

(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2010-61638

(P2010-61638A)

(43) 公開日 平成22年3月18日(2010.3.18)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 17/30 (2006.01)	G 0 6 F 17/30 2 2 0 Z	5 B 0 7 5
	G 0 6 F 17/30 1 7 0 A	
	G 0 6 F 17/30 4 1 9 B	

審査請求 有 請求項の数 22 O L 外国語出願 (全 22 頁)

(21) 出願番号 特願2009-146095 (P2009-146095) (22) 出願日 平成21年6月19日 (2009. 6. 19) (31) 優先権主張番号 200810111482.2 (32) 優先日 平成20年6月26日 (2008. 6. 26) (33) 優先権主張国 中国 (CN) (特許庁注：以下のものは登録商標) 1. フロッピー	(71) 出願人 505418870 エヌイーシー (チャイナ) カンパニー、 リミテッド NEC (China) Co., Ltd. 中華人民共和国 100007 ベイジン , ドングチェングディストリクト, ドング シンチャオ エー22, ナシカング タワ ー ビー, スイート1222 (74) 代理人 100093595 弁理士 松本 正夫 (72) 発明者 ジャオ ユウ 中華人民共和国 100084 ベイジン , チンフア サイエンス パーク, イ ノベーション プラザ, ビルディング エー, 11エフ 最終頁に続く
---	--

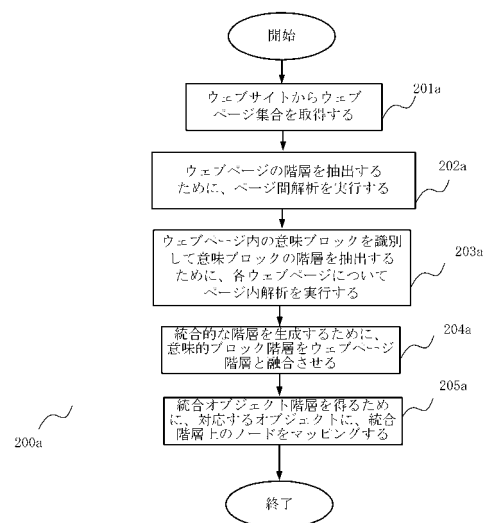
(54) 【発明の名称】 階層構築方法および階層構築システム

(57) 【要約】

【課題】 オブジェクト階層を構築するために方法とシステムを提供する。

【解決手段】 この階層構築方法は、ウェブサイトからウェブページ集合を取得し、ウェブページの階層を抽出するために、取得したウェブページについてページ間解析を行い、ウェブページ中の意味ブロックを識別し、全てのウェブページについて意味ブロックの階層を抽出するために、取得した各ウェブページについてページ内解析を行い、統合的な階層を生成するために、意味ブロックの階層をウェブページの階層と融合する。その後、統合的なオブジェクト階層を生成するために、生成された統合的な階層上のノードを対応するオブジェクトにマッピングする。本発明による階層構築システムおよび方法は、従来と比較し、ページ間解析結果およびページ内解析結果を融合させることにより、より正確でより効率的にオブジェクト階層を構築することが可能である。

【選択図】 図 1 B



【特許請求の範囲】**【請求項 1】**

ウェブサイトからウェブページ集合を取得するステップと、
前記ウェブページの階層を抽出するために、取得した前記ウェブページについてページ間解析を行なうステップと、
前記ウェブページ中の意味ブロックを識別し、全てのウェブページについて意味ブロックの階層を抽出するために、取得した前記各ウェブページについてページ内解析を行なうステップと、
統合的な階層を生成するために、前記意味ブロックの階層を前記ウェブページの階層と融合するステップとを有することを特徴とする階層構築方法。

10

【請求項 2】

統合的なオブジェクト階層を取得するために、統合的な階層上の各ノードを対応するオブジェクトにマッピングするステップをさらに有することを特徴とする請求項 1 に記載の階層構築方法。

【請求項 3】

前記ウェブページによって表わされる前記オブジェクトの階層を取得するために、前記ページ間解析の後に、ウェブページの階層上の各ノードを対応するオブジェクトにマッピングするステップと、
前記意味ブロックによって表わされるオブジェクトの階層を取得するために、前記ページ内解析の後に、前記意味ブロックの階層上の各ノードを対応するオブジェクトにマッピングするステップをさらに有し、
前記融合ステップで、統合的なオブジェクト階層を取得するために、前記ウェブページによって表わされるオブジェクトの階層と前記意味ブロックによって表わされるオブジェクトの階層とを融合することを特徴とする請求項 1 に記載の階層構築方法。

20

【請求項 4】

前記融合ステップが、
互いの競合を解決するために、前記ウェブページの階層と前記意味ブロックの階層を調整するステップと、
統合的な階層を生成するために、前記意味ブロックの階層に従って、前記ウェブページの階層に、前記意味ブロックを仮想ウェブページとして補完するステップを有することを特徴とする請求項 1 に記載の階層構築方法。

30

【請求項 5】

利用者が関心を持つオブジェクトタイプを入力するステップと、
取得した前記ウェブページから、入力された前記オブジェクトタイプを有するオブジェクト関連ウェブページをフィルタリングして取り出すステップとをさらに有し、
前記ページ間解析と前記ページ内解析を前記オブジェクト関連ウェブページ上で実行することを特徴とする請求項 1 に記載の階層構築方法。

【請求項 6】

前記フィルタリングステップが、
取得した前記ウェブページのハイパーリンクから階層型ハイパーリンクを識別するステップと、
識別した前記階層型ハイパーリンクを参照して、前記各ウェブページについて階層型ナビゲーション経路を生成するステップと、
生成した前記階層型ナビゲーション経路をチェックすることにより、前記オブジェクト関連ウェブページを識別するステップとを有することを特徴とする請求項 5 に記載の階層構築方法。

40

【請求項 7】

生成された階層型ナビゲーション経路に沿ってウェブページの言語的コンテンツを収集するステップをさらに含み、
前記チェックステップが、

50

前記オブジェクト関連ウェブページを識別するために、入力された前記オブジェクトタイプに従って収集したウェブページの前記言語コンテンツを問い合わせるステップを含むことを特徴とする請求項 6 に記載の階層構築方法。

【請求項 8】

前記ページ内解析を行なうステップが、
意味ブロックを生成するために、前記各ウェブページについてウェブページ分割を行なうステップと、

全ての前記ウェブページについて前記意味ブロックの階層を抽出するステップと、
前記各意味ブロックについてタイトルを生成するステップを含むことを特徴とする請求項 1 に記載の階層構築方法。

10

【請求項 9】

前記ページ内解析を行なうステップが、
種々のオブジェクト関連ウェブページに導くハイパーリンクのバンドルを含むオブジェクト開始ページを、取得したウェブページから選択するステップと、

前記意味ブロックを生成するために、選択した前記オブジェクト開始ページについてウェブページ分割を行なうステップと、

前記意味ブロックの階層を抽出するステップと、
前記各意味ブロックについてタイトルを生成するステップとを含むことを特徴とする請求項 5 に記載の階層構築方法。

20

【請求項 10】

前記タイトル生成ステップで、
前記タイトルのテキストが前記意味ブロックのリテラルコンテンツ中に含まれていない場合、前記意味ブロックが属するウェブページのページ内コンテキスト及びページ間コンテキストを利用して、前記タイトルを生成することを特徴とする請求項 8 又は請求項 9 に記載の階層構築方法。

【請求項 11】

前記マッピングステップが、
前記対応するオブジェクトのタイトルに各ノードのタイトルをマッピングするステップと、

前記オブジェクトの階層関係にノードの階層関係をマッピングするステップを含むことを特徴とする請求項 2 又は請求項 3 に記載の階層構築方法。

30

【請求項 12】

ウェブサイトから全てのウェブページを取得するウェブページ取得手段と、
前記ウェブページの階層を抽出するために、取得した前記ウェブページについてページ間解析を行なうページ間解析手段と、

前記ウェブページ中の意味ブロックを識別し、全てのウェブページについて意味ブロックの階層を抽出するために、取得した前記各ウェブページについてページ内解析を行なうページ内解析手段と、

統合的な階層を生成するために、前記意味ブロックの階層を前記ウェブページの階層と融合する融合手段とを備えることを特徴とする階層構築システム。

40

【請求項 13】

統合的なオブジェクト階層を取得するために、統合的な階層上の各ノードを対応するオブジェクトにマッピングするマッピング手段をさらに備えることを特徴とする請求項 12 に記載の階層構築システム。

