

(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2019-159404

(P2019-159404A)

(43) 公開日 令和1年9月19日 (2019.9.19)

(51) Int.Cl.	F I	テーマコード (参考)
G06F 16/00 (2019.01)	G06F 17/30 210A	
	G06F 17/30 419B	
	G06F 17/30 350C	

審査請求 未請求 請求項の数 10 O L (全 16 頁)

(21) 出願番号	特願2018-41087 (P2018-41087)	(71) 出願人	500257300
(22) 出願日	平成30年3月7日 (2018.3.7)		ヤフー株式会社
			東京都千代田区紀尾井町1番3号
特許法第30条第2項適用申請有り ・ ウェブサイトの		(74) 代理人	100106909
アドレス: (URL: http://www.anlp.jp/nlp2018/proceedings/link.html)、		弁理士	棚井 澄雄
ウェブサイトの掲載日: 平成30年3月5日、		(74) 代理人	100149548
言語処理学会 第24回年次大会 発表論文集、第412~415頁、		弁理士	松沼 泰史
「ウェブ検索クエリに対する教師なしエンティティリンク」		(74) 代理人	100154852
		弁理士	酒井 太一
		(74) 代理人	100174986
		弁理士	林 康旨
		(72) 発明者	齋藤 智輝
			東京都千代田区紀尾井町1番3号 ヤフー株式会社内

最終頁に続く

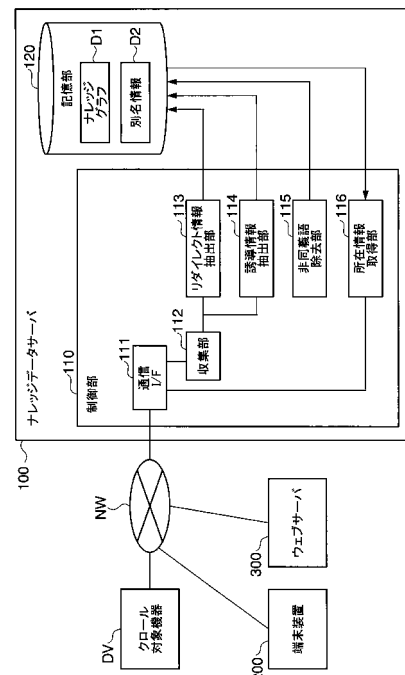
(54) 【発明の名称】 情報処理装置、情報処理方法、およびプログラム

(57) 【要約】

【課題】エンティティと関連付けされた別名情報を収集することが可能な情報処理装置、情報処理方法、およびプログラムを提供すること。

【解決手段】ウェブページのリダイレクト情報を収集する収集部と、前記収集部により収集された前記リダイレクト情報に基づいて、第1エンティティの名前を示す第1ワードに関連付けられた第1ページへのリダイレクト元である、第2ページに関連付けされた第2ワードを、前記第1エンティティの別名情報として抽出する抽出部とを備える情報処理装置。

【選択図】図1



【特許請求の範囲】**【請求項 1】**

ウェブページのリダイレクト情報を収集する収集部と、

前記収集部により収集された前記リダイレクト情報に基づいて、第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページへのリダイレクト元である、第 2 ページに関連付けされた第 2 ワードを、前記第 1 エンティティの別名情報として抽出する抽出部と

を備える情報処理装置。

【請求項 2】

第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページを収集する収集部と、

前記収集部により収集された前記第 1 ページに第 2 ページへの誘導情報がある場合、前記第 1 ワードを、前記第 2 ページに関連付けられた第 2 ワードにより示される第 2 エンティティの別名情報として抽出する抽出部と

を備える情報処理装置。

【請求項 3】

前記別名情報のうち、所定の条件を満たさないワードを除去する除去部をさらに備える

、

請求項 1 または 2 に記載の情報処理装置。

【請求項 4】

前記除去部は、前記第 1 エンティティの名前を示す前記第 1 ワードと関連付けられたウェブページにおいて、別名に関する情報の存在を示唆するキーワードが含まれる一文に含まれていないワードを除去する、

請求項 3 に記載の情報処理装置。

【請求項 5】

前記除去部は、前記第 1 エンティティの名前を示す前記第 1 ワードと関連付けられたウェブページにおいて、前記第 1 エンティティの名前の言い換え表現として示されるワードに含まれていないワードを除去する、

請求項 3 または 4 に記載の情報処理装置。

【請求項 6】

検索ワードと合致するワードが前記別名情報に含まれている場合、前記ワードと関連付けられたエンティティに対応付けられている所在情報を取得する取得部をさらに備える、

請求項 1 から 5 のいずれか一項に記載の情報処理装置。

【請求項 7】

コンピュータが、

ウェブページのリダイレクト情報を収集し、

収集された前記リダイレクト情報に基づいて、第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページへのリダイレクト元である、第 2 ページに関連付けされた第 2 ワードを、前記第 1 エンティティの別名情報として抽出する、

情報処理方法。

【請求項 8】

コンピュータが、

第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページを収集し、

収集された前記第 1 ページに第 2 ページへの誘導情報がある場合、前記第 1 ワードを、前記第 2 ページに関連付けられた第 2 ワードにより示される第 2 エンティティの別名情報として抽出する、

情報処理方法。

【請求項 9】

コンピュータに、

ウェブページのリダイレクト情報を収集させ、

10

20

30

40

50

収集された前記リダイレクト情報に基づいて、第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページへのリダイレクト元である、第 2 ページに関連付けされた第 2 ワードを、前記第 1 エンティティの別名情報として抽出させる、

プログラム。

【請求項 10】

コンピュータに、

第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページを収集させ、

収集された前記第 1 ページに第 2 ページへの誘導情報がある場合、前記第 1 ワードを、前記第 2 ページに関連付けられた第 2 ワードにより示される第 2 エンティティの別名情報として抽出させる、

プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、情報処理装置、情報処理方法、およびプログラムに関する。

【背景技術】

【0002】

従来、ウェブからデータ（文書や画像など）を収集し、収集したデータを自動的にデータベース化するクロウラが知られている。クロウラは、ウェブページ中のリンクを辿って、様々な IP アドレスのウェブページからデータを収集する。クロウラによって収集されたデータは、ウェブ情報データベースに蓄積される。一方、検索エンジンは、ユーザによって入力された検索ワード（以下、「検索クエリ」とも言う）を受信すると、受信した検索ワードに関連するウェブページやアプリページの情報（例えば、URL：Uniform Resource Locator）をウェブ情報データベースから取得し、取得した情報の一覧を検索結果として出力する。

