５が、指定された表示データ以外のデータを画面上において消去する処理を行ってもよい。

【０１１７】

例えば、ユーザが図４に示す画面表示において、矩形の枠で囲った任意のキーワードを指定（例えば、右クリック等）すると、表示データ指定部２０が、指定されたキーワードを含む文書の文書データを表示データの一部として指定し、表示部１５が、指定されたキーワードを含む文書以外の文書の文書データを画面上において消去する。なお、上記において、「表示データの一部」としたのは、表示データ指定部２０が、指定されたキーワードを含む文書の文書データの他に、各矩形の枠で囲ったキーワードや、指定されたキーワードを含む文書が、矩形の枠で囲ったどのキーワードを含んでいるかを示す情報を表示データとして指定する構成を採ることもできるからである。

【０１１８】

また、例えば、ユーザが図４に示す画面表示において、矩形の枠で囲った任意のキーワードを指定（例えば、右クリック等）すると、表示データ指定部２０が、指定されたキーワードを含む文書以外の文書の文書データを表示データの一部として指定し、表示部１５が、指定されたキーワードを含む文書の文書データを画面上において消去する。

【０１１９】

また、本発明の実施の形態においては、例えば、図４または図６に示す画面表示において、ユーザが、矩形の枠で囲った任意のキーワードを指定（例えば、右クリック等）すると、表示データ指定部２０が、指定されたキーワード以外のキーワードが画面表示されている列と同じ列に表示されているデータ（例えば、各文書がどのキーワードを含んでいるかを示す情報）を表示データの一部として指定し、表示部１５が、指定されたキーワードが画面表示されている列と同じ列に表示されているデータを画面上において消去する。

【０１２０】

また、本発明の実施の形態においては、上記の例において、再度同じキーワードが指定されると、表示部１５が、一旦消去された、指定されたキーワードが画面表示されている列についてのデータを、画面表示し直す構成を採ってもよい。

【０１２１】

また、本発明の実施の形態においては、例えば、図４または図６に示す画面表示において、ユーザが、文書データ（例えば、各論文名や著者名についてのデータ）を指定（例えば、右クリック等）すると、表示データ指定部２０が、指定された文書データ以外の文書データが画面表示されている行と同じ行に表示されているデータ（例えば、各文書がどのキーワードを含んでいるかを示す情報）を表示データの一部として指定し、表示部１５が、指定された文書データが画面表示されている行と同じ行に表示されているデータを画面上において消去する。

【０１２２】

また、本発明の実施の形態においては、上記の例において、再度同じ文書データが指定されると、表示部１５が、一旦消去された、指定された文書データが画面表示されている行についてのデータを画面表示し直す構成を採ってもよい。

【０１２３】

本発明は、ｗｅｂ文書の情報検索結果に対しても適用することができる。例えば、任意のｗｅｂサイトから、ユーザがキーワードを指定して、当該キーワード（指定キーワード）を含む文書群を検索したときに、本発明のデータ表示装置１が、検索された文書群中に

含まれる複数のキーワードを抽出して、図２の各ステップに示す処理を行うように構成することもできる。また、上記本発明のｗｅｂ文書の情報検索結果に対する適用例において、データ表示装置１が画面表示するキーワードのうち、情報検索の際にユーザが指定した指定キーワードと同一のキーワードについては、指定キーワードと同一のキーワードであることを示す情報を同一画面上に表示する構成を採ることもできる。