【請求項 14】

前記ページ間解析手段に接続され、前記ウェブページによって表わされる前記オブジェクトの階層を取得するために、前記ページ間解析の後に、ウェブページの階層上の各ノードを対応するオブジェクトにマッピングする第 1 のマッピング手段と、

前記ページ内解析手段に接続され、前記意味ブロックによって表わされるオブジェクトの階層を取得するために、前記ページ内解析の後に、前記意味ブロックの階層上の各ノード

50

ドを対応するオブジェクトにマッピングする第2のマッピング手段をさらに備え、

前記融合手段が、統合的なオブジェクト階層を取得するために、前記ウェブページによって表わされるオブジェクトの階層と前記意味ブロックによって表わされるオブジェクトの階層とを融合することを特徴とする請求項12に記載の階層構築システム。

【請求項15】

前記融合手段が、

互いの競合を解決するために、前記ウェブページの階層と前記意味ブロックの階層を調整する調整ユニットと、

統合的な階層を生成するために、前記意味ブロックの階層に従って、前記ウェブページの階層に、前記意味ブロックを仮想ウェブページとして補完する補完ユニットを備えることを特徴とする請求項12に記載の階層構築システム。

10

【請求項16】

利用者が関心を持つオブジェクトタイプを入力するオブジェクトタイプ入力手段と、

取得した前記ウェブページから、入力された前記オブジェクトタイプを有するオブジェクト関連ウェブページをフィルタリングして取り出すフィルタ手段とをさらに備え、

前記ページ間解析手段と前記ページ内解析手段が、前記ページ間解析と前記ページ内解析を、前記オブジェクト関連ウェブページ上で実行することを特徴とする請求項12に記載の階層構築システム。

【請求項17】

前記フィルタ手段が、

取得した前記ウェブページのハイパーリンクから階層型ハイパーリンクを識別する階層型ハイパーリンク識別ユニットと、

識別した前記階層型ハイパーリンクを参照して、前記各ウェブページについて階層型ナビゲーション経路を生成する階層型ナビゲーション経路生成ユニットと、

生成した前記階層型ナビゲーション経路をチェックすることにより、前記オブジェクト関連ウェブページを識別するオブジェクト関連ウェブページ識別ユニットとを備えることを特徴とする請求項16に記載の階層構築システム。

20

【請求項18】

前記フィルタ手段が、

生成された階層型ナビゲーション経路に沿ってウェブページの言語的コンテンツを収集する収集ユニットをさらに備え、

前記オブジェクト関連ウェブページ識別ユニットが、

前記オブジェクト関連ウェブページを識別するために、入力された前記オブジェクトタイプに従って収集したウェブページの前記言語コンテンツを問い合わせるステップを含むことを特徴とする請求項17に記載の階層構築システム。

30

【請求項19】

前記ページ内解析手段が、

意味ブロックを生成するために、前記各ウェブページについてウェブページ分割を行なうウェブページ分割ユニットと、

全ての前記ウェブページについて前記意味ブロックの階層を抽出する階層抽出ユニットと、

前記各意味ブロックについてタイトルを生成するタイトル生成ユニットを備えることを特徴とする請求項12に記載の階層構築システム。

40

【請求項20】

前記ページ内解析手段が、

種々のオブジェクト関連ウェブページに導くハイパーリンクのバンドルを含むオブジェクト開始ページを、取得したウェブページから選択するオブジェクト開始ページ選択ユニットと、

前記意味ブロックを生成するために、選択した前記オブジェクト開始ページについてウェブページ分割を行なうウェブページ分割ユニットと、

50

前記意味ブロックの階層を抽出する階層抽出ユニットと、
前記各意味ブロックについてタイトルを生成するタイトル生成ユニットとを備えることを特徴とする請求項 16 に記載の階層構築システム。

【請求項 21】

前記タイトル生成ユニットは、
前記タイトルのテキストが前記意味ブロックのリテラルコンテンツ中に含まれていない場合、前記意味ブロックが属するウェブページのページ内コンテキスト及びページ間コンテキストを利用して、前記タイトルを生成することを特徴とする請求項 19 又は請求項 20 に記載の階層構築システム。

【請求項 22】

前記マッピング手段、又は前記第 1 のマッピング手段と前記第 2 のマッピング手段が、
前記対応するオブジェクトのタイトルに各ノードのタイトルをマッピングするタイトルマッピングユニットと、
前記オブジェクトの階層関係にノードの階層関係をマッピングする階層関係マッピングユニットとを備えることを特徴とする請求項 13 又は請求項 14 に記載の階層構築システム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ウェブからドメイン知識 (domain knowledge) を取得するための方法とシステムに関し、特に、ウェブから自動でオブジェクト階層 (object hierarchy) の構築と生成を可能にするシステムおよび方法に関する。

【背景技術】

【0002】

近年、コンピュータは、人々が関心のある情報を見つけ出すのを支援する現代生活の必要不可欠なツールとなっている。増大する莫大な量の多種多様な情報がウェブ上に蓄積されるインターネット時代においては、特に、そうである。コンピュータは計算、記憶あるいは検索のような情報処理においては高速であるが、情報理解における不適格さが、知的情報処理 (intelligent information processing) にとって大きな障害である。そのような問題に対処するために、知的情報処理のための意味的な関連の研究が最近一般的となっている。

例えば、関連する技術が、

ティム・バーナーズ・リー、ジェイムズ・ヘンドラー、オー・ラッシーラ (2001) 等による「セマンティックウェブ、サイエンティフィックアメリカン」、2001 年 5 月、28 - 37 頁 (T. Berners-Lee, J. Hendler, O. Lassila (2001), entitled "The Semantic Web, Scientific American", May 2001, pp. 28-37,) (非特許文献 1)、ナイジェル・シャボット、ティム・バーナーズ・リー、ウエンディー・ホール等による「再考されたセマンティックウェブ」インテリジェントシステムの I E E E、巻 21 号 (3)、96 - 101 頁、2006 年 5 月 / 6 月 (Nigel Shadbolt, Tim Berners-Lee and Wendy

Hall, entitled "The Semantic Web Revisited", IEEE Intelligent Systems 21(3) pp. 96-101, May/June 2006) (非特許文献 2)、Eero Hyvonen による「フィンランドにおけるセマンティックウェブの開始 - ビジョン、技術、調査およびアプリケーション」ヘルシンキ研究所情報技術公開、304 頁 (E. Hyvonen (editor), entitled "Semantic Web

10

20

30

40

50

Kick-Off in Finland - Vision, Technologies, Research, and Applications", HIIT Publications, 2002-001, Helsinki Institute for Information Technology (HIIT), Helsinki, Finland, 304 pp) (非特許文献3)

等に掲載されている。

それらはコンピューターによる情報理解を支援するための形態と技術に集中している。人工知能(AI)および一般的なウェブ情報処理技術の従来分野からの知識表現のための記述論理あるいはフレーム論理のような、ある数学の論理について基づいて、World Wide Web Consortium(W3C)のような標準化組織は、XML、RDF(Resource Description Framework)やOWL(Web Ontology Language)及びルール言語(例えば、ウェブ・ルール言語(Web Rule Language)、ルールマークアップ言語(Rule Markup Language))のような、セマンティック技術の採用を進めるための基礎として役立つ標準を積極的に明示している、

10

また、多くの開発者、企業家および技術者が、意味ベースの知的情報利用の考え方を実現させるために、関連するツールセット、製品、ケーススタディーおよび実際に動作するアプリケーションなどを生み出し展開する段階に入っている。

【0003】

しかしながら、ウェブユーザーに様々な知的情報利用サービスを提供するためにコンピューターの強力な計算性能およびセマンティック関連の標準を利用するには、バックエンドドメイン知識(現在、オントロジはウェブ上の知識表現のための支配された方法である)が重要な役割を果たす。このように、ドメイン知識構築は解決されなければならない重大な問題である。

20

【0004】

現在、主に2種類のドメイン知識が存在する。オントロジと階層である。

【0005】

オントロジは、形式的に用語の関係を規定する文書あるいはファイルである。また、ウェブのための最も代表的な種類のオントロジは、概念分類(taxonomy)と推論規則(inference rule)集合を有する。さらに、概念分類は、オブジェクトのクラスとそれらの間の関係を定義する。例えば、住所は、位置のタイプとして定義され、都市名コードは位置にのみ適用するために定義される。オントロジは、「都市名コードが州コードに対応付けられ、住所がその都市名コードを使用するならば、住所は対応付けられた州コードを持つ。」というような規則を示す。それから、プログラムは、例えば、イサカにあるコーネル大学の住所が、米国のニューヨーク州にあるはずであり、従って、米国標準にフォーマットされるべきであると、容易に推測する。