【0003】

検索エンジンは、ユーザ満足度を向上させるため、ユーザによって入力された検索クエリに関連するデータを検索結果とともに出力するように運用されている場合がある。これを実現するために、テキストデータとエンティティとを互いに対応付けて記憶したデータベースが用いられる場合がある（特許文献 1 参照）。データベースにおいて、ある特定のエンティティと対応付けられているテキストデータが検索クエリとして入力されると、その特定のエンティティが検索結果とともに出力される。

【先行技術文献】

【特許文献】

【0004】

【特許文献 1】特開 2017-76403 号公報

【発明の概要】

【発明が解決しようとする課題】

【0005】

ユーザによって入力される検索クエリは、エンティティの正式名称以外に、該エンティティのあだ名（ニックネーム）、略称、通称、自称などの別名を含む場合がある。これらの別名とエンティティとが対応付けられていない場合、エンティティに関する情報が取得できず、或いはユーザが意図しない不適切なエンティティに関する情報が取得されて検索結果とともに出力される場合がある。

【0006】

本発明は、このような事情を考慮してなされたものであり、エンティティと関連付けられた別名情報を収集することが可能な情報処理装置、情報処理方法、およびプログラムを提供することを目的の一つとする。

【課題を解決するための手段】

【0007】

10

20

30

40

50

本発明の一態様は、ウェブページのリダイレクト情報を収集する収集部と、前記収集部により収集された前記リダイレクト情報に基づいて、第1エンティティの名前を示す第1ワードに関連付けられた第1ページへのリダイレクト元である、第2ページに関連付けられた第2ワードを、前記第1エンティティの別名情報として抽出する抽出部とを備える情報処理装置である。

【発明の効果】

【0008】

本発明の一態様によれば、エンティティと関連付けされた別名情報を収集することができる。

【図面の簡単な説明】

10

【0009】

【図1】本実施形態に係るナレッジデータサーバの使用環境及び構成を示す図である。

【図2】本実施形態に係るナレッジグラフの一例を示す図である。

【図3】本実施形態に係る「野球選手A」のウェブページの一例を示す図である。

【図4】本実施形態に係る検索結果ウェブページの一例を示す図である。

【図5】本実施形態における百科事典ウェブページのリダイレクト処理の一例を説明する図である。

【図6】本実施形態における百科事典ウェブページのリダイレクト情報の一例を示す図である。

【図7】本実施形態における別名情報の一例を示す図である。

20

【図8】本実施形態における収集部により収集された百科事典ウェブページの一例を示す図である。

【図9】本実施形態における非同義語除去部による除去処理の一例を説明する図である。

【図10】本実施形態における非同義語除去部による除去処理の他の例を説明する図である。

【図11】本実施形態のナレッジデータサーバのリダイレクト情報に基づく別名情報抽出動作の一例を示す流れ図である。

【図12】本実施形態のナレッジデータサーバの誘導情報に基づく別名情報抽出動作の一例を示す流れ図である。

【図13】本実施形態のナレッジデータサーバの別名情報の非同義語除去動作の一例を示す流れ図である。

30

【図14】本実施形態のナレッジデータサーバのオンライン動作の一例を示す流れ図である。

【発明を実施するための形態】

【0010】

以下、図面を参照し、本発明の情報処理装置、情報処理方法、及びプログラムの実施形態について説明する。本実施形態においては、情報処理装置が、ナレッジデータサーバの一部を構成するものとして説明する。ナレッジデータサーバは、例えば、収集対象データ（例えば、画像やテキストデータ）を収集し、収集したデータに基づいてナレッジグラフを生成するサーバである。本実施形態において、ナレッジグラフは、エンティティやクラス、プロパティと、エンティティに関連する関連データが記載されているウェブページの所在情報とが関連付けられたデータである。エンティティとは、例えば、ある対象事物の実体（例えば、実世界で存在している物体）を表していてもよいし、ある対象事物の概念（例えば、実世界または仮想世界の中で定義された概念）を表していてもよい。例えば、対象事物が「建物」という概念である場合、エンティティは、「○○塔」や「○○ビルディング」などといった実体を表してよい。また、例えば、対象事物が「経済学」という概念である場合、エンティティは、「ミクロ経済学」や「マクロ経済学」などといった実体のない抽象的な概念を表してよい。以下、実施形態について詳細に説明する。

40

【0011】

<実施形態>

50

< 1 - 1 . ナレッジデータサーバの使用環境 >

図 1 は、本実施形態に係るナレッジデータサーバ 100 の使用環境及び構成を示す図である。ナレッジデータサーバ 100 は、端末装置 200 と、ウェブサーバ 300 と、クローラ対象機器 DV と、ネットワーク NW を介して通信可能に接続される。ネットワーク NW は、ワールドワイドウェブ (World Wide Web) を意味し、インターネットやイントラネットと標準的に用いられる HTML 文書などを利用したシステムである。ネットワーク NW は、無線基地局やプロバイダ装置、専用回線などを更に含んでよい。

【 0012 】

端末装置 200 は、ユーザによって使用される装置であり、例えば、スマートフォンなどの携帯電話、タブレット型コンピュータ、ノート型コンピュータ、デスクトップ型コンピュータなどである。ウェブサーバ 300 は、検索エンジンを用いて検索結果を表示するためのウェブページを生成し、端末装置 200 に提供するサーバである。

【 0013 】

ナレッジデータサーバ 100 は、例えば、制御部 110 と、記憶部 120 とを備える。記憶部 120 は、例えば、RAM (Random Access Memory)、HDD (Hard Disk Drive)、フラッシュメモリ、またはこれらのうち複数組み合わせられたハイブリッド型記憶装置などにより実現される。また、記憶部 120 の一部または全部は、NAS (Network Attached Storage) や外部のストレージサーバなど、ナレッジデータサーバ 100 がアクセス可能な外部装置であってもよい。記憶部 120 には、例えば、ナレッジグラフ D1 と、別名情報 D2 とが記憶される。

【 0014 】

図 2 は、本実施形態に係るナレッジグラフ D1 の一例を示す図である。ナレッジグラフ D1 において記述された事物は、オントロジーによって定義される。オントロジーとは、事物のクラスおよびプロパティを定義したものであり、クラスとプロパティとの間に成り立つ制約を集めたものである。

