(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5417471号
(P5417471)

(45)発行日　平成26年2月12日(2014.2.12)　(24)登録日　平成25年11月22日(2013.11.22)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 140
G06F 17/30 360Z

請求項の数 10　(全 16 頁)

(21)出願番号　特願2012-57240(P2012-57240)
(22)出願日　平成24年3月14日(2012.3.14)
(65)公開番号　特開2013-191046(P2013-191046A)
(43)公開日　平成25年9月26日(2013.9.26)
審査請求日　平成24年9月4日(2012.9.4)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(73)特許権者　301063496
東芝ソリューション株式会社
東京都港区芝浦一丁目１番１号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　國分　智晴
東京都港区芝浦一丁目１番１号　東芝ソリューション株式会社内
(72)発明者　真鍋　俊彦
東京都港区芝浦一丁目１番１号　東芝ソリューション株式会社内

最終頁に続く

(54)【発明の名称】構造化文書管理装置、構造化文書検索方法

(57)【特許請求の範囲】
【請求項１】
見出し、及び本文を含む複数の部分文書を備えた構造化文書を記憶する文書記憶部と、
前記見出しを抽出し、見出しリストを作成する見出し抽出部と、
前記部分文書中の語彙と、前記部分文書と対応する前記見出しとの概念の関連度をそれぞれ計算する関連度計算部と、
検索用キーワードと一致する前記語彙を含む前記部分文書を検索する文書検索部と、
前記検索用キーワードと一致した前記部分文書中の語彙に対する前記関連度が大きい前記見出しを前記関連度が小さい前記見出しより優先して選択する見出し選択部と、
選択された前記見出しを、それぞれ表示見出しとして表示部に表示させる見出し表示制御部と、
を備えることを特徴とする構造化文書管理装置。
【請求項２】
前記見出し選択部は、前記関連度が上位Ｎ個（Ｎは１以上の整数）の前記見出しを選択する
ことを特徴とする請求項１に記載の構造化文書管理装置。
【請求項３】
前記見出し選択部は、前記関連度が所定値以上の前記見出しを選択する
ことを特徴とする請求項１に記載の構造化文書管理装置。
【請求項４】

前記部分文書が、文書内に別の前記部分文書を従属文書として有しており、
前記関連度計算部は、前記従属文書中の前記語彙と、従属元の前記部分文書の前記見出しとの前記関連度を、前記従属文書中の前記語彙と、前記従属文書の前記見出しとの関連度よりも低く計算する
ことを特徴とする請求項１に記載の構造化文書管理装置。
【請求項５】
前記検索用キーワードと一致する前記語彙を含み、かつ前記見出し選択部により選択されなかった前記見出しを含む前記部分文書を、一致する前記語彙の前後の文章を含む態様で、前記表示部に表示させる本文表示制御部と、
をさらに備えることを特徴とする請求項１に記載の構造化文書管理装置。
【請求項６】
前記関連度計算部は、予め記録された概念辞書の語彙間の辞書関連度から、前記見出しと前記構造化文書中の語彙との前記関連度を計算する
ことを特徴とする請求項１に記載の構造化文書管理装置。
【請求項７】
前記見出し表示制御部は、表示した前記見出しが選択されると、選択された前記見出しと対応する前記本文を前記表示部に表示させる、
ことを特徴とする請求項１に記載の構造化文書管理装置。
【請求項８】
前記関連度計算部は、前記見出しが複数の語彙から構成される場合、計算した前記関連度が最も高い前記語彙の前記関連度を前記見出しの前記関連度として設定する
ことを特徴とする請求項１に記載の構造化文書管理装置。
【請求項９】
構造化文書管理装置にて実行される構造化文書検索方法であって、
見出し、及び本文を含む複数の部分文書を備えた構造化文書を記憶する文書記憶ステップと、
文書記憶ステップによる記憶時に、前記見出しを抽出して見出しリストを作成する見出し抽出ステップと、
前記部分文書中の語彙と、前記部分文書と対応する前記見出しとの概念の関連度をそれぞれ計算する関連度計算ステップと、
検索用キーワードと一致する前記語彙を含む前記部分文書を検索する文書検索ステップと、
前記検索用キーワードと一致した前記部分文書中の語彙に対する前記関連度が大きい前記見出しを前記関連度が小さい前記見出しより優先して選択する見出し選択ステップと、
選択された前記見出しを、それぞれ表示見出しとして表示部に表示させる見出し表示ステップと、
を含むことを特徴とする構造化文書検索方法。
【請求項１０】
構造化文書管理装置にて実行される構造化文書検索方法であって、
見出し、及び本文を含む複数の部分文書を備えた構造化文書を記憶する文書記憶ステップと、
文書記憶ステップによる記憶時に、前記見出しを抽出して見出しリストを作成する見出し抽出ステップと、
検索用キーワードと一致する語彙を含む前記部分文書を検索する文書検索ステップと、
前記文書検索ステップにより前記検索用キーワードと一致した前記語彙と、当該語彙が含まれる前記構造化文書と対応する前記見出しとの概念の関連度を計算する関連度計算ステップと、
前記検索用キーワードとの前記関連度が大きい前記見出しを前記関連度が小さい前記見出しより優先して選択する見出し選択ステップと、
選択された前記見出しを、それぞれ表示見出しとして表示部に表示させる見出し表示ス

