

(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2009-282913

(P2009-282913A)

(43) 公開日 平成21年12月3日(2009.12.3)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 17/30 (2006.01)	G O 6 F 17/30 3 4 O A	5 B O 7 5
G 0 6 F 13/00 (2006.01)	G O 6 F 17/30 2 1 O A	
	G O 6 F 17/30 4 1 9 B	
	G O 6 F 13/00 5 4 O E	
	G O 6 F 13/00 5 4 O R	
審査請求 未請求 請求項の数 7 O L (全 13 頁)		

(21) 出願番号	特願2008-136861 (P2008-136861)	(71) 出願人	000004226
(22) 出願日	平成20年5月26日 (2008. 5. 26)		日本電信電話株式会社
			東京都千代田区大手町二丁目3番1号
		(74) 代理人	100070150
			弁理士 伊東 忠彦
		(74) 代理人	100124844
			弁理士 石原 隆治
		(72) 発明者	中渡瀬 秀一
			東京都千代田区大手町二丁目3番1号 日
			本電信電話株式会社内
		(72) 発明者	井沢 味奈子
			東京都千代田区大手町二丁目3番1号 日
			本電信電話株式会社内
		Fターム(参考)	5B075 KK02 NK39 PR08 UU40

(54) 【発明の名称】 個人適応型Web情報検索装置及び方法及びプログラム

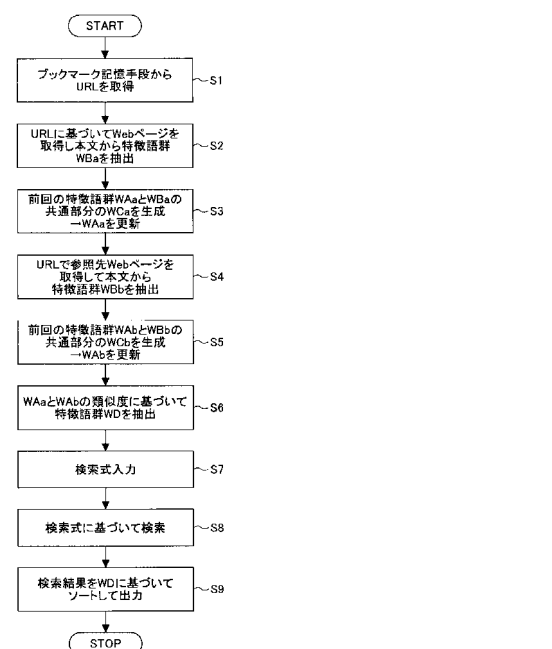
(57) 【要約】

【課題】個人の興味に応じたインターネット情報検索ができるようにするために、個人のプロフィール情報を適切に取得することによって個人適応検索を行う。

【解決手段】本発明は、ある利用者のプロフィール情報となる特徴語の組を当該利用者のブックマークに含まれるURLの指し示すWebページ本文及びそのページが参照するリンク先のWebページからも抽出し、ブックマークの更新による変化に追隨して、プロフィール情報（特徴語群）を更新前、更新後で比較して共通部分となる特徴語群を生成し、プロフィール情報を更新する。更新されたプロフィール情報及び更新されたリンク先のプロフィール情報のある一定以上の類似度を有するプロフィール情報に基づいて、検索結果をソートして出力する。

【選択図】 図1

本発明の原理を説明するための図



【特許請求の範囲】**【請求項 1】**

個人の興味応じた Web ページの情報検索を行うための個人適応型情報検索方法であって、

URL 抽出手段が、ブックマーク記憶手段から、利用者が過去に閲覧し、保存した Web ページのうち有用なものとして保存されている URL を取得する URL 抽出ステップと、

第 1 の特徴語抽出手段が、前記 URL に基づいて Web ページを取得し、該 Web ページの本文から抽出された特徴語群 W A a を第 1 のプロファイル情報記憶手段に格納する特徴語抽出ステップと、

第 1 の更新手段が、前回の特徴語群 W B a が格納されている第 2 のプロファイル情報記憶手段と前記第 1 のプロファイル情報記憶手段に格納されている特徴語群 W A a , W B a とを比較してそれらの共通部分となる特徴語群 W C a を生成し、該第 2 のプロファイル情報記憶手段に格納することにより、該第 2 のプロファイル情報記憶手段の内容を更新する第 1 の更新ステップと、

第 2 の特徴語抽出手段が、取得した前記 URL に基づいて参照先 Web ページを取得し、該参照先 Web ページの本文から抽出された特徴語群 W A b を第 1 の追加プロファイル情報記憶手段に格納する参照先ページ取得ステップと、

第 2 の更新手段が、前回の特徴語群 W B b が格納されている第 2 の追加プロファイル情報記憶手段と前記第 1 の追加プロファイル記憶手段に格納されている W A b 、 W B b とを比較してそれらの共通部分となる特徴語群 W C b を生成し、該第 2 の追加プロファイル情報記憶手段に格納することにより、該第 2 の追加プロファイル情報記憶手段の内容を更新する第 2 の更新ステップと、

