

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2013年9月19日(19.09.2013)



(10) 国際公開番号

WO 2013/136545 A1

- (51) 国際特許分類:
G06F 17/30 (2006.01)
- (21) 国際出願番号: PCT/JP2012/068505
- (22) 国際出願日: 2012年7月20日(20.07.2012)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2012-057240 2012年3月14日(14.03.2012) JP
- (71) 出願人(米国を除く全ての指定国について): 株式会社東芝(Kabushiki Kaisha Toshiba) [JP/JP]; 〒1058001 東京都港区芝浦一丁目1番1号 Tokyo (JP). 東芝ソリューション株式会社(TOSHIBA SOLUTIONS CORPORATION) [JP/JP]; 〒1056691 東京都港区芝浦一丁目1番1号 Tokyo (JP).
- (72) 発明者: および
- (75) 発明者/出願人(米国についてのみ): 國分 智晴(KOKUBU, Tomoharu) [JP/JP]; 〒1058001 東京都港区芝浦一丁目1番1号 株式会社東芝 知的財産部内 Tokyo (JP). 真鍋 俊彦(MANABE, Toshi-

hiko) [JP/JP]; 〒1058001 東京都港区芝浦一丁目1番1号 株式会社東芝 知的財産部内 Tokyo (JP). 仲野 亘(NAKANO, Wataru) [JP/JP]; 〒1058001 東京都港区芝浦一丁目1番1号 株式会社東芝 知的財産部内 Tokyo (JP).

(74) 代理人: 酒井 宏明(SAKAI, Hiroaki); 〒1006020 東京都千代田区霞が関三丁目2番5号 霞が関ビルディング 酒井国際特許事務所 Tokyo (JP).

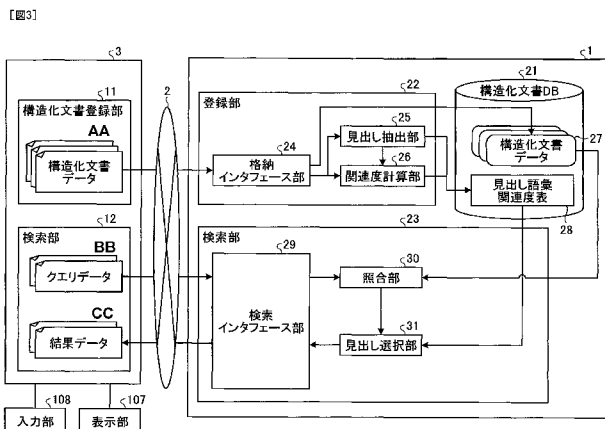
(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW,

[続葉有]

(54) Title: STRUCTURED DOCUMENT MANAGEMENT DEVICE, STRUCTURED DOCUMENT SEARCH METHOD

(54) 発明の名称: 構造化文書管理装置、構造化文書検索方法



- 11 STRUCTURED DOCUMENT REGISTRATION UNIT
12, 23 SEARCH UNIT
21 STRUCTURED DOCUMENT DATABASE
22 REGISTRATION UNIT
24 STORAGE INTERFACE UNIT
25 HEADING EXTRACTION UNIT
26 RELEVANCE COMPUTATION UNIT
27 STRUCTURED DOCUMENT DATA
28 HEADING WORD RELEVANCE TABLE
29 SEARCH INTERFACE UNIT
30 COMPARISON UNIT
31 HEADING SELECTION UNIT
107 DISPLAY UNIT
108 INPUT UNIT
AA STRUCTURED DOCUMENT DATA
BB QUERY DATA
CC RESULT DATA

(57) Abstract: A structured document management device according to an embodiment of the present invention is provided with a document storage unit, a heading extraction unit, a relevance computation unit, a document search unit, a heading selection unit and a heading display unit. The document storage unit stores multiple structured documents. The heading extraction unit extracts the headings of the structured documents, and creates a heading list that contains the extracted headings. The relevance computation unit computes a conceptual relevance between each of multiple words and multiple headings, said words being in the structured documents, said headings corresponding to the structured documents. The document search unit searches for a structured document that contains a word that matches a search keyword. The heading selection unit selects a heading, giving greater precedence to a heading with a large relevance to the word that matches the search keyword, than to a heading with a small relevance to said word. A display control unit displays the heading selected by the heading selection unit, said heading being displayed as a display heading on a display unit.

(57) 要約:

[続葉有]

WO 2013/136545 A1



MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラ
シア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッ
パ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK,
MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR),

OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML,
MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告 (条約第 21 条(3))

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。

明 細 書

発明の名称：構造化文書管理装置、構造化文書検索方法

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に近いものであるかを確認する必要がある、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：E l e m e n t）と呼び、要素はタグ（t a g）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、< d o c >というタグで囲まれたルート要素が存在する。< d o c >要素は、そのドキュメントの文書IDとしてi d = 1が割り当てられている。< d o c >要素は、< t i t l e >要素を持ち、< t i t l e >要素はその構造化文書の見出しを示している。また、< d o c >要素は、5つの< s e c >要素を有している。< s e c >要素は、< d o c >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。< s e c >というタグで囲まれた中には、< s e c t i t l e >要素と、< p a r a >要素とが含まれている。< s e c t i t l e >は、その部分文書の見出しを示すタグである。また、< p a r a >は、その部分文書の説明文を示すタグである。この< s e c t i t l e >、および< p a r a >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@ e i d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@ e i d = 2 0 8にて定義された部分文書が、@ e i d = 2 0 5にて定義された部分文書に含まれており、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< s e c t i t l e >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