30

【0006】

階層は、ノードと、ノードを接続するエッジを含み、時にノードに附随するインスタンスを含んでいる。オントロジと比較して、階層は非常に簡単な形式である。オントロジにおける、クラス、特性、定義および関係のような多くの要素は、階層において無視することが可能である。しかし、階層からそれらの要素を推論するいくつかの方法がある。したがって、階層は、明示的であるが略式の仕様を持った一種の擬オントロジと見なすことができる。

40

【0007】

関連技術として、主に2種類のオントロジ構築(OB)方法が存在する。すなわち、ある原料に基づいたオントロジ構築方法と、ある既存のオントロジに基づいたオントロジ構築方法である。原料ベースのオントロジ構築方法においては、例えば、オントロジは、テキスト、辞書、知識ベース、半構造化データあるいは関係スキーマから構築することが可能である。既存のオントロジベースのオントロジ構築方法においては、テキストあるいは概念のコンテキストを比較することによって、いくつかの既存のオントロジを一つに統合することが可能である。

【0008】

オントロジはセマンティックウェブ(Semantic Web)および関連サービスにとって重大

50

であるが、オントロジが、クラス、クラス定義、クラスの関係、特性などのような人間によってさえ満足するのが難しい多くのコンテンツを通常含んでいるので、形式的オントロジ (formal

ontology) を自動的に構築するのはとにかく困難である。言うまでもなく、オントロジの複雑な構成は、あるリアルタイムウェブ・サービスのような、その大規模構造そして広く普及したアプリケーションを阻害した。さらに、オントロジ統合は、人間の相互作用を通して通常実行される。そのため、階層統合のように簡単に実現することができない。

【 0 0 0 9 】

また、階層構築 (HB) について若干の関連技術が存在する。例えば、特開 2 0 0 1 - 3 4 6 3 5 号公報 (特許文献 1) は、ウェブから階層を構築する方法を開示している。具体的には、1つの用語 (すなわち1つのノード) は各ウェブページから抽出され、また、階層的関係はウェブページ間のリンクに基づいて構築される。全てのページの関係構築する代わりに、この方法では、同じタイプのウェブページについてのみ関係を構築する。例えば、2つの製品ページ間のリンクは保持されるが、製品ページと広告ページ間のリンクは無視される。

また、N. リウ、C. C. ヤン等による、「ウェブサイト・トピック階層生成へのアプローチに基づくリンク分類」(N. Liu, C.C. Yang, entitled "A link classification based

approach to website topic hierarchy generation" (WWW2007)) (非特許文献 4) においては、ページ間リンク構成解析に基づいてウェブサイト中のウェブページ間の階層的関係を抽出する方法が提供されている。そして、各ウェブページをトピック・オブジェクトで包み、トピック階層を構築する。

上述した特許文献 1 及び非特許文献 4 の開示内容は、専ら参照の目的でここに取り入れられる。

【 先行技術文献 】

【 特許文献 】

【 0 0 1 0 】

【 特許文献 1 】 特開 2 0 0 1 - 3 4 6 3 5 号公報

【 非特許文献 】

【 0 0 1 1 】

【 非特許文献 1 】 ティム・バーナーズ・リー、ジェイムズ・ヘンドラー、オー・ラッシーラ (2 0 0 1) 等による「セマンティックウェブ、サイエンティフィックアメリカン」、2 0 0 1 年 5 月、2 8 - 3 7 頁 (T. Berners-Lee, J. Hendler, O. Lassila(2001), entitled "The Semantic Web, Scientific American", May 2001, pp. 28-37,)

【 非特許文献 2 】 ナイジェル・シャボット、ティム・バーナーズ・リー、ウエンディー・ホール等による「再考されたセマンティックウェブ」インテリジェントシステムの I E E E、巻 2 1 号 (3)、9 6 - 1 0 1 頁、2 0 0 6 年 5 月 / 6 月 (Nigel Shadbolt, Tim Berners-Lee and WendyHall, entitled "The Semantic Web Revisited", IEEE Intelligent Systems 21(3) pp.96-101, May/June 2006)

【 非特許文献 3 】 E e r o H y v o n e n による「フィンランドにおけるセマンティックウェブの開始 - ビジョン、技術、調査およびアプリケーション」ヘルシンキ研究所情報技術公開、3 0 4 頁 (E. Hyvonen (editor), entitled "Semantic WebKick-Off in Finland - Vision, Technologies, Research, and Applications", HIITPublications, 2002-001, Helsinki Institute for Information Technology (HIIT), Helsinki, Finland, 304 pp)

【 非特許文献 4 】 N. リウ、C. C. ヤン等による、「ウェブサイト・トピック階層生成へのアプローチに基づくリンク分類」(N. Liu, C.C. Yang, entitled "A link classification based approach to website topic hierarchy generation" (WWW2007))

【 発明の概要 】

【 発明が解決しようとする課題 】

【 0 0 1 2 】

しかしながら、H Bに関する関連技術（特許文献 1 及び非特許文献 1 に記載の技術）において、既存の方法は、オブジェクト／トピックがページ全体によって表わされる場合を考慮するだけであり、また、オブジェクト／トピック中の関係はページ間ハイパーリンク解析によって取得される。しかしながら、オブジェクト／トピック（階層のノード）の一部分だけがページ全体によって表され、オブジェクトの他の部分はウェブページのいくつかの部分によって覆われているだけである。さらに、ページ間のリンク中には階層的関係以外に多くのノイズが存在するので、ページ間関係のみから抽出されたハイパーリンクは、十分に正確とは言えない。

【 課題を解決するための手段 】

10

【 0 0 1 3 】

関連技術における H B 方法の欠陥を考慮し、本発明は、より正確でより有効な方法でウェブサイトからオブジェクト（例えば製品）の階層を自動的に抽出するためになされたものである。

【 0 0 1 4 】

本発明においては、ページ間解析（すなわち、ウェブページの階層の解析）をページ内解析（すなわち、ウェブページ内の意味ブロックの関係についての解析）と統合することによる、ウェブサイトからの自動階層抽出のための統合された方法を提案する。ページ内部の意味ブロック内に包含された階層的関係は、ページ間解析からのみ生ずる不正確な階層を修正するために利用される。

20

【 0 0 1 5 】

より具体的には、本発明の統合的な階層抽出方法は、主に、（ 1 ）ページ間階層解析、（ 2 ）ページ内階層解析、そして（ 3 ）統合階層生成の、 3 つの工程を含んでいる。

【 0 0 1 6 】

ページ間階層解析中、階層は、ウェブサイトの全ページ集合の意味関係解析に基づいて生成される。一方、ネストされたオブジェクトは、ウェブサイトから抜粋され、各トピックをその代表的なページと結び付ける。他方、ウェブページ間の階層関係は、ハイパーリンクとコンテンツの解析を統合する、ハイパーリンクベースの方法あるいはハイブリッド方法で識別される。したがって、オブジェクト階層は、ウェブページ間のオブジェクトとページのペアおよび階層関係の統合により抽出される。

30

【 0 0 1 7 】

その後、ページ内階層解析において、階層は、ウェブページ内部の意味ブロック解析に基づいて生成される。意味ブロック解析は、オブジェクト表示ページへ案内するハイパーリンクのバンドルを有する各ページについて実行される。また、それは、意味ブロック間のハイパーリンクおよび階層的関係を含むネストされた意味ブロックを取り出す。これらのネストされた意味ブロックもオブジェクトとして結び付けられる。したがって、新たなオブジェクト集合の階層は、オブジェクトとページのペア、オブジェクトとブロックのペアおよび意味ブロック間の階層的関係の統合により抽出することができる。

【 0 0 1 8 】

最後に、洗練されたオブジェクト階層が、ページ間解析とページ内解析の結果を融合させることにより生成される。実施の形態において、融合動作は互いに不当な階層的関係を調整し互いに欠けている階層的関係を補完することを含むことが可能である。もちろん、ページ間解析とページ内解析の結果に対する融合動作が説明された実施例への制限されていないと当業者のために考えるのは簡単である。

40

【 0 0 1 9 】

さらに、前述の説明は本発明の原理を簡潔に説明するためにのみ利用されるものであり、本発明を限定するものではない。例えば、前述の例においては、ウェブページとオブジェクトおよび意味ブロックとオブジェクトのマッピング動作は、ページ間解析とページ内解析の工程においてそれぞれ実行されるとして区別されている。しかしながら、他のいくつかの実施の形態では、ウェブページの階層と意味ブロックのネストされた関係（これら

50

はページ間解析とページ内解析の結果として取得される)は、最初に融合され、それから、統合的な階層上のノード(ウェブページあるいは意味的なブロック)を、最終的なオブジェクト階層を得るためにオブジェクトにマッピングすることが可能である。

