

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5540079号
(P5540079)

(45) 発行日 平成26年7月2日(2014.7.2)

(24) 登録日 平成26年5月9日(2014.5.9)

(51) Int.Cl.

F I

G 0 6 F 17/30 (2006.01)

G 0 6 F 17/30 3 2 0 D

請求項の数 16 (全 23 頁)

(21) 出願番号	特願2012-508592 (P2012-508592)	(73) 特許権者	510330264
(86) (22) 出願日	平成22年4月27日 (2010.4.27)		アリババ・グループ・ホールディング・リミテッド
(65) 公表番号	特表2012-525645 (P2012-525645A)		ALIBABA GROUP HOLDING LIMITED
(43) 公表日	平成24年10月22日 (2012.10.22)		英国領、ケイマン諸島、グランド・ケイマン、ジョージ・タウン、ワン・キャピタル・プレイス、フォース・フロア、ピー・オー・ボックス 847
(86) 国際出願番号	PCT/US2010/032581		
(87) 国際公開番号	W02010/126892		
(87) 国際公開日	平成22年11月4日 (2010.11.4)	(74) 代理人	110001243
審査請求日	平成25年3月29日 (2013.3.29)		特許業務法人 谷・阿部特許事務所
(31) 優先権主張番号	200910136206.6		
(32) 優先日	平成21年4月29日 (2009.4.29)		
(33) 優先権主張国	中国 (CN)		

最終頁に続く

(54) 【発明の名称】 知識ベース構築の方法および装置

(57) 【特許請求の範囲】

【請求項 1】

コンピューティング装置を使用する、知識ベース構築の方法であって、
前記コンピューティング装置の基本データ処理層を使用して、ウェブページから文を取得することと、

前記コンピューティング装置のデータマイニング層を使用して、前記取得された文を単語に解析することと、

知識ベースの第1のカテゴリ内の1つまたは複数の代表的な単語を、前記取得された文から解析された前記単語と照合することと、

前記代表的な単語のうちの1つと、前記取得された文から解析された前記単語のうちの1つとの間に一致が存在する場合、前記取得された文中の前記一致された単語に隣接する単語の列を、第1のエントリとして前記第1のカテゴリに追加することと、

前記取得された文から解析された前記単語を、前記知識ベースの第2のカテゴリの第2のエントリと一致するときに、前記第1のカテゴリと前記第2のカテゴリとの間に確立された相関関係が存在するか否かを判定することと、

前記第1のカテゴリと前記第2のカテゴリとの間に確立された相関関係が存在すると判定された場合、前記第1のカテゴリの前記第1のエントリと、前記第2のカテゴリの前記第2のエントリとの間の相関関係を確立することと、を含む、方法。

【請求項 2】

ウェブページから文を取得することは、前記取得された文を、前記取得された文中の句

10

20

読点に基づいて、複数のより短い文に分割することを含み、かつ前記取得された文を解析することが、前記取得された文を解析すること、または前記複数のより短い文を解析することを含む、請求項 1 に記載の方法。

【請求項 3】

前記基本データ処理層が、個々の文の出現の回数を計数することと、

前記データマイニング層が、前記第 1 のエン트리および前記第 1 のエントリに隣接する 1 つまたは複数の前記代表的な単語を有する、任意の文の出現の回数に基づいて、前記第 1 のカテゴリの前記第 1 のエントリの重み付けされた値を確立することと、をさらに含む、請求項 1 に記載の方法。

【請求項 4】

前記データマイニング層は、前記取得された文を分割するように、前記 1 つまたは複数の代表的な単語を含む、解析システムを採用する、請求項 1 に記載の方法。

【請求項 5】

前記知識ベースは、共通単語システムおよび実単語システムを含み、前記共通単語システムおよび前記実単語システムは、それぞれ異なるカテゴリを含み、前記代表的な単語は、前記実単語システムのカテゴリ対応インデックスワードと、前記共通単語システムのカテゴリ対応シードワードとを含み、前記取得された文中の前記一致した単語に隣接する前記単語の列が、前記第 1 のエン트리として前記第 1 のカテゴリに追加されると、前記単語の列が、前記第 1 のカテゴリを含む前記共通単語システムまたは前記実単語システムに追加される、請求項 1 に記載の方法。

【請求項 6】

前記第 1 のカテゴリが、前記共通単語システム内に含まれる前記カテゴリのうちの 1 つである場合、

前記第 1 のエントリを、前記第 1 のカテゴリに対応する前記シードワードとして設定することをさらに含む、請求項 5 に記載の方法。

【請求項 7】

前記第 1 のカテゴリの前記第 1 のエントリと、前記第 2 のカテゴリの前記第 2 のエントリとの間の相関関係を確立することは、

前記第 1 のエン트리および前記第 2 のエントリの文の出現の頻度を得ることと、

前記第 1 のエン트리および前記第 2 のエントリの文の出現の前記頻度が、所定の閾値を超える場合、前記第 1 のエントリと第 2 のエントリとの間の相関関係を確立することと、を含む、請求項 1 に記載の方法。

【請求項 8】

各カテゴリおよび各カテゴリの下各エントリに応じて、それぞれの結果ファイルを作成する、前記データマイニング層と、

複数の結果ファイルを単一の結果ファイルに統合する、前記コンピューティング装置の統合層と、をさらに含む、請求項 1 に記載の方法。

【請求項 9】

個々の文の出現の回数を計数することと、

1 つまたは複数の代表的な単語および前記第 1 のエントリを有する、任意の文の出現の回数に基づいて、前記第 1 のカテゴリの前記第 1 のエントリの重み付けされた値を確立することと、

異なるカテゴリの下各エントリの重み付けされた値を比較することと、

エントリ対応カテゴリを、フィルタにかけることと、をさらに含む、請求項 8 に記載の方法。

【請求項 10】

前記ウェブページからテーブルを取得することと、

前記第 1 のエントリのプロパティとして、前記第 1 のエントリと対になってテーブル内に複数回出現する単語を属性化することと、をさらに含む、請求項 1 に記載の方法。

【請求項 11】

ウェブページから文を取得することは、前記ウェブページから、特殊記号を含む文を取得することを含む、請求項1に記載の方法。

【請求項12】

コンピュータにより実行される情報検索の方法であって、

該コンピュータが、

知識ベースにおいて、1つまたは複数のキーワードに関連するウェブページおよびエン
トリ内の前記1つまたは複数のキーワードに基づいて、ユーザによって入力された検索語
句と関連する、ラベルを識別することと、

前記ラベルに対応する前記ウェブページを検索することと、

前記ユーザに、前記ウェブページまたは前記ウェブページへのリンクを提供することと
、を含む各ステップを実行するように構成され、

前記知識ベースは、

コンピューティング装置の基本データ処理層を使用して、複数のウェブページのうちの
1つから文を取得することと、

前記コンピューティング装置のデータマイニング層を使用して、前記取得された文を単
語に解析することと、

前記知識ベースの第1のカテゴリ内の1つまたは複数の代表的な単語を、前記取得され
た文から解析された前記単語と、照合することと、

前記代表的な単語のうちの1つと、前記取得された文から解析された前記単語のうちの
1つとの間に一致が存在する場合、前記取得された文中の前記一致した単語に隣接する単
語の列を、第1のエントリとして前記第1のカテゴリに追加することと、

前記取得された文から解析された前記単語を、前記知識ベースの第2のカテゴリの第2
のエントリと一致するときに、前記第1のカテゴリと前記第2のカテゴリとの間に、確立
された相関関係が存在するか否かを判定することと、

前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在すると
判定された場合、前記第1のカテゴリの前記第1のエントリと、前記第2のカテゴリの前
記第2のエントリとの間の相関関係を確立することと、によって構築されることを特徴と
する方法。

