(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6834774号
(P6834774)
(45)発行日　令和3年2月24日(2021.2.24)　(24)登録日　令和3年2月8日(2021.2.8)

(51)Int.Cl.　F I
G06F 16/29 (2019.01)　G06F 16/29

請求項の数 1　(全 11 頁)

(21)出願番号　特願2017-101200(P2017-101200)
(22)出願日　平成29年5月22日(2017.5.22)
(65)公開番号　特開2018-195272(P2018-195272A)
(43)公開日　平成30年12月6日(2018.12.6)
審査請求日　令和1年8月23日(2019.8.23)

(73)特許権者　000003207
トヨタ自動車株式会社
愛知県豊田市トヨタ町１番地
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(72)発明者　鈴木　功一
愛知県豊田市トヨタ町１番地　トヨタ自動車株式会社内

審査官　▲はま▼中　信行

最終頁に続く

(54)【発明の名称】情報抽出装置

(57)【特許請求の範囲】
【請求項１】
文書データを取得する文書データ取得部と、
前記文書データから住所の候補文字列を抽出する候補文字列抽出部と、
前記候補文字列に対してジオコーディングを行うことにより、位置情報の取得を試みる位置情報取得部と、
前記位置情報取得部による前記位置情報の取得結果に応じて、前記住所とする文字列を決定する住所文字列決定部と
を備え、
前記位置情報取得部による前記位置情報の取得に成功した場合、前記位置情報取得部による前記位置情報の再取得に失敗するまで、
前記候補文字列抽出部が、前記候補文字列を後方に延長して、前記文書データから前記候補文字列を再抽出し、
前記位置情報取得部が、再抽出された前記候補文字列に対して前記ジオコーディングを行うことにより、前記位置情報の再取得を試み、
前記位置情報取得部による前記位置情報の再取得に失敗した場合、
前記住所文字列決定部が、前記位置情報の再取得に失敗する直前に、前記位置情報の取得に成功した前記候補文字列を、前記住所とする文字列に決定する
情報抽出装置。
【発明の詳細な説明】

【技術分野】
【０００１】
本発明は、情報抽出装置に関する。
【背景技術】
【０００２】
従来より、ＷＥＢクローリングに代表されるように、ＷＥＢページ等の文書データから施設情報（例えば、ＰＯＩ（Point of Interest）名称、住所、電話番号、郵便番号、キーワード等）を抽出し、当該施設情報をデータベースに自動的に蓄積できるようにした技術が知られている。
【０００３】
下記特許文献１には、インターネットに接続されたサーバから文書データを取得し、当該文書データに含まれている住所文字列に、「都道府県」、「市町村」、「町域」、「街区」、「号」のいずれまでが含まれているかによって、当該住所文字列の詳細度（１～５）を決定する技術が開示されている。例えば、住所文字列に「号」までが含まれている場合には、最も高い詳細度「５」が決定されるといった具合である。また、下記特許文献１には、所定の閾値以上の詳細度を有する住所文字列を含む文書データを、データベースに組み込むようにした技術が開示されている。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２０１２－２５６３５６号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、従来の技術では、文書データから抽出する住所文字列の終端を高精度に特定することができない。このため、従来の技術では、住所の途中までしか住所文字列として取得しない場合や、住所に続く住所以外の語句までも住所文字列として取得してしまう場合がある。また、従来の技術では、文書データから複数のＰＯＩ名称の候補文字列が抽出された場合、いずれの候補文字列が実際のＰＯＩ名称であるかを判断することが困難であるため、実際にはＰＯＩ名称ではない候補文字列を、ＰＯＩ名称として誤って抽出してしまう場合がある。
【０００６】
このようなことから、従来、文書データからの施設情報の抽出精度を高めることが困難であった。
【０００７】
本発明は、上述した従来技術の課題を解決するため、文書データからの施設情報の抽出処理を適切に行い、文書データからの施設情報の抽出精度を高めることを目的とする。
【課題を解決するための手段】
【０００８】
本発明の実施形態の情報抽出装置は、文書データを取得する文書データ取得部と、前記文書データから住所の候補文字列を抽出する候補文字列抽出部と、前記候補文字列に対してジオコーディングを行うことにより、位置情報の取得を試みる位置情報取得部と、前記位置情報取得部による前記位置情報の取得結果に応じて、前記住所とする文字列を決定する住所文字列決定部とを備え、前記位置情報取得部による前記位置情報の取得に成功した場合、前記位置情報取得部による前記位置情報の再取得に失敗するまで、前記候補文字列抽出部が、前記候補文字列を後方に延長して、前記文書データから前記候補文字列を再抽出し、前記位置情報取得部が、再抽出された前記候補文字列に対して前記ジオコーディングを行うことにより、前記位置情報の再取得を試み、前記位置情報取得部による前記位置情報の再取得に失敗した場合、前記住所文字列決定部が、前記位置情報の再取得に失敗する直前に、前記位置情報の取得に成功した前記候補文字列を、前記住所とする文字列に決