特徴語合成手段が、前記第 2 のプロファイル情報記憶手段と前記第 2 の追加プロファイル情報記憶手段に格納されている特徴語群 W A a 、 W A b の個々に対して類似度を比較し、所定の類似度以上の特徴語群 W D を抽出する特徴語合成ステップと、

検索結果ソート手段が、入力された検索式により検索された結果を前記特徴語群 W D に基づいて検索結果をソートして出力する検索結果ソートステップと、
を行うことを特徴とする個人適応型 Web 情報検索方法。

【請求項 2】

前記参照先ページ取得ステップにおいて、

取得した参照先の URL とその参照レベル n に基づいて、該 URL の示すソースを取得する処理を参照レベルが、所定のレベル N になるまで繰り返す
請求項 1 記載の個人適応型 Web 情報検索方法。

【請求項 3】

前記検索結果ソートステップにおいて、

前記検索結果に含まれる各文書から文書ベクトルを生成し、

前記特徴語群 W D によるベクトルとの内積を計算して、該内積の大きい順に前記文書をソートする

請求項 1 記載の個人適応型 Web 情報検索方法。

【請求項 4】

個人の興味応じた Web ページの情報検索を行うための個人適応型情報検索装置であって、

利用者により選択された Web ページをブックマークとして格納されているブックマーク記憶手段と、

前記ブックマーク記憶手段から、利用者が過去に閲覧し、保存した Web ページのうち有用なものとして保存されている URL を取得する URL 抽出手段と、

前記 URL に基づいて Web ページを取得するページ取得手段と、

前記 Web ページの本文から抽出された特徴語群 W A a を第 1 のプロファイル情報記憶手段に格納する第 1 の特徴語抽出手段と、

前回の特徴語群 W B a が格納されている第 2 のプロファイル情報記憶手段と前記第 1 のプロファイル情報記憶手段に格納されている特徴語群 W A a , W B a とを比較してそれらの共通部分となる特徴語群 W C a を生成し、該第 2 のプロファイル情報記憶手段に格納することにより、該第 2 のプロファイル情報記憶手段の内容を更新する第 1 の更新手段と、

前記ページ取得手段で取得した W e b ページから参照先の U R L を抽出し、参照先 W e b ページを取得する参照先ページ取得手段と、

前記参照先 W e b ページの本文から抽出された特徴語群 W A b を第 1 の追加プロファイル情報記憶手段に格納する第 2 の特徴語抽出手段と、

前回の特徴語群 W B b が格納されている第 2 の追加プロファイル情報記憶手段と前記第 1 の追加プロファイル記憶手段に格納されている W A b 、 W B b とを比較してそれらの共通部分となる特徴語群 W C b を生成し、該第 2 の追加プロファイル情報記憶手段に格納することにより、該第 2 の追加プロファイル情報記憶手段の内容を更新する第 2 の更新手段と、

前記第 2 のプロファイル情報記憶手段と前記第 2 の追加プロファイル情報記憶手段に格納されている特徴語群 W A a 、 W A b の個々に対して類似度を比較し、所定の類似度以上の特徴語群 W D を抽出する特徴語合成手段と、

入力された検索式により検索手段により検索された結果を前記特徴語群 W D に基づいて検索結果をソートして出力する検索結果ソート手段と、

を有することを特徴とする個人適応型 W e b 情報検索装置。

【請求項 5】

前記ページ取得手段は、

取得した前記 W e b ページの参照先 U R L と参照レベル n を前記参照先ページ取得手段に出力する手段を含み、

前記参照先ページ取得手段は、

取得した前記参照先の U R L と前記参照レベル n に基づいて、該 U R L の示すソースを取得する処理を参照レベルが、所定のレベル N になるまで繰り返す手段を含む

請求項 4 記載の個人適応型 W e b 情報検索装置。

【請求項 6】

前記検索結果ソート手段は、

前記検索結果に含まれる各文書から文書ベクトルを生成する手段と、

前記特徴語群 W D によるベクトルとの内積を計算して、該内積の大きい順に前記文書をソートする手段を含む

請求項 4 記載の個人適応型 W e b 情報検索方法。

【請求項 7】

請求項 4 乃至 6 のいずれか 1 項記載の個人適応型 W e b 情報検索装置を構成する各手段としてコンピュータを機能させるための個人適応型 W e b 情報検索プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、個人適応型 W e b 情報検索装置及び方法及びプログラムに係り、特に、個人のプロファイルによって W e b 検索結果の得点付けを行う場合のプロファイル情報の更新が可能な個人適応型 W e b 情報検索装置及び方法及びプログラムに関する。

【背景技術】

【0002】

従来の個人適応型情報検索技術において、個人の興味内容に応じて作成されたプロファイル情報を用いて、情報検索結果の適合度順位を変更する手法としては以下のようなものがある。

