(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7434867号
(P7434867)

(45)発行日　令和6年2月21日(2024.2.21)

(24)登録日　令和6年2月13日(2024.2.13)

(51)国際特許分類

		FI	
G06F	*16/957(2019.01)*	G06F	16/957
G06F	*16/951(2019.01)*	G06F	16/951
H04L	*67/02 (2022.01)*	H04L	67/02

請求項の数　10　（全16頁）

(21)出願番号　特願2019-223095(P2019-223095)
(22)出願日　令和1年12月10日(2019.12.10)
(65)公開番号　特開2020-98596(P2020-98596A)
(43)公開日　令和2年6月25日(2020.6.25)
審査請求日　令和4年8月9日(2022.8.9)
(31)優先権主張番号　201811549030.2
(32)優先日　平成30年12月18日(2018.12.18)
(33)優先権主張国・地域又は機関
中国(CN)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(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　ジョン・ジョォングアン
中国，１０００２７，ベイジン，チャオヤン ディストリクト，ゴン ティ ベイルゥ ナンバー２エイ，パシフィック センチュリー プレイス，スペース ８，ゲート ６，ユニット ３エフ，３５５ 富士通研究開発中心有限公司内

最終頁に続く

(54)【発明の名称】　ウェブページから情報を抽出する方法、装置及び記憶媒体

(57)【特許請求の範囲】

【請求項１】

ウェブページから情報を抽出する方法であって、コンピュータが、

前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成するステップと、

前記木構造におけるナビゲーションバーノードを決定するステップと、

前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するステップと、

マッチする葉ノードに対応するページにおける情報を抽出するステップと、を実行する、方法。

【請求項２】

前記木構造におけるナビゲーションバーノードを決定するステップは、

前記木構造に出現する回数が所定閾値よりも大きい葉ノードのみを含む非葉ノードを決定するステップと、

前記非葉ノードを並び替えて前記ナビゲーションバーノードを決定するステップと、を含む、請求項１に記載の方法。

【請求項３】

葉ノードの出現回数が所定閾値よりも大きいか否かを決定することは、

前記葉ノードのテキスト及び経路情報の前記木構造における出現回数が前記所定閾値よりも大きいか否かを決定すること、を含む、請求項２に記載の方法。

【請求項４】
前記経路情報は、前記葉ノードからそのｎ番目の先祖ノードまでの経路であり、ｎは正整数である、請求項３に記載の方法。
【請求項５】
ｎは５以上である、請求項４に記載の方法。
【請求項６】
前記非葉ノードを並び替えて前記ナビゲーションバーノードを決定するステップは、
前記非葉ノードの特徴値を計算するステップであって、前記特徴値は、前記非葉ノードによりカバーされる葉ノードの数及び前記回数により決定される、ステップと、
前記非葉ノードのうちの最大の特徴値を有する非葉ノードを前記ナビゲーションバーノードとして決定するステップと、を含む、請求項２に記載の方法。
【請求項７】
マッチする葉ノードに対応するページにおける情報を抽出するステップは、
前記マッチする葉ノードに対応するページに含まれるターゲットノードを決定するステップと、
前記ターゲットノードによりカバーされる各葉ノードのテキストをそれぞれ抽出するステップと、を含む、請求項１乃至６の何れかに記載の方法。
【請求項８】
前記ターゲットノードは、
前記ターゲットノードに含まれる各葉ノードのテキスト及び経路情報の前記木構造における出現回数が所定閾値以下であること、
前記ターゲットノードが、前記木構造に出現する回数が所定閾値よりも大きい葉ノードのみを含む非葉ノードのうちの非葉ノードではないこと、及び
前記ターゲットノードに含まれる全ての葉ノードのテキストの合計長さが該木構造における他の非葉ノードのテキストの合計長さよりも大きいこと、により決定される、請求項７に記載の方法。
【請求項９】
ウェブページから情報を抽出する装置であって、
前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成する木構造生成手段と、
前記木構造におけるナビゲーションバーノードを決定するナビゲーションバーノード決定手段と、
前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するマッチノード決定手段と、
マッチする葉ノードに対応するページにおける情報を抽出する情報抽出手段と、を含む、装置。
【請求項１０】
プログラムを記憶したコンピュータ読み取り可能な記憶媒体であって、前記プログラムがプロセッサにより実行される際に、
ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成するステップと、
前記木構造におけるナビゲーションバーノードを決定するステップと、
前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するステップと、
マッチする葉ノードに対応するページにおける情報を抽出するステップと、を実行させる、記憶媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、自然言語処理に関し、具体的には、複数のウェブページに基づく情報抽出に