定する。
【発明の効果】
【０００９】
文書データからの施設情報の抽出処理を適切に行い、文書データからの施設情報の抽出精度を高めることができる。
【図面の簡単な説明】
【００１０】
【図１】実施形態に係る情報抽出装置の機能構成を示す図である。
【図２】実施形態に係る情報抽出装置による処理の手順を示すフローチャートである。
【図３】実施形態に係るアノテータ処理部によるアノテータ処理の手順を示すフローチャートである。
【図４】実施形態に係るアノテータ処理部による住所取得処理の手順を示すフローチャートである。
【図５】実施形態に係るアノテータ処理部によるナイーブベイズ推定値取得処理の手順を示すフローチャートである。
【発明を実施するための形態】
【００１１】
以下、図面を参照して、本発明の実施形態の情報抽出装置について説明する。
【００１２】
（情報抽出装置１００の機能構成）
図１は、実施形態に係る情報抽出装置１００の機能構成を示す図である。図１に示す情報抽出装置１００は、ＷＥＢクローリングを行うことによって、インターネット上のＷＥＢページ１１０（「文書データ」の一例）からＰＯＩデータ（「施設情報」の一例）を抽出し、当該ＰＯＩデータを施設情報ＤＢ１２０へ登録および更新することが可能な装置である。
【００１３】
図１に示すように、情報抽出装置１００は、クローラ処理部１０１、パーサ処理部１０２、スクレーパ処理部１０３、アノテータ処理部１０４、およびデータリンカ処理部１０５を備える。
【００１４】
クローラ処理部１０１は、ＷＥＢクローリングを行うことにより、インターネット上のＷＥＢサイトからＷＥＢページ１１０を取得し、当該ＷＥＢページ１１０をメモリに格納する。すなわち、クローラ処理部１０１は、「文書データ取得部」としての機能を有する。
【００１５】
パーサ処理部１０２は、クローラ処理部１０１によって取得されたＷＥＢページ１１０に対してパーサ処理を行うことにより、当該ＷＥＢページ１１０から、特定のキーワードによる、特定のＨＴＭＬ（HyperText Markup Language）ファイルの選択を行う。
【００１６】
スクレーパ処理部１０３は、パーサ処理部１０２によって選択されたＨＴＭＬファイルに対してスクレーパ処理を行うことにより、当該ＨＴＭＬファイルから不要部分を削除し、残りの部分を構造体として出力する。
【００１７】
アノテータ処理部１０４は、スクレーパ処理部１０３によって出力された構造体に対してアノテータ処理を行うことにより、当該構造体に含まれるテキストデータを解析し、当該テキストデータから、予め定義された属性値（ＰＯＩ名称、住所、電話番号、郵便番号、キーワード等）を取得する。
【００１８】
特に、アノテータ処理部１０４は、「候補文字列抽出部」、「位置情報取得部」、および「住所文字列決定部」としての機能を有している。すなわち、アノテータ処理部１０４

は、ＷＥＢページ１１０から住所の候補文字列を抽出し、当該候補文字列に対してジオコーディングを行うことにより位置情報の取得を試み、位置情報の取得結果に応じて、住所とする文字列を決定することができる。これにより、アノテータ処理部１０４は、ＷＥＢページ１１０から抽出する住所文字列の終端を高精度に特定することができる。この点については、図４を用いて詳細に説明する。

【００１９】

さらに、アノテータ処理部１０４は、ＷＥＢページ１１０からＰＯＩ名称の候補文字列を抽出し、各候補文字列について、ナイーブベイズ推定値を取得することができる。そして、アノテータ処理部１０４は、ＰＯＩ名称の候補文字列が、所定文字列を含む、または、強調されている場合、その候補文字列のナイーブベイズ推定値を高めることができる。これにより、アノテータ処理部１０４は、実際にＰＯＩ名称である可能性が最も高い文字列を、ＰＯＩ名称として抽出することができる。この点については、図５を用いて詳細に説明する。