【 0015 】

クラスとは、オントロジーにおいて、同じ性質を持つ事物同士を一つのグループにしたものである。事物の性質がどういったものであるのか、すなわち事物がどのクラスに属するのかは、後述するプロパティにより決定される。

【 0016 】

例えば、くちばしを持ち、卵生の脊椎動物であり、前肢が翼になっている、という性質を持つ事物は、「鳥」というクラスに分類される。また、「鳥」というクラスの中で、飛べない、という性質を持つ事物は、例えば、「ペンギン」や「ダチョウ」という、より下位のクラスに分類される。このように、クラスの体系は、上位と下位の関係を有する階層構造となっていてよい。上位のクラスの性質は、下位のクラスに継承される。上述した例では、「鳥」というクラスの、「くちばしを持ち、卵生の脊椎動物であり、前肢が翼になっている」という性質は、「ペンギン」や「ダチョウ」という下位のクラスの性質にも含まれることになる。クラスを識別するためのクラス名は、例えば、「鳥」というクラスであれば、そのクラス名は「鳥」という文字列によって表されてよい。なお、クラス名は、必ずしも意味を表している必要はなく、例えば、「鳥」というクラスであっても、「情報 1」や「C1」といった単なる識別情報を示す文字列が割り当てられてもよい。上述したエンティティ、すなわち事物は、オントロジーにより定義されたクラス体系の中に含まれる、いずれかのクラスに属するものとする。

【 0017 】

プロパティとは、事物の性質や特徴、クラス間の関係を記述する属性である。例えば、プロパティは、「～を体の構成要素としてもつ」という性質や、「～に生息する」という性質を示す属性であってもよいし、「あるクラスが上位クラスであり、あるクラスが下位クラスである」というクラス間の上位下位の関係を示す属性であってもよい。プロパティを識別するためのプロパティ名は、上述したクラス名と同様に、そのプロパティ名自体が意味を表していてもよいし、意味を表していなくてもよい。

10

20

30

40

50

【 0 0 1 8 】

ナレッジグラフ D 1 は、上述したクラスがノードとして表され、上述したプロパティがラベル付き、且つ方向性のあるエッジとして表された有向グラフである。このようなグラフ構造によって、事物についての情報がノードにより、事物間の関係がエッジによって判別可能となる。

【 0 0 1 9 】

図 2 示される例のナレッジグラフ D 1 は、エンティティ E 1 と、エンティティ E 2 とが、「所属チーム」というプロパティ名のクラスに属する。本実施形態において、各エンティティには、各エンティティを識別することが可能な情報（以下、エンティティ識別情報 E I D）と、エンティティ名と、当該エンティティのウェブページの所在情報とが関連付けられる。所在情報とは、ウェブ上の位置を特定するための情報であり、例えば、URL である。所在情報が示すウェブページには、当該エンティティの関連データが記載されている。図 2 に示すナレッジグラフ D 1 において、エンティティ E 1 には、エンティティ識別情報 E I D の「0 0 0 1」と、エンティティ名の「野球選手 A」と、当該エンティティに関連する関連データが記載されているウェブページの所在情報である「http://百科事典ウェブページ/野球選手 A」という URL とが対応付けられている。また、エンティティ E 2 には、エンティティ識別情報 E I D の「0 0 0 2」と、エンティティ名「野球チーム B」と、当該エンティティに関連する関連データが記載されているウェブページの所在情報である「http://百科事典ウェブページ/野球チーム B」という URL とが対応付けられている。以降の説明において、エンティティに関連する関連データをエンティティの関連データと記載する。また、エンティティの関連データが記載されているウェブページを、エンティティのウェブページとも記載する。

【 0 0 2 0 】

図 3 は、本実施形態に係る「野球選手 A」のウェブページの一例を示す図である。「野球選手 A」のウェブページには、「野球選手 A」の関連データが記載されている。

【 0 0 2 1 】

< 1 - 2 . 検索結果ウェブページ >

図 4 は、本実施形態に係る検索結果ウェブページの一例を示す図である。図 4 に示されるように、端末装置 2 0 0 の表示部 2 1 0 には、クエリ入力領域 2 1 1 と、ナレッジパネル 2 2 0 と、検索結果 2 3 0 とが含まれる検索結果ウェブページが表示される。ナレッジパネル 2 2 0 には、検索クエリに対応するエンティティ 2 2 1 と、エンティティ 2 2 1 に関連する画像 2 2 2 と、エンティティ 2 2 1 に関連する関連情報 2 2 3 とが含まれる。ユーザは、端末装置 2 0 0 を用いて、端末装置 2 0 0 の表示部 2 1 0 に表示された検索ページに検索クエリを入力する。検索クエリとは、一つの検索ワードまたは複数の検索ワードの組み合わせである。端末装置 2 0 0 は、ユーザによって入力された検索クエリをウェブサーバ 3 0 0 に送信する。

【 0 0 2 2 】

ウェブサーバ 3 0 0 は、端末装置 2 0 0 から受信した検索クエリをナレッジデータサーバ 1 0 0 に送信する。ナレッジデータサーバ 1 0 0 は、受信した検索クエリに対応するエンティティに関連する情報をウェブサーバ 3 0 0 に送信する。ナレッジデータサーバ 1 0 0 は、例えば、ナレッジグラフ D 1 に基づいて、受信した検索クエリと関連付けられたエンティティに対応付けられた所在情報をウェブサーバ 3 0 0 に送信する。ウェブサーバ 3 0 0 は、受信した所在情報に基づいて、当該所在情報が示すウェブページのうち、エンティティの関連データを抽出する。ウェブサーバ 3 0 0 は、抽出したエンティティの関連データ（この一例では、ナレッジパネル 2 2 0 に示す情報）を含めた検索結果ウェブページを生成する。

【 0 0 2 3 】

図 4 に示される例においては、検索クエリとして「野球選手 A」が入力され、ナレッジパネル 2 2 0 には「野球選手 A」に関連する画像や様々な情報が表示されている。ナレッジパネル 2 2 0 には、例えば、「野球選手 A」の画像 2 2 2 や、「野球選手 A」の生年月

10

20

30

40

50

日や出身地等の関連情報 2 2 3 が表示される。このように、検索結果 2 3 0 だけでなく、ナレッジパネル 2 2 0 を表示部 2 1 0 に表示することで、検索結果に対するユーザ満足度を向上させることができる。