関する。
【背景技術】
【０００２】
インターネットから情報を収集、抽出することは、知識ベースを構築する重要な手段である。例えば、電子商取引会社のウェブページから製品の情報を抽出し、製品の知識ベースを構築することができる。従来の方法は、主に２種類の方法に分類される。
【０００３】
１種類目の方法は、類似の構造を有するページ（例えば、電子商取引会社のウェブサイトの製品リストのページでは、各ページの構造は類似する）の場合は、手動でテンプレートを作成し、或いは教師なし、教師ありの方法によりウェブページに含まれる製品情報の構造テンプレートを学習して、これらの学習により得られた構造テンプレートを用いて他の類似のウェブページを解析してもよい。図１Ａに示すように、携帯電話のページの構造情報を学習することで、図書及び靴の製品情報を抽出してもよい。
【０００４】
２種類目の方法は、単一の構造を有する（非類似の）ページの場合は、図１Ｂに示すように、ウェブページの構造を動的に解析し、キーワードのリストにより関連情報のウェブページにおける位置を特定し、値を抽出してもよい。
【発明の概要】
【発明が解決しようとする課題】
【０００５】
以下は、本発明の態様を基本的に理解させるために、本発明の簡単な概要を説明する。なお、この簡単な概要は、本発明を網羅的な概要ではなく、本発明のポイント又は重要な部分を意図的に特定するものではなく、本発明の範囲を意図的に限定するものではなく、後述するより詳細的な説明の前文として、単なる概念を簡単な形で説明することを目的とする。
【０００６】
本発明は、ウェブページから情報を抽出する方法、装置及び記憶媒体を提供する。
【課題を解決するための手段】
【０００７】
本発明の１つの態様では、ウェブページから情報を抽出する方法であって、前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成するステップと、前記木構造におけるナビゲーションバーノードを決定するステップと、前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するステップと、マッチする葉ノードに対応するページにおける情報を抽出するステップと、を含む、方法を提供する。
【０００８】
本発明のもう１つの態様では、ウェブページから情報を抽出する装置であって、前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成する木構造生成手段と、前記木構造におけるナビゲーションバーノードを決定するナビゲーションバーノード決定手段と、前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するマッチノード決定手段と、マッチする葉ノードに対応するページにおける情報を抽出する情報抽出手段と、を含む、装置を提供する。
【０００９】
本発明の他の態様では、対応するコンピュータプログラムコード、コンピュータ読み取り可能な記憶媒体及びコンピュータプログラムプロダクトをさらに提供する。
【００１０】
本発明に係るウェブページから情報を抽出する方法及び装置によれば、ホームページのＵＲＬ（ユニフォームリソースロケータ）に基づいて、同一のドメイン名において分布している複数のウェブページから必要な情報を抽出することができる。

【００１１】
以下は図面を参照しながら本発明の好ましい実施形態を詳細に説明することにより、本発明の上記及び他の利点はより明確になる。
【図面の簡単な説明】
【００１２】
本開示の上記及び他の利点及び特徴を理解させるために、以下は図面を参照しながら本開示の具体的な実施形態を詳細に説明する。図面及び以下の詳細な説明は本明細書に含まれ、本明細書の一部を構成する。同一の機能及び構造を有する素子は同一の符号で示される。なお、これらの図面は単なる本開示の典型的な例を説明するためのものであり、本開示の範囲を限定するものではない。
【図１Ａ】類似の構造を有するウェブページの例を示す図である。
【図１Ｂ】単一の構造を有するウェブページの情報抽出の例を示す図である。
【図２Ａ】複数のページの情報抽出の例を示す図である。
【図２Ｂ】本発明の方法の全体的な流れの例を示す図である。
【図３】本発明の実施形態に係るウェブページから情報を抽出する方法の流れを示すフローチャートである。
【図４Ａ】ナビゲーションバーノードに対応するＨＴＭＬ構造及びＤｏｍ木構造の例を示す図である。
【図４Ｂ】情報抽出の例を示す図である。
【図５】本発明の実施形態に係るウェブページから情報を抽出する装置の例を示すブロック図である。
【図６】本発明の実施形態に係る方法及び／又は装置を実現可能な汎用パーソナルコンピュータの例示的な構成を示すブロック図である。
【発明を実施するための形態】
【００１３】
以下、図面を参照しながら本発明の例示的な実施例を詳細に説明する。説明の便宜上、明細書には実際の実施形態の全ての特徴が示されていない。なお、実際に実施する際に、開発者の具体的な目標を実現するために、特定の実施形態を変更してもよい、例えばシステム及び業務に関する制限条件に応じて実施形態を変更してもよい。また、開発作業が非常に複雑であり、且つ時間がかかるが、本公開の当業者にとって、この開発作業は単なる例の作業である。
【００１４】
なお、本発明を明確にするために、図面には本発明に密に関連する装置の構成要件及び／又は処理のステップのみが示され、本発明と関係のない細部が省略されている。
【００１５】
上述したように、インターネットから情報を収集、抽出することは、知識ベースを構築する重要な手段である。図１Ａ及び図１Ｂに示す従来技術は、一部の要求を満たすことができるが、依然として限界がある。
【００１６】
図２Ａに示すように、ｈｔｔｐ：／／ｏｗｔｗａｒｅ.ｃｏｍは会社のホームページのＵＲＬであり、製品、協力会社、連絡先などの会社の情報は異なるページに分布し、３つのページの主要情報の所在する部分も類似の構造を有しない。
【００１７】
ホームページのＵＲＬのみが既知である場合、従来方法は、このような複数のページに分布している情報を抽出することができない。一方、通常ホームページのＵＲＬは容易に入手できる。このため、ホームページのＵＲＬ情報を拡張して他の情報を抽出する方法は、依然として解決すべき問題である。
【００１８】
従来技術に存在する問題を解決するために、本発明は、ホームページＵＲＬのみが既知である場合、（１）関連情報を含む他のページを自動的に拡張し、（２）各関連ページか

ら主要情報を含む位置を取得し、（３）異なる属性タイプを有するページについて個別の情報抽出を行うことができる、複数のページに基づく情報抽出方法を提供する。

【００１９】

図２Ｂは本発明の方法の全体的な流れの例を示す図である。図２Ｂに示すように、本発明に係る方法は、主に以下の３つの部分を含む。

【００２０】

（１）ホームページを拡張することで複数のページの集合を取得する。

【００２１】

（２）統計的方法を用いてウェブページの集合に対して統計的な分類を行い、ナビゲーションバーノード（ｎａｖｉｇａｔｉｏｎ　ｂａｒ　ｎｏｄｅ）を取得し、そして、キーワード辞書を用いてナビゲーションバーノードに含まれる葉ノードのテキストのマッチングを行い、マッチするノード情報に基づいて抽出すべきページを取得する。