I D 109、102、106、112および115で示される部分文書に対して、@e i d=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I D 2の構造化文書においては、要素I D 202、205、および211で示される部分文書に対して、@e i d=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I D 208で示される部分文書に対しては、@e i d=206、および209の2つの見出しが抽出される。文書I D 2の構造化文書においては、要素I D 208で示される部分文書の見出しとして、自身の<s e c>タグで囲われた@e i d=209の見出しだけではなく、親階層における@e i d=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<s e c>要素内の子階層にて<s e c>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においては、見出し@e i d=206を含む部分文書@e i d=205にとって、部分文書@e i d=208が従属文書に相当し、一方、部分文書@e i d=208にとって、部分文書@e i d=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D B 21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l e n g t h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l e n g t h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l e n g t hを用いて表され、 $1 / (\text{距離} \text{ l e n g t h} + 1)$ 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* = 115 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 DB 21 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、 $1 / (\text{距離 } length + 1)$ 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ができる。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、@e i d = 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、@e i d = 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か・・・」が、@e i d = 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 1 2 ～図 1 4 を用いて説明する。図 1 2 は、構造化文書の登録時の処理の流れを示している。図 1 2 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 1 1 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 2 4 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 S 1 0 1）。次いで、見出し抽出部 2 5 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 S 1 0 2）。そして、見出し抽出部 2 5 は、抽出した見出しから見出しリストを作成し（ステップ S 1 0 3）、構造化文書 DB 2 1 に記憶する（ステップ S 1 0 4）。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 1 3 から説明する。図 1 3 に示されるように、関連度計算部 2 6 は、構造化文書 DB 2 1 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 S 2 0 1）。次いで、関連度計算部 2 6 は、選択した見出しから語彙を抽出する（ステップ S 2 0 2）。次いで、関連度計算部 2 6 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 S 2 0 3）。関連度計算部 2 6 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 S 2 0 4）。次いで、関連度計算部 2 6 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 S 2 0 5）。そして、関連度計算部 2 6 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 S 2 0 6）。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 S 2 0 7）、処理が完了した場合（ステップ S 2 0 7 : 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 S 2 0 7 : 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 3 1 によって見出しが選択される処理の流れを、図 1 4 を用いて説明する。見出し選択部 3 1 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ID1、および文書ID2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

- 2 1 構造化文書ＤＢ
- 2 2 登録部
- 2 3 検索部
- 2 4 格納インタフェース部
- 2 5 見出し抽出部
- 2 6 関連度計算部
- 2 7 構造化文書データ
- 2 9 検索インタフェース部
- 3 0 照合部
- 3 1 見出し選択部
- 1 0 5 媒体駆動装置
- 1 0 6 通信制御装置
- 1 0 7 表示部
- 1 0 8 入力部
- 1 0 9 バスコントローラ
- 1 1 0 記憶媒体

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、

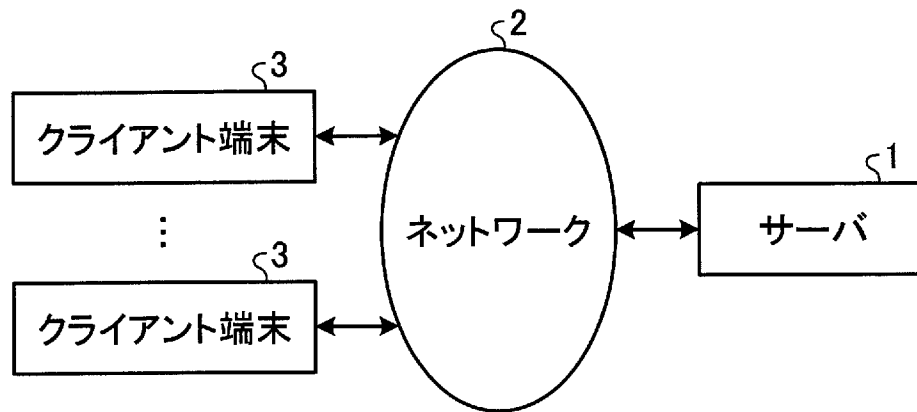
検索用キーワードと一致する前記語彙を含む前記部分文書を検索す
る文書検索ステップと、