テップと、
を含むことを特徴とする構造化文書検索方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明の実施形態は、構造化文書管理装置、構造化文書検索方法に関する。
【背景技術】
【０００２】
従来、電子データを構造化文書として生成し、情報の共有化を容易にしたり、より効率的に情報を検索できるようにしたりする技術が知られている。例えば、ＨＴＭＬ(Hyper Text Markup Language)では、文書の構成要素、例えば文書の見出し、本文、リスト構造などをタグ(tag)で記載することにより、文書の構造を表現することができる。また、目的に応じて文書構造を示すタグを独自に定義することができるＸＭＬ(Extensible Markup Language)も利用されるようになっている。このような構造化文書に対して検索を行う場合、タグによってどういうデータが文書中のどの位置に存在するのかを把握しやすくなり、検索性を向上させることができる。
【０００３】
こうした、構造化文書を検索した結果を表示する方法としては、検索結果の文章から自動的に要約を生成して表示する文書要約技術が知られている。文書要約技術の代表的な技術としてＫＷＩＣ(KEYWORD IN CONTEXT)要約技術が知られており、ＫＷＩＣでは検索対象の文書中から検索用キーワードを含むテキストの前後所定文字数抜き出して表示する。
【０００４】
また、構造化文書を検索した結果を表示する方法としては、検索に用いたキーワードと一致した語彙を含む文書に対応した見出しを検索結果として表示する方法が知られている。
【先行技術文献】
【特許文献】
【０００５】
【特許文献１】特開２００２－２７８９７２号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
しかしながら、見出しを検索結果として表示する場合、仮に検索用キーワードと文書中の語彙とが一致していたとしても、見出しが検索用キーワードとは関連度の低いものであった場合、利用者はその情報を自分が探している情報であると認識できない。その場合、利用者は実際にその文章を読んで、自分が探したい内容と近いものであるかを確認する必要があり、より一層の検索の利便性の向上が求められていた。
【０００７】
本発明は、上記に鑑みてなされたものであって、検索時の利便性を向上できる構造化文書管理装置を提供することにある。
【課題を解決するための手段】
【０００８】
上述した課題を解決し、目的を達成するために、実施形態の構造化文書管理装置は、文書記憶部と、見出し抽出部と、関連度計算部と、文書検索部と、見出し選択部と、見出し表示部と、を備える。文書記憶部は、複数の構造化文書を記憶する。見出し抽出部は、構造化文書の見出しを抽出し、抽出した見出しを含む見出しリストを作成する。関連度計算部は、構造化文書中の語彙と、構造化文書と対応する見出しとの概念の関連度をそれぞれ計算する。文書検索部は、検索用キーワードと一致する語彙を含む構造化文書を検索する。見出し選択部は、検索用キーワードと一致した語彙に対する関連度が大きい見出しを、関連度が小さい見出しより優先して選択する。表示制御部は、見出し選択部により選択さ

れた見出しを、表示見出しとして表示部に表示させる。

【図面の簡単な説明】

【０００９】

【図１】図１は、構造化文書管理システムのシステム構築例を示す模式図である。

【図２】図２は、サーバおよびクライアント端末のモジュール構成図である。

【図３】図３は、第１の実施形態のサーバおよびクライアント端末の概略構成を示すブロック図である。

【図４】図４は、第１の実施形態の構造化文書の１例を示す図である。

【図５】図５は、第１の実施形態の構造化文書の１例を示す図である。

【図６】図６は、第１の実施形態の見出しリストの１例を示す図である。

【図７】図７は、第１の実施形態の概念辞書の一例を示す図である。

【図８】図８は、第１の実施形態の語彙間の関連度を示すデータ図である。

【図９】図９は、第１の実施形態の見出しに対する本文中の語彙との関連度を示す図である。

【図１０】図１０は、第１の実施形態の検索結果の表示の仕方の一例を示す図である。

【図１１】図１１は、第１の実施形態の検索結果の表示の仕方の変形例を示す図である。

【図１２】図１２は、第１の実施形態の構造化文書を登録する際の処理の流れを示すフロー図である。

【図１３】図１３は、第１の実施形態の見出しに対する本文中の語彙との関連度を計算する処理の流れを示すフロー図である。

【図１４】図１４は、第１の実施形態の検索時において検索結果として表示する見出しを決定する処理の流れを示すフロー図である。

【図１５】図１５は、第２の実施形態の検索時において検索結果として表示する見出しを決定する処理の流れを示すフロー図である。

【発明を実施するための形態】

【００１０】

（第１の実施形態）

以下に、本発明にかかる構造化文書管理装置の第１の実施形態を図面に基づいて詳細に説明する。図１は、第１の実施形態にかかる構造化文書管理システムのシステム構築例を示す模式図である。ここでは、実施形態の構造化文書管理システムとして、図１に示すように、構造化文書管理装置であるサーバコンピュータ（以下、サーバという。）１に、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）等のネットワーク２を介して、クライアントコンピュータ（以下、クライアント端末という。）３が複数台接続されたサーバクライアントシステムを想定する。

【００１１】

図２は、サーバ１およびクライアント端末３のモジュール構成図である。サーバ１およびクライアント端末３は、例えば、通常のコンピュータを利用したハードウェア構成を有している。すなわち、サーバ１およびクライアント端末３は、情報処理を行うＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）１０１、ＢＩＯＳなどを記憶した読出し専用メモリであるＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）１０２、各種データを書き換え可能に記憶するＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）１０３、各種データベースとして機能するとともに各種のプログラムを格納するＨＤＤ（Ｈａｒｄ　Ｄｉｓｃ　Ｄｒｉｖｅ）１０４、記憶媒体１１０を用いて情報を保管したり外部に情報を配布したり外部から情報を入手するためのＣＤ－ＲＯＭドライブ等の媒体駆動装置１０５、ネットワーク２を介して外部の他のコンピュータと通信により情報を伝達するための通信制御装置１０６、処理経過や結果等を操作者に表示するＣＲＴ（Ｃａｔｈｏｄｅ　Ｒａｙ　Ｔｕｂｅ）やＬＣＤ（Ｌｉｑｕｉｄ　Ｃｒｙｓｔａｌ　Ｄｉｓｐｌａｙ）等の表示部１０７、並びに操作者がＣＰＵ１０１に命令や情報等を入力するためのキーボードやマウス等の入力部１０８等を備えた構成であり、これらの各部間で送受信されるデータをバスコントローラ１０９が調停して動作する。