【0020】

本発明の階層構築方法は、ウェブサイトからウェブページ集合を取得するステップと、ウェブページの階層を抽出するために、取得したウェブページについてページ間解析を行なうステップと、ウェブページ中の意味ブロックを識別し、全てのウェブページについて意味ブロックの階層を抽出するために、取得した各ウェブページについてページ内解析を行なうステップと、統合的な階層を生成するために、意味ブロックの階層をウェブページの階層と融合するステップとを有する。

10

【0021】

本発明の階層構築システムは、ウェブサイトから全てのウェブページを取得するウェブページ取得手段と、ウェブページの階層を抽出するために、取得したウェブページについてページ間解析を行なうページ間解析手段と、ウェブページ中の意味ブロックを識別し、全てのウェブページについて意味ブロックの階層を抽出するために、取得した各ウェブページについてページ内解析を行なうページ内解析手段と、統合的な階層を生成するために、意味ブロックの階層をウェブページの階層と融合する融合手段とを備える。

【発明の効果】

【0022】

まず、本発明がオントロジではなく階層に注目しているので、ドメイン知識構築の多くの現実の事例を処理することが可能になる。さらに、本発明は、ウェブサイトにおける既存の非公式、半形式上の知識の再使用を容易にし、できるだけ世界/ドメインの共通の理解を反映することが可能である。

20

【0023】

さらに、本発明に適用される統合オブジェクト階層抽出方法は、ページ間解析ベースの方法あるいはページ内解析ベースの方法のいずれかより、高精度で階層を取得することが可能である。ページ間解析およびページ内解析の結果は、互いに調整し補完することが可能である。

【0024】

また、本発明において適用したページ内解析が、オブジェクト表示ページ(ページ間解析中に識別されるページ)に導くハイパーリンクのバンドルを有するページについてのみ実行することが可能であるので、ウェブページの全てのページについてページ内解析を実行するより、非常に効率の高い処理が可能である。

30

【0025】

本発明の上述した特徴と他の特徴及び効果は、添付図面と組み合わせた以下の説明からより明らかになると思われる。本発明の範囲が、ここに説明された例あるいは特定の実施の形態に限定されないことは言うまでもない。

【図面の簡単な説明】

【0026】

本発明の前述した特徴及び他の特徴は、添付図面と共に以下の説明を参照することでよりよく理解されるであろう。

40

【図1A】本発明の第1の実施の形態による統合オブジェクト階層構築システム100aの内部構成例を示すブロック図である。

【図1B】図1Aにおいて示した統合オブジェクト階層構築システム100aの動作を説明するためのフローチャートである。

【図2A】本発明の第2の実施の形態による統合オブジェクト階層構築システム100bの内部構成例を示すブロック図である。

【図2B】図2Aにおいて示した統合オブジェクト階層構築システム100bの動作を説明するためのフローチャートである。

50

【図 3 A】本発明の第 3 の実施の形態による統合オブジェクト階層構築システム 1 0 0 c の内部構成例を示すブロック図である。

【図 3 B】図 3 A において示した統合オブジェクト階層構築システム 1 0 0 c の動作を説明するためのフローチャートである。

【図 4】本発明の第 3 の実施の形態による統合オブジェクト階層構築システム 1 0 0 c に含まれるオブジェクト関連ウェブページの識別のためのフィルタ手段 3 0 2 のさらに詳細な内部構成例を示すブロック図である。

【図 5】ページ内階層解析を実行するページ内解析手段 1 0 3 の内部構成例を示すブロック図である。

【図 6】意味ブロックタイトル抽出の処理、および融合処理とマッピング処理について説明する概略図である。

【図 7】本発明による統合オブジェクト階層構築システムの中に含まれる融合手段とマッピング手段のさらに詳細な内部構成例を示すブロック図である。

【図 8】本発明を実現するために利用されるコンピュータ・システムの概略を示すブロック図である。

【発明を実施するための形態】

【0 0 2 7】

以下、本発明の実施の形態について、添付図面を参照して説明する。以下に述べる実施の形態は例として説明するものであり、本発明の範囲を制限するものとして考慮されるべきでないことは言うまでもない。

【0 0 2 8】

本発明は、知識抽出、管理および利用のためのシステムおよび方法についてなされている。特に、本発明は、例えば、ウェブサイトのウェブページ集合を考慮することにより、非常に正確で効率的なオブジェクト階層抽出を実現する方法およびシステムを提供する。もちろん、本発明の適応例が以下に説明する実施例に限定されるものでなく、他の知識源からのドメイン知識の解析および管理に同様に利用することが可能であることは当業者であれば理解できるであろう。

【0 0 2 9】

まず、図 1 A は、本発明の第 1 の実施の形態による統合オブジェクト階層構築システム 1 0 0 a の内部構成例を示すブロック図であり、図 1 B は、図 1 A に示す統合オブジェクト階層構築システム 1 0 0 a の動作について説明するフローチャートである。

図 1 A において示すように、統合オブジェクト階層構築システム 1 0 0 a の中核部分は、オブジェクト階層構築モジュール 1 0 a にある。オブジェクト階層構築モジュール 1 0 a は、ウェブページ記憶装置 1 0 8 から、ウェブサイトのウェブページ集合を取得し、処理の後、ウェブサイトに対するオブジェクト階層 L を構築し、その後、オブジェクト階層 L をオブジェクト階層記憶装置 1 0 9 に格納する。図示しないウェブサイトクローラアプリケーション (website crawling application) は、インターネットから 1 つ以上のウェブサイトのウェブページ集合をダウンロードし、階層抽出のためにウェブページ記憶装置 1 0 8 に取得したウェブページを格納する。ウェブページ解析モジュール 1 1 0 は、ウェブページ中のハイパーリンク情報を抽出し、ハイパーリンク記憶装置 1 1 1 に抽出した情報を格納するために、ウェブページ記憶装置 1 0 8 内のウェブページを解析するのに利用される。

図示のように、オブジェクト階層構築モジュール 1 0 a は、ウェブページ取得手段 1 0 1 と、ページ間解析手段 1 0 2 と、ページ内解析手段 1 0 3 と、融合手段 1 0 4 およびマッピング手段 1 0 5 とを含んでいる。これらの構成要素に加えて、オブジェクト階層構築モジュール 1 0 a は、ページ間解析結果を格納するためのウェブページ階層記憶装置 1 0 6 およびページ内解析結果を格納するための意味的ブロック記憶装置 1 0 7 を含んでいる。

【0 0 3 0】

図 1 B のフローチャートを参照すると、まずステップ 2 0 1 a において、ウェブページ

10

20

30

40

50

取得手段 101 が、ウェブサイトからウェブページ集合を取得する。例えば、ウェブページ取得手段 101 は、ウェブサイトの全てのウェブページを取得することが可能である。そして、ページ間解析手段 102 およびページ内解析手段 103 は、ハイパーリンク記憶装置 111 に格納されたウェブページのハイパーリンク情報を参照して、取得したウェブページについてページ間解析とページ内解析をそれぞれ実行する。そして、ページ間解析結果として抽出されるウェブページの階層をウェブページ階層記憶装置 106 に格納する。また、ページ内解析結果として抽出される意味ブロック、意味ブロックの階層および意味ブロックのタイトルを全て意味ブロック記憶装置 107 に格納する（ステップ 202 a およびステップ 203 a）。その後、ステップ 204 a において、融合手段 104 は、統合的な階層を生成するためにウェブページ階層と意味的ブロック階層を融合させる。ステップ 205 a において、その後、マッピング手段 105 は、統合的なオブジェクト階層を得るために、対応するオブジェクトに、統合的な階層上のノード（ウェブページあるいは意味的ブロック）をマッピングする。その統合的なオブジェクト階層は、オブジェクト階層記憶装置 109 に格納される。後述するように、階層のマッピングはオブジェクトのタイトルにノードのタイトルをマッピングすること、オブジェクトの階層関係にノードの階層関係をマッピングすることを含む。

最終的に生成された統合オブジェクト階層は、オブジェクト（例えば、製品）関連であり、各ノードによって表わされるオブジェクトは、ウェブページあるいはウェブページ内の意味ブロックである。

【0031】

オブジェクト階層記憶装置 109 に格納された様々なウェブサイトに対するオブジェクト階層は、後で、種々の階層関連アプリケーション（図示せず）によって利用される。階層関連アプリケーションは、様々なウェブサイトから抽出された階層を統合し、整合させるための階層統合アプリケーションのようなものである。