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、

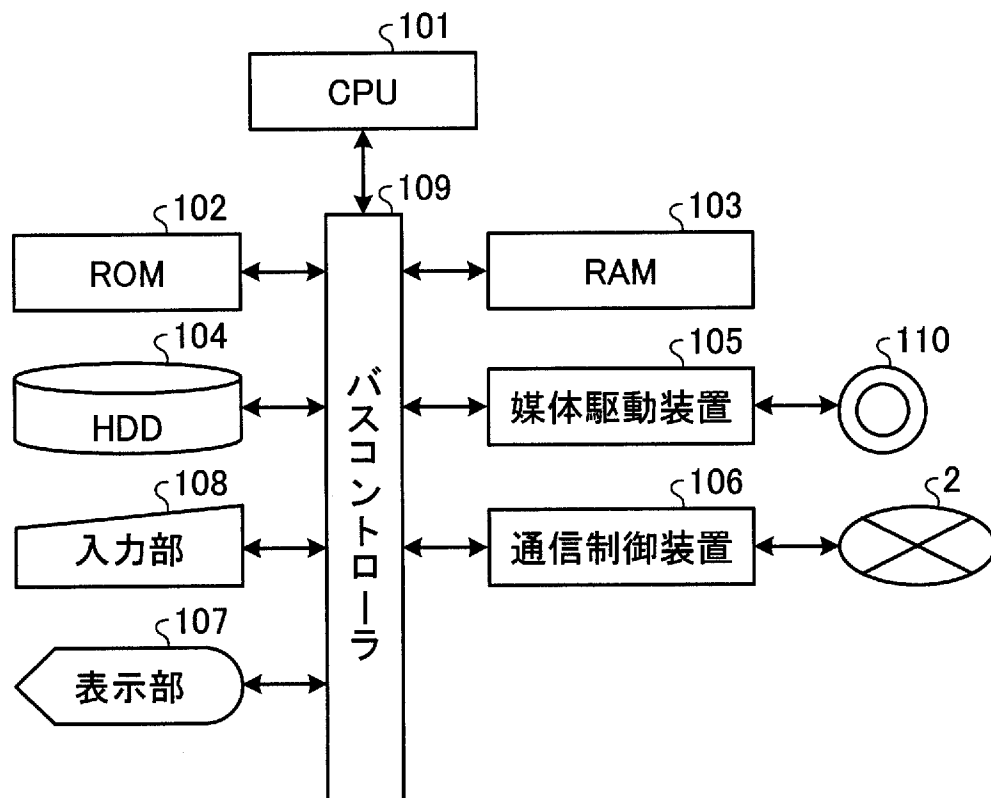
前記検索用キーワードとの前記関連度が大きい前記見出しを前記関
連度が小さい前記見出しより優先して選択する見出し選択ステップと
、

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。

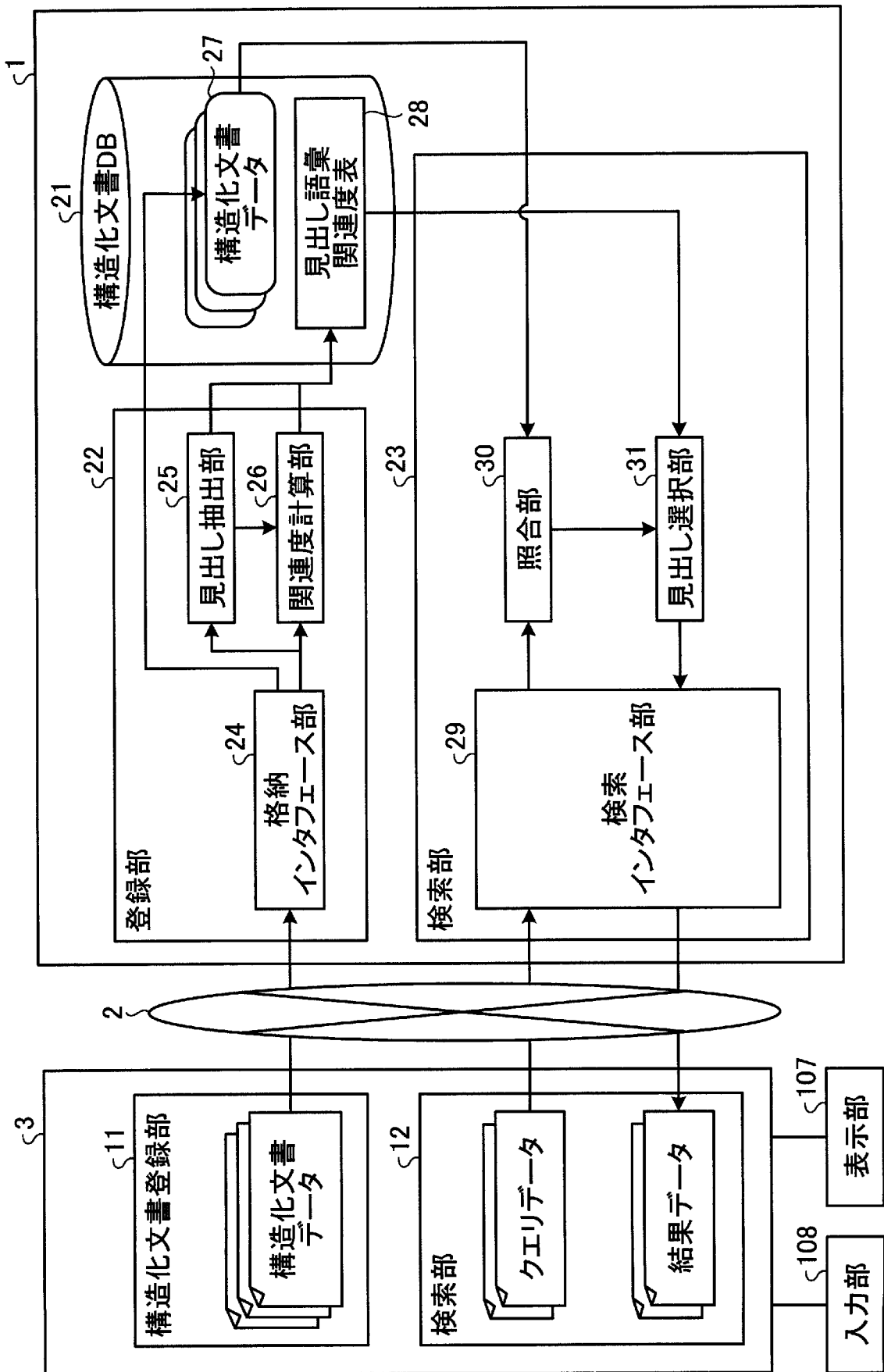
[図1]