【００２２】

（３）抽出すべきページの情報タイプに応じて、異なる解析器を用いて抽出を行う。

【００２３】

以下は、図３、図４Ａ及び図４Ｂを参照しながら、本発明の実施形態に係るウェブページから情報を抽出する方法を詳細に説明する。

【００２４】

図３は本発明の実施形態に係るウェブページから情報を抽出する方法の流れを示すフローチャートである。

【００２５】

まず、ステップ３０１において、ウェブページ及びその全ての拡張ウェブページにおける該ウェブページのドメイン名を含む各ページについて木構造を生成する。具体的には、本実施形態では、図２Ａに示すＵＲＬを一例にすると、会社ホームページＵＲＬはu_{root}＝ｈｔｔｐ：／／ｗｗｗ．ｏｗｔｗａｒｅ．ｃｏｍ／であり、抽出すべき情報は該会社の他の属性、例えば製品、連絡先などである。

【００２６】

まず、クローラー（ｃｒａｗｌｅｒ）を用いてu_{root}に対応するＨＴＭＬページp_{root}をクロールし、ページを解析してそれに含まれる全てのＵＲＬの集合ｕ＝［u_0，u_1，u_2，…，u_n］を取得する。ページに含まれるＵＲＬが該会社に関連する場合があり、関連しない場合もあり、例えば広告や外部リンクなどの場合もあると考慮すると、特定のルールに従って一部のＵＲＬの集合ｕ’＝［u_0’，u'_1，u'_2，…，u'_n］を選択し、ここで、u'_iにはｄｏｍａｉｎ（u_{root}）が含まれ、ｄｏｍａｉｎ（ＵＲＬ）はＵＲＬトップレベルドメイン名を抽出する操作であり、例えばｄｏｍａｉｎ（u_{root}）＝ｗｗｗ．ｏｗｔｗａｒｅ．ｃｏｍ。このように、同一のドメイン名を有する全てのＵＲＬ、例えばｈｔｔｐ：／／ｗｗｗ.ｏｗｔｗａｒｅ.ｃｏｍ／ｉｎｄｅｘ.ｐｈｐ／ｚｈ／ｐｒｏｄｕｃｔｓ／を保持することができる。

【００２７】

好ましくは、u'_iに対応するページp_iは他のＵＲＬ情報を含む可能性があると考慮するため、p_iをさらに拡張してもよい。各ｐｉについて、同様のルールでＵＲＬ及び対応するページを拡張し、毎回の拡張の後に同一のＵＲＬ及びページを併合する。拡張のプロセスはｎ回だけ繰り返してもよい。一定の数のページを取得でき、且つページの数が多すぎないように、通常ｎ＝２にしてもよい。これによって、同一のドメイン名を有するページの集合ｐ＝［＜p_0，u_0＞，＜p_1，u_1＞，＜p_2，u_2＞，…，＜p_n，u_n＞］を取得でき、ここで、p_iはウェブページを表し、u_iはウェブページに対応するＵＲＬを表す。

【００２８】

次に、ステップ３０２において、木構造におけるナビゲーションバーノードを決定する。具体的には、本実施形態では、集合ｐからナビゲーションバーノードを取得する。上述したように、目的は、集合ｐから該会社情報を含むページ、例えば製品、連絡先などを取

得することである。通常、ナビゲーションバーノードにおけるリンクにより、これらの情報に対応するページを取得できる。ナビゲーションバーノードを情報アンカーとして選択する主な理由は３つある。

【００２９】

（１）情報は正確である。ナビゲーションバーノードに含まれるリンクが指向するページは、会社の紹介と見なすことができる。例えば、「製品とサービス」に対応するページは該会社の製品を紹介し、「連絡先」は会社の住所、電話番号などの情報のページにリンクする。ウェブページにおける他の部分に出現するリンクは、必ずしも該会社の情報を説明するものではなく、他の会社の紹介や広告などの情報である可能性がある。

【００３０】

（２）情報は全面的である。ナビゲーションバーノードは基本的に該会社に関連する全ての情報を含み、ナビゲーションバーノードを取得すると、関連情報を含む全てのページを取得でき、これは後続の情報抽出に非常に役に立つ。

【００３１】

（３）比較的に取得しやすい。異なるウェブページは異なる構造を有する可能性があるが、ナビゲーションバーノードの様式は殆ど同じである。このような共通性により、ウェブ構造からナビゲーションバーノードの位置を正確に見つけることができる。

【００３２】

以下は、ナビゲーションバーノードの決定方法を例示的に説明する。

【００３３】

上記の３つの特徴により、各ページp_i（$p_i \in p$）におけるノードを計数することで、頻繁に出現するノードを取得してもよい。これらのノードにはナビゲーションバーノードが含まれるため、特徴値に基づいてこれらの頻繁に出現するノードを並び替えることでナビゲーションバーノードを取得してもよい。具体的な方法は以下の通りである。

【００３４】

図４Ａに示すように、集合ｐにおける各ページp_iについて、まずp_iをＤｏｍ木の構造に変換する。

【００３５】

Ｄｏｍ木における各葉ノード$node_i$について、$node_i$の経路パターン$path_i$を取得し、$path_i$は、該葉ノードに対応するテキストと、ｎ番目の先祖ノードまでの経路により構成される。実際の経験によると、殆どのページでは、ｎは５以上の整数値であってもよい。例えば、ナビゲーションバーノード「連絡先」について、ｎ＝５の場合は、$path_i$＝「ｕｌ＿ｌｉ＿ｕｌ＿ｌｉ＿ａ＿連絡先」を取得できる。

【００３６】

次に、各$path_i$の文書頻度df_i、即ち$path_i$が異なる文書に出現する回数を算出する。統計により経路頻度辞書ｎｏｄｅ＿ｐａｔｔｅｒｎ＿ｄｉｃｔｉｏｎａｒｙ｛＜$path_1$，df_1＞，…，＜$path_n$，df_n〉｝を取得してもよく、ここで、df_i＞ｔ、ｔは次のように設定された閾値である。

【数１】

$$t = \begin{cases} |p| * 0.2 & if |p| \geq 20 \\ |p| * 0.3 & if |p| \geq 10 \\ 3 & else \end{cases}$$