【0032】

図 2 A および 2 B は、本発明の第 2 の実施の形態による統合オブジェクト階層構築システム 100 b とその動作処理を示す。第 1 の実施の形態によるシステム 100 a と比較すると、第 2 の実施の形態においては、マッピング手段 105 が融合手段 104 の前に置かれ、ページ間解析とページ内解析のそれぞれに対する 2 つのマッピング手段（第 1 のマッピング手段 105 1 と第 2 のマッピング手段 105 2）として構成されている。第 1 のマッピング手段 105 1 は、ページ間解析手段 102 の後に置かれ、ウェブページによって表わされるオブジェクトの階層を構築するために、対応するオブジェクトに、ページ間解析結果として取得されるウェブページの階層上のノード（すなわち、ウェブページ）をマッピングする。第 2 のマッピング手段 105 2 は、ページ内解析手段 103 の後に置かれ、意味的なブロックによって表わされるオブジェクトの階層を構築するために、対応するオブジェクトに、ページ内解析結果として取得される意味的なブロックの階層上のノード（すなわち、意味的なブロック）をマッピングする。その後、ウェブページによって表わされるオブジェクトの階層、および意味ブロックによって表わされるオブジェクトの階層は、第 1 のマッピング手段 105 1 および第 2 のマッピング手段 105 2 から、融合動作のための融合手段 104 に出力される。融合手段 104 において、2 つの階層が、統合的なオブジェクト階層 L を生成するために融合される。第 1 の実施の形態と同様に、統合的なオブジェクト階層 L は、オブジェクト階層記憶装置 109 に格納される。

【0033】

図 2 B は、図 2 A に示す統合オブジェクト階層構築システム 100 b の動作を説明するためのフローチャートである。図 1 B と比較して、第 1 と第 2 の実施の形態間の違いが、第 1 と第 2 マッピングステップ 203 b および 205 b にあることが分かる。さらに、ウェブページ - オブジェクトマッピング処理および意味ブロック - オブジェクトマッピング処理が、ページ間解析とページ内解析において既に実行されているので、融合ステップ 206 b の後、統合的なオブジェクト階層 L が直ちに生成される。

【0034】

第 1 の実施の形態と類似する図 2 A に示される他の構成要素及びおよび図 2 B において示される他のステップに関しては、それらの説明を分かりやすくするために説明を省略する。

【 0 0 3 5 】

さらに、図 3 A および 3 B はより有効な他の実施の形態を示している。本発明の目的がオブジェクトに関連する階層を生成することであるので、ページ間解析中に、ウェブページ取得手段 1 0 1 によって取得されたウェブページの集合からオブジェクト関連のウェブページをまず検索し、次に階層関係を判定するために分析し処理する必要があるオブジェクト関連のウェブページだけを検索することが重要である。詳細については、図 3 A および 3 に示されている。図 3 A は、本発明の第 3 の実施の形態による統合オブジェクト階層構築システム 1 0 0 c の内部構成例を示すブロック図であり、図 3 B は、図 3 A において示すシステム 1 0 0 c の動作を説明するためのフローチャートである。

【 0 0 3 6 】

図 1 に示す第 1 の実施の形態と比較すると、第 1 と第 2 実施の形態に類似する構成要素に加えて、図 3 A に示すシステム 1 0 0 c におけるオブジェクト階層構築モジュール 1 0 c は、オブジェクトタイプ入力手段 3 0 1 およびフィルタ手段 3 0 2 を含んでいる。まず図 3 B のフローチャートを参照すると、第 1 と第 2 実施の形態と同様に、まずステップ 2 0 1 c で、ウェブページ取得手段 1 0 1 は、ウェブページ記憶装置 1 0 8 からウェブサイトのウェブページの集合を取得する。ステップ 2 0 2 c で、利用者は関心のあるオブジェクトタイプをオブジェクトタイプ入力手段 3 0 1 によって入力する。その後、フィルタ手段 3 0 2 は、ウェブページ取得手段 1 0 1 によって取得したウェブページから利用者が関心を持つオブジェクトタイプを有するウェブページを、オブジェクト関連のウェブページとして取り出すフィルタリングを行う（ステップ 2 0 3 c）。ステップ 2 0 4 c で、オブジェクト関連のウェブページの階層を抽出するために、ページ間解析手段 1 0 2 は、フィルタリングして得たオブジェクト関連のウェブページについてだけページ間解析を実行する。同様に、ページ内解析について、ページ内解析手段 1 0 3 は、ページ内意味ブロック解析を行うために、オブジェクト関連のウェブページにハイパーリンクのバンドルを有するページだけを選択することが可能である（ステップ 2 0 5 c）。次に、第 1 の実施の形態と同様に、融合手段 1 0 4 は、統合的な階層を生成するために、ステップ 2 0 4 c で構築されたウェブページの階層と、ステップ 2 0 5 c で構築された意味的なブロックの階層とを融合させる（ステップ 2 0 6 c）。そして、ステップ 2 0 7 c で、マッピング手段 1 0 5 は、統合的なオブジェクト階層を構築するために、対応するオブジェクトに、統合的な階層上の各ノードをマッピングする。その後、処理が終了する。

【 0 0 3 7 】

図 3 A において示すシステムは、図 1 A で示した第 1 の実施の形態のシステムに基づいているが、対応オブジェクトタイプ入力手段 3 0 1 およびフィルタ手段 3 0 2 をシステム 1 0 0 b に追加すれば、図 2 A に示す第 2 の実施の形態に第 3 の実施の形態の技術原理を同様に適用することが可能であることは当業者にとって明白である。

【 0 0 3 8 】

図 4 は、オブジェクト関連ウェブページを識別するためのフィルタ手段 3 0 2 のさらに詳細な内部構成例を示すブロック図である。図示のように、この例において、フィルタ手段 3 0 2 は、階層型ハイパーリンク識別ユニット 4 0 1、階層型ナビゲーション経路生成ユニット 4 0 2、オブジェクト関連ウェブページ識別ユニット 4 0 3 および収集ユニット 4 0 4 を含む。この実施例において、オブジェクト関連ウェブページのフィルタリングは、階層型ナビゲーション経路（HNP：hierarchical navigation path）ベースの方法で実行することができる。もちろん、この HNP 方法は、あくまで一例として説明している。オブジェクト関連ページのフィルタリングを実行するために、他の適切な既存方法を採用することが可能であることは、当業者にとって容易に理解できる。

【 0 0 3 9 】

基本的に、HNP は特定のウェブサイトと関連付けられる。それは、ウェブサイトのル

10

20

30

40

50

ートページから行き先ページに利用者の操作を案内するための仮のナビゲーション経路を構成するウェブページ間の階層的関係を有するそれらのハイパーリンクのマルチステップを意味する。HNPの構造ハイパーリンク（ここでは、階層的ハイパーリンク（HL：hierarchical

hyperlink）と称する）は、ピア・ツー・ピア勧告を示唆する参照ハイパーリンクと異なると共に、あるページから別のページに対してただショートカットを提供する単なるナビゲーションハイパーリンクとも異なる。代わりに、HLはウェブページ構築のために利用され、ウェブページ間のある種の階層的関係（例えば、全体と一部あるいは親と子）を埋め込む。親ページの意味は、連続するHL（すなわち、HNP）に沿って子ページに継承される。このように、HNPは、その行き先ウェブページのコンテンツ上に意味のある指標を与える。

10

【0040】

図4を参照すると、階層型ハイパーリンク識別ユニット401は、ウェブサイト内のすべてのハイパーリンクからHLを識別するために利用される。

一例として、階層型ハイパーリンク識別ユニット401は、単なるナビゲーションハイパーリンク（すなわち、HLに付随する雑音情報（例えば、直接又は間接の兄弟間のハイパーリンクや上向きのハイパーリンク））を除去するためのアルゴリズムを採用する。

このアルゴリズムは、1)URL構文の解析と、2)意味的なハイパーリンク解析の2つのステップを含んでいる。

ステップ1は、URL文法（すなわち、ハイパーリンクのソースと行き先のウェブページ間に階層的関係があると識別できる`http://[host]/[path]/[file]#[fragment]`に包含される情報）を利用する。

20

その後、意味的なハイパーリンク解析のためのステップ2で、ウェブページ集合P1内のウェブページが同じリンク集から生じ、これらのページが共通の下向きのページ集合P2を有する場合、P1が同じ階層的レベルの兄弟ページである確率が高く、かつ、P2が、P1（P1内のページは互いにリンクされている）内に含まれ、あるいはP1と同じ親ページを共有する可能性が高いという、規則が採用される。

従って、P1からP2へのハイパーリンクは非HLであると見なされる。ここで、リンク集とは、1つのウェブページ中に同じレイアウトとプレゼンテーションプロパティを持つリンクの集合を意味する。それは、通常、ページの意味的なブロックのうちの1つを表わす。