【 0 0 2 4 】

< 1 - 3 . ナレッジデータサーバ 1 0 0 の構成 >

図 1 に戻り、制御部 1 1 0 は、例えば、C P U (Central Processing Unit) などのハードウェアプロセッサがプログラム (ソフトウェア) を実行することにより実現される。制御部 1 1 0 は、例えば、通信 I / F (Interface) 1 1 1 と、収集部 1 1 2 と、リダイレクト情報抽出部 1 1 3 (抽出部) と、誘導情報抽出部 1 1 4 (抽出部) と、非同義語除去部 1 1 5 (除去部) と、所在情報取得部 1 1 6 とを備える。また、これらの構成要素のうち一部または全部 (内包する記憶部を除く) は、L S I (Large Scale Integration) や A S I C (Application Specific Integrated Circuit) 、 F P G A (Field-Programmable Gate Array) 、 G P U (Graphics Processing Unit) などのハードウェア (回路部 ; circuitry を含む) によって実現されてもよいし、ソフトウェアとハードウェアの協働によって実現されてもよい。通信 I / F 1 1 1 は、ネットワーク NW を介した各機能部と、他の機器 (例えば、クロールの対象機器 (以下、クロール対象機器 D V) 、端末装置 2 0 0 、及びウェブサーバ 3 0 0) との通信を仲介する。

10

【 0 0 2 5 】

収集部 1 1 2 は、所定のウェブページに関する収集対象データを、ネットワーク NW を介してクロール対象機器 D V から収集する。収集部 1 1 2 による収集処理の対象となるウェブページとは、例えば、利用者によって編集可能な百科事典のウェブページ (以下、百科事典ウェブページ) である。

20

【 0 0 2 6 】

収集部 1 1 2 は、いわゆるクローラプログラムにより実現される。収集対象データは、ネットワーク NW 上 (クロール対象機器 D V の記憶領域内) にあり、ブラウザで閲覧可能なデータである。なお、収集対象データは、ブラウザに限らず、アプリケーションプログラムによって再生されるデータであってもよい。収集対象データは、例えば、H T M L (HyperText Markup Language) のソースを示すテキストデータを含むウェブページである。収集部 1 1 2 は、百科事典ウェブページのうち、例えば、ナジッレグラフ D 1 に含まれるエンティティのそれぞれの名前と対応するテキストデータが記載されたウェブページが存在する場合、そのすべてを収集する。

30

【 0 0 2 7 】

また、収集対象データは、例えば、百科事典ウェブページに関するリダイレクト情報である。図 5 は、本実施形態における百科事典ウェブページのリダイレクト処理の一例を説明する図である。図 6 は、本実施形態における百科事典ウェブページのリダイレクト情報の一例を示す図である。例えば、百科事典ウェブページにおいて、ユーザが端末装置 2 0 0 を操作して「野球選手 A」の別名である「別名 a」を検索クエリとして入力した場合、端末装置 2 0 0 には入力された「別名 a」と関連付けされたウェブページ P 1 は表示されることなく、「野球選手 A」と関連付けされたウェブページ P 2 へのリダイレクト (転送) が行われ、端末装置 2 0 0 にはこのウェブページ P 2 が表示される。すなわち、ウェブページ P 1 はリダイレクト元のページであり、リダイレクト先のページとしてウェブページ P 2 が設定されたページ (リダイレクト先としてウェブページ P 2 の URL が設定された転送用ページ) である。図 6 に示されるように、リダイレクト情報は、上記のようなリダイレクト元のページと関連付けされたワードと、リダイレクト先のページと関連付けされたワードとが対応付けされたリストを含む。図 6 には、例えば、リダイレクト元のワード「別名 a」と、リダイレクト先のワード「野球選手 A」とが対応付けされたデータが含まれる。

40

【 0 0 2 8 】

尚、収集部 1 1 2 によるクロール処理を行わずに、例えば、ナレッジデータサーバ 1 0 0 の運用担当者が運用端末のブラウザを操作して手動で収集対象データを収集し、ナレッ

50

ジデータサーバ１００に設けられた入力部（図示しない）を介して収集対象データを入力するようにしてもよい。

【００２９】

リダイレクト元のワードは、リダイレクト先のワードの同義語であり、これら２つのワードは同一のエンティティを示していることが想定される。すなわち、リダイレクト元のワードは、リダイレクト先のワードの別名であることが想定される。このため、リダイレクト情報抽出部１１３は、収集部１１２により収集された収集対象データに含まれるリダイレクト情報から、リダイレクト元のワードおよびリダイレクト先のワードを抽出し、リダイレクト元のワードを、リダイレクト先のワードの別名として、別名情報Ｄ２に登録する。図７は、本実施形態における別名情報Ｄ２の一例を示す図である。図７に示される例では、エンティティである「野球選手Ａ」の別名１として「別名ａ」、別名２として「別名ｂ」、別名３として「別名ｃ」が登録されている。

10

【００３０】

誘導情報抽出部１１４は、収集部１１２により収集された収集対象データに含まれる百科事典ウェブページのＨＴＭＬのソースを示すテキストデータから誘導情報を取得し、抽出した誘導情報に基づいて別名情報を抽出し、別名情報Ｄ２に登録する。図８は、本実施形態における収集部１１２により収集された百科事典ウェブページの一例を示す図である。図８に示される百科事典ウェブページは、「キャラクタＣ」と関連付けされたページであり、「キャラクタＣ（アニメ）」および「キャラクタＣ（テレビドラマ）」のそれぞれと関連付けされたページへの誘導情報を含んでいる。これらの「キャラクタＣ（アニメ）」および「キャラクタＣ（テレビドラマ）」はいずれもエンティティとなりうるものである。この場合、「キャラクタＣ」のワードは、エンティティとしての「キャラクタＣ（アニメ）」および「キャラクタＣ（テレビドラマ）」の別名（例えば、略称）である可能性がある。そこで、誘導情報抽出部１１４は、あるウェブページと関連付けされたワード（図８に示される例では「キャラクタＣ」）を、該ウェブページに含まれる誘導情報に含まれるワード（図８に示される例では「キャラクタＣ（アニメ）」および「キャラクタＣ（テレビドラマ）」）の別名として、別名情報Ｄ２に登録する。