【００２０】

データリンカ処理部１０５は、データリンカ処理を行うことにより、アノテータ処理部１０４により取得された各属性値をＰＯＩデータとして、当該ＰＯＩデータに対して、ジオコーディング、ＰＯＩマスタとの名寄せ等を行い、当該ＰＯＩデータを施設情報ＤＢ１２０に対して登録または更新する。

【００２１】

なお、情報抽出装置１００の各機能は、例えば、各種情報処理装置（例えば、サーバ、パーソナルコンピュータ等）において、各種記憶装置（例えば、ＲＯＭ（Read Only Memory）、フラッシュメモリ等）に記憶されたプログラムを、コンピュータ（例えば、ＣＰＵ（Central Processing Unit）等）が実行することにより、実現される。

【００２２】

（情報抽出装置１００による処理の手順）

図２は、実施形態に係る情報抽出装置１００による処理の手順を示すフローチャートである。図２の処理は、例えば、情報抽出装置１００にスケジュール設定されることにより、情報抽出装置１００によって定期的（例えば、１日毎）に実行される。

【００２３】

まず、クローラ処理部１０１が、ＷＥＢクローリングを行うことにより、インターネット上のＷＥＢサイトからＷＥＢページ１１０を取得し、当該ＷＥＢページ１１０をメモリに格納する（ステップＳ２０１）。

【００２４】

次に、パーサ処理部１０２が、ステップＳ２０１で取得されたＷＥＢページ１１０に対してパーサ処理を行うことにより、当該ＷＥＢページ１１０から、特定のキーワードによる、特定のＨＴＭＬファイルの選択を行う（ステップＳ２０２）。

【００２５】

次に、スクレーパ処理部１０３が、ステップＳ２０２で選択されたＨＴＭＬファイルに対してスクレーパ処理を行うことにより、当該ＨＴＭＬファイルから不要部分を削除し、残りの部分（すなわち、ＰＯＩデータの抽出対象とするテキストデータ。例えば、口コミ情報等）を構造体として出力する（ステップＳ２０３）。

【００２６】

次に、アノテータ処理部１０４が、ステップＳ２０３で出力された構造体に対してアノテータ処理を行うことにより、当該構造体に含まれるテキストデータを解析し、当該テキストデータから、予め定義された属性値（ＰＯＩ名称、住所、電話番号、郵便番号、キーワード等）を取得する（ステップＳ２０４）。なお、アノテータ処理部１０４によるアノテータ処理の詳細については、図３を用いて後述する。

【００２７】

次に、データリンカ処理部１０５が、データリンカ処理を行うことにより、ステップＳ２０４で取得された各属性値をＰＯＩデータとして、当該ＰＯＩデータに対して、ジオコ

ーディング、ＰＯＩマスタとの名寄せ等を行い、当該ＰＯＩデータを施設情報ＤＢに対して登録または更新する（ステップＳ２０５）。そして、情報抽出装置１００は、図２に示す一連の処理を終了する。

【００２８】

（アノテータ処理部１０４によるアノテータ処理の手順）

図３は、実施形態に係るアノテータ処理部１０４によるアノテータ処理の手順を示すフローチャートである。図３は、図２にフローチャートにおけるステップＳ２０４のアノテータ処理を詳細に説明するものである。図３の処理には、スクレーパ処理部１０３から出力された構造体（ＨＴＭＬ構造）が入力される。

【００２９】

まず、アノテータ処理部１０４は、構造体に含まれるテキストデータに対して、形態素解析を行い、当該テキストデータを、複数の形態素（単語、品詞等）単位に分割する（ステップＳ３０１）。

【００３０】

次に、アノテータ処理部１０４は、ステップＳ３０１で複数の形態素に分割されたテキストデータの中から、郵便番号および電話番号を取得する（ステップＳ３０２）。例えば、アノテータ処理部１０４は、「‐」および数字からなる所定のフォーマットの文字列（例えば、「ｘｘｘ－ｘｘｘｘ」）や、直前に「郵便番号」、「〒」等が存在する文字列を、郵便番号として取得する。また、例えば、アノテータ処理部１０４は、「‐」および数字からなる所定のフォーマットの文字列（例えば、「ｘｘｘ－ｘｘｘｘ－ｘｘｘｘ」）や、直前に「電話番号」、「ＴＥＬ」等が存在する文字列を、電話番号として取得する。