【００１２】
このようなサーバ１およびクライアント端末３では、ユーザが電源を投入するとＣＰＵ１０１がＲＯＭ１０２内のローダーというプログラムを起動させ、ＨＤＤ１０４よりＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）というコンピュータのハードウェアとソフトウェアとを管理するプログラムをＲＡＭ１０３に読み込み、このＯＳを起動させる。このようなＯＳは、ユーザの操作に応じてプログラムを起動したり、情報を読み込んだり、保存を行ったりする。ＯＳのうち代表的なものとしては、Ｗｉｎｄｏｗｓ（登録商標）、ＵＮＩＸ（登録商標）等が知られている。これらのＯＳ上で動作するプログラムをアプリケーションプログラムと呼んでいる。なお、アプリケーションプログラムは、所定のＯＳ上で動作するものに限らず、後述の各種処理の一部の実行をＯＳに肩代わりさせるものであってもよいし、所定のアプリケーションソフトやＯＳなどを構成する一群のプログラムファイルの一部として含まれているものであってもよい。
【００１３】
ここで、サーバ１は、アプリケーションプログラムとして、構造化文書管理プログラムをＨＤＤ１０４に記憶している。この意味で、ＨＤＤ１０４は、構造化文書管理プログラムを記憶する記憶媒体として機能する。また、一般的には、サーバ１のＨＤＤ１０４にインストールされるアプリケーションプログラムは、ＣＤ－ＲＯＭやＤＶＤなどの各種の光ディスク、各種光磁気ディスク、フレキシブルディスクなどの各種磁気ディスク、半導体メモリ等の各種方式のメディア等の記憶媒体１１０に記録されて提供される。このため、ＣＤ－ＲＯＭ等の光情報記録メディアやＦＤ等の磁気メディア等の可搬性を有する記憶媒体１１０も、構造化文書管理プログラムを記憶する記憶媒体となり得る。さらには、構造化文書管理プログラムは、例えば通信制御装置１０６を介して外部から取り込まれ、ＨＤＤ１０４にインストールされてもよい。
【００１４】
サーバ１は、ＯＳ上で動作する構造化文書管理プログラムが起動すると、この構造化文書管理プログラムに従い、ＣＰＵ１０１が各種の演算処理を実行して各部を集中的に制御する。一方、クライアント端末３は、ＯＳ上で動作するアプリケーションプログラムが起動すると、このアプリケーションプログラムに従い、ＣＰＵ１０１が各種の演算処理を実行して各部を集中的に制御する。サーバ１およびクライアント端末３のＣＰＵ１０１が実行する各種の演算処理のうち、実施形態の構造化文書管理システムにおいて特徴的な処理について、以下に説明する。
【００１５】
図３は、第１の実施形態におけるサーバ１およびクライアント端末３の概略構成を示すブロック図である。図３に示すように、クライアント端末３は、アプリケーションプログラムにより実現される機能構成として、構造化文書登録部１１と、検索部１２とを備える。
【００１６】
構造化文書登録部１１は、入力部１０８から入力された構造化文書データやクライアント端末３のＨＤＤ１０４に予め記憶された構造化文書データを、後述するサーバ１の構造化文書データベース（構造化文書ＤＢ）２１に登録するためのものである。この構造化文書登録部１１は、登録すべき構造化文書データとともに格納要求をサーバ１に送信する。
【００１７】
検索部１２は、ユーザにより入力部１０８から入力された指示に従って、構造化文書ＤＢ２１から所望のデータを検索するための検索用キーワードなどが記述されたクエリデータを作成し、当該クエリデータを含む検索要求をサーバ１へ送信する。また、検索部１２は、サーバ１から送信された当該検索要求に対応する結果データを受け取り、これを表示部１０７に表示する。
【００１８】
一方、サーバ１は、構造化文書管理プログラムにより実現される機能構成として、登録部２２と、検索部２３とを備える。また、サーバ１は、ＨＤＤ１０４などの記憶装置を利

用した構造化文書ＤＢ２１を備える。
【００１９】
登録部２２は、クライアント端末３からの格納要求を受けて、クライアント端末３から送信された構造化文書データを構造化文書ＤＢ２１に格納する処理を行う。登録部２２は、格納インタフェース部２４と、見出し抽出部２５と、関連度計算部２６とを備える。
【００２０】
格納インタフェース部２４は、構造化文書データの入力を受け付けて、構造化文書データを構造化文書ＤＢ２１に格納するために、クライアント端末３から送信された構造化文書データを構文解析する。そして、格納インタフェース部２４は、データ中に出現する要素に、要素間で出現順序が比較可能な識別子（以下、要素ＩＤという。）を付与した上で、要素ＩＤが付与された構造化文書データを構造化文書ＤＢ２１（構造化文書データ記憶手段）に格納する。なお、要素ＩＤはクライアント端末３側で予め構造化文書に手動で付与しておいてもよい。
【００２１】
図４は、この要素ＩＤが付与された構造化文書データの一例を示したものである。構造化文書データを記述するための代表的な言語としてＸＭＬ（Ｅｘｔｅｎｓｉｂｌｅ Ｍａｒｋｕｐ Ｌａｎｇｕａｇｅ）が挙げられる。図４に示す構造化文書データは、ＸＭＬで記述されたものである。ＸＭＬでは、文書構造を構成する個々のパーツを「要素」（エレメント：Ｅｌｅｍｅｎｔ）と呼び、要素はタグ（ｔａｇ）を使って記述する。具体的には、要素の始まりを示すタグ（開始タグ）と、終わりを示すタグ（終了タグ）の２つのタグでデータを挟み込んで、１つの要素を表現している。なお、開始タグと終了タグで挟み込まれたテキストデータは、当該開始タグと終了タグで表された１つの要素に含まれるテキスト要素である。
【００２２】
図４では、＜ｄｏｃ＞というタグで囲まれたルート要素が存在する。＜ｄｏｃ＞要素は、そのドキュメントの文書ＩＤとしてｉｄ＝１が割り当てられている。＜ｄｏｃ＞要素は、＜ｔｉｔｌｅ＞要素を持ち、＜ｔｉｔｌｅ＞要素はその構造化文書の見出しを示している。また、＜ｄｏｃ＞要素は、５つの＜ｓｅｃ＞要素を有している。＜ｓｅｃ＞要素は、＜ｄｏｃ＞要素によって規定される構造化文書と親子関係にある構造化文書であり、本実施形態においては部分文書と呼ぶ。＜ｓｅｃ＞というタグで囲まれた中には、＜ｓｅｃｔｉｔｌｅ＞要素と、＜ｐａｒａ＞要素とが含まれている。＜ｓｅｃｔｉｔｌｅ＞は、その部分文書の見出しを示すタグである。また、＜ｐａｒａ＞は、その部分文書の説明文を示すタグである。この＜ｓｅｃｔｉｔｌｅ＞、および＜ｐａｒａ＞で定義されてテキストが「本文」に相当する。それぞれのタグには@ｅｉｄという形式で要素ＩＤが付与されている。
【００２３】
また、図５も同様に構造化文書の一例を示している。図５においても、図４の構造化文書と同じ構造を有しているが、要素ＩＤである@ｅｉｄ＝２０８にて定義された部分文書が、@ｅｉｄ＝２０５にて定義された部分文書中に含まれており、親子関係の階層となっている。
【００２４】
見出し抽出部２５は、格納インタフェース部２４から受理した構造化文書から見出しを抽出して、抽出した見出しをリスト化する。見出しを抽出する際には、構造化文書中の＜ｓｅｃｔｉｔｌｅ＞要素によって囲まれたテキストが見出しであると認識される。図６は、文書ＩＤ　１、および文書ＩＤ　２の２つの構造化文書において見出しをリスト化したデータの一例を示している。図６に示されるように、文書ＩＤ　１の構造化文書においては、要素ＩＤ　１０９、１０２、１０６、１１２および１１５で示される部分文書に対して、@ｅｉｄ＝１１０、１０３、１０７、１１３および１１６が、それぞれ見出しとして抽出される。
【００２５】