【0003】

通常の情報検索において利用者が使用した検索式、その検索結果に表示された文字列やその検索結果のうちユーザが保存した文書の内容文に含まれる単語からストップワードを

10

20

30

40

50

除いて得られる語や高頻度の語（特徴語）の組を当該利用者のプロフィール情報とし、それと情報検索の結果として得られる文書の文書ベクトルとの類似度順位を検索結果の順位とする（例えば、非特許文献1参照）。

【非特許文献1】野美山浩「個人適応型情報検索システム：個人の興味を学習する階層記憶モデルとその協調的フィルタリングへの適用」情報処理学会研究報告、情報学基礎研究会報告 IPSJ SIG Notes Vol. 96, No.70 pp. 49-56 (1996)

【発明の開示】

【発明が解決しようとする課題】

【0004】

しかしながら、上記従来の方法には以下のような問題点がある。

10

【0005】

上記の方法では、検索結果全体には利用者要求に適合しないWebページも多く含まれるので、そこから抽出される特徴語には利用者の興味に合致しない特徴語も多く抽出されてしまう。

【0006】

また、Webページの保存は通常、永続的なWebサイトについては行われていないことが多い。そのため、上述した従来の方法では、その対象Webページからプロフィール情報として適切な特徴語を抽出することが困難であるため、それを用いた検索結果の個人適応化も適切に行われていない。

20

【0007】

本発明は、上記の点に鑑みなされたもので、個人の興味に応じたインターネット情報検索ができるようにするために、個人のプロフィール情報を適切に取得することによって個人適応検索を行うことが可能な個人適応型Web情報検索装置及び方法及びプログラムを提供することを目的とする。

【課題を解決するための手段】

【0008】

図1は、本発明の原理を説明するための図である。

【0009】

本発明（請求項1）は、個人の興味に応じたWebページの情報検索を行うための個人適応型情報検索方法であって、

30

URL抽出手段が、ブックマーク記憶手段から、利用者が過去に閲覧し、保存したWebページのうち有用なものとして保存されているURLを取得するURL抽出ステップ（ステップ1）と、

第1の特徴語抽出手段が、URLに基づいてWebページを取得し、該Webページの本文から抽出された特徴語群WAaを第1のプロファイル情報記憶手段に格納する特徴語抽出ステップ（ステップ2）と、

第1の更新手段が、前回の特徴語群WBaが格納されている第2のプロファイル情報記憶手段と第1のプロファイル情報記憶手段に格納されている特徴語群WAa、WBaとを比較してそれらの共通部分となる特徴語群WCaを生成し、該第2のプロファイル情報記憶手段に格納することにより、該第2のプロファイル情報記憶手段の内容を更新する第1の更新ステップ（ステップ3）と、

40

第2の特徴語抽出手段が、取得したURLに基づいて参照先Webページを取得し、該参照先Webページの本文から抽出された特徴語群WAbを第1の追加プロフィール情報記憶手段に格納する参照先ページ取得ステップ（ステップ4）と、

第2の更新手段が、前回の特徴語群WBbが格納されている第2の追加プロフィール情報記憶手段と第1の追加プロフィール記憶手段に格納されているWAb、WBbとを比較してそれらの共通部分となる特徴語群WCbを生成し、該第2の追加プロフィール情報記憶手段に格納することにより、該第2の追加プロフィール情報記憶手段の内容を更新する第2の更新ステップ（ステップ5）と、

特徴語合成手段が、第2のプロファイル情報記憶手段と第2の追加プロフィール情報記

50

憶手段に格納されている特徴語群W A a、W A bの個々に対して類似度を比較し、所定の類似度以上の特徴語群W Dを抽出する特徴語合成ステップ（ステップ6）と、

検索結果ソート手段が、検索式が入力されると（ステップ7）検索された結果を特徴語群W Dに基づいて検索結果をソートして出力する検索結果ソートステップ（ステップ8）と、を行う。

【0010】

また、本発明（請求項2）は、参照先ページ取得ステップ（ステップ4）において、

取得した参照先のURLとその参照レベルnに基づいて、該URLの示すソースを取得する処理を参照レベルが、所定のレベルNになるまで繰り返す。

【0011】

10

また、本発明（請求項3）は、検索結果ソートステップ（ステップ8）において、

検索結果に含まれる各文書から文書ベクトルを生成し、

特徴語群W Dによるベクトルとの内積を計算して、該内積の大きい順に文書をソートする。

【0012】

図2は、本発明の原理構成図である。

【0013】

本発明（請求項4）は、個人の興味応じたWebページの情報検索を行うための個人適応型情報検索装置であって、

利用者により選択されたWebページをブックマークとして格納されているブックマーク記憶手段1と、

20

ブックマーク記憶手段1から、利用者が過去に閲覧し、保存したWebページのうち有用なものとして保存されているURLを取得するURL抽出手段2と、

URLに基づいてWebページを取得するページ取得手段3と、

Webページの本文から抽出された特徴語群W A aを第1のプロファイル情報記憶手段9に格納する第1の特徴語抽出手段4と、

前回の特徴語群W B aが格納されている第2のプロファイル情報記憶手段8と第1のプロファイル情報記憶手段に格納されている特徴語群W A a、W B aとを比較してそれらの共通部分となる特徴語群W C aを生成し、該第2のプロファイル情報記憶手段8に格納することにより、該第2のプロファイル情報記憶手段8の内容を更新する第1の更新手段10と、