【００３７】

ページ数｜ｐ｜の最終結果への影響を低減するために、閾値ｔを段階的に設定する。

【００３８】

経路頻度辞書を取得した後、集合ｐにおける各p_iに対応するＤｏｍ木構造に対して２回目の走査を行い、今回は、各非葉ノードｎｏｄｅ$_i$について、それによりカバーされる全てのＮＵＬＬでない葉ノードの集合がｃ＝［c_0，c_1，c_2，…，c_n］となると仮定すると、各c_iについて、ｐａｔｈ$_i$（c_i）が経路頻度辞書ｎｏｄｅ＿ｐａｔｔｅｒｎ＿ｄｉｃｔｉｏｎａｒｙに存在する場合、該ｎｏｄｅ$_i$の情報を記録する。最後に、候補辞書ｃａｎｄｉｄａｔｅ＿ｐａｔｔｅｒｎ＿ｄｉｃｔｉｏｎａｒｙ｛＜$path_1$，［df_1，cn_1］＞，…，＜$path_n$，［df_n，cn_n］＞｝を取得してもよく、ここで、ｐａｔｈ$_i$は非葉ノードｎｏｄｅ$_i$から先祖ノードまでの経路情報を表し、ｄｆ$_i$は文書頻度を表し、ｃｎ$_i$はｎｏｄｅ$_i$によりカバーされる全てのＮＵＬＬでない葉ノードの数を表す。葉ノードのｐａｔｈ$_i$とは異なって、非葉ノードのｐａｔｈ$_i$はテキスト情報を含まない。図４Ａにおける３に示すように、「連絡先」からｕｌノードまでの対応する経路はｕｌ＿ｌｉ＿ｕｌ＿ｄｉｖ＿ｄｉｖであり、ｎ＝５となる。

【００３９】

最後に、（ｃｎ＊ｄｆ／｜ｐ｜）の値に従って候補辞書ｃａｎｄｉｄａｔｅ＿ｐａｔｔｅｒｎ＿ｄｉｃｔｉｏｎａｒｙを並び替え、最大値に対応する経路をナビゲーションバーノード経路テンプレートとして取得し、該最大値に対応する経路における先祖ノードをナビゲーションバーノードとして決定してもよい。該会社のホームページの下にある所定のＨＴＭＬページについて、該テンプレートを用いてナビゲーションバーノードの位置を特定してもよい。

【００４０】

なお、上記の統計的方法を用いてナビゲーションバーノードを決定することは、単なるナビゲーションバーノードの決定方法の一例である。本発明は、これに限定されず、他の適切な方法を用いてナビゲーションバーノードを決定してもよい。

【００４１】

次に、ステップ３０３において、ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定する。具体的には、本実施形態では、ステップ３０２においてナビゲーションバーノードが取得された後に、該ナビゲーションバーノードによりカバーされる各ＮＵＬＬでない葉ノードについて、辞書ｋｅｙｗｏｒｄ＿ｄｉｃｔを用いて葉ノードに対応するテキストのマッチングを行う。辞書ｋｅｙｗｏｒｄ＿ｄｉｃｔには、例えば「製品紹介」、「連絡先」などの所定のキーワードが含まれる。葉ノードがキーワードにマッチする場合、対応するＨＴＭＬ要素から「ｈｒｅｆ」属性を検索してもよく、その属性値は対応するウェブページのＵＲＬである。例えば、図４Ａにおける「連絡先」ノードに対応するＨＴＭＬ要素には次のリンクが含まれる。

【００４２】

ｈｒｅｆ＝ｈｔｔｐ：／／ｗｗｗ.ｏｗｔｗａｒｅ.ｃｏｍ／ｉｎｄｅｘ.ｐｈｐ／ｚｈ／ａｂｏｕｔ／ｃｏｎｔａｃｔ－ｕｓ／

従って、集合ｐから関連情報を含むウェブページの集合p'＝［＜p_0'，u'_0，t'_0＞，＜p'_1，u'_1，t'_1＞，＜p'_2，u'_2，t'_2＞，…，＜p'_n，u'_n，t'_n＞］を選択してもよく、ここで、p'_i及びu'_iは上記の定義されたp_i及びu_iと同じであり、t'_iは、該ページに対応するタイプ、例えば製品、人物、連絡先などを表す。これによって、ページの異なるタイプに応じて、異なる解析器を選択して抽出を行うことができる。

【００４３】

各p'_iについて、まず、ＨＴＭＬページを前処理する必要がある。前処理の目的は、まずページにおける主要情報を抽出することである。このプロセスは共通のものであり、ウェブページのタイプt'とは関係がない。抽出された結果は、後で抽出を行う時の入力としてもよい。図４Ｂの（１）に示すように、元のＨＴＭＬページには多くの内容が含まれているが、実線の枠で示される部分のみが必要な内容であり、ナビゲーションバーノード、サイドリスト、ラベルＦｏｏｔｅｒなどの要素を含む他の部分を全て除去する必要があり、除去しないと、抽出時にノイズデータの影響を受けやすくなる。

【００４４】

ステップ３０２において生成された経路頻度辞書ｎｏｄｅ＿ｐａｔｔｅｒｎ＿ｄｉｃｔｉｏｎａｒｙ及び候補辞書ｃａｎｄｉｄａｔｅ＿ｐａｔｔｅｒｎ＿ｄｉｃｔｉｏｎａｒｙを考慮すると、以下の方法を用いてナビゲーションバーノードによりカバーされる１つ又は複数のキーワードにマッチする葉ノードを決定してもよい。

【００４５】

集合ｐ’ｉにおける非葉ノードｎｏｄｅiについて、それによりカバーされる全てのＮＵＬＬでない葉ノードの集合がｃ＝［ｃ０，ｃ１，ｃ２，…，ｃｎ］であると仮定すると、次の３つの条件が同時に満たされた場合、ｎｏｄｅｉが１つ又は複数のキーワードにマッチする葉ノードを含むターゲット内容ノードであると決定してもよい。