【００３１】

次に、アノテータ処理部１０４は、住所取得処理を実行することにより、ステップＳ３０１で複数の形態素に分割されたテキストデータの中から、住所を取得する（ステップＳ３０３）。住所取得処理の詳細については、図４を用いて後述する。

【００３２】

次に、アノテータ処理部１０４は、ステップＳ３０１で複数の形態素に分割されたテキストデータのうち、<title>タグが付されている部分と、<h>タグが付されている部分とのそれぞれに対して、ＰＯＩ名称の候補文字列の抽出を試みる（ステップＳ３０４）。<title>タグおよび<h>タグは、ＰＯＩ名称が設定されている可能性が高いからである。但し、これに限らず、アノテータ処理部１０４は、これ以外のタグが付されている部分についても、ＰＯＩ名称の候補文字列の抽出を試みるようにしてもよい。

【００３３】

次に、アノテータ処理部１０４は、ステップＳ３０４で抽出されたＰＯＩ名称の候補文字列から、ＰＯＩ名称として不要と思われる部分を除去する（ステップＳ３０５）。さらに、アノテータ処理部１０４は、頻出語フィルタ処理を行うことにより、ステップＳ３０４で抽出されたＰＯＩ名称の候補文字列の中から、ＰＯＩ名称である可能性の高い候補文字列を抽出する（ステップＳ３０６）。例えば、アノテータ処理部１０４は、ＰＯＩ名称の候補文字列に、予め学習しておいたＰＯＩ名称の頻出語が含まれている場合、その候補文字列がＰＯＩ名称である可能性が高いと判断する。

【００３４】

次に、アノテータ処理部１０４は、ナイーブベイズ推定値取得処理を実行することにより、ステップＳ３０６で抽出された各候補文字列に対して、ナイーブベイズ推定値を取得する（ステップＳ３０７）。ナイーブベイズ推定値取得処理の詳細については、図４を用いて後述する。そして、アノテータ処理部１０４は、図３に示す一連の処理を終了する。

【００３５】

（アノテータ処理部１０４による住所取得処理の手順）

図４は、実施形態に係るアノテータ処理部１０４による住所取得処理の手順を示すフローチャートである。図４は、図３にフローチャートにおけるステップＳ３０３の住所取得処理を詳細に説明するものである。なお、図４の処理には、複数のテキストデータを含む

構造体が入力される。これに応じて、アノテータ処理部１０４は、図４の処理を、構造体に含まれるテキストデータ毎に実行する。

【００３６】

まず、アノテータ処理部１０４は、構造体に含まれるテキストデータから、住所を含むと推定される文書を抽出する（ステップＳ４０１）。例えば、アノテータ処理部１０４は、住所に関する特定のキーワード（例えば、「住所：」、「県」、「市」等）を含む文書（例えば、「この度ついに××県××市××〇丁目〇番地〇号に新規オープンしました」等）を抽出する。

【００３７】

次に、アノテータ処理部１０４は、ステップＳ４０１で抽出された文書が、住所の書式（例えば、「××県××市」等）に合致する文字列を含んでいるか否かを判断する（ステップＳ４０２）。

【００３８】

ステップＳ４０２において、住所の書式に合致する文字列を含んでいないと判断された場合（ステップＳ４０２：Ｎｏ）、アノテータ処理部１０４は、図４に示す一連の処理を終了する。

【００３９】

一方、ステップＳ４０２において、住所の書式に合致する文字列を含んでいると判断された場合（ステップＳ４０２：Ｙｅｓ）、アノテータ処理部１０４は、住所の書式に合致すると判断された文字列の長さが、１２８文字未満であるか否かを判断する（ステップＳ４０３）。

【００４０】

ステップＳ４０３において、住所の書式に合致すると判断された文字列の長さが、１２８文字未満ではないと判断された場合（ステップＳ４０３：Ｎｏ）、アノテータ処理部１０４は、図４に示す一連の処理を終了する。