30

ページ取得手段5で取得したWebページから参照先のURLを抽出し、参照先Webページを取得する参照先ページ取得手段13と、

参照先Webページの本文から抽出された特徴語群W A bを第1の追加プロファイル情報記憶手段17に格納する第2の特徴語抽出手段11と、

前回の特徴語群W B bが格納されている第2の追加プロファイル情報記憶手段16と第1の追加プロファイル記憶手段17に格納されているW A b、W B bとを比較してそれらの共通部分となる特徴語群W C bを生成し、該第2の追加プロファイル情報記憶手段16に格納することにより、該第2の追加プロファイル情報記憶手段16の内容を更新する第2の更新手段18、

40

第2のプロファイル情報記憶手段8と第2の追加プロファイル情報記憶手段16に格納されている特徴語群W A a、W A bの個々に対して類似度を比較し、所定の類似度以上の特徴語群W Dを抽出する特徴語合成手段19と、

入力された検索式により検索手段21により検索された結果を特徴語群W Dに基づいて検索結果をソートして出力する検索結果ソート手段22と、を有する。

【0014】

また、本発明（請求項5）は、ページ取得手段5において、

取得したWebページの参照先URLと参照レベルnを参照先ページ取得手段13に出力する手段を含み、

参照先ページ取得手段13は、

50

取得した参照先のURLと参照レベルnに基づいて、該URLの示すソースを取得する処理を参照レベルが、所定のレベルNになるまで繰り返す手段を含む。

【0015】

また、本発明（請求項6）は、検索結果ソート手段22において、検索結果に含まれる各文書から文書ベクトルを生成する手段と、特徴語群WDによるベクトルとの内積を計算して、該内積の大きい順に文書をソートする手段を含む。

【0016】

本発明（請求項7）は、請求項4乃至6のいずれか1項記載の個人適応型Web情報検索装置を構成する各手段としてコンピュータを機能させるための個人適応型Web情報検索プログラムである。

【発明の効果】

【0017】

上述のように本発明によれば、利用者がインターネット検索途中に一時的に閲覧した検索結果ページ全体や保存したWebページの本文から抽出するという従来の特徴語抽出法に比べ、ブックマークに含まれるURLの指し示すWebページ及び、そのページが参照する別のWebページから特徴語を抽出することで、より利用者に個人適合したプロフィール情報を作成できるようになり、その結果、それを用いた検索結果の適合度順位付けがより個人適応化される。

【0018】

また、その後ブックマークが更新され、内容が変化してもそれに応じて変化する特徴語群を更新前後の特徴語群の比較から適切に計算することによってプロフィール情報の内容を適切に更新することができる。

【発明を実施するための最良の形態】

【0019】

以下、図面と共に本発明の実施の形態を説明する。

【0020】

図3は、本発明の一実施の形態における個人適応型インターネット情報検索装置の構成を示す。

【0021】

同図に示す装置は、ブックマーク記憶部1、URL抽出部2、タイミング制御部3、URLフィルタリング部4、基本HTML取得部5、本文抽出部6、特徴語取得部7、基本特徴語保存部A8、基本特徴語保存部B9、比較更新部10、リンク取得部11、URLフィルタリング部12、HTML取得部13、本文抽出部14、特徴語取得部15、追加特徴語保存部A16、追加特徴語保存部B17、比較更新部18、特徴語合成部19、検索式入力部20、検索エンジン部21、検索結果ソート部22、結果表示部23から構成される。

【0022】

上記の構成における動作を以下に示す。

【0023】

個人適応型インターネット情報検索装置は、利用者が過去に保存したWebページのうち有用なものとしてブックマークに保存されているURLなど、Webページの情報をインターネットブラウザのブックマークから取得し、そのURLのWebページ及びそのページ参照先Webページ本文中からそれらの特徴語群を抽出して、それらをプロフィール情報とする。さらに、ブックマークがその後、更新された場合には更新前後の参照先Webページから得られる特徴語群を比較してプロフィール情報をより適切なものに更新する。

【0024】

そして検索の際には、情報検索結果をこのプロフィール情報中の特徴語群によって個人適応化した適合度順位にソートする。

10

20

30

40

50

【 0 0 2 5 】

図 4 は、本発明の一実施の形態における全体処理のフローチャートである。

【 0 0 2 6 】

予め利用者は情報検索に先立って自分にとって興味のある Web ページをブックマーク記憶部 1 に格納しておくものとする。

【 0 0 2 7 】

ステップ 1 0 1) URL 抽出部 2 は、ブックマーク記憶部 1 に保存されている URL を抽出して、URL フィルタリング部に送る。この URL 抽出を行うタイミングはタイミング制御部 3 により指定される利用者の指定する時点、もしくは X 時間おきなど予めスケジューリングされたタイミングに基づいて指定される。