【０１２４】

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】

【０１２５】

【図１】本発明のシステム構成の一例を示す図である。

【図２】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。

【図３】各キーワードの頻度とスコアの例を示す図である。

【図４】画面表示例を示す図である。

【図５】各キーワードの内部頻度と内部スコアの例を示す図である。

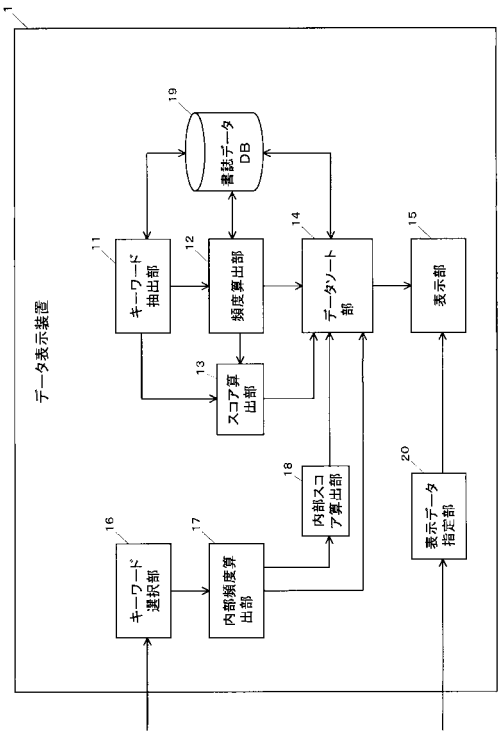
【図６】画面表示例を示す図である。

【符号の説明】

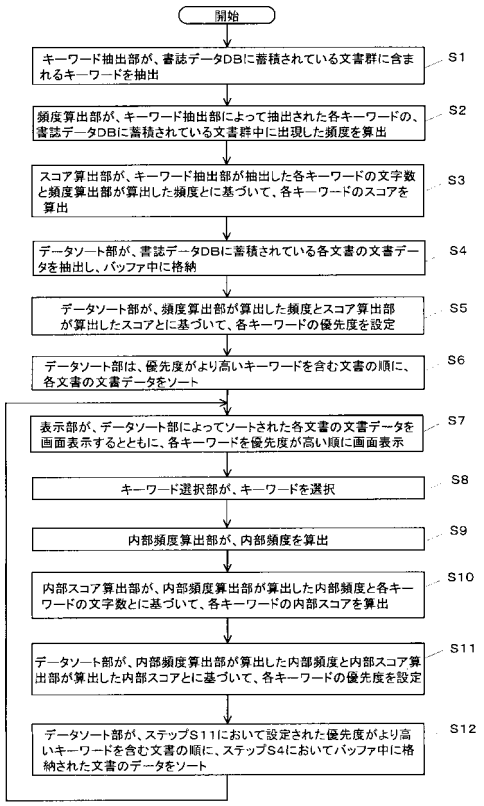
【０１２６】

１　データ表示装置
１１　キーワード抽出部
１２　頻度算出部
１３　スコア算出部
１４　データソート部
１５　表示部
１６　キーワード選択部
１７　内部頻度算出部
１８　内部スコア算出部
１９　書誌データＤＢ
２０　表示データ指定部

【図１】



【図２】



【図３】

	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	…
スコア	120	60	48	40	36	32	42	28	24	20	16	12	…
頻度	20	15	12	10	9	8	7	7	6	5	4	3	…

【図４】

	1	2	3	4	5	6	7	8	9	10	11	12	…
（論文名）（著者名）（頻度1）（頻度2）	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	…
1 A 逗孔 尺度	日本語	解析	情報	自動	翻訳	表現					要約		
2 B 次元 放送 ……	日本語	解析	情報	自動	翻訳			抽出	手法	名詞			
3 C	日本語	解析	情報					抽出					
4 D	日本語	解析	情報		翻訳								
5 E	日本語	解析			翻訳	表現	モデル						
6 F	日本語	解析		自動	翻訳	表現	モデル						
7 G	日本語	解析	情報			表現		抽出	手法	名詞	要約	検索	
8 H	日本語	解析	情報					抽出					
9 I	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
10 J	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
11 K	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
12 L	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
13 M	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
14 N	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
15 O	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
16 P	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
17 Q	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
18 R	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
19 S	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
20 T	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
21 U	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
22 V	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
23 W	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
24 X	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
25 Y	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
26 Z	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
27 A−1 a−1	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
28 B−1 b−1	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	
29 C−1 c−1 …	日本語	解析	情報	自動	翻訳	表現	モデル	抽出	手法	名詞	要約	検索	

【図５】

	情報	解析	自動	日本語	表現	翻訳	モデル	抽出	手法	名詞	要約	検索	…
内部スコア	48	40	32	42	24	20	24	16	12	8	8	4	…
内部頻度	12	10	8	7	6	5	4	4	3	2	2	1	…

【図６】

	1	2	3	4	5	6	7	8	9	10	11	12	…
（論文名）（著者名）（頻度1）（頻度2）	情報	解析	自動	日本語	表現	翻訳	モデル	抽出	手法	名詞	要約	検索	
1 A a	情報	解析	自動	日本語	表現	翻訳					要約		
2 B b	情報	解析	自動	日本語		翻訳							
3 C c …	情報	解析	自動	日本語									
4 U u …	情報	解析	自動		表現	翻訳	モデル	抽出		名詞	要約		
5 V v	情報	解析	自動		表現			抽出	手法			検索	
6 Z z	情報	解析	自動			翻訳	モデル		手法				
7 W w	情報	解析	自動	日本語				抽出	手法	名詞			
8 D d …	情報	解析		日本語	表現			抽出					
9 E e	情報	解析			表現	翻訳	モデル						
10 Y y	情報	解析		日本語	表現	翻訳	モデル						
11 I i	情報		自動	日本語	表現	翻訳							
12 J j	情報			日本語									
13 F f		解析		日本語									
14 G g		解析		日本語									
15 H h		解析		日本語									
16 X x		解析			表現	翻訳	モデル	抽出	手法	名詞			
17 A−1 a−1		解析											
18 K k			自動	日本語		翻訳	モデル						
19 B−1 b−1			自動					抽出	手法	名詞	要約	検索	
20 L l				日本語	表現			抽出					
21 M m				日本語					手法				
22 N n				日本語						名詞	要約		
23 O o				日本語								検索	
24 P p				日本語									
25 Q q				日本語									
26 R r				日本語									
27 S s				日本語									
28 T t				日本語									
29 C−1 c−1 …						翻訳	モデル						
…													

フロントページの続き

(56)参考文献 特開２００５－０５６１２５（ＪＰ，Ａ）
特開２００１－１４２８８７（ＪＰ，Ａ）
特開平１１－０２５１０８（ＪＰ，Ａ）
特開２００４－０２１７６３（ＪＰ，Ａ）
特開２００５－０１１３０１（ＪＰ，Ａ）
松田　透，統計的確率に基づくキーワード重要度算出モデル，情報処理学会研究報告　Ｖｏｌ．９６　Ｎｏ．８７，日本，社団法人情報処理学会，１９９６年　9月１３日，第96巻　第87号，123～128

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０