

(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2007-140639

(P2007-140639A)

(43) 公開日 平成19年6月7日(2007.6.7)

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

審査請求 未請求 請求項の数 11 O L (全 20 頁)

(21) 出願番号	特願2005-330009 (P2005-330009)	(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
(22) 出願日	平成17年11月15日(2005.11.15)	(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 NK06 PQ02 PQ46 PQ72 PR04 UU06

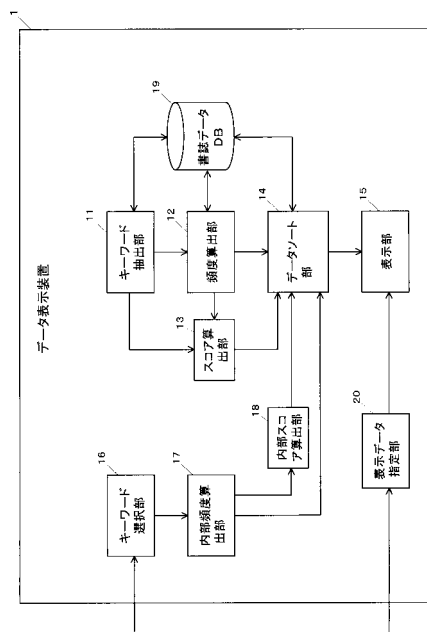
(54) 【発明の名称】 データ表示装置、データ表示方法およびデータ表示プログラム

(57) 【要約】

【課題】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を可能とする。

【解決手段】データ表示装置1は、書誌データDB19内の文書群中の各キーワードの、文書群中に出現した頻度とスコアとを算出し、頻度とスコアとに基づき、各キーワードについて優先度を設定し、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して画面表示する。また、データ表示装置1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が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。

【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、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、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

10

20

30

40

50

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、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、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。

10

【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

20

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、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

30

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するステップと、

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ステップと、

40

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テップと、

前記算出されたスコアを、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ステップと、

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ステップと、

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ステップとを有する

ことを特徴とするデータ表示方法。

50

【請求項 1 1】

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、データ表示方法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、データ表示方法およびデータ表示プログラムの提供を目的とする。

【課題を解決するための手段】

【0007】

10

20

30

40

50

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。 10

【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。 20

【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。 30

【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、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、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。 40

【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。 50

【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。

10

【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。

【0016】

また、本発明は、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するステップと、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ステップと、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テップと、前記算出されたスコアを、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ステップと、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ステップと、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ステップとを有することを特徴とする。

20

【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。

30

【発明の効果】

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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のウェイト法を用いて、各キーワードのスコアを算出する。

【0030】

50

(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】

$$\text{Score}(D) = \left(\text{tf}(w, D) \times \log\left(\frac{N}{\text{df}(w)}\right) \right)$$

10

上記の式において、wは、ユーザが入力するキーワード、 $\text{tf}(w, D) \times \log\left(\frac{N}{\text{df}(w)}\right)$ をwに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、 $\text{tf}(w, D)$ は、文書Dでのwの出現回数であり、 $\text{df}(w)$ 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、書誌データベース19に蓄積されている文書群として、 $\text{tf}(w, D)$ を算出する。また、例えば、書誌データベース19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 $\text{df}(w)$ の意味の説明において記述した「全文書」として、 $\text{df}(w)$ を算出する。

20

【0034】

そして、算出された $\text{tf}(w, D)$ と $\log\left(\frac{N}{\text{df}(w)}\right)$ との積を、各キーワードwのスコアとして算出する。

【0035】

(Okapiのウェイト法を用いたスコアの算出方法)

一般に、Okapiのウェイト法(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

30

【0037】

【数1】

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

【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、 $\text{tf}(w, D)$ は、文書Dでのwの出現回数であり、 $\text{df}(w)$ 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、 $\text{length}(D)$ 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。

40

【0039】

Okapiのウェイト法の本発明への適用に当たっては、例えば、上記文書Dを、書誌データベース19に蓄積されている文書群として、

【0040】

【数 2】

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

【0041】

を算出する。算出された値を t f 項とする。

【0042】

また、例えば、書誌データ D B 1 9 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 d f (w) の意味の説明において記述した「全文書」として、 $\log (N / d f (w))$ を算出する。算出された $\log (N / d f (w))$ を i d f 項とする。そして、算出された t f 項と i d f 項との積を、各キーワード w のスコアとして算出する。