10

【 0 0 2 8 】

ステップ 1 0 2) URL フィルタリング部 4 は、URL にフィルタリングを行い、不要な URL は除去し、そうでないものは基本 HTML 取得部 5 に送る。フィルタリングの具体的方法には、例えば、予め作成した URL ブラックリストに一致する URL を除去する方法や、URL 文字列パターンによるマッチングルールに基づいて方法などを用いる。

【 0 0 2 9 】

なお、本ステップの処理については図 5 で詳述する。

【 0 0 3 0 】

ステップ 1 0 3) 基本 HTML 取得部 5 は、指定された URL の Web ページ (基本ページ) をインターネットから取得し、そのページの HTML ソースを本文抽出部 6 とリンク抽出部 1 1 へ送る。このとき、参照レベルの情報を付加する。

20

【 0 0 3 1 】

ステップ 1 0 4) 本文抽出部 6 では、基本ページの HTML ソースからタグなどを取り除いて文章の本文だけを抽出して、特徴語取得部 7 に送る。

【 0 0 3 2 】

ステップ 1 0 5) 特徴語抽出部 7 では、抽出された本文を解析して単語を抽出し、それからストップワードを除去した後、特徴語抽出手法によって、それらの単語から特徴語を選び出す。そして、基本特徴語保存部 B 9 に保存する。但し、ここで用いる特徴語抽出手法としては T F I D F 法のような既存の方法であれば何でもよい。

【 0 0 3 3 】

ステップ 1 0 6) 比較更新部 1 0 において、基本特徴語保存部 B 9 から特徴語群 W B を取得し、もし、基本特徴語保存部 A 8 に何も保存されていない場合には、取得した特徴語群 W B を基本特徴語保存部 A 8 に保存する。しかし、既に、基本特徴語保存部 A 8 に特徴語群が存在する場合はその特徴語群 W A を取得し、W A , W B の両者を比較してそれらの共通部分となる特徴語群 W C を生成し、W C で基本特徴語保存部 A 8 の内容を更新する。

30

【 0 0 3 4 】

ステップ 1 0 7) リンク抽出部 1 1 において、ステップ 1 0 3 において、基本 HTML 取得部 5 から送られた対象となる HTML ソースからアンカータグによって示されるリンク先 URL を取得し、URL フィルタリング部 1 2 に送る。なお、本ステップの処理については図 6 で詳述する。

40

【 0 0 3 5 】

ステップ 1 0 8) URL フィルタリング部 1 2 では、URL フィルタリング部 4 と同様にリンク先 URL をフィルタリングし、得られた URL を HTML 取得部 1 3 に送る。

【 0 0 3 6 】

なお、本ステップの処理については図 5 で詳述する。

【 0 0 3 7 】

ステップ 1 0 9) HTML 取得部 1 3 は、指定された URL の Web ページをインターネットから取得し、そのページの HTML ソースを本文抽出部 1 4 に送り、また、その URL が基本ページから予め決められた N リンク以内の URL であれば、その HTML ソ

50

ースをリンク抽出部 11 に送る。ここでは、リンク抽出部 11、URL フィルタリング部 12、HTML 取得部 13 で基本ページから N リンク以内の Web ページを反復取得しているが、同等の処理結果を得るためにこれらのブロックを N 段連続して構成したものでよい。当該ステップの処理については図 7 で詳述する。

【0038】

ステップ 110) 本文抽出部 14 では、HTML 取得部 13 で取得された HTML ソースからタグなどを取り除いて文章の本文だけを抽出して特徴語取得部 15 に送る。

【0039】

ステップ 111) 特徴語抽出部 15 では抽出された本文を解析して単語を抽出し、それからストップワードを除去した後、特徴語抽出手法によって、それらの単語から特徴語を選び出し、追加特徴語保存部 B 17 に格納する。但し、ここで用いる特徴語抽出手法としては T F I D F 法のような既存の方法であれば何を用いてもよい。

10

【0040】

ステップ 112) 比較更新部 18 において、追加特徴語保存部 B 17 から特徴語群 W B を取得し、もし追加特徴語保存部 A 16 に何も保存されていない場合には、取得した特徴語群を追加特徴語保存部 A 16 に保存する。しかし、既に追加特徴語保存部 A 16 に特徴語群が存在する場合は、その特徴語群 W A を取得し、W A、W B の両者を比較してそれらの共通部分となる特徴語群 W C を生成し、W C で追加特徴語保存部 A 16 の内容を更新する。

【0041】

20

ステップ 113) 特徴語合成部 19 では、基本特徴語保存部 A 8 の特徴語群 W A を取得し、検索結果ソート部 17 に送る。追加特徴語保存部 A 16 に含まれる特徴語も取得し、それら個々に対して特徴語群 W A 中の全ての語との類似度を比較し、W A のいずれかの特徴語とある一定の類似度以上があるものは検索結果ソート部 17 に送る。但し、ここで用いる類似度としては余弦類似度など既存の何を用いてもよい。当該ステップの詳細な処理は図 8 で詳述する。