[図2]



[図3]



[図4]

サンプル文書1

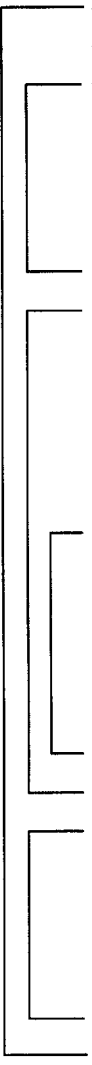
```

<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">
      無線機能を実線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

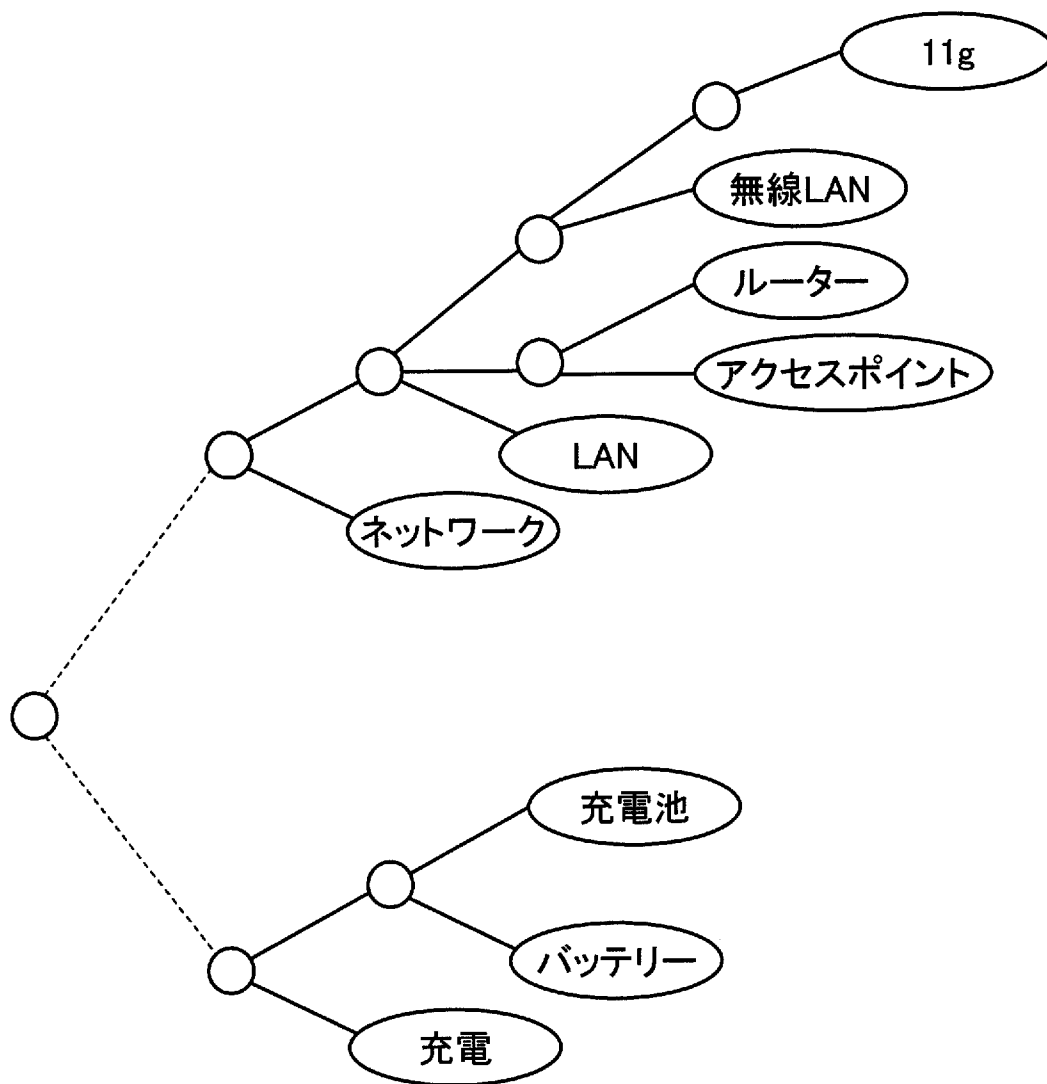