20

【００３１】

非同義語除去部１１５は、リダイレクト情報抽出部１１３および誘導情報抽出部１１４の少なくとも１つにより抽出されて記憶部１２０に記憶された別名情報Ｄ２における非同義語の除去処理（クレンジング）を行う。リダイレクト情報抽出部１１３および誘導情報抽出部１１４により抽出されたワードは、エンティティの別名ではない（同義関係ではない）ワードを含む可能性がある。そこで、非同義語除去部１１５は、別名情報Ｄ２のうち、所定の条件を満たす別名のみを残し、その他の別名は除去することで、データのクレンジングを行う。非同義語除去部１１５は、例えば、あるエンティティを示す名前（以下、エンティティ名）と関連付けられたウェブページ（例えば、百科事典ウェブページ）において、所定のキーワードを含む一文に含まれている別名は残し、その他の別名は除去する。所定のキーワードとは、別名の記載の存在を示唆するワードである。所定のキーワードとは、例えば、「自称」、「通称」、「別名」、「略称」などの名詞のキーワードや、「呼ばれる」、「称する」、「とも言う」などの文末などに現れるキーワードである。

30

40

【００３２】

図９は、本実施形態における非同義語除去部１１５による除去処理の一例を説明する図である。図９に示される例において、非同義語除去部１１５は、「略称」というキーワードを含む一文「略称は別名ｂ。」に含まれる「別名ｂ」は別名情報Ｄ２に残すように処理を行う。また、非同義語除去部１１５は、「呼ばれる」というキーワードを含む一文「野球選手Ａは、別名ｃと呼ばれることもある。」に含まれる「別名ｃ」は別名情報Ｄ２に残すように処理を行う。

【００３３】

また、非同義語除去部１１５は、例えば、あるエンティティ名と関連付けられたウェブページ（例えば、百科事典ウェブページ）において、このエンティティ名の言い換え表現

50

として示されるワードに含まれている別名は残し、その他の別名は除去する。図 10 は、本実施形態における非同義語除去部 115 による除去処理の他の例を説明する図である。図 10 に示される例において、非同義語除去部 115 は、「リンゴ酸」の後に括弧書きで記載されたワードである「林檎酸」、「リンゴさん」、「malic acid」は別名情報 D2 に残すように処理を行う。尚、所在情報取得部 116 は、百科事典ウェブページ以外に、ニュースページなどの様々なページを上記の除去処理に利用してもよい。

【0034】

図 1 に戻り、所在情報取得部 116 は、ナレッジグラフ D1 および別名情報 D2 に基づいて、ウェブサーバ 300 から受信した検索クエリに対応するエンティティに対応付けられている所在情報を取得する。所在情報取得部 116 は、ナレッジグラフ D1 に含まれるエンティティ名の中に、受信した検索クエリと合致するエンティティ名が存在するか否かを判定する。所在情報取得部 116 は、受信した検索クエリと合致するエンティティ名が存在すると判定した場合、該エンティティに対応付けられている所在情報を取得する。また、所在情報取得部 116 は、受信した検索クエリと合致するエンティティ名がナレッジグラフ D1 に存在しないと判定した場合、別名情報 D2 に別名として登録されている別名の中に、受信した検索クエリと合致する別名が存在するか否かを判定する。所在情報取得部 116 は、受信した検索クエリと合致する別名が別名情報 D2 に存在すると判定した場合、該別名と関連付けられたエンティティに対応付けられている所在情報を取得する。このとき、検索クエリと、エンティティ名または別名とが完全一致する場合だけでなく、部分一致または意味的に同一とみなされる程度の相違がある場合も、対応するエンティティ名または別名とみなしてよい。

10

20

【0035】

次に、所在情報取得部 116 は、取得した所在情報をウェブサーバ 300 に送信する。所在情報取得部 116 は、受信した検索クエリと合致するエンティティおよび別名が存在しないと判定した場合、所在情報が存在しない旨を示す情報をウェブサーバ 300 に送信する。

【0036】

< 1 - 4 . リダイレクト情報に基づく別名情報抽出動作 >

図 11 は、本実施形態のナレッジデータサーバ 100 のリダイレクト情報に基づく別名情報抽出動作の一例を示す流れ図である。ナレッジデータサーバ 100 は、例えば、日次、週次等の任意のタイミングで実施されるバッチ処理により別名情報抽出動作を行う。

30

【0037】

まず、収集部 112 は、例えば、百科事典ウェブページに関するリダイレクト情報を、ネットワーク NW を介してクロール対象機器 DV から収集する (S10)。

【0038】

次に、リダイレクト情報抽出部 113 は、収集部 112 により収集されたリダイレクト情報から、リダイレクト元のワードおよびリダイレクト先のワードを抽出する (S12)。次に、リダイレクト情報抽出部 113 は、抽出したワードのうち、リダイレクト元のワードを、リダイレクト先のワードの別名として、別名情報 D2 に登録する (S14)。以上により、本フローチャートの処理を終了する。

40

【0039】

< 1 - 5 . 誘導情報に基づく別名情報抽出動作 >

図 12 は、本実施形態のナレッジデータサーバ 100 の誘導情報に基づく別名情報抽出動作の一例を示す流れ図である。ナレッジデータサーバ 100 は、例えば、日次、週次等の任意のタイミングで実施されるバッチ処理により別名情報抽出動作を行う。

【0040】

まず、収集部 112 は、例えば、ナレッジグラフ D1 に含まれるエンティティ名のそれぞれと関連付けられた百科事典ウェブページの HTML のソースを示すテキストデータを、ネットワーク NW を介してクロール対象機器 DV から収集する (S20)。

【0041】

50

次に、誘導情報抽出部 114 は、収集部 112 により収集された百科事典ウェブページの HTML のソースを示すテキストデータから誘導情報を取得し、取得した誘導情報に基づいて別名情報を抽出する (S22)。次に、誘導情報抽出部 114 は、百科事典ウェブページと関連付けされたワードを、該百科事典ウェブページに含まれる誘導情報に含まれるワードの別名として、別名情報 D2 に登録する (S24)。以上により、本フローチャートの処理を終了する。

【0042】

< 1 - 6 . 別名情報のクレンジング動作 >

図 13 は、本実施形態のナレッジデータサーバ 100 の別名情報の非同義語除去動作の一例を示す流れ図である。ナレッジデータサーバ 100 は、例えば、日次、週次等の任意のタイミングで実施されるバッチ処理により非同義語除去動作を行う。