【００４１】

一方、ステップＳ４０３において、住所の書式に合致すると判断された文字列の長さが、１２８文字未満であると判断された場合（ステップＳ４０３：Ｙｅｓ）、アノテータ処理部１０４は、住所の書式に合致すると判断された文字列を候補文字列とし、当該候補文字列に対してジオコーディングを実施する（ステップＳ４０４）。例えば、アノテータ処理部１０４は、特定の機関から提供されたジオコーディング用のＡＰＩ（Application Programming Interface）を実行することにより、候補文字列に対応する位置情報（経度および緯度）を取得する。

【００４２】

そして、アノテータ処理部１０４は、ジオコーディングによる位置情報の取得に成功したか否かを判断する（ステップＳ４０５）。ステップＳ４０５において、ジオコーディングによる位置情報の取得に成功したと判断された場合（ステップＳ４０５：Ｙｅｓ）、アノテータ処理部１０４は、位置情報の取得に成功した候補文字列をメモリに格納し（ステップＳ４０６）、候補文字列を後方に延長して（ステップＳ４０７）、ステップＳ４０１で抽出された文書から、候補文字列を再抽出する（ステップＳ４０８）。そして、アノテータ処理部１０４は、ステップＳ４０４に処理を戻す。

【００４３】

なお、アノテータ処理部１０４は、例えば、ステップＳ４０７による候補文字列の延長を、住所に関する特定の語句単位（例えば、都道府県、市区町村、番地等）で行う。例えば、現在の候補文字列が「××県」であった場合において、その次の語句が「××市」であった場合、アノテータ処理部１０４は、「××県××市」を新たな候補文字列としてもよい。ここで、アノテータ処理部１０４は、現在の候補文字列の次に、住所に関する特定の語句ではない品詞が存在する場合、候補文字列を、その品詞まで延長してもよい。例えば、現在の候補文字列が「××県××市××〇丁目〇番地〇号」であった場合において、その次の品詞が「に」であった場合、アノテータ処理部１０４は、「××県××市××〇

丁目〇番地〇号に」を新たな候補文字列としてもよい。

【００４４】

一方、ステップＳ４０５において、ジオコーディングによる位置情報の取得に失敗したと判断された場合（ステップＳ４０５：Ｎｏ）、位置情報の取得に成功した候補文字列がメモリに格納されているか否かを判断する（ステップＳ４０９）。

【００４５】

ここで、「ジオコーディングによる位置情報の取得に失敗した場合」とは、実際に、ジオコーディングのＡＰＩにてエラーが発生した場合に限らず、例えば、候補文字列の一部の文字列から位置情報が取得された場合（すなわち、候補文字列が完全一致しなかった場合）も含む。例えば、ＡＰＩによっては、「××県××市××〇丁目〇番地〇号に」を入力した場合に、エラーが発生せずに、位置情報の取得が可能な一部の文字列「××県××市××〇丁目〇番地〇号」から、位置情報を取得する場合がある。この場合、アノテータ処理部１０４は、「ジオコーディングによる位置情報の取得に失敗した」と判断するようにしてもよい。

【００４６】

また、ＡＰＩによっては、候補文字列との一致度を示す信頼度を返す場合がある。この場合、例えば、アノテータ処理部１０４は、直前の候補文字列から信頼度が上昇した場合または直前の候補文字列と信頼度が同一の場合、「ジオコーディングによる位置情報の取得に成功した」と判断し、直前の候補文字列から信頼度が低下した場合、「ジオコーディングによる位置情報の取得に失敗した」と判断するようにしてもよい。

【００４７】

ステップＳ４０９において、位置情報の取得に成功した候補文字列がメモリに格納されていないと判断された場合（ステップＳ４０９：Ｎｏ）、アノテータ処理部１０４は、図４に示す一連の処理を終了する。

【００４８】

一方、ステップＳ４０９において、位置情報の取得に成功した候補文字列がメモリに格納されていると判断された場合（ステップＳ４０９：Ｙｅｓ）、アノテータ処理部１０４は、メモリに格納されている候補文字列が、区、地番、および枝番を含むか否かを判断する（ステップＳ４１０）。ステップＳ４１０において、メモリに格納されている候補文字列が、区、地番、および枝番を含まないと判断された場合（ステップＳ４１０：Ｎｏ）、アノテータ処理部１０４は、図４に示す一連の処理を終了する。