また、文書ＩＤ　２の構造化文書においては、要素ＩＤ　２０２、２０５、および２１１で示される部分文書に対して、@ｅｉｄ＝２０３、２０６、および２１２が、それぞれ見出しとして抽出される。また、要素ＩＤ　２０８で示される部分文書に対しては、@ｅｉｄ＝２０６、および２０９の２つの見出しが抽出される。文書ＩＤ　２の構造化文書においては、要素ＩＤ　２０８で示される部分文書の見出しとして、自身の＜ｓｅｃ＞タグで囲われた@ｅｉｄ＝２０９の見出しだけではなく、親階層における@ｅｉｄ＝２０６の見出しも抽出される。本実施形態において、従属文書とは、親階層の部分文書を定義する＜ｓｅｃ＞要素内の子階層にて＜ｓｅｃ＞要素にて定義された部分文書である。図５に示される構造化文書おいては、見出し@ｅｉｄ＝２０６を含む部分文書@ｅｉｄ＝２０５にとって、部分文書@ｅｉｄ＝２０８が従属文書に相当し、一方、部分文書@ｅｉｄ＝２０８にとって、部分文書@ｅｉｄ＝２０５は、従属元の部分文書に相当する。

【００２６】

見出し抽出部２５は、生成した見出しリストを構造化文書ＤＢ２１に記憶するとともに、見出しリストを関連度計算部２６へと引き渡す。関連度計算部２６は、見出し抽出部２５によって抽出された見出しと、対応する部分文書中に含まれる語彙との関連度を計算する。関連度の計算にあたっては、図７にて示される概念辞書が用いられる。概念辞書は、概念の上下構造に基づき、それぞれの概念がどれくらい近似したものであるかを示している。例えば、図７における「ルーター」と「アクセスポイント」は、同じノードから分岐した同じ階層に位置しており、その概念上の距離ｌｅｎｇｔｈは「１」として示される。また、親ノードと子ノードとの概念的な距離ｌｅｎｇｔｈも「１」として示される。図８は、概念辞書に予め設定された辞書関連度に基づき語彙間の関連度を計算した表である。関連度は概念的な距離ｌｅｎｇｔｈを用いて表され、１／（距離ｌｅｎｇｔｈ＋１）によって計算され、距離ｌｅｎｇｔｈが５以上のものは０として示している。

【００２７】

関連度計算部２６は、それぞれの見出しから語彙を抽出し、本文中の語彙との間で関連度を計算する。語彙の抽出の仕方は、既存の方法を用いることができ、テキスト中から語彙を認識して抽出する。例えば、@ｅｉｄ＝１１６にて定義された「無線ＬＡＮのトラブルシューティング」という見出しからは、語彙として「ＬＡＮ、無線ＬＡＮ」の２語彙が抽出される。一方、この部分文書の@ｅｉｄ＝１１５で定義される本文からは、「ＬＡＮ、無線ＬＡＮ、ルーター、アクセスポイント」の語彙が抽出される。この場合、見出し中の語彙それぞれに対する各語彙の関連度が計算される。語彙「ＬＡＮ」に対する「ＬＡＮ、無線ＬＡＮ、ルーター、アクセスポイント」の関連度は順に「１．０、０．３３３、０．３３３、０．３３３」となり、語彙「無線ＬＡＮ」に対する「ＬＡＮ、無線ＬＡＮ、ルーター、アクセスポイント」の関連度は順に「０．３３３、１．０、０．２５、０．２５」となる。この場合、各語彙に対して関連度が大きい語彙の値が優先されるため、@ｅｉｄ＝１１６に対する@ｅｉｄ＝１５の部分文書中の語彙の関連度は、「１．０、１．０、０．３３３、０．３３３」となる。関連度計算部２６は、それぞれの見出しと部分文書との組み合わせに対してこの計算を行い、計算結果を図９で示す、見出し語彙関連度表として、構造化文書ＤＢ２１に記憶する。なお、関連度の計算の際に、例えば文書ＩＤ　２の見出しである@ｅｉｄ＝２０６のように、子階層の部分文書との間で関連度を計算する場合は、同じ階層の部分文書との間で関連度を計算する場合と比較して、その関連度が少なく計算され、本実施形態においては、１／（距離ｌｅｎｇｔｈ＋１）を１／２にした値となる。このように構造化文書の階層の深さが深いほど関連度を小さくしていく。

【００２８】

図３へと戻り、検索部２３の機能構成について説明する。検索部２３は、検索インタフェース部２９と、照合部３０と、見出し選択部３１とを備えている。

【００２９】

検索インタフェース部２９は、検索用キーワードの入力を受け付けて、受け付けた検索用キーワードを含むクエリデータにより指定された検索用キーワードと一致する語彙を含むデータを得るために照合部３０を呼び出す。