【請求項13】

コンピュータにより実行される情報検索の方法であって、

該コンピュータが、

知識ベースのエントリを使用して、ユーザによって入力された検索語句を解析すること
と、

前記検索語句から解析された単語を、前記知識ベースの前記エントリと照合することと
、

前記検索語句から解析された単語との一致を有するエントリに関連する、前記知識ベー
スのエントリを識別することと、

前記検索語句を、前記検索語句から解析された単語との一致を有する前記エントリに関
連する、前記知識ベースのエントリで更新することと、

前記更新された検索語句に基づいて、検索を実施することと、を含む各ステップを実行
するように構成され、

前記知識ベースは、

コンピューティング装置の基本データ処理層を使用して、ウェブページから文を取得す
ることと、

前記コンピューティング装置のデータマイニング層を使用して、前記取得された文を単
語に解析することと、

前記知識ベースの第1のカテゴリ内の1つまたは複数の代表的な単語を、前記取得され
た文から解析された前記単語と、照合することと、

前記代表的な単語のうちの1つと、前記取得された文から解析された前記単語のうちの
1つとの間に一致が存在する場合、前記取得された文中の一致した単語に隣接する単語の

10

20

30

40

50

列を、第1のエントリとして前記第1のカテゴリに追加することと、

前記取得された文から解析された前記単語を、前記知識ベースの第2のカテゴリの第2のエントリと一致するときに、前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在するか否かを判定することと、

前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、前記第1のカテゴリの前記第1のエントリと、前記第2のカテゴリの前記第2のエントリとの間の相関関係を確立することと、によって構築されることを特徴とする方法。

【請求項14】

知識ベースを構築する、コンピューティング装置であって、当該コンピューティング装置は、

ウェブページから、1つまたは複数の文を取得する、基本データ処理モジュールと、
前記ウェブページから取得された前記1つまたは複数の文を解析する、データマイニングモジュールと、を備え、

前記データマイニングモジュールは、さらに、

前記知識ベースの第1のカテゴリ内の、1つまたは複数の代表的な単語を、前記取得された文から解析された前記単語と照合し、

前記代表的な単語のうちの1つと、前記取得された文から解析された前記単語のうちの1つとの間に一致が存在する場合、前記取得された文中の一致した単語に隣接する単語の列を、第1のエントリとして前記第1のカテゴリに追加し、

前記取得された文から解析された前記単語を、前記知識ベースの第2のカテゴリの第2のエントリと一致するときに、前記第1のカテゴリと、前記第2のカテゴリとの間に、確立された相関関係が存在するか否かを判定し、

前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、前記第1のカテゴリの前記第1のエントリと、前記第2のカテゴリの前記第2のエントリとの間の相関関係を確立する、コンピューティング装置。

【請求項15】

検索プログラムであって、

該検索プログラムはコンピュータを、

ユーザによって入力された検索語句に対応するラベルを識別する、第1のクエリーモジュールと、

前記ラベルに対応するウェブページを識別する、第2のクエリーモジュールと、

前記ユーザに、前記ウェブページまたは前記ウェブページへのリンクを提供する、インターフェースモジュールと、

1つまたは複数のキーワードに関連する、知識ベースの前記ウェブページおよびエントリの、前記1つまたは複数のキーワードに基づいて、前記ウェブページに対応するラベルを作成する、ラベル作成モジュールと、
して機能させるように構成され、

前記知識ベースは、

コンピューティング装置の基本データ処理層を使用して、複数のウェブページのうちの1つから文を取得することと、

前記コンピューティング装置のデータマイニング層を使用して、前記取得された文を単語に解析することと、

前記知識ベースの第1のカテゴリ内の1つまたは複数の代表的な単語を、前記取得された文から解析された前記単語と、照合することと、

前記代表的な単語のうちの1つと、前記取得された文から解析された前記単語のうちの1つとの間に一致が存在する場合、前記取得された文中の前記一致した単語に隣接する単語の列を、第1のエントリとして前記第1のカテゴリに追加することと、

前記取得された文から解析された前記単語を、前記知識ベースの第2のカテゴリの第2のエントリと一致するときに、前記第1のカテゴリと前記第2のカテゴリとの間に、確立

10

20

30

40

50

された相関関係が存在するか否かを判定することと、

前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、前記第1のカテゴリの前記第1のエントリと、前記第2のカテゴリの前記第2のエントリとの間の相関関係を確立することと、によって構築されることを特徴とする検索プログラム。

【請求項16】

検索プログラムであって、

該検索プログラムはコンピュータを、

ユーザ入力検索語句を、知識ベースのエントリに基づいて、単語に解析する、解析モジュールと、

前記検索語句から解析された単語を、前記知識ベースの前記エントリと照合する、照合モジュールと、

前記検索語句から解析された単語との一致を有するエントリに関連する、前記知識ベースのエントリを識別する、クエリーモジュールと、

前記検索語句から解析された単語との一致を有する前記エントリに関連する、前記知識ベースのエントリで、前記検索語句を更新する、更新モジュールと、

前記更新された検索語句に基づいて、検索を実施する、検索モジュールと、して機能させるように構成され、

前記知識ベースは、

コンピューティング装置の基本データ処理層を使用して、ウェブページから文を取得することと、

前記コンピューティング装置のデータマイニング層を使用して、前記取得された文を単語に解析することと、

前記知識ベースの第1のカテゴリ内の1つまたは複数の代表的な単語を、前記取得された文から解析された前記単語と、照合することと、

前記代表的な単語のうちの1つと、前記取得された文から解析された前記単語のうちの1つとの間に一致が存在する場合、前記取得された文中の一致した単語に隣接する単語の列を、第1のエントリとして前記第1のカテゴリに追加することと、

前記取得された文から解析された前記単語を、前記知識ベースの第2のカテゴリの第2のエントリと一致するときに、前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在するか否かを判定することと、

前記第1のカテゴリと前記第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、前記第1のカテゴリの前記第1のエントリと、前記第2のカテゴリの前記第2のエントリとの間の相関関係を確立することと、によって構築されることを特徴とする検索プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

(関連出願)

本出願は、2009年4月29日に出願された、発明の名称が「METHOD AND APPARATUS OF KNOWLEDGE BASE BUILDING」である、中国特許出願第200910136206.6号の優先権の利益を主張し、参照によってその全体が本明細書に組み込まれる。

【0002】

本出願は、コンピュータと通信の分野に関連し、より具体的には、知識ベースを構築するための、方法および装置に関する。

【背景技術】

【0003】

広く使用されているコンピュータおよびネットワークに関連する技術において、資料の検索は、主な特徴である。多くのユーザの間で、すべての利用可能な情報源から、どのよ

10

20

30

40

50

うにユーザの探している情報を検索するかということは、共通の関心事である。したがって、さまざまな検索技術が開発されてきた。

【0004】

主要な検索技術のうちの1つは、キーワード検索である。ユーザは1つまたは複数のキーワードを検索語句として入力し、検索エンジンは、その検索語句を含むウェブページを識別するように、検索語句に基づいて検索を実行する。しかし、単語が複数の意味を有する場合がしばしばあり、また、1つの単語は、異なる産業または異なる分野において、さまざまな解釈または用途を有する場合がある。単語の、考えられる意味のすべてがユーザに関係するわけではないため、無関係な意味に基づいた検索で出現したウェブページは、ユーザにとって役に立たない可能性がある。How-net等のウェブサイトの存在は、そのような問題を部分的に対処するように思われる。

10

【0005】

How-netでは、1つの単語または句が複数の概念を含み、複数の概念のそれぞれに基づいて、複数の検索が行われる。そのような検索の結果は、より正確である傾向にある。

【0006】

しかし、既存のHow-netは、手作業で構築および整理され、したがって、頻度の高い（最も一般的な）コンテンツのみを対象にする傾向がある。したがって、それは、ネットワークの限定された包含範囲を有する。さらに、ウェブの急速な発展に伴い、ウェブ上で利用可能となる情報量の速度は、How-netの手動更新の速度をはるかに上回る。そのため、How-netを使用する検索結果も、最善ではない傾向にある。

20

【発明の概要】

【0007】

本開示は、知識ベースを構築するための方法および装置の、例示的な実装形態を提供する。本方法および装置は、知識ベースの自動作成を実装して、そのような知識ベースの精度を向上させるように、使用することができる。

【0008】

一態様において、方法は、コンピューティング装置の基本データ処理層を使用して、ウェブページから文を取得する。取得された文は、コンピューティング装置のデータマイニング層を使用して、単語に解析される。知識ベースの第1のカテゴリ内の、1つまたは複数の代表的な単語は、取得された文から解析された単語と照合される。代表的な単語のうちの1つと、取得された文から解析された単語のうちの1つとの間に一致が存在する場合、取得された文中の一致した単語に隣接する単語の列が、第1のエントリとして第1のカテゴリに追加される。取得された文から解析された単語を、知識ベースの第2のカテゴリの第2のエントリと照合するときに、第1のカテゴリと第2のカテゴリとの間に、確立された相関関係が存在するか否かが判定される。第1のカテゴリと第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間の相関関係が確立される。