【００４９】

一方、ステップＳ４１０において、メモリに格納されている候補文字列が、区、地番、および枝番を含むと判断された場合（ステップＳ４１０：Ｙｅｓ）、アノテータ処理部１０４は、メモリに格納されている候補文字列の長さが、所定文字数未満であるか否かを判断する（ステップＳ４１１）。ステップＳ４１１において、メモリに格納されている候補文字列の長さが、所定文字数未満ではないと判断された場合（ステップＳ４１１：Ｎｏ）、アノテータ処理部１０４は、図４に示す一連の処理を終了する。

【００５０】

一方、ステップＳ４１１において、メモリに格納されている候補文字列の長さが、所定文字数未満であると判断された場合（ステップＳ４１１：Ｙｅｓ）、アノテータ処理部１０４は、メモリに格納されている候補文字列を、住所とする文字列に決定する（ステップＳ４１２）。そして、アノテータ処理部１０４は、図４に示す一連の処理を終了する。

【００５１】

（アノテータ処理部１０４によるナイーブベイズ推定値取得処理の手順）

図５は、実施形態に係るアノテータ処理部１０４によるナイーブベイズ推定値取得処理の手順を示すフローチャートである。図５は、図３にフローチャートにおけるステップＳ３０７の処理を詳細に説明するものである。なお、図５の処理には、複数のＰＯＩ名称の候補文字列を含む候補リストが入力される。これに応じて、アノテータ処理部１０４は、図５の処理を、候補リストに含まれるＰＯＩ名称の候補文字列毎に実行する。

【００５２】
まず、アノテータ処理部１０４は、ＰＯＩ名称の候補文字列をナイーブベイズ分類器にかけることにより、ＰＯＩ名称としての確からしさの推定を行う（ステップＳ５０１）。そして、アノテータ処理部１０４は、ステップＳ５０１で推定されたＰＯＩ名称としての確からしさを示す、ナイーブベイズ推定値を取得する（ステップＳ５０２）。このナイーブベイズ推定値は、「１．００」を最大値とするものである。
【００５３】
次に、アノテータ処理部１０４は、ＰＯＩ名称の候補文字列が、所定文字列を含むか、または、強調されているか否かを判断する（ステップＳ５０３）。所定文字列とは、例えば、「株式会社」、「店」等、ＰＯＩ名称である可能性を高める文字列である。このため、情報抽出装置１００には、予め、複数の所定の文字列が、メモリ等に予め登録されている。また、ＰＯＩ名称の候補文字列が強調されている場合とは、例えば、ＰＯＩ名称の候補文字列に強調タグ（例えば、、、<b>、<font>等）が付されている場合等である。
【００５４】
ステップＳ５０３において、ＰＯＩ名称の候補文字列が、所定文字列を含まず、且つ、強調されていないと判断された場合（ステップＳ５０３：Ｎｏ）、アノテータ処理部１０４は、ステップＳ５０５へ処理を進める。
【００５５】
一方、ステップＳ５０３において、ＰＯＩ名称の候補文字列が、所定文字列を含む、または、強調されていると判断された場合（ステップＳ５０３：Ｙｅｓ）、アノテータ処理部１０４は、ステップＳ５０２で取得されたナイーブベイズ推定値を所定倍（例えば、１．２５倍）する（ステップＳ５０４）。但し、これによりナイーブベイズ推定値が「１．００」を超える場合は、ナイーブベイズ推定値を「１．００」とする。そして、アノテータ処理部１０４は、ステップＳ５０５へ処理を進める。
【００５６】
ステップＳ５０５では、アノテータ処理部１０４は、ナイーブベイズ推定値を、ＰＯＩ名称の候補文字列に対応付けて、候補リストに保存する。そして、アノテータ処理部１０４は、図５に示す一連の処理を終了する。
【００５７】
以上説明したように、本実施形態に係る情報抽出装置１００によれば、ＷＥＢページ１１０から抽出した候補文字列に対してジオコーディングを行うことにより、位置情報の取得を試み、当該位置情報の取得結果に応じて、住所とする文字列を決定することができる。特に、本実施形態に係る情報抽出装置１００によれば、位置情報の取得に成功した場合、位置情報の再取得に失敗するまで、候補文字列を後方に延長して、位置情報の再取得を試みることができる。これにより、本実施形態に係る情報抽出装置１００によれば、ＷＥＢページ１１０から抽出する住所文字列の終端を高精度に特定することができる。したがって、本実施形態に係る情報抽出装置１００によれば、ＷＥＢページ１１０からのＰＯＩデータの抽出精度を高めることができる。
【００５８】
また、本実施形態に係る情報抽出装置１００によれば、ＷＥＢページ１１０から抽出したＰＯＩ名称の候補文字列が所定文字列を含むか、または、強調されている場合、そのＰＯＩ名称としての確からしさを示すナイーブベイズ推定値を高めることができる。このため、本実施形態に係る情報抽出装置１００によれば、例えば、ＷＥＢページ１１０から複数のＰＯＩ名称の候補文字列が抽出された場合であっても、これら複数のＰＯＩ名称の候補文字列の中から、ナイーブベイズ推定値に基づいて、実際のＰＯＩ名称である可能性が最も高い文字列を抽出することができる。したがって、本実施形態に係る情報抽出装置１００によれば、ＷＥＢページ１１０からのＰＯＩデータの抽出精度を高めることができる。
【００５９】