【００３０】
照合部３０は、構造化文書ＤＢ２１へとアクセスし、構造化文書データ２７からクエリデータにより指定された検索用キーワードを含む構造化文書を検索し、検索用キーワードと一致する語彙を含む部分文書の一覧を見出し選択部３１へと送る。例えば、検索用キーワードが「無線ＬＡＮ」である場合、部分文書として、文書ＩＤ　１の@ｅｉｄ＝１０９、１０２、１０６、１１２、１１５、および文書ＩＤ　２の@ｅｉｄ＝２０２、２０５、２０８、２１１がヒットし、この検索結果が見出し選択部３１へと送られる。
【００３１】
見出し選択部３１は、検索用キーワードと一致した語彙に対して関連度が大きい見出しを、関連度が小さい見出しよりも優先して選択し、この選択結果を検索インタフェース部２９へと引き渡す。関連度が大きい見出しを優先する方法としては、関連度が低い見出しは選択しないようにしたり、関連度が上位の見出しのみを選択したりするような方法が考えられる。具体的には、まず、見出し選択部３１は、ヒットした部分文書それぞれの見出しの検索用キーワードと一致する語彙に対する関連度を見出し語彙関連度表から調べる。上述の「無線ＬＡＮ」という検索用キーワードに対しては、関連度が０より大きい見出しは、文書ＩＤ　１では@ｅｉｄ＝１１０、１１６であり、見出し選択部３１はこれらの関連度を取得する。見出し選択部３１は、この取得した関連度のうち上位Ｎ個、例えば２個を選択し、検索結果に表示見出しとして表示する見出しを選択する。この場合、文書ＩＤ　１の部分文書の要素ＩＤ@ｅｉｄ＝１０９と対応した見出し@ｅｉｄ＝１１０と、部分文書の要素ＩＤ@ｅｉｄ＝１１５と対応した見出し@ｅｉｄ＝１１６と、が選択される。また、文書ＩＤ　２の部分文書の要素ＩＤ@ｅｉｄ＝２０５と対応した見出し@ｅｉｄ＝２０６と、部分文書の要素ＩＤ@ｅｉｄ＝２０８と対応した見出し@ｅｉｄ＝２０９と、が選択される。見出し選択部３１は、この選択結果を検索インタフェース部２９へと送る。
【００３２】
検索インタフェース部２９は、見出し選択部３１から受け取った見出しを、表示部１０７に対して、表示させるように出力する。図１０は、表示部に表示された検索結果画面の一例を示している。図１０に示されるように、検索インタフェース部２９は、文書ＩＤ　１のタイトルである「パソコン取扱説明書」を表示した下に、表示見出しである「ネットワーク接続」と「無線ＬＡＮのトラブルシューティング」の２つの表示見出しを表示させるよう処理を行う。また、検索インタフェース部２９は、文書ＩＤ　２のタイトルである「携帯端末取扱説明書」を表示した下に、表示見出しである「ネットワーク設定」、および「アクセスポイントの設定」を表示させる。利用者はこの表示された表示見出しを選択することで、この表示見出しと対応付けられた本文を閲覧することができる。
【００３３】
なお、この表示画面の別の例としては図１１で示す態様となるようにすることできる。図１１においては、検索インタフェース部２９は、見出し選択部３１から送られた見出し以外の見出しについては、検索用キーワードと一致する語彙の前後の文も表示するようにしている。図１１に示されるように、タイトルである「パソコン取扱説明書」の下に、@ｅｉｄ＝１０２の部分文書中の本文である「無線ＬＡＮとは無線通信を利用してデータの・・・」が、@ｅｉｄ＝１０６の部分文書中の本文である「無線機能を無線ＬＡＮオン／オフボタンで有効にしてか・・・」が、@ｅｉｄ＝１１２の部分文書中の本文である「対策のためパスワード設定や、無線ＬＡＮの暗号化設定などを備えており・・・」が、それぞれ表示されている。検索用キーワードと一致する語彙を含む前後それぞれ何文字を抽出するかは適宜変更可能である。このようにすることで、見出しの語彙と、検索用キーワードと一致する語彙との関連度が低いため、表示見出しからでは利用者がその部分文書中に検索用キーワードが含まれているか否かわかりにくい文書であっても、利用者は文章から内容を把握することができるようになる。本実施形態では、検索インタフェース部２９が、見出し表示制御部、および本文表示制御部に相当する。
【００３４】

以上に示した本実施形態における構造化文書の登録、および検索の処理の流れを図１２～図１４を用いて説明する。図１２は、構造化文書の登録時の処理の流れを示している。図１２の処理は例えばクライアント端末３の構造化文書登録部１１から構造化文書を登録する旨の指示が出されたときに処理がスタートする。まず、格納インタフェース部２４は、クライアント端末３から送られた構造化文書の読み込みを行う（ステップＳ１０１）。次いで、見出し抽出部２５は、読み込んだ構造化文書から見出しを抽出する（ステップＳ１０２）。そして、見出し抽出部２５は、抽出した見出しから見出しリストを作成し（ステップＳ１０３）、構造化文書ＤＢ２１に記憶する（ステップＳ１０４）。そして、処理を終了する。

【００３５】

次いで、見出しと本文中の語彙との関連度を計算する処理の流れを図１３から説明する。図１３に示されるように、関連度計算部２６は、構造化文書ＤＢ２１に記憶された見出しリストからデータ１行分の見出しを選択する（ステップＳ２０１）。次いで、関連度計算部２６は、選択した見出しから語彙を抽出する（ステップＳ２０２）。次いで、関連度計算部２６は、見出しと対応する本文、ここでは＜ｓｅｃｔｉｔｌｅ＞と＜ｐａｒａ＞で定義されたテキストの中から、語彙を抽出する（ステップＳ２０３）。関連度計算部２６は、見出し中の語彙と、部分文書中の語彙との間で関連度を計算する。（ステップＳ２０４）。次いで、関連度計算部２６は、見出し中に語彙が複数ある場合に、それぞれの語彙との関連度のうち高いほうの値を見出しの関連度として設定する（ステップＳ２０５）。そして、関連度計算部２６は、見出し語彙関連度表の該当する部分文書と見出しとの組み合わせのデータの「見出し語彙関連度」の項目へ関連度のデータを追加する（ステップＳ２０６）。最後に、全ての見出しについて関連度を計算する処理が完了したか否かの判定がなされ（ステップＳ２０７）、処理が完了した場合（ステップＳ２０７：Ｙｅｓ）、一連の処理を終了し、処理が完了していない場合（ステップＳ２０７：Ｎｏ）、次の行の見出しについて同様の処理を繰り返す。

【００３６】

次に、検索時に見出し選択部３１によって見出しが選択される処理の流れを、図１４を用いて説明する。見出し選択部３１は、検索用キーワードと一致した語彙を含む構造化文書を取得する（ステップＳ３０１）。次いで、見出し選択部３１は、取得した構造化文書中で、検索用キーワードと一致した語彙を含む部分文書の見出しに対する、当該キーワードに対する関連度を見出し語彙関連度表から取得する（ステップＳ３０２）。見出し選択部３１は、全ての一致語彙を含む部分文書に対して関連度を取得したか否かの判定を行い（ステップＳ３０３）、全て取得済みである場合（ステップＳ３０３：Ｙｅｓ）、一致した語彙を含む部分文書の見出しを関連度に基づき降順でソートする（ステップＳ３０４）。一方、全ての部分文書に対する関連度が取得できていないと判定された場合（ステップＳ３０３：Ｎｏ）、ステップＳ３０２の処理を繰り返す。見出し選択部３１は、関連度の上位Ｎ個の見出しを選択し、構造化文書中の出現順でソートする（ステップＳ３０５）。そして、見出し選択部３１は、全ての構造化文書（本実施形態では、文書ＩＤ　１、および文書ＩＤ　２の２つの文書）において、見出しの選択が終了したか否かを判定し（ステップＳ３０６）、終了した場合は（ステップＳ３０６：Ｙｅｓ）、ステップＳ３０５でソートして選択した見出しを表示見出しとして検索インタフェース部２９へと送り（ステップＳ３０７）、処理を終了する。全ての構造化文書での見出しの選択が終了していない場合は（ステップＳ３０６：Ｎｏ）。ステップＳ３０１からの処理を繰り返し、別の構造化文書を取得する。