30

【0009】

ウェブページから文を取得することは、取得された文を、取得された文中の句読点に基づいて、複数のより短い文に分割することを含んでもよい。さらに、取得された文を解析することは、取得された文を解析すること、または複数のより短い文を解析することを含んでもよい。

40

【0010】

本方法は、さらに、基本データ処理層を使用して、個々の文の出現の回数を計数してもよく、データマイニング層を使用して、第1のカテゴリの第1のエントリの重み付けした値を、第1のエントリ、および第1のエントリに隣接する1つまたは複数の代表的な単語を有する、任意の文の出現の回数に基づいて、確立してもよい。

【0011】

データマイニング層は、取得された文を分割するように、1つまたは複数の代表的な単

50

語を含む、解析システムを採用してもよい。

【0012】

知識ベースは、共通単語システムおよび実単語システムを含んでもよい。共通単語システムおよび実単語システムは、それぞれ異なるカテゴリを含んでもよい。代表的な単語は、実単語システムのカテゴリ対応インデックスワードおよび共通単語システムのカテゴリ対応シードワードを含んでもよい。取得された文中の一致した単語に隣接する単語の列が、第1のエントリとして第1のカテゴリに追加されると、単語の列は、第1のカテゴリを含む、共通単語システムまたは実単語システムに追加されてもよい。第1のカテゴリが、共通単語システム内に含まれるカテゴリのうちの1つである場合、第1のエントリは、第1のカテゴリに対応するシードワードとして設定されてもよい。

10

【0013】

第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間の相関関係を確立することは、第1のエントリおよび第2のエントリを有する文の出現の頻度を得ることと、第1のエントリおよび第2のエントリを有する文の出現の頻度が、所定の閾値を超える場合、第1のエントリと第2のエントリとの間の相関関係を確立することを含んでもよい。

【0014】

データマイニング層は、各カテゴリおよび各カテゴリの下各エントリに応じて、それぞれの結果ファイルを作成することができる。コンピューティング装置の統合層は、複数の結果ファイルを、単一の結果ファイルに統合することができる。個々の文の出現の回数は、計数される。第1のカテゴリの第1のエントリの重み付けされた値は、1つまたは複数の代表的な単語および第1のエントリを有する、任意の文の出現の回数に基づいて、確立されられてもよい。異なるカテゴリの下各エントリの重み付けされた値は、比較されてもよい。エントリ対応カテゴリは、フィルタにかけることができる。

20

【0015】

本方法は、さらに、ウェブページからテーブルを取得してもよく、第1のエントリのプロパティとして、第1のエントリと対になってテーブル内に何度も出現する単語をアトリビュートしてもよい。

【0016】

ウェブページから文を取得することは、ウェブページから特殊記号を含む文を取得することを含むことができる。

30

【0017】

別の態様では、情報検索の方法は、ウェブページ内の1つまたは複数のキーワードと、知識ベース内の1つまたは複数のキーワードに関連するエントリとに基づいて、ユーザによって入力された検索語句と一致する、ラベルを識別することと、ラベルに対応するウェブページを検索することと、ユーザに、このウェブページまたはこのウェブページへのリンクを提供することを含む。

【0018】

知識ベースは、コンピューティング装置の基本データ処理層を使用して、ウェブページから文を取得することと、取得された文を、コンピューティング装置のデータマイニング層を使用して、単語に解析することと、知識ベースの第1のカテゴリ内の、1つまたは複数の代表的な単語を、取得された文から解析された単語と照合することと、代表的な単語のうちの1つと、取得された文から解析された単語のうちの1つとの間に一致が存在する場合、取得された文中の一致した単語に隣接する単語の列を、第1のエントリとして第1のカテゴリに追加することと、取得された文から解析された単語を、知識ベースの第2のカテゴリの第2のエントリと照合するときに、第1のカテゴリと第2のカテゴリとの間に、確立された相関関係が存在するか否かを判定することと、第1のカテゴリと第2のカテゴリとの間に確立された相関関係が存在すると判定された場合、第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間の相関関係を確立することによって、構築されてもよい。

40

50

【0019】

さらなる別の態様では、情報検索の方法は、知識ベースのエントリを使用して、ユーザによって入力された検索語句を解析することと、検索語句から解析された単語を、知識ベースのエントリと照合することと、検索語句から解析された単語との一致を有するエントリに関連する、知識ベースのエントリを識別することと、検索語句を、検索語句から解析された単語との一致を有するエントリに関連する、知識ベースのエントリで更新することと、更新した検索語句に基づいて、検索を実行することとを含む。

【0020】

知識ベースは、コンピューティング装置の基本データ処理層を使用して、ウェブページから文を取得することと、コンピューティング装置のデータマイニング層を使用して、取得された文を単語に解析することと、知識ベースの第1のカテゴリ内の1つまたは複数の代表的な単語を、取得された文から解析された単語と照合することと、代表的な単語のうちの1つと、取得された文から解析された単語のうちの1つとの間に一致が存在する場合、取得された文中の一致した単語に隣接する単語の列を、第1のエントリとして第1のカテゴリに追加することと、取得された文から解析された単語を、知識ベースの第2のカテゴリの第2のエントリと照合するときに、第1のカテゴリと第2のカテゴリとの間に、確立された相関関係が存在するか否かを判定することと、第1のカテゴリと第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間の相関関係を確立することとによって、構築されてもよい。

【0021】

一態様では、知識ベースを構築するコンピューティング装置は、ウェブページから、1つまたは複数の文を取得する、基本データ処理モジュールと、ウェブページから取得された1つまたは複数の文を解析する、データマイニングモジュールとを含む。データマイニングモジュールは、さらに、知識ベースの第1のカテゴリ内の、1つまたは複数の代表的な単語を、取得された文から解析された単語と照合し、代表的な単語のうちの1つと、取得された文から解析した単語のうちの1つとの間に一致が存在する場合、取得された文中の一致した単語に隣接する単語の列を、第1のエントリとして第1のカテゴリに追加し、取得された文から解析された単語を、知識ベースの第2のカテゴリの第2のエントリと照合するときに、第1のカテゴリと、第2のカテゴリとの間に、確立された相関関係が存在するか否かを判定し、第1のカテゴリと、第2のカテゴリとの間に、確立された相関関係が存在すると判定された場合、第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間の相関関係を確立する。

【0022】

一態様では、検索エンジンは、ユーザによって入力された検索語句に対応するラベルを識別する、第1のクエリーモジュールと、ラベルに対応するウェブページを識別する、第2のクエリーモジュールと、ユーザに、ウェブページまたはウェブページへのリンクを提供する、インターフェースモジュールと、ウェブページの1つまたは複数のキーワードおよびこれらの1つまたは複数のキーワードに関連する知識ベースのエントリに基づいて、ウェブページに対応するラベルを作成する、ラベル作成モジュールとを含む。

【0023】

別の態様では、検索エンジンは、知識ベースのエントリに基づいて、ユーザによって入力された検索語句を解析する、解析モジュールと、検索語句から解析された単語を、知識ベースのエントリと照合する、照合モジュールと、検索語句から解析された単語との一致を有するエントリに関連する、知識ベースのエントリを識別する、クエリーモジュールと、検索語句から解析された単語との一致を有するエントリに関連する、知識ベースのエントリで、検索語句を更新する、更新モジュールと、更新した検索語句に基づいて、検索を実施する、検索モジュールとを含む。

【図面の簡単な説明】

【0024】

【図 1 A】本開示の一実施形態による、コンピューティング装置の図である。

【図 1 B】本開示の一実施形態による、ネットワークシステムの図である。

【図 1 C】本開示の一実施形態による、知識ベースの作成の流れ図である。

【図 2】本開示の、別の実施形態による、知識ベースの作成の流れ図である。

【図 3】本開示の一実施形態による、ウェブページのスキーマを分析する場合の、情報検索の流れ図である。

【図 4】本開示の一実施形態による、ユーザの意図を分析する場合の、情報検索の流れ図である。

【図 5】本開示の、別の実施形態による、コンピューティング装置の図である。

【図 6】本開示の一実施形態による、検索エンジンのブロック図である。

10

【図 7】本開示の、別の実施形態による、検索エンジンのブロック図である。

【発明を実施するための形態】

【0025】

本開示は、ウェブページに出現する単語を分析する技術を説明する。ウェブページからの文中の単語であり、かつ知識ベース内のカテゴリに追加される単語は、そのカテゴリ下のエントリと見なされる。カテゴリ間の相関関係に基づいて、対となって現れるエントリ間の相関関係もまた、確立される。これは、知識ベースの自動構築を可能にし、したがって、処理における手作業での資源の必要性を回避する。