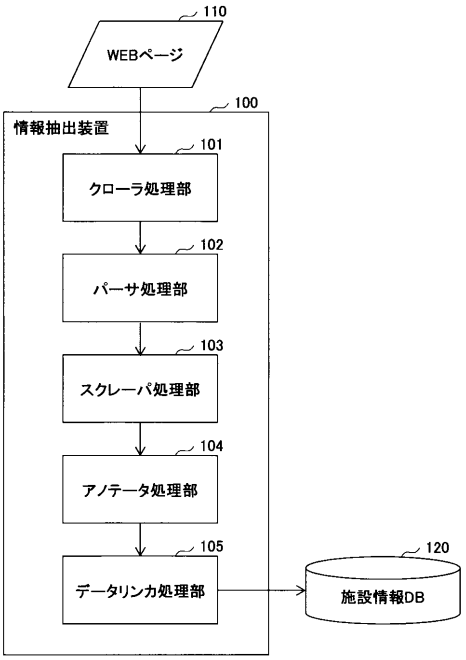
以上、本発明の好ましい実施形態について詳述したが、本発明はこれらの実施形態に限定されるものではなく、特許請求の範囲に記載された本発明の要旨の範囲内において、種々の変形又は変更が可能である。

【符号の説明】

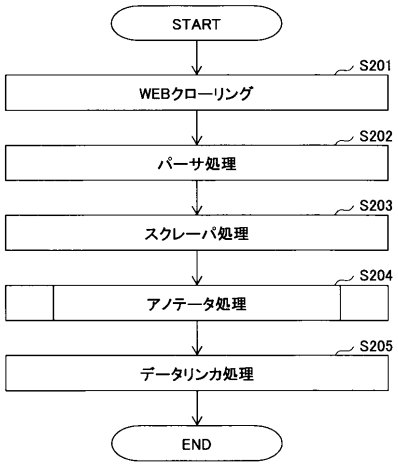
【００６０】

１００　情報抽出装置
１０１　クローラ処理部（文書データ取得部）
１０２　パーサ処理部
１０３　スクレーパ処理部
１０４　アノテータ処理部（候補文字列抽出部、位置情報取得部、住所文字列決定部）
１０５　データリンカ処理部
１１０　ＷＥＢページ
１２０　施設情報ＤＢ