10

【0043】

まず、収集部 112 は、例えば、非同義語除去処理の対象とするエンティティ名と関連付けられた百科事典ウェブページの HTML のソースを示すテキストデータを、ネットワーク NW を介してクロール対象機器 DV から収集する (S30)。

【0044】

次に、非同義語除去部 115 は、別名情報 D2 から、上記のクレンジング処理の対象としたエンティティの別名情報を抽出する (S32)。次に、非同義語除去部 115 は、抽出した別名が、収集部 112 により収集された百科事典ウェブページにおいて所定のキーワードを含む一文に含まれているか否かを判定する (S34)。非同義語除去部 115 は、抽出した別名が、所定のキーワードを含む一文に含まれていると判定した場合、この別名は適切なもの (エンティティの同義語である) であると判定して除去処理の対象とすることなく、本フローチャートの処理を終了する。

20

【0045】

一方、非同義語除去部 115 は、抽出した別名が、所定のキーワードを含む一文に含まれていないと判定した場合、抽出した別名が、百科事典ウェブページにおいて上記のエンティティ名の言い換え表現として示されるワードに含まれているか否かを判定する (S36)。非同義語除去部 115 は、抽出した別名が、上記のエンティティ名の言い換え表現として示されるワードに含まれていると判定した場合、この別名は適切なもの (エンティティの同義語である) であると判定して除去処理の対象とすることなく、本フローチャートの処理を終了する。

30

【0046】

一方、非同義語除去部 115 は、抽出した別名が、上記のエンティティを示す名前の言い換え表現として示されるワードに含まれていないと判定した場合、この別名は不適切なもの (エンティティの同義語ではない) と判定して、この別名を別名情報 D2 から除去する (S38)。以上により、本フローチャートの処理を終了する。

【0047】

< 1 - 7 . オンライン動作 >

図 14 は、本実施形態のナレッジデータサーバ 100 のオンライン動作の一例を示す流れ図である。まず、所在情報取得部 116 は、ウェブサーバ 300 から検索クエリを受信する (S40)。

40

【0048】

次に、所在情報取得部 116 は、ナレッジグラフ D1 に含まれるエンティティ名の中に、受信した検索クエリと合致するエンティティ名が存在するか否かを判定する (S42)。所在情報取得部 116 は、受信した検索クエリと合致するエンティティ名が存在すると判定した場合、該エンティティに対応付けられている所在情報を取得し、ウェブサーバ 300 に送信し (S44)、本フローチャートの処理を終了する。

【0049】

一方、所在情報取得部 116 は、受信した検索クエリと合致するエンティティ名がナレッジグラフ D1 に存在しないと判定した場合、別名情報 D2 に別名として登録されている

50

別名の中に、受信した検索クエリと合致する別名が存在するか否かを判定する（S 4 6）。所在情報取得部 1 1 6 は、受信した検索クエリと合致する別名が別名情報 D 2 に存在すると判定した場合、該別名と関連付けられたエンティティに対応付けられている所在情報を取得し、ウェブサーバ 3 0 0 に送信し（S 4 4）、本フローチャートの処理を終了する。

【0 0 5 0】

一方、所在情報取得部 1 1 6 は、受信した検索クエリと合致する別名が存在しないと判定した場合、所在情報が存在しない旨を示す情報をウェブサーバ 3 0 0 に送信し（S 4 8）、本フローチャートの処理を終了する。

【0 0 5 1】

以上、説明したように、本実施形態のナレッジデータサーバ 1 0 0 は、ウェブページのリダイレクト情報を収集する収集部 1 1 2 と、前記収集部 1 1 2 により収集された前記リダイレクト情報に基づいて、第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページへのリダイレクト元である、第 2 ページに関連付けされた第 2 ワードを、前記第 1 エンティティの別名情報として抽出するリダイレクト情報抽出部 1 1 3 とを備えるこれによって、エンティティと関連付けされた別名情報を収集することができる。

【0 0 5 2】

また、本実施形態のナレッジデータサーバ 1 0 0 は、第 1 エンティティの名前を示す第 1 ワードに関連付けられた第 1 ページを収集する収集部 1 1 2 と、前記収集部 1 1 2 により収集された前記第 1 ページに第 2 ページへの誘導情報がある場合、前記第 1 ワードを、前記第 2 ページに関連付けられた第 2 ワードにより示される第 2 エンティティの別名情報として抽出する誘導情報抽出部 1 1 4 とを備えるこれによって、エンティティと関連付けされた別名情報を収集することができる。

【0 0 5 3】

また、本実施形態のナレッジデータサーバ 1 0 0 によれば、エンティティの別名情報を収集し、エンティティ名と別名とを同義語として総合的に管理することで、エンティティを適切に評価することができる。例えば、エンティティの検索数の実績値に応じて検索結果ページ内での表示位置を変更するような制御（ランキング）を行っている場合には、エンティティ名および別名の双方に対する検索結果に基づいてより正確な制御を行うことができる。

【0 0 5 4】

以上、本発明を実施するための形態について実施形態を用いて説明したが、本発明はこうした実施形態に何等限定されるものではなく、本発明の要旨を逸脱しない範囲内において種々の変形及び置換を加えることができる。

【符号の説明】

【0 0 5 5】

1 0 0 ... ナレッジデータサーバ
1 1 0 ... 制御部
1 1 1 ... 通信 I / F
1 1 2 ... 収集部
1 1 3 ... リダイレクト情報抽出部
1 1 4 ... 誘導情報抽出部
1 1 5 ... 非同義語除去部
1 1 6 ... 所在情報取得部
1 2 0 ... 記憶部
2 0 0 ... 端末装置
2 1 0 ... 表示部
2 1 1 ... クエリ入力領域
2 2 0 ... ナレッジパネル
2 2 1 ... エンティティ