【0026】

一実施形態では、知識ベースは、1つまたは複数のカテゴリを含む。各カテゴリは、それぞれの対応するエントリおよび代表的な単語を有する。1つのエントリは、1つまたは複数のカテゴリに対応してもよく、異なるカテゴリに対して異なる重み付けを有してもよい。エントリもまた、対応するプロパティを有することができる。さらに、カテゴリ間およびエントリ間に相関関係が確立されてもよい。たとえば、「製品」のカテゴリは、「携帯電話」という対応するエントリ、ならびに、「販売」、「モデル」、「ブランド」、および「機能性」等の代表的な単語を有してもよい。エントリ「携帯電話」は、機能性、サイズ、バッテリータイプ、等のプロパティを有してもよい。一実施形態では、カテゴリ、各カテゴリに対応する代表的な単語、およびカテゴリ間の相関関係は、知識ベース内にあらかじめ設定される。知識ベースが増大すると、エントリ、エントリ間の相関関係、およびエントリのプロパティが追加されることになる。

20

【0027】

【表 1】

表 1. エントリとカテゴリとの間の相関関係の例

エントリ	重み付けの合計 (すべてのカテゴリにおける重み付けの合計)	対応するカテゴリ (このカテゴリにおけるエントリのそれぞれの重み付け)
アップル	340,000	果物(100,000)、ラップトップコンピュータ(100,000)、携帯電話(100,000)、衣料品(40,000)
.....		

30

【0028】

【表 2】

表 2. エントリおよびそのエントリの対応するプロパティの例

エントリ	プロパティ
携帯電話	サイズ
	バッテリータイプ

40

【0029】

【表 3】

表 3. エントリ間の相関関係の例

エントリ	関連エントリ
携帯電話	Nokia
	

【0030】

【表 4】

表 4. カテゴリ間の相関関係の例

カテゴリ	関連カテゴリ
製品	ブランド
	

10

【0031】

【表 5】

表 5. カテゴリおよびそのカテゴリの対応する代表的な単語

カテゴリ	代表的な単語
製品	販売
	

20

【0032】

表 5 に示される「販売」に加えて、カテゴリ「製品」に対応するかもしれない、他の代表的な単語は、たとえば、「モデル」、「ブランド」等を含む。別の例として、カテゴリ「映画とテレビ」は、「監督」、「主演男優」、「主演女優」、「リリース」等の代表的な単語を含むかもしれない。一実施形態では、各カテゴリの代表的な単語は、それぞれのカテゴリの特性に基づいて、あらかじめ設定される、すなわち既定である。

【0033】

一実施形態では、表 1 ～ 表 5 のデータを保存するために、テキスト文書、表、データベース、または他の適切な手段を用いてもよい。表 1 ～ 表 5 は、例として提供されており、相関関係を変えることなく異なる方法で組み合わせられてもよいことを理解されたい。

30

【0034】

図 1 A に示されるように、一実施形態では、開示される知識ベースを構築するコンピューティング装置は、基本データ処理層、データマイニング層、統合層、および利用層を含んでもよい。あるいは、これらの機能的な層は、異なるコンピューティング装置内に実装されてもよい。これらの異なるコンピューティング装置は、サーバおよび／またはクライアント端末装置としてもよく、図 1 B に示されるように、ネットワークを形成することができる。たとえば、基本データ処理層は、クライアント 11 内に実装されてもよく、データマイニング層は、サーバ 12 内に実装されてもよく、統合層は、サーバ 12 またはサーバ 13 内に実装されてもよく、かつ利用層は、クライアント 14 内に実装されてもよい。他の実施形態では、クライアント 11、サーバ 12、サーバ 13、およびクライアント 14 に加えて、他のサーバおよびクライアントが存在してもよい。

40

【0035】

基本データ処理層は、ウェブページから文を取得する。取得された文は、ウェブページのコンテンツからの文とすることができる。データマイニング層は、取得された文のそれぞれを、単語に解析して、知識ベース内のカテゴリ、たとえば第 1 のカテゴリの代表的な単語を、文から解析された単語と照合する。代表的な単語と、文から解析された単語との間に良好な一致が存在する場合、文から解析された、一致した単語に隣接する単語および

50

／または記号の列は、第1のエントリとして、第1のカテゴリに追加される。文から解析された単語を、知識ベースの第2のカテゴリの第2のエントリと照合するとき、第1のカテゴリと、第2のカテゴリとの間に相関関係が確立されたか否かについての判定が行われる。第1のカテゴリと、第2のカテゴリとの間に相関関係が存在する事象では、第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間に、相関関係が確立される。つまり、第2のカテゴリの第2のエントリは、第1のカテゴリの第1のエントリの対応するエントリとして、追加される場合がある。同様に、第1のカテゴリの第1のエントリは、第2のカテゴリの第2のエントリの対応するエントリとして追加される場合がある。当業者は、上記に説明される第1のカテゴリおよび第2のカテゴリは、任意の2つのカテゴリとなる場合があることを理解するであろう。便宜のため、および2つのカテゴリを区別するために、それらは第1のカテゴリおよび第2のカテゴリと称される。同様に、第1のエントリおよび第2のエントリも、任意の2つのエントリとなる場合がある。

10

【0036】

コンピューティング装置はまた、図1Aに示されるように、統合層および利用層も含んでもよい。統合層は、データマイニング層によって作り出された、さまざまなカテゴリのための結果ファイルを、単一の結果ファイルに統合する。利用層は、データの利用を可能にする。

【0037】

説明のため、および例として、データマイニング層は、カテゴリ1、カテゴリ2、およびカテゴリ3のための、以下の結果ファイルを作り出す。

20

【0038】

【表6】

結果ファイル1	結果ファイル2	結果ファイル3
カテゴリ1	カテゴリ2	カテゴリ3
エントリ1 100	エントリ1 50	エントリ1 80
エントリ2 50	エントリ2 100	エントリ2 8
	エントリ3 80	エントリ3 100

30

【0039】

統合層は、これらの3つの結果ファイルを、下の表6に示されるように単一の結果ファイルに統合する。

【0040】

【表7】

表6. 統合後の結果ファイルの例

カテゴリ エントリ 重み付け	カテゴリ 1	カテゴリ 2	カテゴリ 3
エントリ 1	100	50	80
エントリ 2	50	100	8
エントリ 3	0	80	100

40

【0041】

表6において、「0」は、エントリと、カテゴリとの間に相関関係がないことを示す。

【0042】

図1Cは、一実施形態による、以下のステップを含む、知識ベース構築の一般的な過程

50

100を示す。

【0043】

101において、コンピューティング装置内の基本データ処理層は、ウェブページから文を取得する。

【0044】

102において、コンピューティング装置のデータマイニング層は、文を解析、すなわち分割する。

【0045】

103において、データマイニング層は、知識ベースの第1のカテゴリに対応する代表的な単語を、文から解析された単語と照合する。

【0046】

知識ベースの構築の開始において、カテゴリ、および各カテゴリに対応する代表的な単語は、定義され、かつ確立される必要がある。知識ベースの構築が継続すると、新たなエントリが知識ベースに追加されるにつれて、代表的な単語は更新される。

【0047】

104において、代表的な単語と、文から解析された単語との間に良好な一致が存在する場合、データマイニング層は、文中の一致した単語に隣接する単語および／または記号の列を、第1のエントリとして、第1のカテゴリに追加する。

【0048】

105において、文から解析された単語を、知識ベースの第2のカテゴリの第2のエントリと照合するとき、データマイニング層は、第1のカテゴリと、第2のカテゴリとの間に相関関係が確立されたか否かを判定する。第1のカテゴリと、第2のカテゴリとの間に相関関係が存在する事象において、データマイニング層は、第1のカテゴリの第1のエントリと、第2のカテゴリの第2のエントリとの間に、相関関係を確立する。