【数２】

- ✓ $path(node_i) \notin candidate\_pattern\_dictionary$，
- ✓ $path(c_i) \notin node\_pattern\_dictionary$，$\forall c_i \in c$，且つ
- ✓ $\sum text\_len(c_i) > \sum text\_len(c_j)$

【００４６】

ここで、ｃｉはｎｏｄｅｉによりカバーされるＮＵＬＬでない葉ノードであり、ｃｊはｎｏｄｅｊによりカバーされるＮＵＬＬでない葉ノードであり、ｉ≠ｊとなり、ｔｅｘｔ＿ｌｅｎ（＊）は葉ノードに対応するテキストの長さを表す。言い換えれば、ｎｏｄｅｉによりカバーされる全てのＮＵＬＬでない葉ノードのテキストの合計長さは、他のノードｎｏｄｅｊによりカバーされる全てのＮＵＬＬでない葉ノードのテキストの合計長さよりも大きい。

【００４７】

上記の３つの条件を同時に満たすノードｎｏｄｅｉが決定されると、所定のキーワードにマッチする葉ノードが決定されることを意味する。

【００４８】

最後に、ステップ３０４において、マッチする葉ノードに対応するページにおける情報を抽出する。具体的には、本実施形態では、上記３つの条件を満たすノードｎｏｄｅｉが決定された後、該ノードによりカバーされる葉ノードに含まれる情報を抽出してもよい。

【００４９】

好ましくは、その各葉ノードを独立した属性抽出空間としてもよく、図４Ｂにおける（２）及び（３）に示すように、各ノード＜ｄｉｖ　ｃｌａｓｓ＝“ｐａｎｅｌ－ｇｒｉｄ－ｃｅｌｌ”…＞を独立した属性空間とする。これによって、属性値の境界を決定することができ、即ち、各値はセクション{{…}}からの値のみである。例えば、人物情報を抽出する場合、セクション{{…}}に含まれる情報は同一の人物を表すためのものであり、異なる{{…}}の情報は異なる人物を表すと見なしてもよいため、抽出エラーを回避することができる。

【００５０】

好ましくは、抽出範囲が決定された後、ｐ’ｉのタイプｔ’ｉに応じて、異なる解析器、例えばエンティティ認識器（ＮＥＲ）、固有名詞認識器、数値認識器などを選択して特定情報の抽出を行ってもよい。図４Ｂの（３）では、固有名詞認識器の結果の例を示している。

【００５１】

なお、以上は会社ホームページに基づいて関連情報を抽出することを説明しているが、本発明はこれに限定されず、必要に応じて任意のウェブページの任意の情報の抽出に拡張されてもよい。

【００５２】

上記の方法は、コンピュータ実行可能なプログラムにより完全に実現されてもよいし、ハードウェア及び／又はファームウェアを用いて部分的又は完全に実現されてもよい。ハ

ードウェア及び／又はファームウェアにより実現される場合、又はコンピュータ実行可能なプログラムがプログラムを実行可能なハードウェア装置にロードされる場合、後述するウェブページから情報を抽出する装置が実現される。以下は、上述した詳細な内容を省略し、これらの装置の概要を説明する。なお、これらの装置は上記の方法を実行することができるが、上記方法は後述する装置の構成部を採用し、或いは構成部により実行されるものに限定されない。

【００５３】

図５は本発明の実施形態に係るウェブページから情報を抽出する装置５００の例を示すブロック図である。装置５００は、木構造生成部５０１、ナビゲーションバーノード決定部５０２、マッチノード決定部５０３及び情報抽出部５０４を含む。木構造生成部５０１は、ウェブページ及びその全ての拡張ウェブページにおける該ウェブページのドメイン名を含む各ページについて木構造を生成する。ナビゲーションバーノード決定部５０２は、該木構造におけるナビゲーションバーノードを決定する。マッチノード決定部５０３は、該ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定する。情報抽出部５０４は、マッチする葉ノードに対応するページにおける情報を抽出する。

【００５４】

図５に示すウェブページから情報を抽出する装置５００は図３に示す方法に対応する。よって、ウェブページから情報を抽出する装置５００の詳細は、図３におけるウェブページから情報を抽出する方法について説明において既に詳細に説明され、ここでその説明を省略する。

【００５５】

上記処理及び装置はソフトウェア及び／又はファームウェアにより実現されてもよい。ソフトウェア及び／又はファームウェアにより実施されている場合、記憶媒体又はネットワークから専用のハードウェア構成を有するコンピュータ（例えば図６示されている汎用パーソナルコンピュータ６００）に上記方法を実施するためのソフトウェアを構成するプログラムをインストールしてもよく、該コンピュータは各種のプログラムがインストールされている場合は各種の機能などを実行できる。

【００５６】

図６は本発明の実施形態に係る方法及び／又は装置を実現可能な汎用パーソナルコンピュータの例示的な構成を示すブロック図である。図６において、中央処理部（ＣＰＵ）６０１は、読み出し専用メモリ（ＲＯＭ）６０２に記憶されているプログラム、又は記憶部６０８からランダムアクセスメモリ（ＲＡＭ）６０３にロードされたプログラムにより各種の処理を実行する。ＲＡＭ６０３には、必要に応じて、ＣＰＵ６０１が各種の処理を実行するに必要なデータが記憶されている。ＣＰＵ６０１、ＲＯＭ６０２、及びＲＡＭ６０３は、バス６０４を介して互いに接続されている。入力／出力インターフェース６０５もバス６０４に接続されている。

【００５７】

入力部６０６（キーボード、マウスなどを含む）、出力部６０７（ディスプレイ、例えばブラウン管（ＣＲＴ）、液晶ディスプレイ（ＬＣＤ）など、及びスピーカなどを含む）、記憶部６０８（例えばハードディスクなどを含む）、通信部６０９（例えばネットワークのインタフェースカード、例えばＬＡＮカード、モデムなどを含む）は、入力／出力インターフェース６０５に接続されている。通信部６０９は、ネットワーク、例えばインターネットを介して通信処理を実行する。必要に応じて、ドライバ６１０は、入力／出力インターフェース６０５に接続されてもよい。取り外し可能な媒体６１１は、例えば磁気ディスク、光ディスク、光磁気ディスク、半導体メモリなどであり、必要に応じてドライバ６１０にセットアップされて、その中から読みだされたコンピュータプログラムは必要に応じて記憶部６０８にインストールされている。