10

20

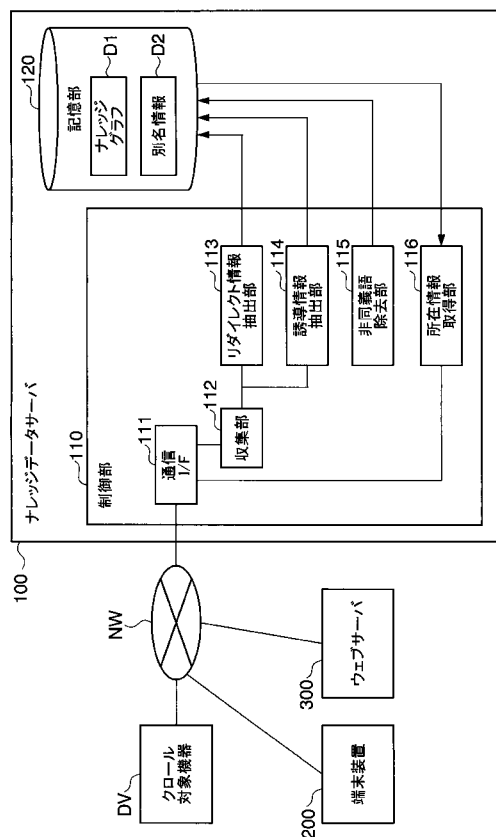
30

40

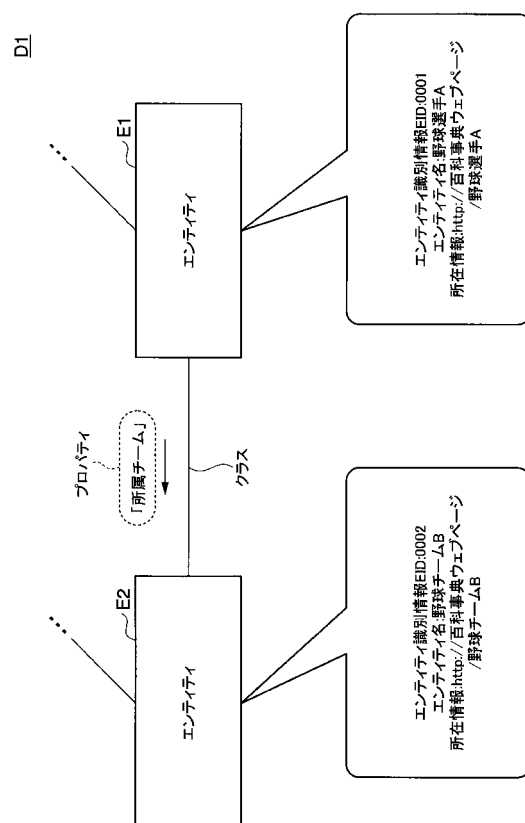
50

```
2 2 2 ... 画像
2 2 3 ... 関連情報
2 3 0 ... 検索結果
3 0 0 ... ウェブサーバ
D 1 ... ナレッジグラフ
D 2 ... 別名情報
R 1 ... リダイレクト情報
```

【 図 1 】



【圖 2】

2

【図 3】

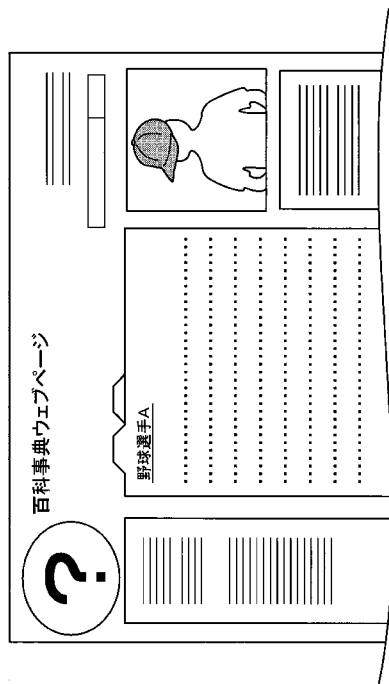


図3

【図 4】

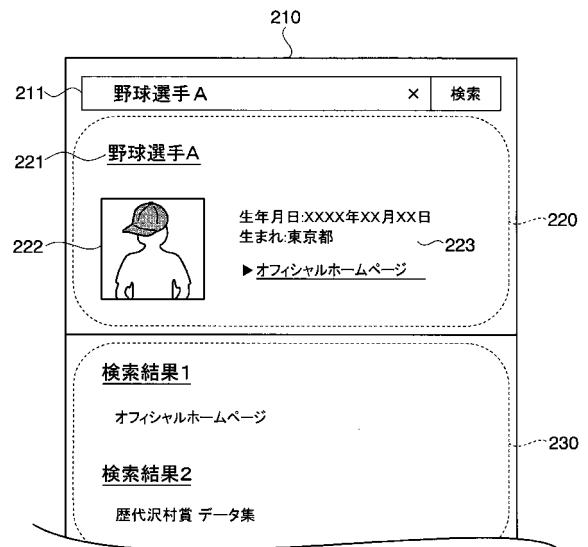


図4

【図 5】

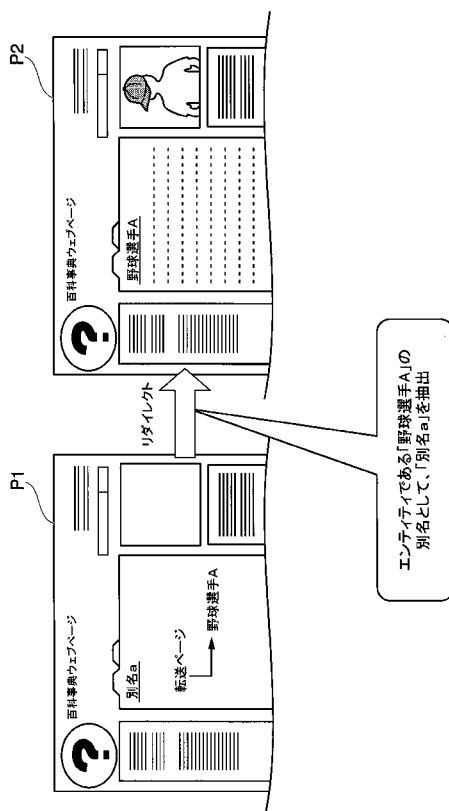


図5

【図 6】

R1	
リダイレクト元のワード	リダイレクト先のワード
別名a	野球選手A
別名b	野球選手A
別名c	野球選手A
別名d	野球選手B
...	...

図6

【図 7】

D2				
エンティティ	別名1	別名2	別名3	別名4
野球選手A	別名a	別名b	別名c	...
野球選手B	別名d	別名e	—	—
キャラクターC (アニメ)	キャラクターC	—	—	—
キャラクターC (映画)	キャラクターC	—	—	—
リンゴ酸	林檎酸	リンゴさん	malic acid	—
...	...	...	...	...