【0049】

知識ベースを構築するための、本明細書に説明される過程は、知識ベースを更新するために使用することができ、かつ定期的に繰り返すことができる。

【0050】

図2は、一実施形態による、以下のステップを含む、知識ベース構築の詳細な過程200を示す。

【0051】

201において、データ処理層は、ウェブページから文を取得する。具体的には、データ処理層は、単文および句、ならびに文の出現の頻度、すなわち、ウェブページ上の同じ文の出現率を取得する。ウェブページ上のテキストメッセージは、テキストメッセージから得られる文中の句読点によって、前もって保存して、後に収集することができる。

【0052】

文は、単文、句、または長文である可能性がある。単文は、終止符、疑問符、または感嘆符の前の文であって、文の単語間に他の句読点を有さない文を意味する。句は、コンマまたはセミコロンを末尾に使用するものであって、句の単語間に他の句読点を有さないものを意味する。長文は、その間に1つまたは複数のコンマまたはセミコロンを有する、終止符、疑問符、または感嘆符の前の文を意味する。長文が検索される場合、句読点にしたがって、多くの短い句に分割される。文がより長く、内容がより複雑になるにつれて、より容易に分析するために、多くの句に分割され、したがって、さらにより正確な結果を得ることになる。たとえば、検索される文は、「本店出售AA牌BB1型号的手机」であってもよい。

【0053】

202において、データマイニング層は、解析システムを使用して、取得された文を解析する。たとえば、文「本店出售AA牌BB1型号的手机」は解析後、「本店、出售、AA、牌、BB1、型号、手机」となる。このカテゴリに対応する単語は、解析システムに追加することができ、これは、文を分割するために使用される。

10

20

30

40

50

【0054】

解析すること、すなわち分割を達成するのは、容易ではない。たとえば、「诺基亚」という語句は、小規模で基本的な用語集しか含まない傾向がある、従来の解析システムを使用する場合、容易に解析されない場合がある。通常、従来の解析システムは、最新の外国語または翻訳を有さない。従来の解析システムが、単語を照合する手段を有しない場合、従来の解析システムは、分割の単位として、未知の単語の個々の文字を使用する。したがって、「诺基亚」という語句は、「诺、基、亚」として解析される可能性がある。「诺基亚」という語句が解析システムに追加されると、「诺基亚」という語句は良好な一致となることができる。つまり、「诺基亚」という語句は、1つの完全な単語として解析される。

10

【0055】

203において、データマイニング層は、第1のカテゴリの代表的な単語を、解析された単語と照合する。代表的な単語と、文から解析された単語とが、一貫して一致する場合、この一致は、この文について良好であると考えられ、良好に一致した単語は、保持される。第1のカテゴリでは、不一致の文は、落とされる。不一致の文は、他のカテゴリの代表的な単語と照合するために再利用することができる。

【0056】

204において、マイニング層は、良好な一致が、知識ベースにまだ含まれていない未知の単語を有するかどうかを判断する。(下記に説明されるステップ205に続くが)良好な一致が知識ベースにまだ含まれていない未知の単語を有しない場合には、文の終わりに、過程200は、他の良好な一致が知識ベース内にまだ含まれていない未知の単語を有するかどうかを判断するために、さらに継続することができる。未知の単語が含まれていない場合、過程200は、それでもなお、他のカテゴリの代表的な単語を、それぞれの文からそれらを解析した後で得られる単語と照合することができる。次いで、ステップ203が繰り返される。

20

【0057】

205において、マイニング層は、文中の、良好に一致した単語に隣接する未知の単語および／または記号の列を、第1の層に追加される第1のエントリとして見なす。列は、多数の未知の単語を含む場合がある。たとえば、「新片满城尽带黄金甲」(英語訳「the new movie Curse of the Golden Flower」)(日本語訳「新作映画王妃の紋章」)という句に対する文は、代表的な単語と照合するために、「新／片／满／城／尽／带／黄金／甲」となるように、個別の文字または語句に解析され、そこで「满／城／尽／带／黄金／甲」は未知の単語である。「满城尽带黄金甲」という句は、単語「片」に隣接する未知の列と考えられ、これは、独立し、かつ完全な単語として扱われる。

30

【0058】

206において、データマイニング層は、解析システムを更新するために、第1のエントリを解析システムに追加することができる。更新された解析システムは、単語を容易に解析しない場合がある。たとえば、句「新片满城尽带黄金甲」に再び遭遇した場合、解析システムは、句を1つの単語、「新片满城尽带黄金甲」として扱い、たとえば、「新／片／满／城／尽／带／黄金／甲」に、解析しないであろう。

40

【0059】

207において、データマイニング層は、第1のエントリの出現の頻度と、それらがその中に配置された文中の隣接する代表的な単語とに基づいて、第1のカテゴリにおける第1のエントリの重み付けを提供する。たとえば、取得した文の出現の頻度の計数で、文1

50

において出現する、第1のエントリBB1および代表的な単語「型号」の回数は、1000である。それらが文2内で出現する回数は100であり、文3内では、出現の回数は10である。したがって、重み付けは、 $f(1000) + f(100) + f(10)$ である。これらのそれぞれは、たとえば、10を底とする対数関数などの、重み付けの関数としての、それぞれの文における出現の頻度である。

【0060】

208において、データマイニング層は、文における、第1のカテゴリの第1のエントリおよび第2のカテゴリの第2のエントリの出現頻度を取得する。それにより、第1のカテゴリと、第2のカテゴリとの間の相関関係が確立される。

【0061】

209において、この頻度がデフォルトの相関関係の閾値を超えると、データマイニング層は、第1のエントリと第2のエントリとの間の関係を確立する。一実施形態では、第1のエントリに対する追加の相関関係を確立するように、ステップ208を繰り返すことができる。相関関係の閾値を通して、過程200は、誤記による相関関係内のエラーを、フィルタにかけて除去することができる。たとえば、既に確立された、カテゴリ「モデル」と、カテゴリ「ブランド」との間の相関関係では、「BB1」と「AA」との間の相関関係を確立することができる。

【0062】

一実施形態では、ステップ206、207、および208は、3つの個別の過程であり、完全に連続的な実装を有せず、また同時に実装することもできる。

【0063】

一実施形態では、知識ベースは、共通単語システムおよび実単語システムを含む。実単語システムに含まれる単語は、インデックスワードに対応し、共通単語システムに含まれる単語は、シードワードに対応する。共通単語システムに含まれるエントリは、主に、場所の名前等のように変化することの少ない、決まった単語である。実単語システムに含まれるエントリは、個人の名前および映画の題名等のように、高い頻度で更新される単語である。共通単語システムと、実単語システムとの間の違いは、それぞれのシステムに含まれるカテゴリに依存する。実単語システム内のインデックスワードは、対応するカテゴリの下のエントリに含まれない。共通単語システム内のシードワードは、対応するカテゴリの下のエントリに属する。共通単語システム下および実単語システム下のカテゴリは、異なる更新の周期を使用することができる。共通単語システムの更新の周期は、実単語システムの更新の周期よりも長い周期とすることができる。

【0064】

表7および表8は、それぞれ、共通単語システムの実例および実単語システムの実例を示す。

【0065】

【表8】

表7. 共通単語システムの例

共通単語システム		
カテゴリ11	カテゴリ12	……

【0066】

【表9】

表8. 実単語システムの例

実単語システム		
カテゴリ21	カテゴリ22	……

【0067】

未知の列が、第1のエントリとして、第1のカテゴリに追加されると、第1のエントリとしての未知の列は、第1のカテゴリが属する（共通単語システムまたは実単語システムのいずれにおいても）システムに追加される。第1のカテゴリが、共通単語システム内のカテゴリである場合、第1のエントリは、第1のカテゴリに対応するシードワードである可能性がある。