【００５８】

ソフトウェアにより上記処理を実施する場合、ネットワーク、例えばインターネット、

又は記憶媒体、例えば取り外し可能な媒体６１１を介してソフトウェアを構成するプログラムをインストールする。

【００５９】

なお、これらの記憶媒体は、図６に示されている、プログラムを記憶し、機器と分離してユーザへプログラムを提供する取り外し可能な媒体６１１に限定されない。取り外し可能な媒体６１１は、例えば磁気ディスク（フロッピーディスク（登録商標）を含む）、光ディスク（光ディスク－読み出し専用メモリ（ＣＤ－ＲＯＭ）、及びデジタル多目的ディスク（ＤＶＤ）を含む）、光磁気ディスク（ミニディスク（ＭＤ）（登録商標））及び半導体メモリを含む。或いは、記憶媒体は、ＲＯＭ６０２、記憶部６０８に含まれるハードディスクなどであってもよく、プログラムを記憶し、それらを含む機器と共にユーザへ提供される。

【００６０】

本発明は、対応するコンピュータプログラムコード、機器が読み取り可能な命令コードが記憶されているコンピュータプログラムプロダクトをさらに提供する。該命令コードは、機器により読み取られ、実行される際に、上記の本発明の実施例に係る方法を実行することができる。

【００６１】

それに応じて、本発明は、機器が読み取り可能な命令コードを含むプログラムプロダクトが記録されている記憶媒体をさらに含む。該記憶媒体は、フロッピーディスク、光ディスク、光磁気ディスク、メモリカード、メモリスティック等を含むが、これらに限定されない。

【００６２】

また、上述の各実施例を含む実施形態に関し、更に以下の付記を開示する。

（付記１）

ウェブページから情報を抽出する方法であって、

前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成するステップと、

前記木構造におけるナビゲーションバーノードを決定するステップと、

前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するステップと、

マッチする葉ノードに対応するページにおける情報を抽出するステップと、を含む、方法。

（付記２）

統計的方法を用いて前記ナビゲーションバーノードを決定する、付記１に記載の方法。

（付記３）

前記木構造におけるナビゲーションバーノードを決定するステップは、

前記木構造に出現する回数が所定閾値よりも大きい葉ノードのみを含む非葉ノードを決定するステップと、

前記非葉ノードを並び替えて前記ナビゲーションバーノードを決定するステップと、を含む、付記２に記載の方法。

（付記４）

葉ノードの出現回数が所定閾値よりも大きいか否かを決定することは、

前記葉ノードのテキスト及び経路情報の前記木構造における出現回数が前記所定閾値よりも大きいか否かを決定すること、を含む、付記３に記載の方法。

（付記５）

前記経路情報は、前記葉ノードからそのｎ番目の先祖ノードまでの経路であり、ｎは正整数である、付記４に記載の方法。

（付記６）

ｎは５以上である、付記５に記載の方法。

（付記７）

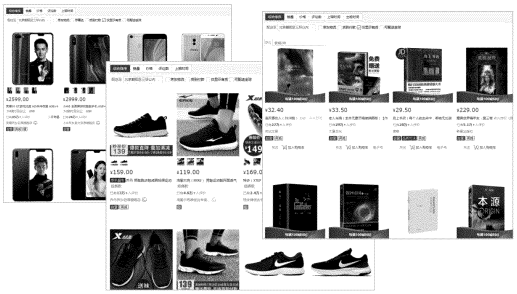
前記非葉ノードを並び替えて前記ナビゲーションバーノードを決定するステップは、
前記非葉ノードの特徴値を計算するステップであって、前記特徴値は、前記非葉ノードによりカバーされる葉ノードの数及び前記回数により決定される、ステップと、
前記非葉ノードのうちの最大の特徴値を有する非葉ノードを前記ナビゲーションバーノードとして決定するステップと、を含む、付記３に記載の方法。
（付記８）
前記特徴値は、前記非葉ノードによりカバーされる葉ノードの数と前記回数との積の、前記ウェブページのドメイン名を含むページの総数に対する比率である、付記７に記載の方法。
（付記９）
マッチする葉ノードに対応するページにおける情報を抽出するステップは、
前記マッチする葉ノードに対応するページに含まれるターゲットノードを決定するステップと、
前記ターゲットノードによりカバーされる各葉ノードのテキストをそれぞれ抽出するステップと、を含む、付記１乃至８の何れかに記載の方法。
（付記１０）
前記ターゲットノードは、
前記ターゲットノードに含まれる各葉ノードのテキスト及び経路情報の前記木構造における出現回数が前記所定閾値以下であること、
前記ターゲットノードが、前記木構造に出現する回数が所定閾値よりも大きい葉ノードのみを含む非葉ノードのうちの非葉ノードではないこと、及び
前記ターゲットノードに含まれる全ての葉ノードのテキストの合計長さが該木構造における他の非葉ノードのテキストの合計長さよりも大きいこと、により決定される、付記９に記載の方法。
（付記１１）
前記ターゲットノードによりカバーされる各葉ノードのテキストをそれぞれ抽出するステップは、
前記ターゲットノードに対応するページのタイプに応じて、異なる解析器を選択して抽出を行うステップ、を含む、付記９に記載の方法。
（付記１２）
前記ターゲットノードの各葉ノードを独立した属性抽出空間とする、付記１１に記載の方法。
（付記１３）
前記解析器は、エンティティ認識器、固有名詞認識器又は数値認識器である、付記１１に記載の方法。
（付記１４）
決定されたナビゲーションバーノードの経路情報を用いて前記ウェブページ及びその全ての拡張ウェブページにおけるナビゲーションバーノードを決定する、付記１乃至８の何れかに記載の方法。
（付記１５）
ＵＲＬトップレベルドメイン名を抽出することにより、前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含むページを決定する、付記１乃至８の何れかに記載の方法。
（付記１６）
前記木構造は、ＨＴＭＬ文書オブジェクトモデル（ＤＯＭ）である、付記１乃至８の何れかに記載の方法。
（付記１７）
前記キーワードは、所定のキーワードである、付記１乃至８の何れかに記載の方法。
（付記１８）
前記拡張ウェブページをｎ回だけ拡張して前記ウェブページのドメイン名を含むページ