```
<doc @id="2">  
  <title @eid="201">携帯端末説明書</title>  
  <sec @eid="202">  
    <sectitle @eid="203">バッテリー</sectitle>  
    <para @eid="204">  
      無線LAN機能を使っているときは充電できません。  
    </para>  
  </sec>  
  <sec @eid="205">  
    <sectitle @eid="206">ネットワーク設定</sectitle>  
    <para @eid="207">  
      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]

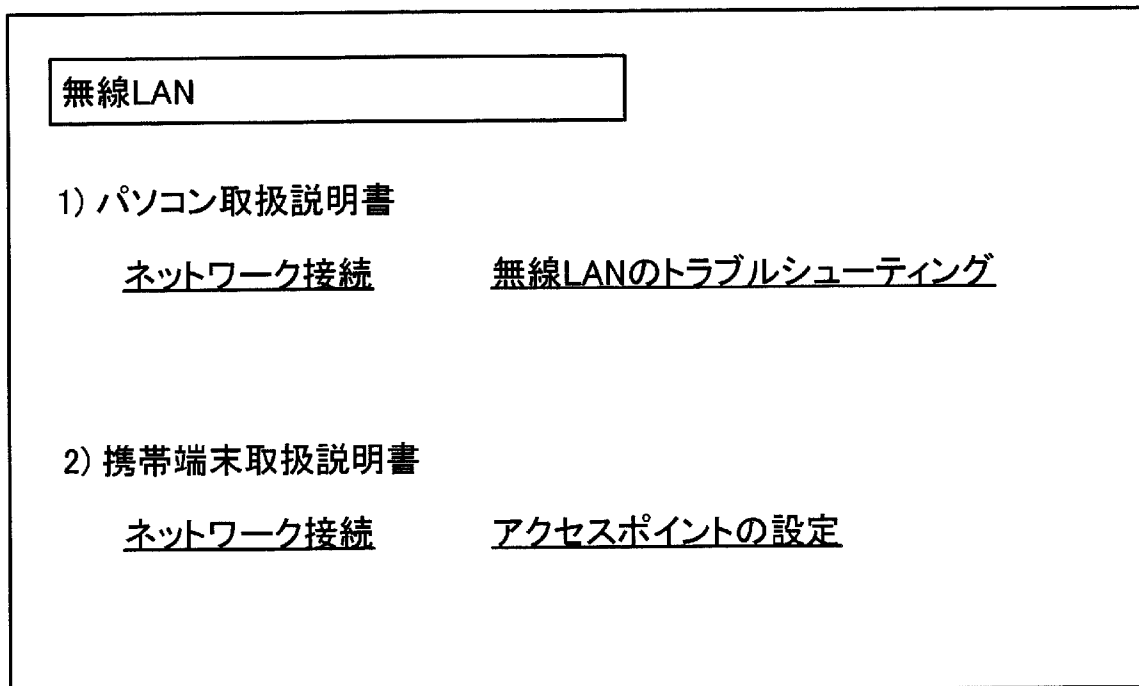


[illegible]

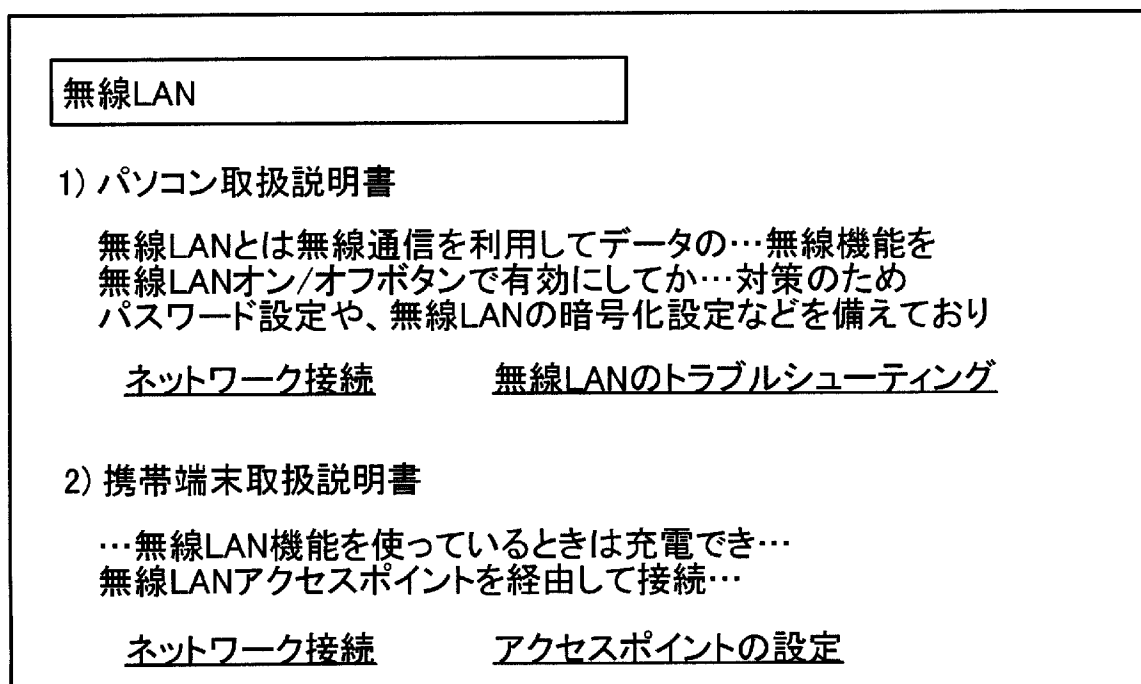
[図9]