【0068】

マイニング層はまた、特徴的な記号に基づいて、未知の列が、第1のカテゴリ内のエントリに対応するかどうか判断することができる。特徴的な記号は、所与のカテゴリに関連する、句読点等の、たとえば、角括弧、コンマ、著作記号およびその他を含む。たとえば、カテゴリが映画またはテレビである場合、基本データ処理層は、著作記号を有する文を得てもよく、マイニング層は、映画のカテゴリ内の対応するインデックスワードと、著作記号を有する文中の単語とを照合するであろう。良好な一致が存在する場合、次いで、著作記号で引用された単語（すなわち、未知の列）は、映画（またはテレビ）のカテゴリの下でのエントリとなる。丸括弧内の単語は、通常、英語の固有名詞（括弧の前の単語）であり、コンマの前および後の単語は、通常、同じカテゴリに属する。

【0069】

データマイニング層はまた、第1のエントリのためのプロパティを設定することもできる。一実施形態では、データ処理層は、ウェブページからテーブルを取得する。データマイニング層は、所与の単語を第1のエントリのプロパティとし、これは、そのような単語が第1のエントリと対となって何度もテーブル内に出現するときになされる。たとえば、第1のエントリは製品とすることができる。これは、通常、製品の出自、製造業者、サイズ、モデル（または仕様）を列挙するテーブルの形態である。たとえば、多くの種類および多くのタイプの製造業者が存在する場合があるが、「製造業者」という単語は、第1のエントリと対となって何度も出現する。そのような事例では、「製造業者」という単語は、第1のエントリのプロパティとされる。

【0070】

データマイニング層は、カテゴリを1つずつ分析して、各カテゴリごとに、それぞれの結果ファイルを作成する。この結果ファイルは、カテゴリ、カテゴリの対応するエントリ、およびカテゴリの各エントリの重み付けを含んでもよい。知識ベースは通常1つのカテゴリのみを有することはないことを考慮すると、統合層を介して多数の結果ファイルを1つの結果ファイルに一体化することができる。

【0071】

統合層は、対応するエントリのカテゴリをフィルタにかけることができる。データマイニング層は、代表的な単語を伴う未知の列の出現の結果、未知の列を、所与の代表的な単語に対応するカテゴリに追加する。フィルタリングが、代表的な単語と共に出現する未知の列の頻度のみに基づく場合、フィルタリングにおけるエラーが発生する場合がある。たとえば、少ない頻度で出現するが、それでもやはり正しい、いくつかの一般的でない単語が存在する場合がある。その一方で、高い頻度で出現するにもかかわらず、恐らく誤記に起因する、かかる一般的な単語としては特定の文中に出現するのは誤りかもしれない、いくつかの一般的な単語が存在する場合がある。そのような問題は、データマイニング層によって理解されない可能性があるため、統合層によるフィルタリングが必要である。一実施形態では、統合層は、所与のエントリの、そのエントリに対応するさまざまなカテゴリにおける、個別の重み付けを比較する。比較が特定の条件に適合する場合、このエントリがこれらのカテゴリに追加されることが正しいと見なされる。比較が特定の条件に適合しない場合、エントリと、エントリが正しくなく追加されたカテゴリとの間の相関関係は、取り消される。比較を行うための、多くの方法が存在する。一実施形態では、最も大きい重み付けと、ゼロでない最も小さい重み付けとが比較され、最小重み付けと最大重み付けとの比率が、第1の閾値よりも低い場合、最小重み付けはゼロに設定されて、それぞれのエントリと、最小重み付けに対応するカテゴリとの相関関係が取り消される。あるいは、所与のエントリに対するゼロでない最小重み付けが、エントリの総重み付け（エントリの

10

20

30

40

50

重み付けの合計)と比較され、総重み付けに対する最小非ゼロ重み付けの比率が、第2の閾値よりも低い場合、最小非ゼロ重み付けされたものはゼロに設定されて、それぞれのエン트리と、最小非ゼロ重み付けに対応するカテゴリとの間の相関関係は取り消される。

【0072】

知識ベースは、さまざまな分野で 사용할 ことができる。たとえば、知識ベースは、より良い検索結果を得るために、検索エンジンにサービスを提供して、ユーザの目的を分析するように使用することができる。別の例として、知識ベースは、ユーザに、連想させる情報を提供することによって、ユーザにプロンプトを提供することができる。したがって、いくつかの実施形態では、知識ベースは、アプリケーション層も含み、検索の実行が、アプリケーション層を活用する1つの方法である。

10

【0073】

図3は、ウェブページのスキーマを分析する場合の情報検索の方法300を示す。

【0074】

301において、ユーザによって入力された検索語句から解析された単語に基づいて、解析された単語は、一致する単語、またはラベルを得るように、検索語句と比較される。

【0075】

302において、一致した単語に対応するウェブページが得られる。

【0076】

303において、得られたウェブページ、または得られたウェブページへのリンクがユーザに提供される。ここで、一致した単語、またはラベルは、ウェブページの1つまたは複数のキーワードと、その1つまたは複数のキーワードに関連する、知識ベースのエントリとに基づいて得られる、新たな検索語である。

20

【0077】

ラベルを得ることの過程は、ウェブページからキーワードを抽出することと、キーワードを知識ベース内のエントリと照合することと、良好に一致したエントリに関連する、関連エントリを得ることと、キーワードおよび関連エントリに基づいて、ラベルを得ることとを含む。この方法で得られるラベルは、ウェブページのコンテンツをより正確に反映することができ、したがって、ラベルを通して、ユーザは、より満足のいく検索結果を得ることができる。たとえば、ウェブページのコンテンツが、句「N78携帯電話の販売」を含む場合に、ユーザが検索語句「诺基亚」(英語で「Nokia」を意味する)を入力すると、既存の検索技術の下では、このウェブページが見つけれられない可能性が高い。なぜならこれは、このウェブページが、句「Nokia」および「Nokia」の同義語のどちらも含まないからである。しかし、開示される知識ベースを用いて、開示される技術を使用すると、「N78」は「Nokia」というブランドのモデルであり、したがって、ユーザが、実際にNokiaの携帯電話のモデルN78を検索している場合、ユーザに提供される検索結果は、より正確になりかもしれない。

30

【0078】

図4は、ユーザの意図を分析する場合の情報検索の過程400を示す。

40

【0079】

401において、ユーザによって入力された検索語句は、知識ベース内のエントリに基づいて解析される。この事例では、検索語句は、多数の単語を有する文、単語、または句とすることができる。たとえば、ユーザは、検索語句「在哪儿能买到BB1」（英語で、「at what place can BB1 be purchased」を意味する）（日本語で、「どこの場所でBB1が購入できるか」を意味する）を入力する場合がある。解析後、検索語句は、次の単語／句に分割することができる。在、哪儿、能、买到、BB1（英語で、「at」、「what place」、「can」、「purchased」、および「BB1」を意味する）（日本語で、「で」、「どこの場所」、「できる」、「購入」、および「BB1」を意味する）。

10

【0080】

402において、検索語句から解析された単語／句を、良好な一致を有するエントリまたは複数のエントリを識別するために、知識ベースのエントリと照合する。たとえば、「購入」は、「売買」のカテゴリの下のエントリであり、それに対して「BB1」は、「モデル」のカテゴリの下のエントリである。

【0081】

403において、知識ベースに基づいて、良好な一致を有するエントリに関連するエントリが得られる。たとえば、「BB1」は、エントリ「AA」および「携帯電話」に関連し、ここで「AA」は「ブランド」のカテゴリに対応し、「携帯電話」は「製品」のカテゴリに対応する。

20

【0082】

404において、検索語句は、関連エントリに基づいて更新される。たとえば、更新された語句は、ユーザの目的をより正確に反映する、「AAブランドの携帯電話購入、モデルはBB1」とすることができる。

【0083】

405において、ウェブページのキーワードおよび更新された検索語句に照合される。具体的には、図3に関連して説明されたラベルと、更新された検索語句が照合されて、良好に一致したラベルに対応するウェブページが識別される。

【0084】

30

406において、識別されたウェブページまたはそのようなウェブページへのリンクが、検索結果としてユーザに提供または提示され、それにより情報検索が達成される。一実施形態では、ウェブページ、またはウェブページへのリンクがユーザに提示される順序は、ウェブページのそれぞれの、ラベルとキーワードとの間の良好な一致の程度に依存する場合がある。最も一致するカテゴリおよびエントリを有するウェブページは、最も良好な一致を有するウェブページであると見なされる。