【0042】

ステップ 114) 検索式入力部 20 において、利用者から検索式の入力を受け付けて、得られた検索式を検索エンジン部 21 に送る。

【0043】

30

ステップ 115) 検索エンジン部 21 では、検索式入力部 20 で得られた検索式によってインターネット検索を行い、その検索結果を検索結果ソート部 17 に送る。

【0044】

ステップ 116) 検索結果ソート部 17 では、検索エンジン部 21 から得られる検索結果に含まれる各文書から文書ベクトルを生成し、特徴語合成部 19 から得られる特徴語群によるベクトルとの内積を計算して、その内積の大きさの順に元の文書をソートし、その結果を結果表示部 23 に送る。

【0045】

ステップ 117) 最後に結果表示部 23 では検索結果ソート部 22 によって得られた検索結果を利用者に表示する。

40

【0046】

次に、上記のステップ 102、108 における URL のフィルタリングについて説明する。以下ではブラックリスト方式の例を説明する。

【0047】

図 5 は、本発明の一実施の形態における URL フィルタリング処理のフローチャートである。以下では、URL フィルタリング部 4 の例を示すが、リンク先の URL のフィルタリングを行う URL フィルタリング部 12 も同様の処理を行う。

【0048】

ステップ 201) URL 抽出部 2 から URL を取得する。

【0049】

50

ステップ 202) 取得した URL を除去すべき URL の登録されたブラックリストデータベース (図示せず) と照合する。

【0050】

ステップ 203) 照合の結果、URL がブラックリストにマッチすれば、本処理終了する。マッチしなければステップ 204 に移行する。

【0051】

ステップ 204) 当該 URL を基本 HTML 取得部 5 に出力する。

【0052】

次に、上記のステップ 107 のリンク取得部 11 における動作を説明する。

【0053】

図 6 は、本発明の一実施の形態におけるリンク取得処理のフローチャートである。

【0054】

ステップ 301) リンク取得部 11 は、基本 HTML 取得部 5 から HTML 文書とその参照レベル $n = k$ を取得する。

【0055】

ステップ 302) HTML 文書中の全てのアンカータグについてその HREF 属性を URL として抽出する。

【0056】

ステップ 303) それぞれの URL に対してその参照レベル n を $k + 1$ として出力する。

【0057】

次に、上記のステップ 109 の HTML 取得部 13 の動作を説明する。

【0058】

図 7 は、本発明の一実施の形態における HTML 取得部の処理を示すフローチャートである。

【0059】

ステップ 401) URL フィルタリング部 12 から URL とその参照レベルを取得する。

【0060】

ステップ 402) URL の指し示す Web ページのソースを取得する。

【0061】

ステップ 403) 参照レベル n と N (予め設定された参照レベルの最大値) とを比較し、 $n < N$ であればページソースと参照レベル n とをリンク取得部 11 に出力してステップ 404 に移行する。 $n = N$ であればステップ 404 に移行する。

【0062】

ステップ 404) ページソースを本文抽出部 14 に出力する。

【0063】

次に、上記のステップ 113 の特徴語合成部 19 の処理を説明する。

【0064】

図 8 は、本発明の一実施の形態における特徴語合成処理のフローチャートである。

【0065】

ステップ 501) 基本特徴語保存部 A8 から基本特徴語 $W_1, \dots, W_n$ を取得し、これらを出力する

ステップ 502) ループ変数 k, i を 1 に初期化する。

【0066】

ステップ 503) $i > m$ (追加特徴語数の最大値) であれば終了する。そうなければステップ 504 に移行する。

【0067】

ステップ 504) $k > n$ (基本特徴語数の最大値) であればステップ 509 に移行する。そうでなければステップ 505 に移行する。

10

20

30

40

50

【 0 0 6 8 】

ステップ 5 0 5) 基本特徴語 W_k と追加特徴語 T_i との類似度である S_{ki} を計算する。

【 0 0 6 9 】

ステップ 5 0 6) 類似度 S_{ki} が予め決められた閾値 C 以上であるならば、ステップ 5 0 7 に移行する。そうでなければ、ステップ 5 0 8 に移行する。

【 0 0 7 0 】

ステップ 5 0 7) 追加特徴語 T_i を出力してステップ 5 0 9 に移行する。

【 0 0 7 1 】

ステップ 5 0 8) $k = k + 1$ とし、ステップ 5 0 4 に移行する。

10

【 0 0 7 2 】

ステップ 5 0 9) $i = i + 1$ とし、ステップ 5 0 3 に移行する。

【 0 0 7 3 】

上記の図 3 に示す構成要素の動作をプログラムとして構築し、個人適応型インターネット情報検索装置として利用されるコンピュータにインストールして実行させる、または、ネットワークを介して流通させることが可能である。

【 0 0 7 4 】

また、構築されたプログラムをハードディスクや、フレキシブルディスク・CD-ROM等の可搬記憶媒体に格納し、コンピュータにインストールする、または、配布することが可能である。