を取得し、ｎは２以上の整数である、付記１乃至８の何れかに記載の方法。
（付記１９）
ウェブページから情報を抽出する装置であって、
前記ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成する木構造生成手段と、
前記木構造におけるナビゲーションバーノードを決定するナビゲーションバーノード決定手段と、
前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するマッチノード決定手段と、
マッチする葉ノードに対応するページにおける情報を抽出する情報抽出手段と、を含む、装置。
（付記２０）
プログラムを記憶したコンピュータ読み取り可能な記憶媒体であって、前記プログラムがプロセッサにより実行される際に、
ウェブページ及びその全ての拡張ウェブページにおける前記ウェブページのドメイン名を含む各ページについて木構造を生成するステップと、
前記木構造におけるナビゲーションバーノードを決定するステップと、
前記ナビゲーションバーノードによりカバーされる、１つ又は複数のキーワードにマッチする葉ノードを決定するステップと、
マッチする葉ノードに対応するページにおける情報を抽出するステップと、を実行させる、記憶媒体。
【００６３】
なお、用語「含む」、「有する」又は他の任意の変形は、排他的に含むことに限定されず、一連の要素を含むプロセス、方法、物又は装置は、これらの要素を含むことだけではなく、明示的に列挙されていない他の要素、又はこのプロセス、方法、物若しくは装置の固有の要素を含む。また、さらなる制限がない限り、用語「１つの…を含む」より限定された要素は、該要素を含むプロセス、方法、物又は装置に他の同一の要素が存在することを排除しない。
【００６４】
以上は図面を参照しながら本発明の好ましい実施例を説明しているが、上記実施例及び例は例示的なものであり、制限的なものではない。当業者は、特許請求の範囲の主旨及び範囲内で本発明に対して各種の修正、改良、均等的なものに変更してもよい。これらの修正、改良又は均等的なものに変更することは本発明の保護範囲に含まれるものである。