30

上述したアルゴリズムは、単に、階層的ハイパーリンク識別の一例として利用されるものであり、本発明を制限するものではない。

【0041】

ウェブサイト中の全てのHLを識別した後、階層型ナビゲーション経路生成ユニット402は、ウェブサイト中の各ウェブドキュメントについてHNPを生成する。

同時に、収集ユニット404は、そのHNPに沿って、URL、アンカー・テキストおよびウェブページタイトルを含むHNP内の言語コンテンツを収集する。

【0042】

そして、ナビゲーション経路が、階層型ナビゲーション経路生成ユニット402によって生成された後、オブジェクト関連ウェブページ識別ユニット403は、オブジェクト関連ウェブページを検索するかあるいはオブジェクトに無関係のウェブページをフィルタリングするために、予め入力されたオブジェクトタイプ名あるいはその同義語でHNPのテキストノードを問い合わせることにより、経路問合せを行なう。例えば、利用者が会社ウェブサイトから製品ウェブページを抽出したい場合、HNPは、「製品」や「サービス」等のキーワードなどで問い合わせる。あるページのHNPのいくつかのノードがこれらのキーワードを含んでいると、HNPが目的ページの正確に意味のあるコンテキストを含んでいるので、そのページを可能性のあるオブジェクト関連ウェブページと見なすことができる。そのようなオブジェクト関連ウェブページを一連のネストされたオブジェクトの代表的なページと見なすことができるかもしれない。また、対応するウェブページのタイト

40

50

ル、および対応するウェブページに導くハイパーリンクのアンカー・テキストからオブジェクトの名前を要約することが可能である。

【0043】

オブジェクト関連ウェブページがフィルタ手段302によってフィルタリングされた後、これらのオブジェクト関連ウェブページは、ページ間解析とページ内解析のために、ページ間解析手段102およびページ内解析手段103に供給される。

【0044】

以上、本発明の第1、第2および第3の実施の形態に従って、統合的オブジェクト階層構築システムおよび方法の全体構成および原理について、添付図面を参照して説明した。3つの特徴、すなわちページ間階層解析（ページ間解析手段102）、ページ内階層解析（ページ内解析手段103）、統合的なオブジェクト階層の生成（第1の実施の形態における融合手段104およびマッピング手段105、あるいは第2の実施の形態における融合手段104、第1のマッピング手段105.1および第2のマッピング手段105.2）が、上述したシステムの重要な技術的特徴であることが理解できるであろう。

これらの特徴は、後でより詳細に説明する。

【0045】

まず、ページ間階層解析（すなわち、ページ間解析手段102の動作）に関しては、当業者によって周知の様々な方法を利用することにより実現することが可能である。例えば、オブジェクト関連ウェブページの処理の場合には、階層型ハイパーリンク識別ユニット401によって識別された階層的ハイパーリンクを利用することが可能である。その結果、2つのオブジェクト関連ウェブページが、連続する階層的ハイパーリンクによってリンクされるなら、それらは親子ペアと見なされ、それらの間の階層的関係が格納される。もちろん、当業者によって知られているように、先行技術には、本発明に適用することが可能な多くのページ間解析方法が存在する。ウェブページの階層を抽出するために、利用者は実際の適用条件に従って適切な方法を選択することが可能である。

【0046】

ページ内階層解析に関しては、上述したように、ページ内解析手段103が、各ウェブページをいくつかのネストされた意味ブロックに分割し、かつ意味ブロックの階層を抽出するために利用される。ページ内階層解析処理についても、当業者によって周知の様々な方法を利用することによって実現することが可能である。ここで、ページ内階層解析の例を図5を参照して説明する。

【0047】

図5は、ページ内階層解析を実行するページ内解析手段103の内部構成例を示すブロック図である。図示のように、この実施例において、ページ内解析手段103は、オブジェクト開始ページ選択ユニット501と、ウェブページ分割ユニット502と、階層抽出ユニット503と、タイトル生成ユニット504とを含む。

【0048】

まず、オブジェクト開始ページ選択ユニット501は、ウェブページ取得手段101によって取得されたウェブページからオブジェクト開始ページを選択する。オブジェクト開始ページとは、様々なオブジェクト関連ウェブページに導くハイパーリンクの束を含んでいるページである。その後、ウェブページ分割ユニット502は、ページからネストされた意味ブロックを生成するために、選択されたオブジェクト開始ページについてウェブページ分割を行なう。さらに効率を向上させるためには、ウェブページ分割ユニット502は、以下の階層抽出のために、オブジェクト関連ウェブページに導くハイパーリンクを含む意味ブロックだけを選択することも可能である。ウェブページ分割については、DOMパターン反復ベースの方法（DOM pattern repetition based method）あるいはビジョンレイアウトベースの方法（vision layout based method）のようないくつかの既存の方法によって実現することが可能である。

これらの既存の方法の詳細はここでは説明しない。意味ブロックの分割の後、階層抽出ユニット503は、意味ブロックの階層を抽出する。その後、タイトル生成ユニット5

10

20

30

40

50

04は、個々の意味ブロックについてタイトルを生成する。

【0049】

1例として、意味ブロックのタイトル生成は、個々の意味ブロックに対するタイトルを、意味ブロックのブロックが配置されるページについてのページ内コンテキストおよび意味ブロックのブロック内部の下向きリンクのあて先ページについてのページ間コンテキストの両方の分析および合成で識別するハイブリッドコンテキストベースの方法によって実現することが可能である。例えば、図6にはその一例を示している。この例において、2つの意味ブロックが、セキュリティ製品ウェブページで、「アンチウイルス」と「アンチスパム」として分割されている。ここでは、意味ブロック「アンチスパム」の点線で囲んだタイトルを抽出する必要がある。意味ブロックのタイトルについて、意味ブロックのリテラルコンテンツからそのテキストを直接抽出することが可能であれば、タイトルを簡単に取得することが可能である。しかしながら、そのようなテキストが存在しないか、テキストがイメージに埋め込まれている場合、意味ブロックのタイトルを要約するために、ページ内コンテキストとページ間コンテキストの両方を利用する。例えば、図6において、意味ブロック「アンチスパム」のタイトルを要約するために、ページ内コンテキスト（意味ブロック内のハイパーリンク「サーバ」および「クライアント」のアンカーテキスト）およびページ間コンテキスト（それら2つのハイパーリンクの行き先ページ「サーバアンチスパム製品リストページ」と「クライアントアンチスパム製品リストページ」のタイトル）の両方を利用することが可能である。

【0050】

図5に戻ると、最後に、分割された意味ブロック、意味ブロックの抽出された階層および意味ブロックの生成されたタイトルはすべて、意味的ブロック記憶装置107に格納される。

【0051】

ページ間階層解析およびページ内階層解析が行われた後、融合手段104は、統合的な階層を生成するためにページ間解析結果およびページ内解析結果を融合させる。図7は、融合手段とマッピング手段の内部構成の詳細を示すブロック図である。図7に示す実施例において、融合手段は、キャリブレーションユニット701および補完ユニット702を含んでいる。キャリブレーションユニット701は、衝突を解決するためにウェブページの階層と意味ブロックの階層を相互に調整するように構成されている。補完ユニット702は、統合階層を生成するために、意味ブロックの階層に従ってウェブページの階層に仮想ウェブページとして意味ブロックを補完するように構成されている。キャリブレーションユニット701については、多くの既存の階層統合方法を様々な階層間の調整を実現するために利用することができる。したがって、ここではその詳細は説明しない。一方、本発明の目的がオブジェクト階層を取得することであり、かつ、多くのオブジェクトがページ全体とは別に、ページの一部（例えば、意味ブロック）によって表わされるので、意味ブロック結果（すなわち、ページ内解析結果）から、ページ間階層解析によって生成されたオブジェクト階層にそのようなオブジェクトおよび他のオブジェクトとの関係を補完すべきである。例えば、図6に示す例において、ページ間解析によって生成されたウェブページの階層は、意味ブロック「アンチスパム」によって表わされるオブジェクトを考慮していない。しかし、融合処理の後、統合階層L'においては、この意味ブロックが他の2つのオブジェクト関連ウェブページ（つまり、「サーバアンチスパム製品リストページ」と「クライアントアンチスパム製品リストページ」）に対するハイパーリンクを含んでいるので、意味ブロック「アンチスパム」が、新たなノードとして、ウェブページ階層に補完されている。

【0052】

最後に、融合手段104によって生成された統合階層L'は、マッピング手段105によって対応する統合オブジェクト階層にマッピングされる。図7に示すように、この実施例において、マッピング手段105は、タイトルマッピングユニット703と階層関係マッピングユニット704を含んでいる。タイトルマッピングユニット703は、ウェブペ