【００３７】

以上に示した本実施形態の構造化文書管理装置においては、検索に用いたキーワードと一致する語彙を含む部分文書が存在していた場合、検索用キーワードとの関連度が高い見出しを優先して表示させることとしたため、利用者は表示見出しから自分の求めている情報がその文書に含まれているかどうかを容易に判断することができるようになる。表示見出しを利用する場合、文章をわざわざ利用者が読んでその文章が求めている内容に近いか

どうかを判断する必要がなく、構造化文書のどの位置に欲しい情報が存在するかを迅速に把握可能となる。

【００３８】

なお、関連度が上位Ｎ個の見出しを選択するのではなく、関連度が所定値以上の見出しを見出し選択部３１が選択するようにしてもよい。また、関連度が、上位Ｎ個であり、かつ所定値以上の見出しを見出し選択部３１が選択するようにしてもよい。

【００３９】

また、表示見出しを表示部に表示させる際に、構造化文書中の表示順でソートしたり、上位のものから先に表示させたりといった構成は必須ではない。

【００４０】

また、見出しや本文を定義するタグの種類は本実施形態のものに限定されず、自由に定義することができる。

【００４１】

（第２の実施形態）

次に、本発明の構造化文書管理装置の第２の実施形態について図１５に基づき説明する。第２の実施形態においては、部分文書の見出しと本文中の語彙との関連度を構造化文書の登録時に予め計算して登録しておくのではなく、利用者が検索した際にキーワードと一致した語彙を含む部分文書のみ関連度を計算する点で異なっている。

【００４２】

図１５は、検索時に見出しを選択する処理の流れを示したフロー図である。図１５に示されるように、見出し選択部３１は、検索用キーワードと一致した語彙を含む構造化文書を取得する（ステップＳ４０１）。次いで、関連度計算部２６は、取得した構造化文書のうち、検索用キーワードと一致した語彙を含む部分文書を１つ選択し、その対応する見出しと検索用キーワードとの関連度を計算する（ステップＳ４０２）。この際の計算の方法については、第１の実施形態にて示した見出しと、本文中の語彙との間で関連度を計算する方法と同様である。

【００４３】

見出し選択部３１は、検索用キーワードと一致した語彙を含む全ての部分文書の見出しに対して関連度の計算が終了したか否かの判定を行い（ステップＳ４０３）、全て計算済みである場合（ステップＳ４０３：Ｙｅｓ）、検索用キーワードが一致する語彙を含む部分文書の見出しを関連度に基づき降順でソートする（ステップＳ４０４）。一方、検索用キーワードと一致した語彙を含む全ての部分文書に対する関連度が計算できていないと判定された場合（ステップＳ４０３：Ｎｏ）、ステップ４０２の処理を繰り返す。見出し選択部３１は、関連度の上位Ｎ個の見出しを選択し、構造化文書中のその見出しの出現順でソートする（ステップＳ４０５）。そして、見出し選択部３１は、全ての構造化文書（本実施形態では、文書ＩＤ　１、および文書ＩＤ　２の２つの文書）において、見出しの選択が終了したか否かを判定し（ステップＳ４０６）、終了した場合は（ステップＳ４０６：Ｙｅｓ）、ステップＳ３０５でソートして選択した見出しを表示見出しとして検索インタフェース部２９へと送り（ステップＳ４０７）、処理を終了する。全ての構造化文書での見出しの選択が終了していない場合は（ステップＳ４０６：Ｎｏ）。ステップＳ４０１からの処理を繰り返す。

【００４４】

本実施形態においては、事前に見出しと本文中の語彙との関連度を計算しておく必要がないため、計算結果を記憶していく記憶容量が確保できないときであっても、本発明を利用することができるようになる。また、関連度を計算する対象も、検索用のキーワードと一致した語彙を含む部分文書中の、当該検索用キーワードと見出し間における関連度のみでよいため、計算にかかる時間も抑制することができる。

【００４５】

なお、本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は

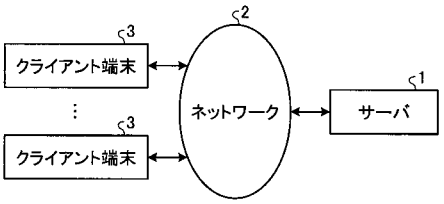
、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、請求の範囲に記載された発明とその均等の範囲に含まれる。

【符号の説明】

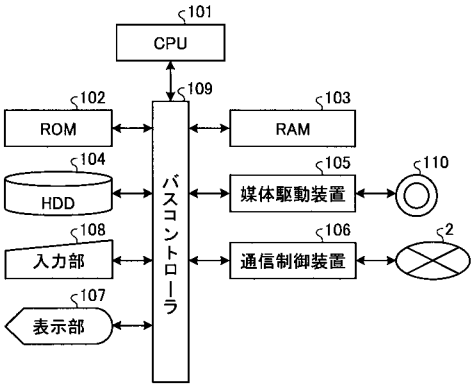
【００４６】

１　サーバ
２　ネットワーク
３　クライアント端末
１１　構造化文書登録部
１２　検索部
２１　構造化文書ＤＢ
２２　登録部
２３　検索部
２４　格納インタフェース部
２５　見出し抽出部
２６　関連度計算部
２７　構造化文書データ
２９　検索インタフェース部
３０　照合部
３１　見出し選択部
１０５　媒体駆動装置
１０６　通信制御装置
１０７　表示部
１０８　入力部
１０９　バスコントローラ
１１０　記憶媒体