【図面】

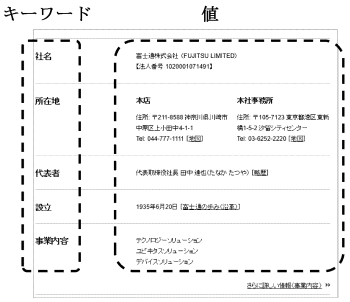
【図１Ａ】

類似の構造を有するウェブページの例を示す図

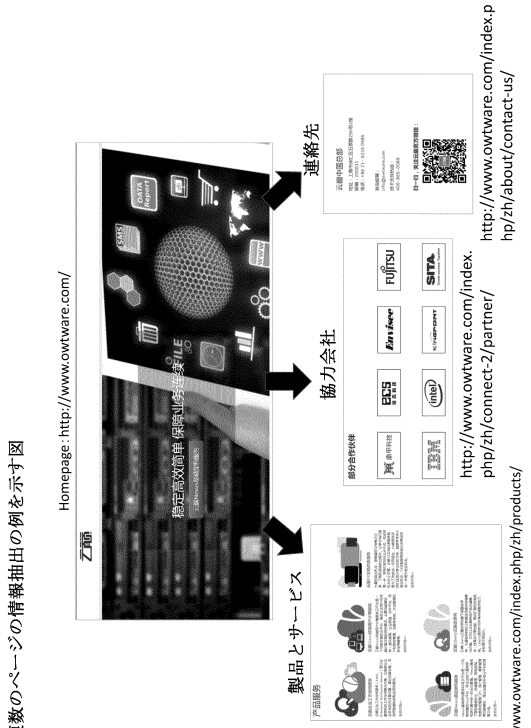


【図１Ｂ】

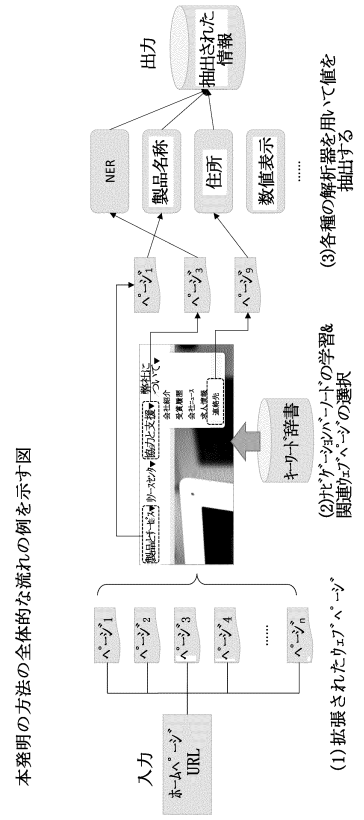
単一の構造を有するウェブページの情報抽出の例を示す図



【図２Ａ】

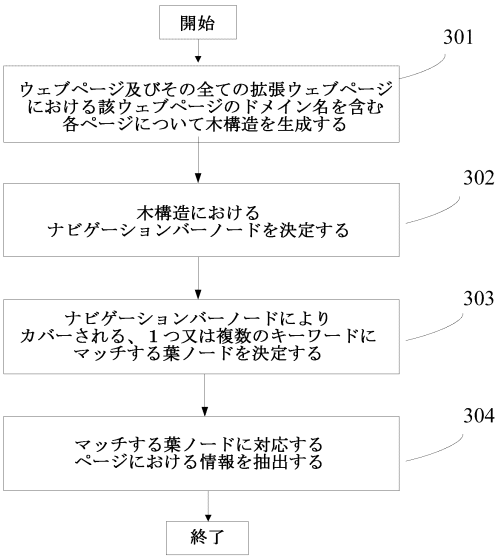


【図２Ｂ】

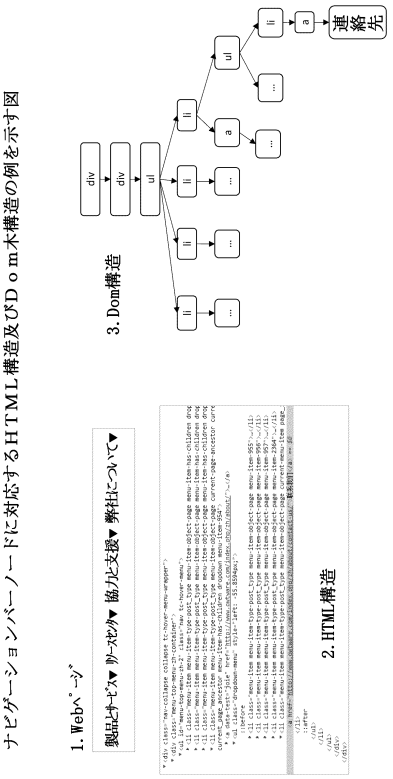


【図３】

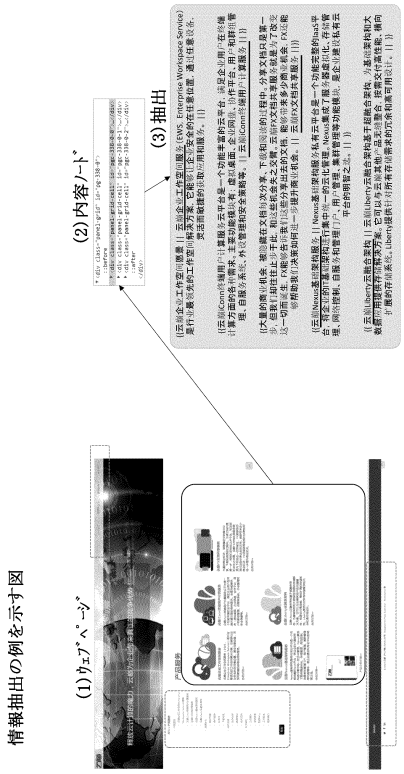
本発明の実施形態に係るウェブページから情報を抽出する方法の流れを示すフローチャート



【図４Ａ】

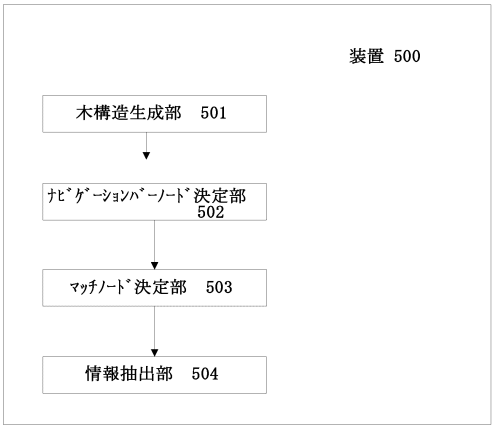


【図４Ｂ】

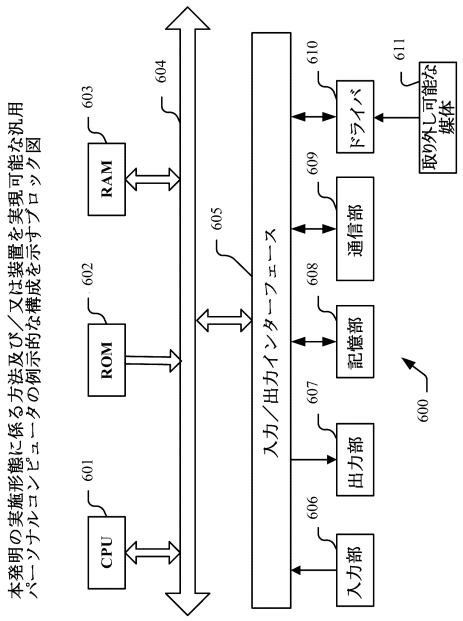


【図５】

本発明の実施形態に係るウェブページから情報を抽出する装置の例を示すブロック図



【図６】


本発明の実施形態に係る方法及び／又は装置を実現可能な汎用
パーソナルコンピュータの例示的な構成を示すブロック図
CPU
601
ROM
602
RAM
603
604
605
入力／出力インターフェース
606
入力部
607
出力部
608
記憶部
609
通信部
610
ドライバ
611
取り外し可能な
媒体
600

フロントページの続き

(72)発明者　孟 遥
中国，１０００２７，ベイジン，チャオヤン ディストリクト，ゴン ティ ベイ ルゥ ナンバー２ エイ，パシフィック センチュリー プレイス，スペース ８，ゲート ６，ユニット ３エフ，３５５ 富士通研究開発中心有限公司内
(72)発明者　孫 俊
中国，１０００２７，ベイジン，チャオヤン ディストリクト，ゴン ティ ベイ ルゥ ナンバー２ エイ，パシフィック センチュリー プレイス，スペース ８，ゲート ６，ユニット ３エフ，３５５ 富士通研究開発中心有限公司内

審査官　甲斐 哲雄
(56)参考文献　特開２００９－０４２９０８（ＪＰ，Ａ）
特開２０１６－２０１１１２（ＪＰ，Ａ）
中国特許出願公開第１０５０６９１０７（ＣＮ，Ａ）
中国特許出願公開第１０３８２３８２４（ＣＮ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１６／００－１６／９５８
Ｈ０４Ｌ　６７／０２