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	

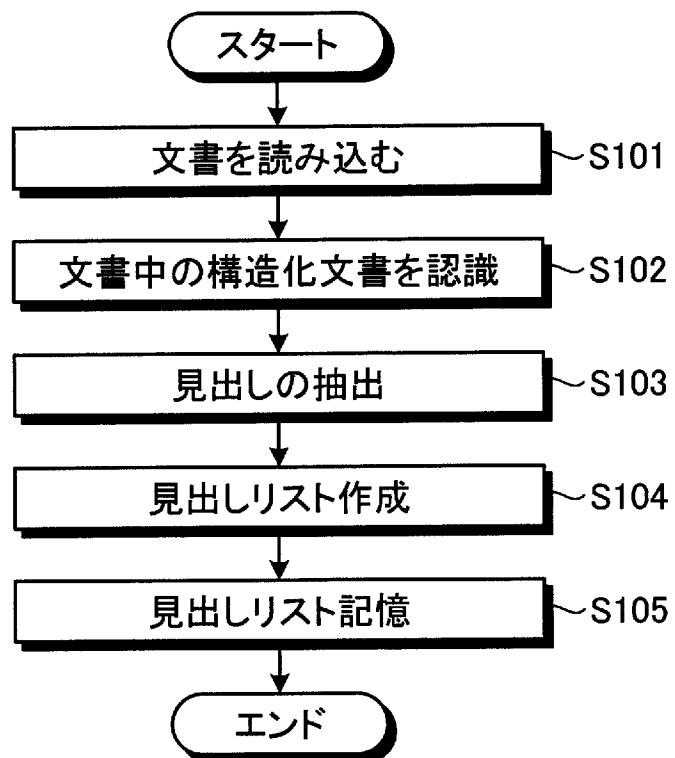
[図10]



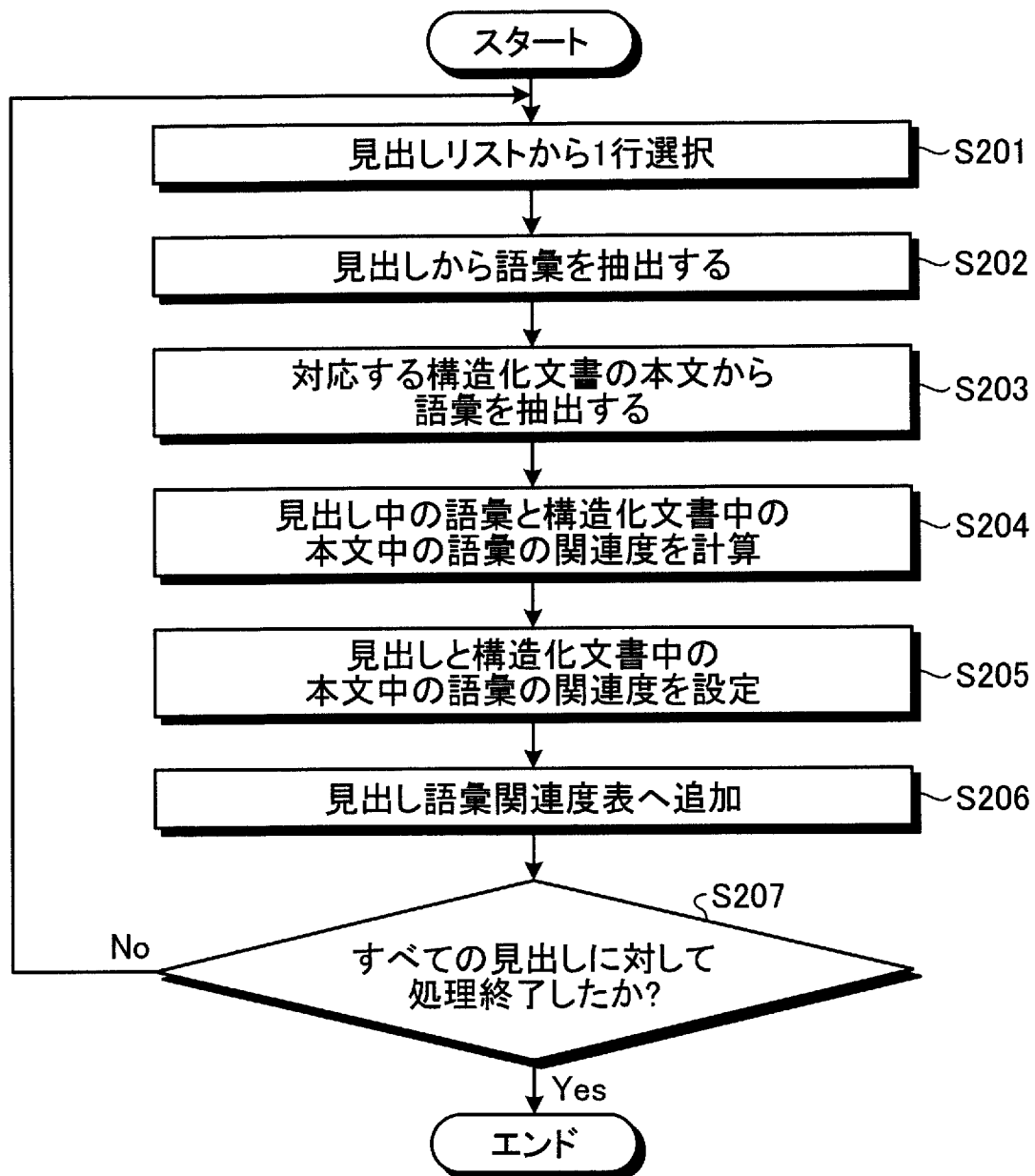
[図11]