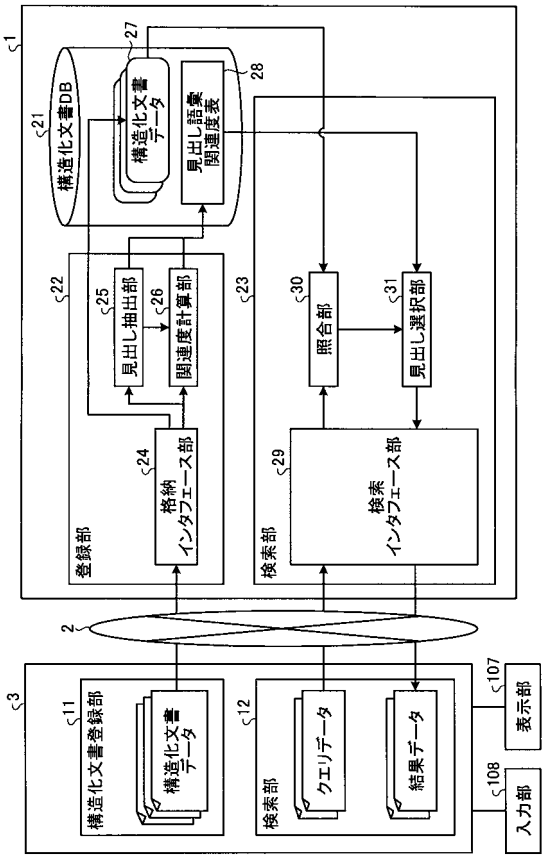
【図１】



【図２】



【図３】



【図４】

サンプル文書1

```
<doc @id="1">
<title @eid="101">パソコン取扱説明書</title>
<sec @eid="109">
<sectitle @eid="110">ネットワーク接続</sectitle>
<para @eid="111">...無線LANを利用する場合はルータに接続されているコンピュータ
で無線LANルータのセットアップを実行します。...</para>
</sec>
<sec @eid="102">
<sectitle eid="103">用語の定義</sectitle>
<para @eid="104">無線LAN</para>
<para @eid="105">無線LANとは無線通信を利用してデータの送受信を行うLANシステ
ムのことである。ワイヤレスLAN(Wireless LAN、WaveLAN[1])、もしくはそれを略して
WLANとも呼ばれる。</para>
</sec>
<sec @eid="106">
<sectitle @eid="107">起動と終了</sectitle>
<para @eid="108">
無線機能を無線LANオン/オフボタンで有効にしてからパソコンを起動します。...
</para>
</sec>
<sec @eid="112">
<sectitle @eid="113">注意事項</sectitle>
<para @eid="114">本製品はセキュリティ対策のためパスワード設定や、無線LANの暗
号化設定などを備えておりますが、完全なセキュリティ保護を保証するモノではありません
</para>
</sec>
<sec @eid="115">
<sectitle @eid="116">無線LANのトラブルシューティング</sectitle>
<para @eid="117">コンピュータがルータまたはアクセスポイントから離れすぎている
...</para>
</sec>
</doc>
```

【図５】

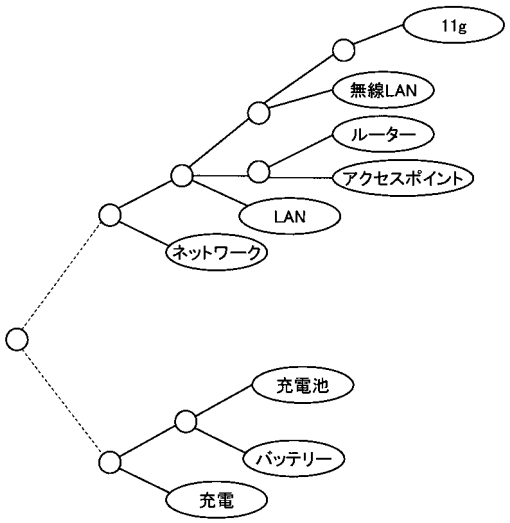
サンプル文書2

```
<doc @id="2">
<title @eid="201">携帯端末説明書</title>
<sec @eid="202">
<sectitle @eid="203">バッテリー</sectitle>
<para @eid="204">
無線LAN機能を使っているときは充電できません。
</para>
</sec>
<sec @eid="205">
<sectitle @eid="206">ネットワーク設定
</sectitle>
<para @eid="207">
無線LANは11gに対応しています。
</para>
<sec @eid="208">
<sectitle @eid="209">アクセスポイントの設定</sectitle>
<para @eid="210">
...利用する無線LANアクセスポイントを「11gを利用する」に設定してく
ださい.....
</para>
</sec>
</sec>
<sec @eid="211">
<sectitle @eid="212">アップデート</sectitle>
<para @eid="213">
無線LANアクセスポイントを経由して接続します。
</para>
</sec>
</doc>
```

【図６】

文書番号	部分文書要素番号	見出し要素番号
1	109	110
1	102	103
1	106	107
1	112	113
1	115	116
2	202	203
2	205	206
2	208	206
2	208	209
2	211	212

【図７】



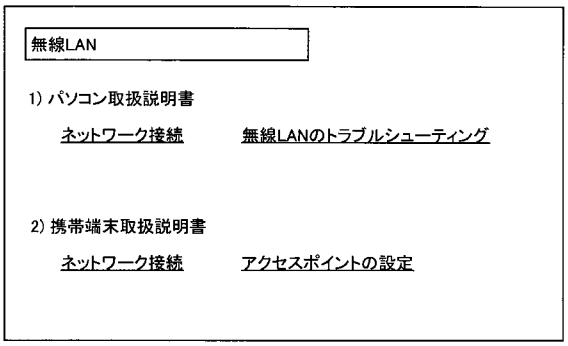
【図８】

	ネットワーク	LAN	アクセスポイント	ルーター	無線LAN	11g	充電	バッテリー	充電池
ネットワーク	1.000	0.333	0.250	0.250	0.250	0.200	0.000	0.000	0.000
LAN	–	1.000	0.333	0.333	0.333	0.250	0.000	0.000	0.000
アクセスポイント	–	–	1.000	0.500	0.250	0.200	0.000	0.000	0.000
ルーター	–	–	–	1.000	0.250	0.200	0.000	0.000	0.000
無線LAN	–	–	–	–	1.000	0.333	0.000	0.000	0.000
11g	–	–	–	–	–	1.000	0.000	0.000	0.000
充電	–	–	–	–	–	–	1.000	0.333	0.333
バッテリー	–	–	–	–	–	–	–	1.000	0.500
充電池	–	–	–	–	–	–	–	–	1.000