図7

【図 8】

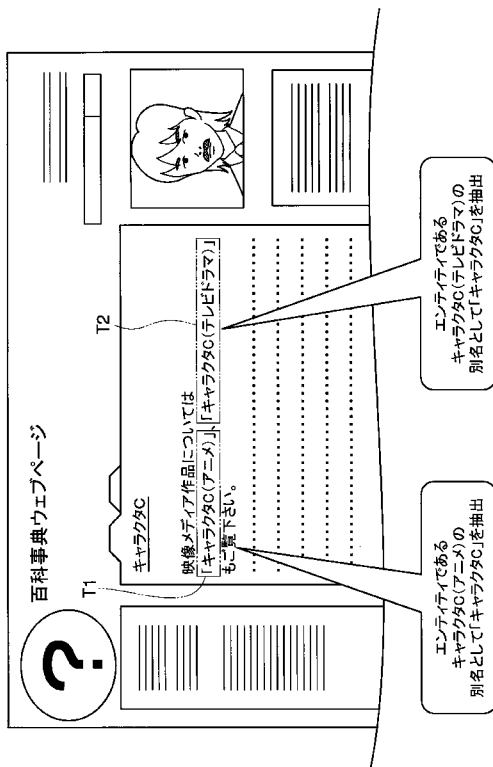


図 8

【図 9】

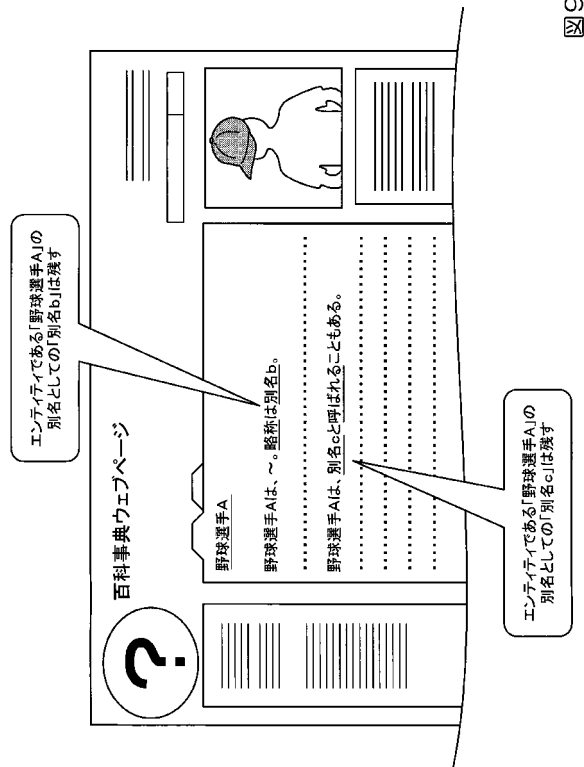


図 9

【図 10】

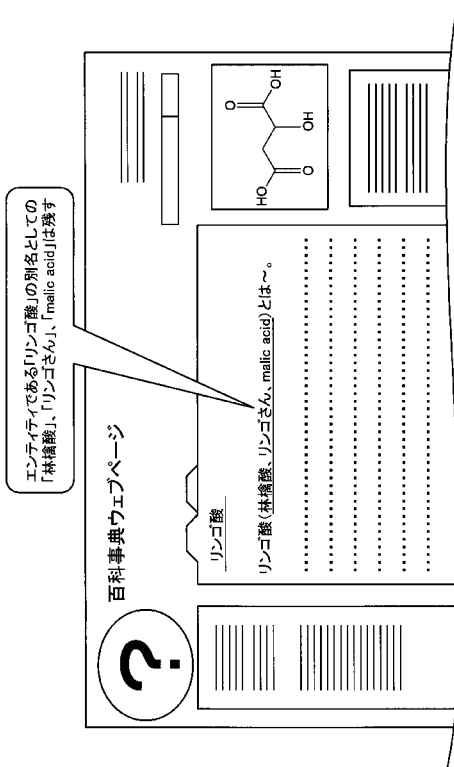


図 10

【図 11】

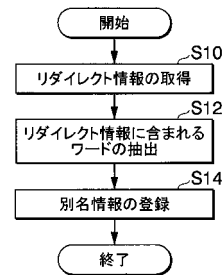


図 11

【図 12】

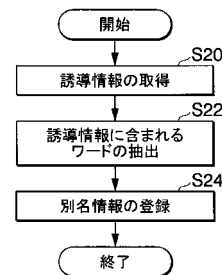


図 12

【 図 1 3 】

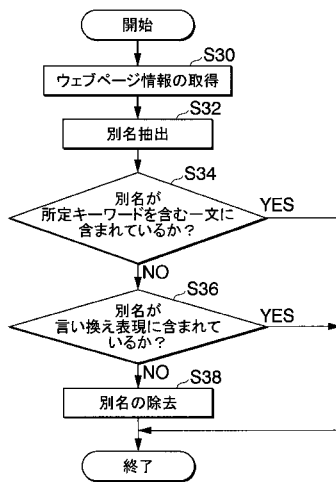


図13

【 図 1 4 】

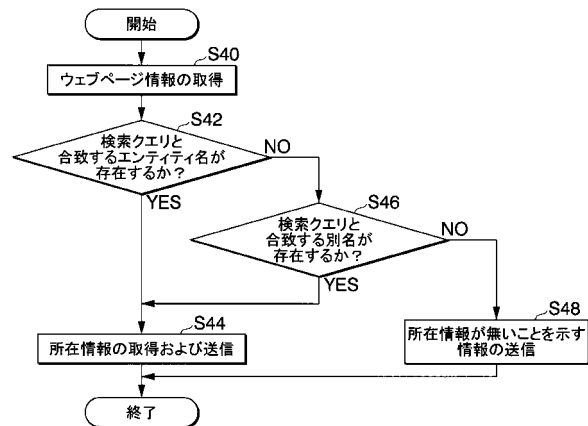


図14

フロントページの続き

- (72)発明者 豊田 樹生
東京都千代田区紀尾井町 1 番 3 号 ヤフー株式会社内
- (72)発明者 夜久 真也
東京都千代田区紀尾井町 1 番 3 号 ヤフー株式会社内
- (72)発明者 野口 正樹
東京都千代田区紀尾井町 1 番 3 号 ヤフー株式会社内
- (72)発明者 岩澤 宏希
東京都千代田区紀尾井町 1 番 3 号 ヤフー株式会社内
- (72)発明者 萩原 健
東京都千代田区紀尾井町 1 番 3 号 ヤフー株式会社内