ージあるいはノードによって表わされる意味ブロックのタイトルを、対応するオブジェクトのタイトルにマッピングするように構成されている。階層関係マッピングユニット704は、ウェブページあるいはノードによって表わされる意味ブロックの階層関係を、対応するオブジェクトの階層関係にマッピングするように構成されている。その後、マッピング手段105によって生成された統合的なオブジェクト階層は、他の階層関係に適用するためにオブジェクト階層記憶装置109に格納される。

【0053】

図8は、本発明を実現するために利用されるコンピュータシステム800の概略ブロック図である。図示のように、コンピュータシステム800は、CPU801、ユーザインターフェース802、周辺装置803、メモリ805、外部記憶装置806および上記構成要素を互いに接続する内部バス804を含んでいる。メモリ805は、さらに、ウェブサイトクロール取得モジュール、オブジェクト階層構築モジュール、階層関連アプリケーションモジュール、ウェブページ解析モジュール、オペレーティング・システム(OS)等を含んでいる。本発明は、主にオブジェクト階層構築モジュール、例えば、図1A、図2Aおよび図3Aに示される、オブジェクト階層構築モジュール10a、10bおよび10cに関連している。ウェブサイトクロール取得モジュールは、ネットワークからのウェブページを取得し、ウェブページ記憶装置にそれらを格納するために利用される。ウェブページ解析モジュールは、ウェブページのハイパーリンク関係を抽出するために取得したウェブページを解析する。抽出されたハイパーリンク関係は、ハイパーリンク記憶装置の中に格納される。外部記憶装置806は、ウェブページ記憶装置108、ハイパーリンク記憶装置111、ウェブページ階層記憶装置106、意味的ブロック記憶装置107およびオブジェクト階層記憶装置109のような、本発明に関する様々なデータベースを含んでいる。

【0054】

上記のように、第1、第2及び第3の実施の形態による統合オブジェクト階層構築システムおよび方法について、添付図面を参照して説明した。本発明の方法及びシステムは、関連技術と比較して、以下のような効果を有する。

【0055】

まず、本発明がオントロジではなく階層に注目しているので、ドメイン知識構築の多くの現実の事例を処理することが可能になる。さらに、本発明は、ウェブサイトにおける既存の非公式、半形式上の知識の再使用を容易にし、できるだけ世界/ドメインの共通の理解を反映することが可能である。

【0056】

さらに、本発明に適用される統合オブジェクト階層抽出方法は、ページ間解析ベースの方法あるいはページ内解析ベースの方法のいずれかより、高精度で階層を取得することが可能である。ページ間解析およびページ内解析の結果は、互いに調整し補完することが可能である。

【0057】

また、本発明において適用したページ内解析が、オブジェクト表示ページ(ページ間解析中に識別されるページ)に導くハイパーリンクのバンドルを有するページについてのみ実行することが可能であるので、ウェブページの全てのページについてページ内解析を実行するより、非常に効率の高い処理が可能である。

【0058】

本発明の特定の実施の形態について、上記のように添付の図面を参照して説明した。しかしながら、本発明は、添付の図面中で示される特定の構成および処理に限定されない。上記の実施の形態において、いくつかの特定のステップは具体例として示されかつ説明されている。しかしながら、本発明の方法処理はこれらの特定のステップに限定されない。当業者は、これらのステップを変更し、修正し、補足することが可能であり、あるいは、いくつかのステップの順序を、本発明の精神および本質的な機能から外れずに変更することが可能であることを理解するだろう。

10

20

30

40

50

【 0 0 5 9 】

本発明の要素は、ハードウェア、ソフトウェア、ファームウェアあるいはその組合せにおいて実装することが可能であり、システム、サブシステム、コンポーネントあるいはサブコンポーネントにおいて利用することが可能である。ソフトウェアの中で実施された場合、本発明の要素は、必要なタスクを実行するためのプログラム、あるいはコードセグメントである。プログラムまたはコードセグメントは、コンピュータ読み取り可能な媒体に格納するか、あるいは伝送ケーブルが通信リンク上の搬送波に包含されたデータ信号によって送信することが可能である。コンピュータ読み取り可能な媒体には、情報を格納するか転送することが可能であるすべての媒体を含む。コンピュータ読み取り可能な媒の具体例は、電子回路、半導体記憶装置、ROM、フラッシュ・メモリー、消去可能ROM (EROM)、フロッピー・ディスク、CD-ROM光ディスク、ハードディスク、光ファイバー媒体、無線周波数(RF)リンクなどを含む。コードセグメントは、インターネット、イントラネットなどのようなコンピュータネットワークを経由してダウンロードすることも可能である。

10

【 0 0 6 0 】

以上、特定の実施の形態を参照して本発明を説明したが、本発明は、図面中で示される上記の特定の実施の形態および特定の構成に限定されない。例えば、示されたいくつかの構成要素は、1つの構成要素としてお互いと組み合わせるかもしれない。あるいは、1つの構成要素はいくつかのサブコンポーネントに分割されるかもしれないし、他の既知の構成要素も加えられるかもしれない。動作処理も実施例において示されるものに限定されない。当業者は、本発明が、本発明の精神および本質的な機能から外れずに、他の特定の形態で実装可能であることを理解するだろう。従って、現在の実施の形態は、全ての点において例示でありかつ限定的でないとして考慮されるべきである。本発明の範囲は、前述の説明によってではなく添付された請求項によって示される。また、したがって、請求項と同等の意味と範囲の内で生ずる変更は全て本発明の範囲に包含される。

20

【 符号の説明 】

【 0 0 6 1 】

1 0 0 a : 統合オブジェクト階層構築システム

1 0 a : オブジェクト階層構築モジュール

1 0 8 : ウェブページ記憶装置

30

1 0 9 : オブジェクト階層記憶装置

1 1 0 : ウェブページ解析モジュール

1 1 1 : ハイパーリンク記憶装置

1 0 1 : ウェブ・ページ取得手段

1 0 2 : ページ間解析手段

1 0 3 : ページ内解析手段

1 0 4 : 融合手段

1 0 5 : マッピング手段

1 0 6 : ウェブページ階層記憶装置

1 0 7 : 意味ブロック記憶装置

40

1 0 0 b : 統合オブジェクト階層構築システム

1 0 b : オブジェクト階層構築モジュール

1 0 5 1 : 第1のマッピング手段

1 0 5 2 : 第2のマッピング手段

1 0 0 c : 統合オブジェクト階層構築システム

1 0 c : オブジェクト階層構築モジュール

3 0 1 : オブジェクトタイプ入力手段

3 0 2 : フィルタ手段

4 0 1 : 階層型ハイパーリンク識別ユニット

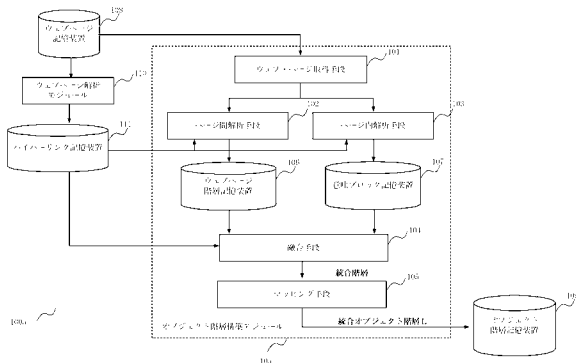
4 0 2 : 階層型ナビゲーション経路生成ユニット

50

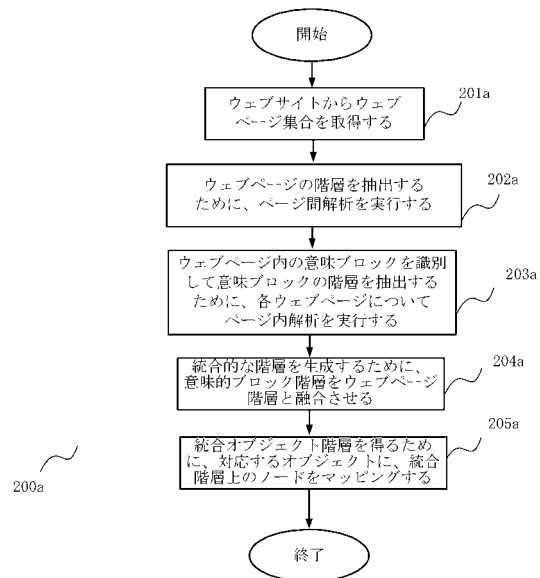
4 0 3 : オブジェクト関連ウェブページ識別ユニット
4 0 4 : 収集ユニット
5 0 1 : オブジェクト開始ページ選択ユニット
5 0 2 : ウェブページ分割ユニット
5 0 3 : 階層抽出ユニット
5 0 4 : タイトル生成ユニット
7 0 1 : キャリブレーションユニット
7 0 2 : 補完ユニット
7 0 3 : タイトルマッピングユニット
7 0 4 : 階層関係マッピングユニット
8 0 0 : コンピュータシステム
8 0 1 : C P U
8 0 2 : ユーザインターフェース
8 0 3 : 周辺装置
8 0 4 : 内部バス
8 0 5 : メモリ
8 0 6 : 外部記憶装置