【0085】

エントリは、複数のカテゴリに対応する場合がある。「アップル」を例にとると、「アップル」は、「果物」のカテゴリの下のエントリ、「衣料品」のカテゴリの下のエントリ、またはさらには「電気製品ブランド」の下のエントリとなる可能性がある。したがって、検索語句の更新およびウェブページの更新の過程において、さまざまなカテゴリに基づいて、追加の検索語句が得られる場合がある。ユーザの意図に最も近い検索語句は、さまざまな更新された検索語句の中から識別されることとなり、これを達成するための多くの方法が存在する。たとえば、カテゴリに対応する、最大の重み付けを有するエントリを判定することができる。知識ベースにおいて、最大の重み付けを有するカテゴリに対応するエントリに基づいて、良好に一致したエントリに関連するエントリが得られる。さらに、これらの関連エントリに基づいて、ユーザによって入力された検索語句が更新される。あるいは、解析後に得られる単語と、多数のカテゴリに対応する代表的な単語とが照合される。知識ベースを通して、および、良好に一致した代表的な単語（複数可）に対応するカテゴリに応じて、そのようなカテゴリに対応するエントリに関連するエントリを得ること

40

50

ができる。検索語句は、得られたエントリに基づいて更新することができる。

【0086】

開示される知識ベースはさらに、ユーザが情報の普及を望む場合、ユーザにプロンプトを提供することができる。たとえば、ユーザが、携帯電話に関連する販売情報を公開することを望む場合において、ユーザが製品フィールド内に「携帯電話」と入力して、良好な一致が存在する場合に、ユーザに、「携帯電話」に関連するエントリならびに「携帯電話」というエントリのプロパティ等のプロンプトが提供または提示されてもよい。その後、ユーザは、プロンプトされた情報をクリックすることによって、他の入力フィールドを完了することができる。そのようにして、ユーザ体験が高まる一方で、操作プロセスは簡易化される。

10

【0087】

上記の説明は、当業者が、開示される知識ベースを構築する手順、およびそのような知識ベースを使用して情報検索を達成する手順を理解することを可能にする。実際の実装は、一装置によって実施することができ、そのような装置の説明が下に説明される。

【0088】

図5は、本開示の一実施形態による、コンピューティング装置500を示す。開示される知識ベースを構築するために使用されるコンピューティング装置の各層は、機能モジュールを用いて実装することができる。したがって、コンピューティング装置は、基本データ処理モジュール501およびデータマイニングモジュール502を含む。

【0089】

基本データ処理モジュール501、すなわちコンピューティング装置500の基本データ処理層は、ウェブページから文を得るために使用される。

20

【0090】

データマイニングモジュール502、すなわちコンピューティング装置500のデータマイニング層は、得られた文を解析するために使用される。データマイニングモジュール502は、知識ベースの第1のカテゴリに対応する代表的な単語を、解析から得られる単語と照合する。少なくとも1つの解析された単語が良好に一致する場合、文中の一致した単語に隣接する未知の単語および／または記号の列は、第1のエントリとして扱われ、第1のカテゴリに追加されるであろう。文中の単語が、第2のカテゴリの第2のエントリと一致する場合、データマイニング層502は、第1のカテゴリと、第2のカテゴリとの間に相関関係が存在するか否かを判定する。相関関係が存在する場合、第1のエントリと、第2のエントリとの間の相関関係が確立される。データマイニングモジュール502はまた、エントリのためのプロパティ／複数のプロパティを確立することもでき、それぞれのカテゴリごとの結果ファイルを作成することもできる。

30

【0091】

コンピューティング装置500は、さらに、統合モジュール503（すなわち、統合層）および利用モジュール504（すなわち、利用層）を備える。統合モジュール503は、データマイニングモジュール502からの結果ファイルを、1つの結果ファイルに統合し、かつエントリに対応するカテゴリをフィルタにかける。

【0092】

利用モジュール504は、さまざまな種類のアプリケーションを提供する。検索エンジンは、利用モジュール504のアプリケーションユニットのうちの1つである。

40

【0093】

図6は、本開示の一実施形態による、検索エンジン600を示す。検索エンジン600は、第1のクエリーモジュール601、第2のクエリーモジュール602、インターフェースモジュール603、およびラベル作成モジュール604を含む。

【0094】

第1のクエリーモジュール601は、ユーザによって入力された検索語句に対応するラベルを得る。第2のクエリーモジュール602は、ラベルに対応するウェブページを得る。インターフェースモジュール603は、ユーザに、ウェブページまたはウェブページへ

50

のリンクを提供する。ラベル作成モジュール604は、ウェブページの1つまたは複数のキーワードと、その1つまたは複数のキーワードに関連する知識ベースのエントリとに基づいて、ウェブページに対応するラベルを作成する。

【0095】

図7は、本開示の別の実施形態による、検索エンジン700を示す。検索エンジン700は、解析モジュール701、照合モジュール702、クエリーモジュール703、更新モジュール704、および検索モジュール705を含む。

【0096】

解析モジュール701は、知識ベースのエントリに基づいて、ユーザによって入力された検索語句を解析する。照合モジュール702は、検索語句から解析された単語を、知識ベースのエントリと照合する。クエリーモジュール703は、検索語句から解析された単語との一致を有するエントリに関連する、知識ベースのエントリを識別する。更新モジュール704は、検索語句から解析された単語との一致を有するエントリに関連する、知識ベースのエントリを有する検索語句を更新する。検索モジュール705は、更新された検索語句に基づいて検索を行う。さらに、検索モジュール705は、ウェブページの文を、更新されたキーワードと照合して、ユーザに、キーワードとの良好な一致を有するウェブページまたはウェブページへのリンクを提供する。一実施形態では、良好な一致を有する、複数のウェブページが存在する場合、検索モジュール705は、ユーザに一致を有するウェブページ、または、たとえば、最も良好な一致を有するウェブページから、良好の程度が最も低い一致を有するウェブページの順序である、降順で、そのようなウェブページへのリンクを提供してもよい。

【0097】

検索エンジン600および検索エンジン700は、それぞれ、図6および図7に示される特徴および機能性を含む、単一の検索エンジンの一部とすることができる。第1のクエリーモジュール601および第2のクエリーモジュール602は、検索モジュール705と同等であり、これは更新された検索語句に基づいて、ウェブページを検索するために、更新された検索語句に対応するラベルを取得する。検索エンジン700はまた、インターフェースモジュール603を含んでもよく、インターフェースモジュール603は、ユーザから検索語句を受け取り、ユーザに、検索から識別されたウェブページ（複数可）またはウェブページ（複数可）へのリンク（複数可）を提供する。

【0098】

説明の便宜のため、例示的なコンピューティング装置または検索エンジンの特徴および機能は、さまざまなモジュールとして説明される。当然ながら、さまざまな実施形態において、本明細書に説明される任意のモジュールの特徴および機能は、ソフトウェアまたはハードウェアの1つまたは複数の事例に実装されてもよい。

【0099】

開示されるコンピューティング装置、検索エンジン、およびそれらのモジュールは、ソフトウェアおよび／またはハードウェアを使用して実装されてもよい。ソフトウェアで実装される場合、ソフトウェアは、フロッピー（登録商標）ディスク、ハードディスク、CD-ROM、およびフラッシュメモリ等の、1つまたは複数のコンピュータ可読媒体に格納されてもよい。開示される方法、知識ベース、および検索エンジンは、ネットワークシステムの1つまたは複数のネットワークコンピュータに実装されてもよい。

【0100】

本開示の実装形態は、文中の単語と、知識ベース内の印をつけられた単語とを照合するであろう。良好に一致する単語に基づいて、知識ベース内のカテゴリに対して、未知の単語が判定され、かつそのカテゴリの下のエントリとして見なされる。また、カテゴリ内の相関関係に基づいて、知識ベースを更新するために、文中に出現するエントリ間に相関関係が構築される。本開示の実装形態はまた、未知の単語および良好に一致した印をつけられた単語の出現の頻度に基づいて、対応するカテゴリ下で未知の単語の重み付けを設定する。それはまた、知識ベースの各フィールドごとに、より多くの情報を提供するために、