20

【 0 0 7 5 】

なお、本発明は、上記の実施の形態に限定されることなく、特許請求の範囲内において種々変更・応用が可能である。

【 産業上の利用可能性 】

【 0 0 7 6 】

本発明は、インターネットの Web 検索技術に適用可能である。

【 図面の簡単な説明 】

【 0 0 7 7 】

【 図 1 】 本発明の原理を説明するための図である。

【 図 2 】 本発明の原理構成図である。

30

【 図 3 】 本発明の一実施の形態における個人適応型 Web 情報検索装置の構成図である。

【 図 4 】 本発明の一実施の形態における全体処理のフローチャートである。

【 図 5 】 本発明の一実施の形態における URL フィルタリング処理のフローチャートである。

【 図 6 】 本発明の一実施の形態におけるリンク取得部の処理を示すフローチャートである。

【 図 7 】 本発明の一実施の形態における HTML 取得部の処理を示すフローチャートである。

【 図 8 】 本発明の一実施の形態における特徴語合成部の処理を示すフローチャートである。

40

【 符号の説明 】

【 0 0 7 8 】

- 1 ブックマーク記憶手段、ブックマーク記憶部
- 2 URL 抽出手段、URL 抽出部
- 3 タイミング制御部
- 4 URL ファイルタリング部
- 5 ページ取得手段
- 6 本文抽出部
- 7 第 1 の特徴語抽出手段、特徴語取得部
- 8 第 2 のプロファイル情報記憶部、基本特徴語保存部 A

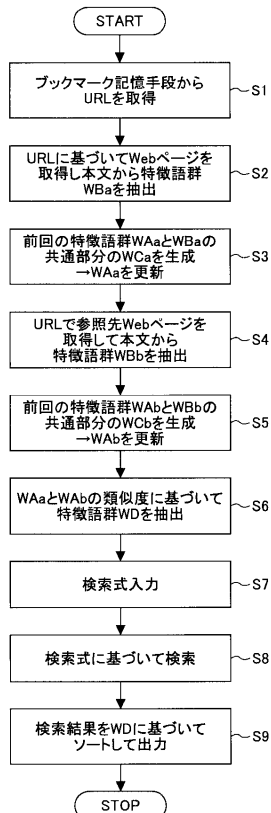
50

- 9 第1のプロファイル情報記憶手段、基本特徴語保存部 B
 10 第1の更新手段、比較更新部
 11 第2の特徴語抽出手段、リンク取得部
 12 URLフィルタリング部
 13 参照先ページ取得手段、HTML取得部
 14 本文抽出部
 15 特徴語取得部
 16 第2の追加プロファイル情報記憶手段、追加特徴語保存部 A
 17 第1の追加プロファイル情報記憶手段、追加特徴語保存部 B
 18 第2の更新手段、比較更新部
 19 特徴語合成手段、特徴語合成部
 20 検索式入力部
 21 検索手段、検索エンジン
 22 検索結果ソート手段、検索結果ソート部
 23 結果出力部