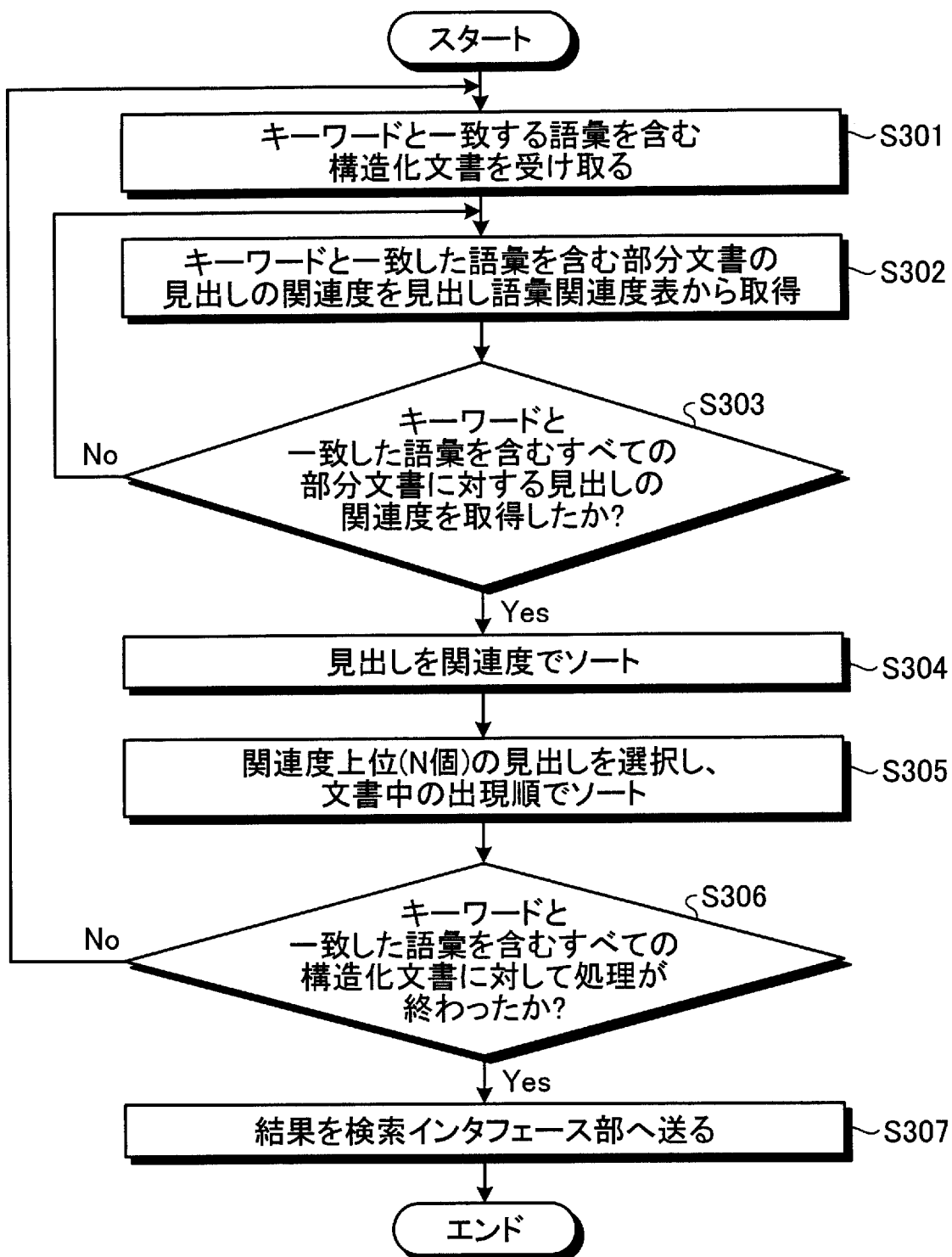
[図12]