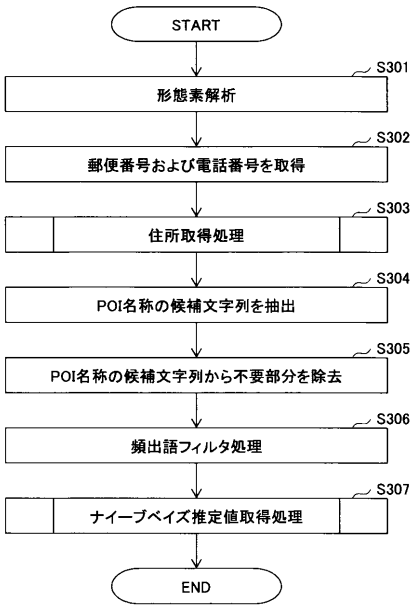
【図１】



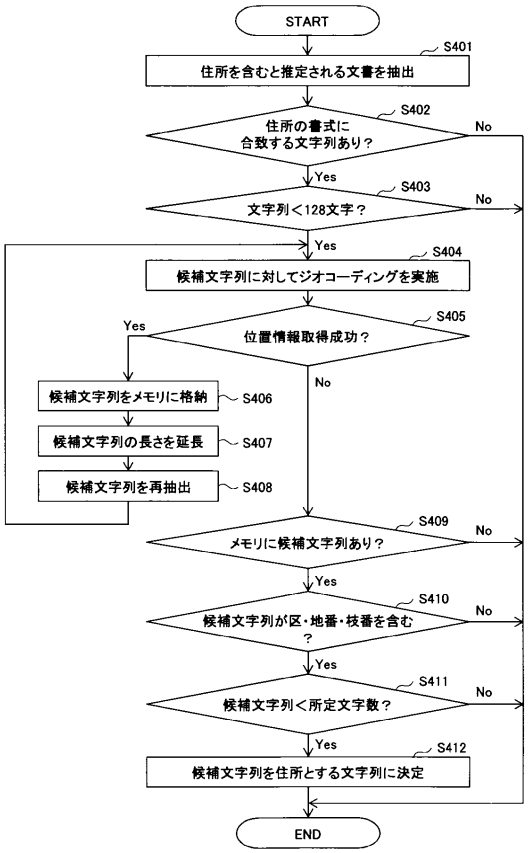
【図２】



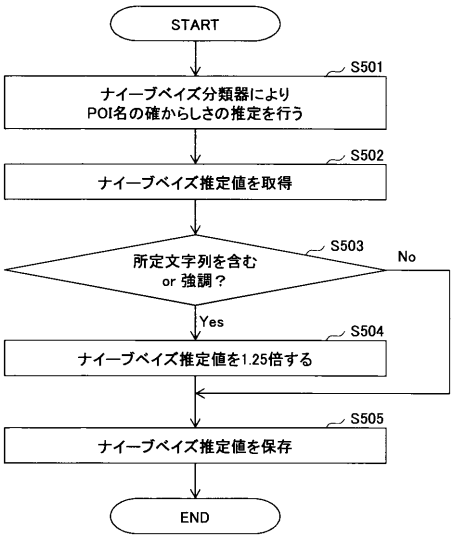
【図３】


START
S301
形態素解析
S302
郵便番号および電話番号を取得
S303
住所取得処理
S304
POI名称の候補文字列を抽出
S305
POI名称の候補文字列から不要部分を除去
S306
頻出語フィルタ処理
S307
ナイーブベイズ推定値取得処理
END

【図４】


START
S401
住所を含むと推定される文書を抽出
S402
住所の書式に合致する文字列あり？
No
Yes
S403
文字列＜128文字？
No
Yes
S404
候補文字列に対してジオコーディングを実施
S405
位置情報取得成功？
Yes
No
候補文字列をメモリに格納
S406
候補文字列の長さを延長
S407
候補文字列を再抽出
S408
S409
メモリに候補文字列あり？
No
Yes
S410
候補文字列が区・地番・枝番を含む？
No
Yes
S411
候補文字列＜所定文字数？
No
Yes
S412
候補文字列を住所とする文字列に決定
END

【図５】


START
S501
ナイーブベイズ分類器により
POI名の確からしさの推定を行う
S502
ナイーブベイズ推定値を取得
S503
所定文字列を含む
or 強調？
No
Yes
S504
ナイーブベイズ推定値を1.25倍する
S505
ナイーブベイズ推定値を保存
END

フロントページの続き

(56)参考文献 特表２００９－５０６４５９（ＪＰ，Ａ）
特開２００４－０８６２７２（ＪＰ，Ａ）
特開２００７－１７９３２９（ＪＰ，Ａ）
特開２０１６－０９１３１５（ＪＰ，Ａ）
特開平１０－２４０７１０（ＪＰ，Ａ）
特開２００６－０６４４４３（ＪＰ，Ａ）
福田拓真，力宗幸男，Ｇｅｏマイクロフォーマットを用いた住所自動検出・地図表示システムの開発，電子情報通信学会技術研究報告，社団法人電子情報通信学会，２０１０年 1月１４日，第１０９巻，第３７９号，ｐ．９３～９８
石田武久，外２名，Ｗｅｂ上のイラストマップを実地図に重ね合わせるシステムの試作，電子情報通信学会技術研究報告，社団法人電子情報通信学会，２０１１年１２月 9日，第１１１巻，第３６１号，ｐ．４３～４８

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １６／００ － １６／９５８
Ｇ０６Ｆ ４０／００ － ４０／１９７