10

【0043】

データソート部 1 4 は、書誌データ D B 1 9 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 1 4 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 1 4 は、頻度算出部 1 2 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 1 3 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。

20

【0045】

データソート部 1 4 は、頻度算出部 1 2 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 1 4 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 1 3 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 1 4 は、スコア算出部 1 3 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。

30

【0047】

各キーワードの優先度は、後述する表示部 1 5 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 1 4 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 1 4 は、後述する内部頻度算出部 1 7 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 1 8 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。

40

【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 1 4 は、後述する内部頻度算出部 1 7 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 1 8 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 1 4 は、後述する内部スコア算出部 1

50

8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。

10

【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。

20

【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。

30

【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。

40

【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のウェイト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

50

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。

10

【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、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。

20

【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。

30

【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。

40

【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 DB 19 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】

50

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・力行促音便	基本形

EOS

10

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、英語の品詞タグ付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】

20

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、求め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のウェイト法を用いて算出されるスコアを用いてもよい。

30

（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、単名詞の左右に接続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、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。

40

【0076】

文献（4）：中川裕志、森辰則、湯本紘彰：“出現頻度と接続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】

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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、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

10

20

30

40

50

。

【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。

10

【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。

20

【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)。

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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

10

20

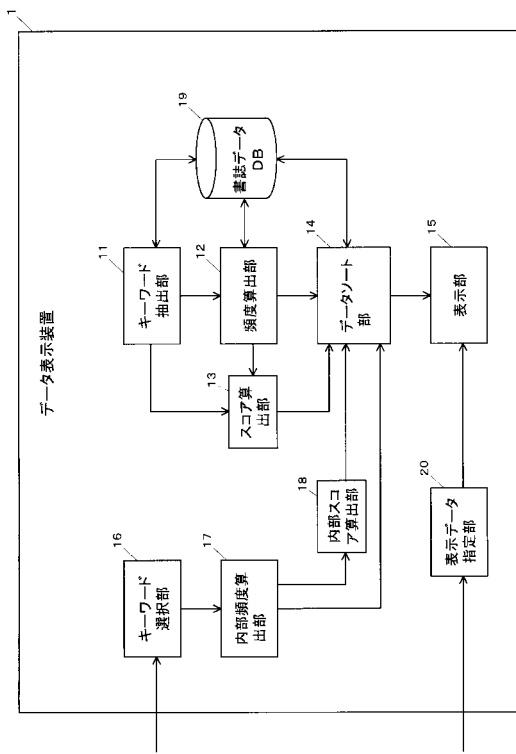
30

40

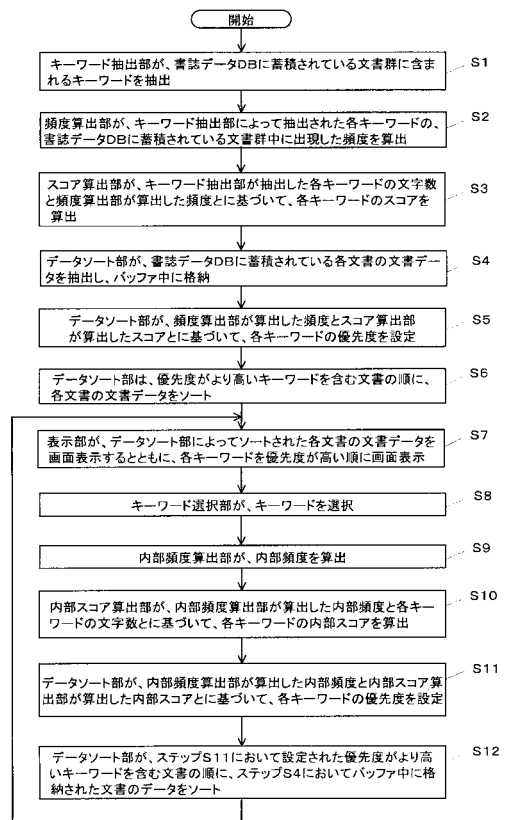
50

- 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	...
入力	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 】

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

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