10

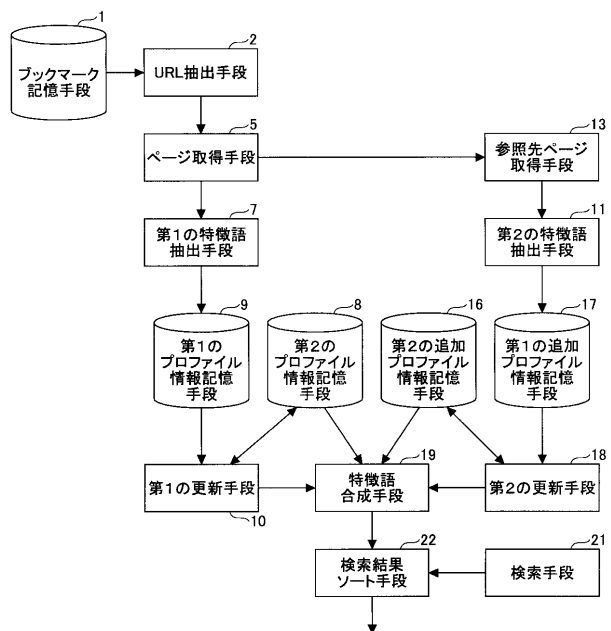
【図1】

本発明の原理を説明するための図



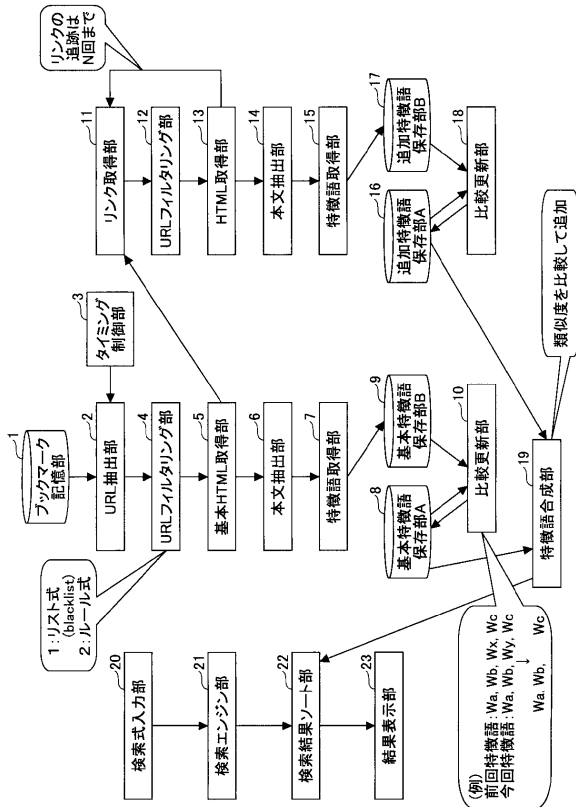
【図2】

本発明の原理構成図



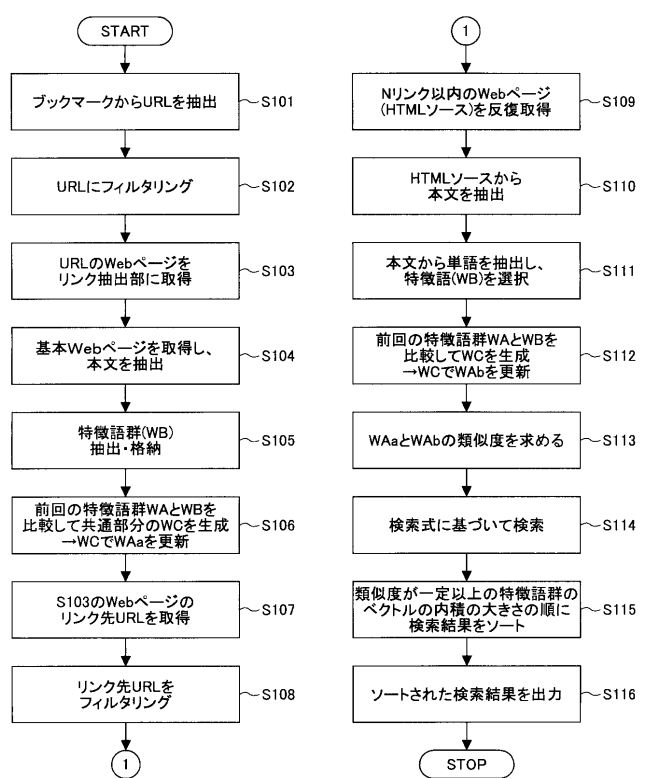
【図 3】

本発明の一実施の形態における個人適応型
インターネット情報検索装置の構成図



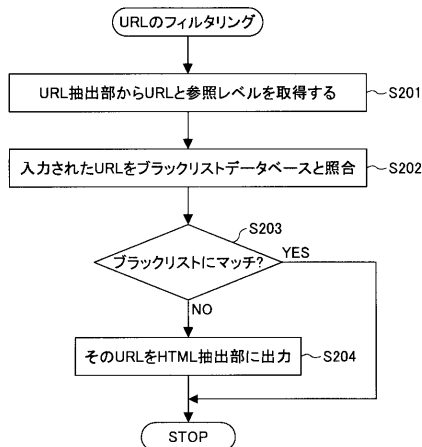
【図 4】

本発明の一実施の形態における全体処理のフローチャート



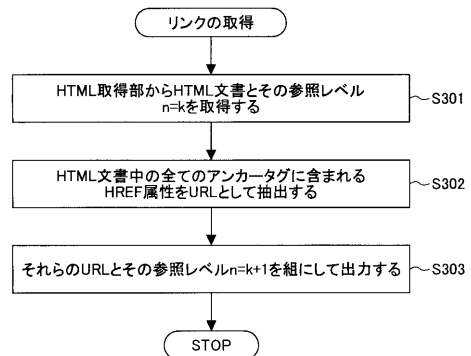
【図 5】

本発明の一実施の形態におけるURLフィルタリング処理のフローチャート



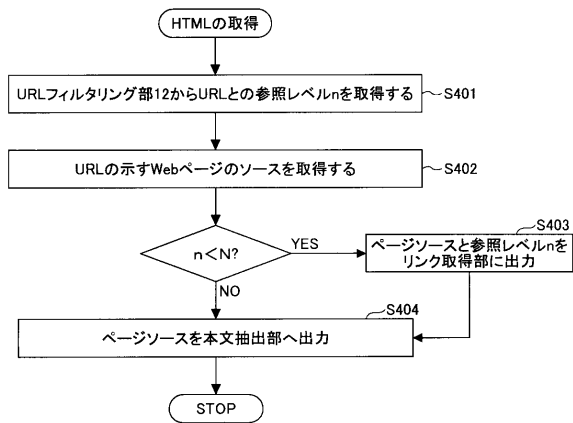
【図 6】

本発明の一実施の形態におけるリンク取得部の処理を示すフローチャート



【 図 7 】

本発明の一実施の形態におけるHTML取得部の処理を示すフローチャート



【 図 8 】

本発明の一実施の形態における特徴語合成部の処理を示すフローチャート