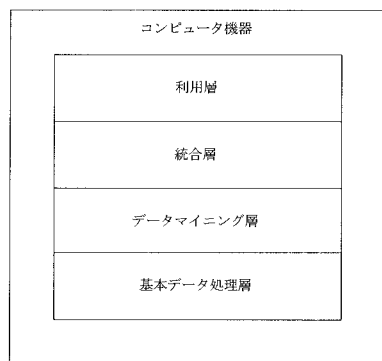
ウェブページのフォーム内の未知の単語の出現を通して、未知の単語のプロパティも設定する。同時に、本開示の実装形態は、ユーザの意図に関して、より正確となるために、知識ベースを通して、ユーザによって入力された検索単語を更新するために使用される。それはまた、より正確な検索結果を有するために、更新された検索語句に基づいて検索する。また、実装形態は、ウェブページがユーザの目的をより正確に示すように、知識ベースを通して、ウェブページに対して、主題のタグを設定する。それはまた、より正確な検索結果を達成するために、タグと更新された検索単語とを照合する。

【0101】

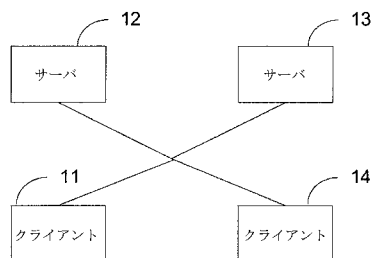
当然ながら、当業者は、本開示の精神および範囲を逸脱することなく、多くの異なる方法で、本開示を変更または修正することができる。したがって、本開示が、本開示の特許請求の範囲およびそれらの等価物の範囲に含まれる、すべての修正および変形を対象とすることが意図される。

10

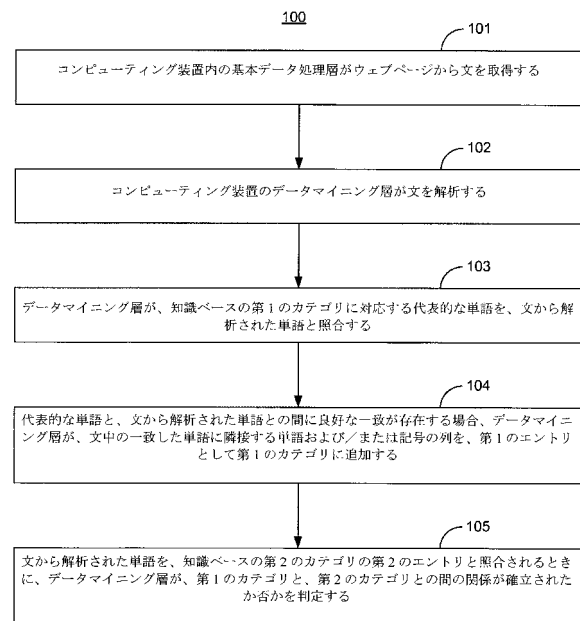
【図1A】



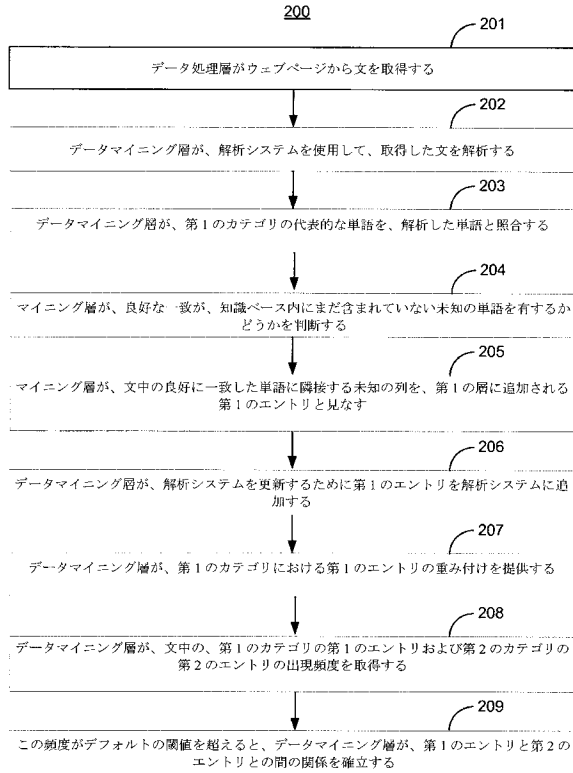
【図1B】



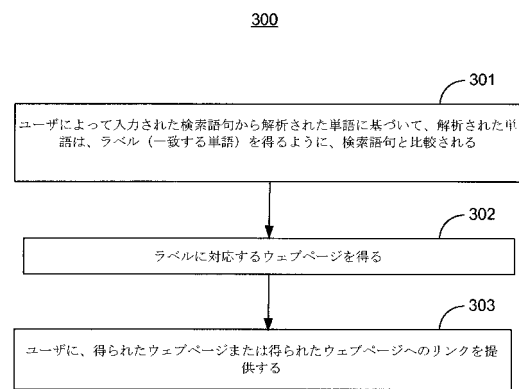
【図1C】



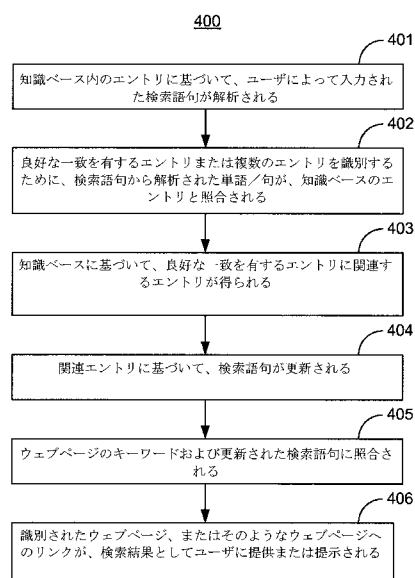
【図 2】



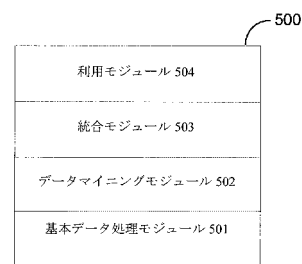
【図 3】



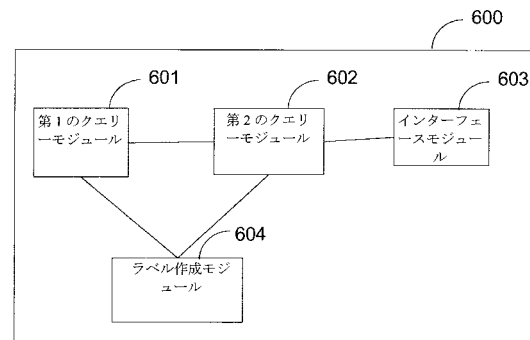
【図 4】



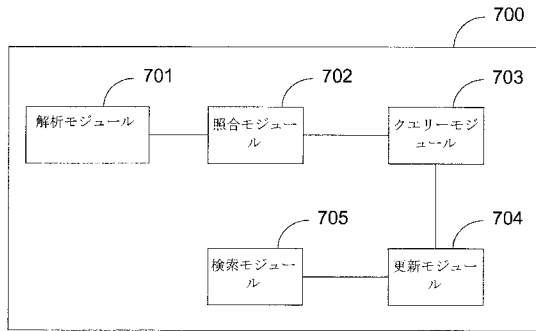
【図 5】



【図 6】



【図 7】



フロントページの続き

- (72)発明者 ホウ レイ
中華人民共和国 310013 ハンチョウ ウェン アル ロード ナンバー 391 ザ ウ
ェスト レイク インターナショナル プラザ オブ エスアンドティー ビルディング エー
10/エフ
- (72)発明者 チン ジーシェン
中華人民共和国 310013 ハンチョウ ウェン アル ロード ナンバー 391 ザ ウ
ェスト レイク インターナショナル プラザ オブ エスアンドティー ビルディング エー
10/エフ
- (72)発明者 チェン ウェイ
中華人民共和国 310013 ハンチョウ ウェン アル ロード ナンバー 391 ザ ウ
ェスト レイク インターナショナル プラザ オブ エスアンドティー ビルディング エー
10/エフ
- (72)発明者 チャン チン
中華人民共和国 310013 ハンチョウ ウェン アル ロード ナンバー 391 ザ ウ
ェスト レイク インターナショナル プラザ オブ エスアンドティー ビルディング エー
10/エフ

審査官 早川 学

- (56)参考文献 特開平05-298371 (JP, A)
特開2008-225560 (JP, A)
特開2006-178671 (JP, A)

- (58)調査した分野(Int.Cl., DB名)
G06F 17/30