10

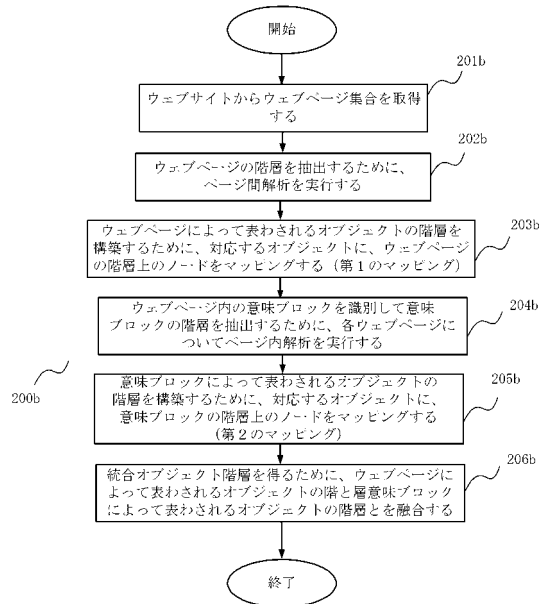
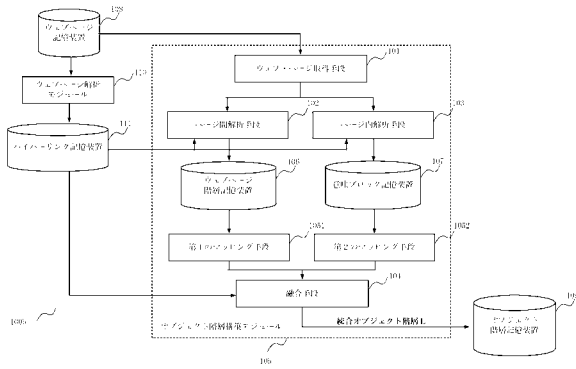
【 図 1 A 】



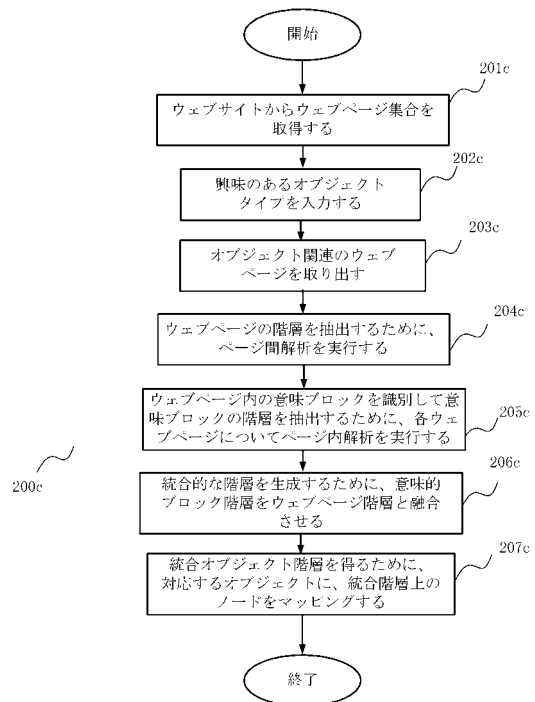
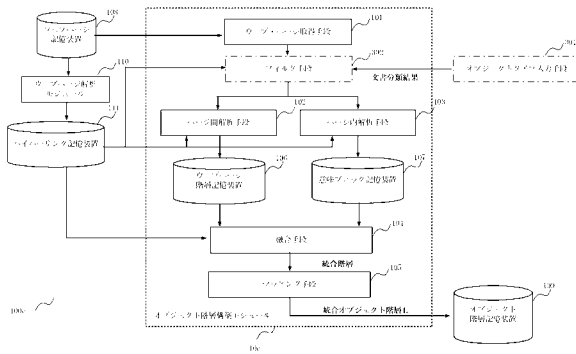
【 図 1 B 】



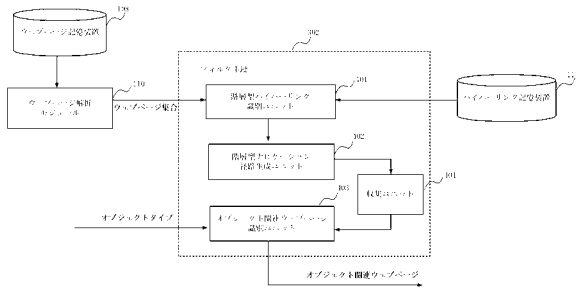
【 図 2 B 】



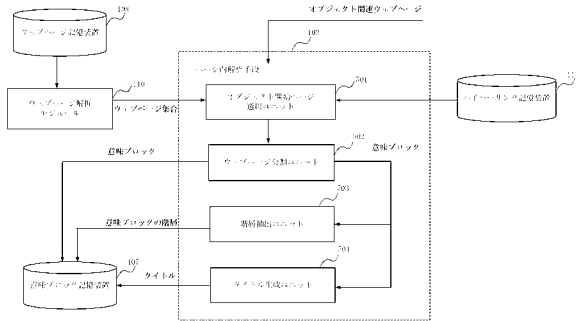
【 ㄨ 3 B 】



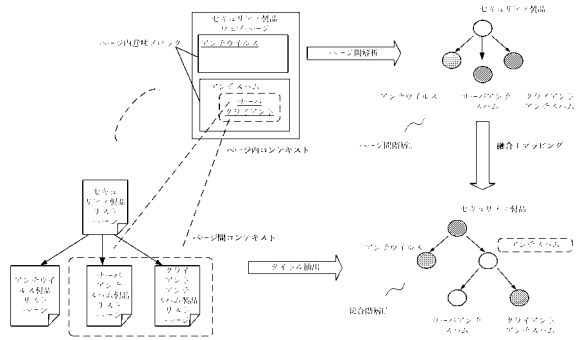
【図 4】



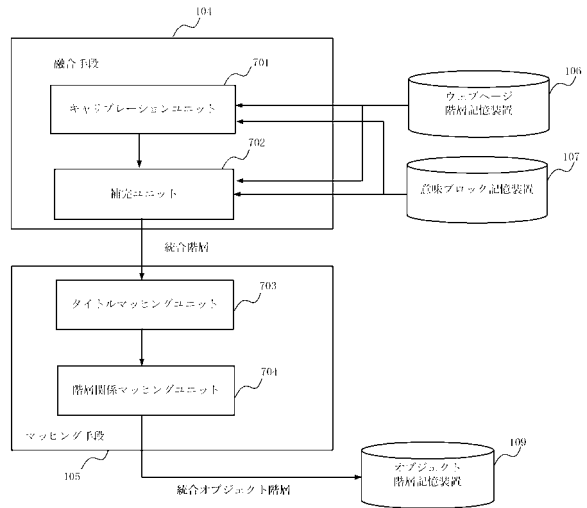
【図 5】



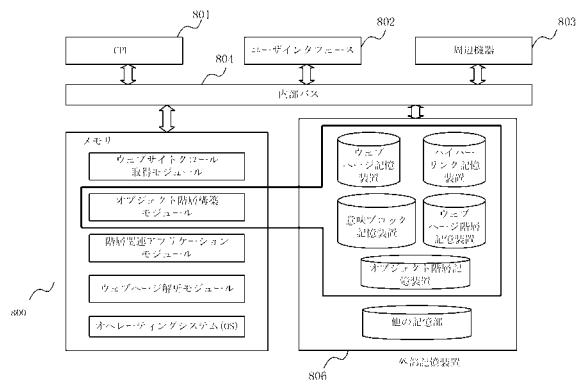
【図 6】



【図 7】



【図 8】



フロントページの続き

(72)発明者 リイ ジェンチャン

中華人民共和国 100084 ベイジン, チンファ サイエンス パーク, イノベーション
プラザ, ビルディング エー, 11エフ

Fターム(参考) 5B075 KK02 ND03 ND36 UU40

【外国語明細書】
2010061638000001.pdf