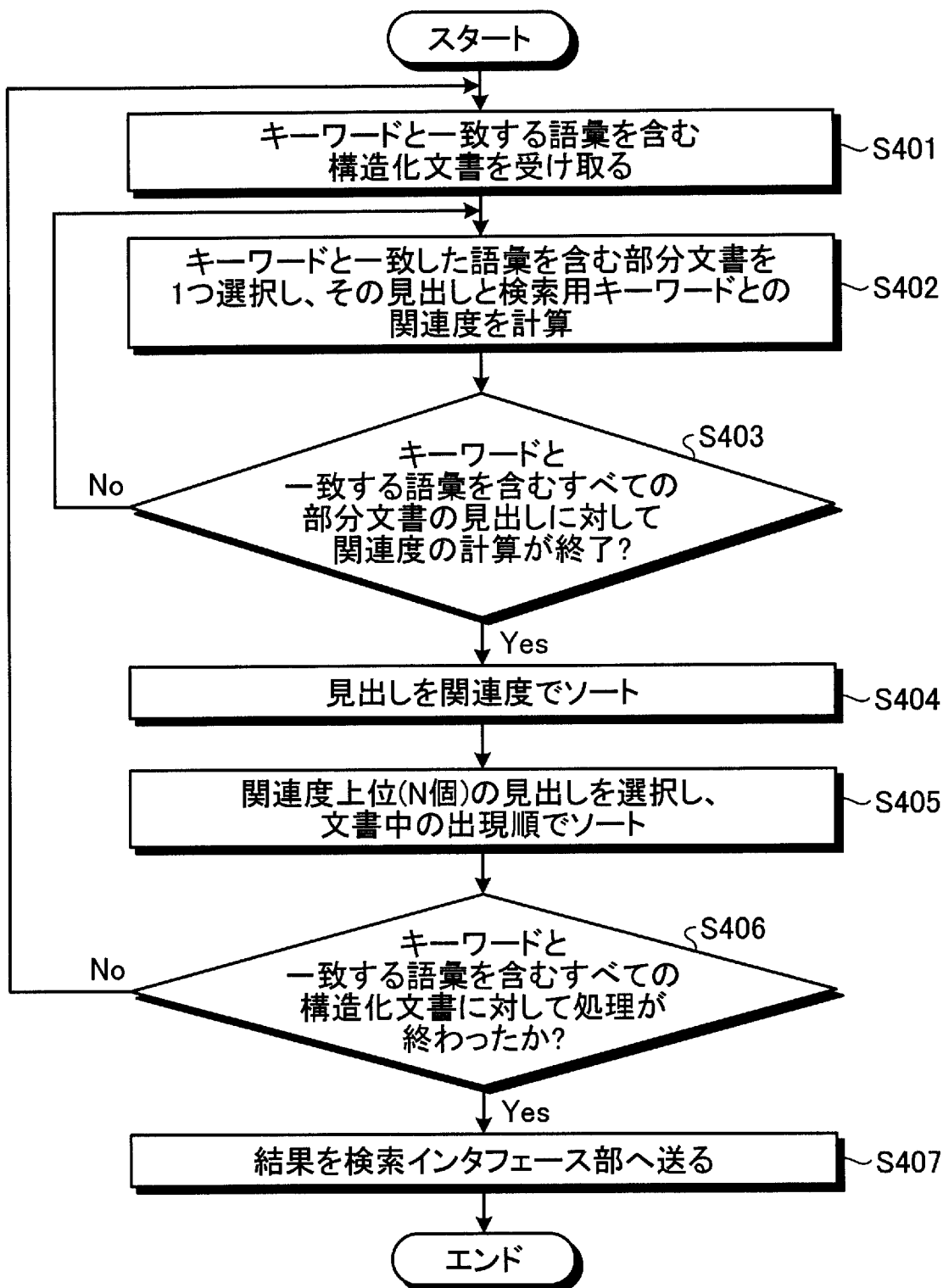
[図13]



[図14]



[図15]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/068505

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2012

Kokai Jitsuyo Shinan Koho 1971-2012 Toroku Jitsuyo Shinan Koho 1994-2012

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2004-126770 A (Toshiba Corp.), 22 April 2004 (22.04.2004), entire text; all drawings (Family: none)	1-10
A	JP 2006-195667 A (Toshiba Corp.), 27 July 2006 (27.07.2006), entire text; all drawings (Family: none)	1-10
A	JP 2008-146209 A (Justsystem Corp.), 26 June 2008 (26.06.2008), entire text; all drawings (Family: none)	1-10

☒ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
21 August, 2012 (21.08.12)Date of mailing of the international search report
28 August, 2012 (28.08.12)Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/068505

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2003-242175 A (Ricoh Co., Ltd.), 29 August 2003 (29.08.2003), entire text; all drawings (Family: none)	1-10

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06F17/30(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（I P C））

Int.Cl. G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 1 2 年
日本国実用新案登録公報	1 9 9 6 - 2 0 1 2 年
日本国登録実用新案公報	1 9 9 4 - 2 0 1 2 年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2004-126770 A（株式会社東芝）2004.04.22, 全文, 全図（ファミリーなし）	1-10
A	JP 2006-195667 A（株式会社東芝）2006.07.27, 全文, 全図（ファミリーなし）	1-10
A	JP 2008-146209 A（株式会社ジャストシステム）2008.06.26, 全文, 全図（ファミリーなし）	1-10

☒ C 欄の続きにも文献が列挙されている。☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

2 1 . 0 8 . 2 0 1 2

国際調査報告の発送日

2 8 . 0 8 . 2 0 1 2

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

鈴木 和樹

5 M

3 2 5 2

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 9 9

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2003-242175 A (株式会社リコー) 2003.08.29, 全文, 全図 (ファミリーなし)	1-10