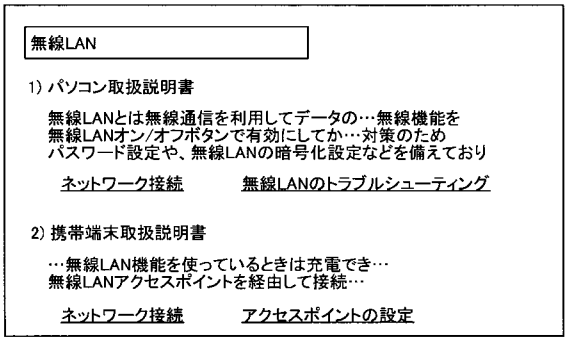
【図９】

文集番号	部分文書要素番号	見出し要素番号	見出し-語彙関連度
1	109	110	ネットワーク:1.0,無線LAN:0.25,LAN:0.333,ルーター:0.25
1	102	103	
1	106	107	
1	112	113	
1	115	116	LAN:1.0,無線LAN:1.0,ルーター:0.333,アクセスポイント:0.333
2	202	203	充電:0.5
2	205	206	ネットワーク:1.0,無線LAN:0.25,LAN:0.333,アクセ スポイント:0.25,11g:0.20
2	208	206	無線LAN:0.125,LAN:0.167,アクセスポイント:0.125,11g:0.10
2	208	209	アクセスポイント:1.0,無線LAN:0.25,LAN:0.33,11g:0.20
2	211	212	

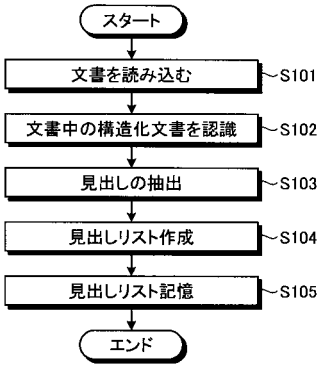
【図１０】



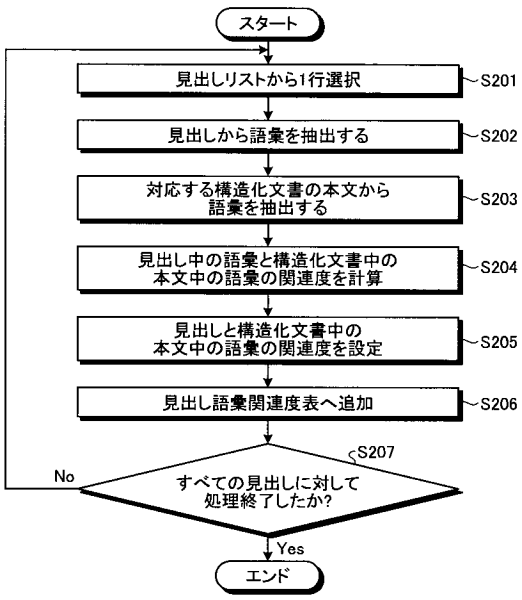
【図１１】



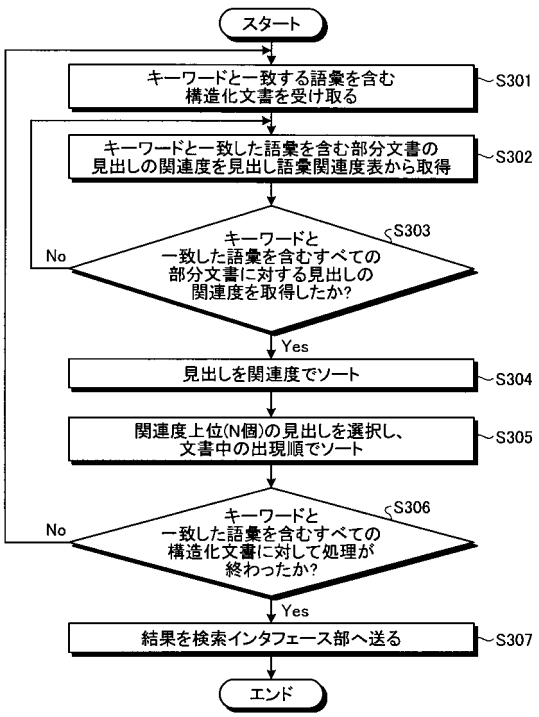
【図１２】



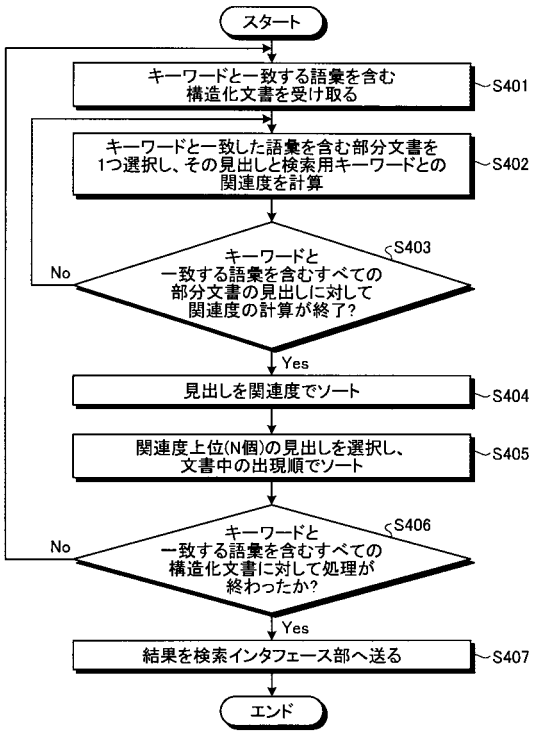
【図１３】



【図１４】



【図１５】


スタート
キーワードと一致する語彙を含む
構造化文書を受け取る
S401
キーワードと一致した語彙を含む部分文書を
1つ選択し、その見出しと検索用キーワードとの
関連度を計算
S402
キーワードと
一致する語彙を含むすべての
部分文書の見出しに対して
関連度の計算が終了?
S403
No
Yes
見出しを関連度でソート
S404
関連度上位(N個)の見出しを選択し、
文書中の出現順でソート
S405
キーワードと
一致する語彙を含むすべての
構造化文書に対して処理が
終わったか?
S406
No
Yes
結果を検索インタフェース部へ送る
S407
エンド

フロントページの続き

(72)発明者　仲野　亘
東京都港区芝浦一丁目１番１号　東芝ソリューション株式会社内

審査官　鈴木　和樹

(56)参考文献　特開２００４－１２６７７０（ＪＰ，Ａ）
特開２００６－１９５６６７（ＪＰ，Ａ）
特開２００８－１４６２０９（ＪＰ，Ａ）
特開２００３－２４２１７５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０